MODERN STATISTICAL APPROACHES FOR RANDOMIZED EXPERIMENTS UNDER INTERFERENCE

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Alex Chin
March 2019

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Johan Ugander, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Julia Palacios

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Guenther Walther

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

This thesis presents new methodology for handling interference in randomized experiments. Interference, a phenomenon in which individuals interact with each other, is widely prevalent in the social and natural sciences, and has major implications for how experiments are optimally designed and analyzed.

I first provide an introduction to interference, including examples and a relevant brief history of causal inference. Next, I demonstrate how researchers can use Stein's method to establish limiting distributional results for estimators under interference. The modern tools afforded by Stein's method allow one to analyze certain regimes of arbitrarily dense interference, which goes beyond the analysis capabilities of existing tools.

In the subsequent chapter, I develop new model-based, adjustment estimators for estimating the global average treatment effect. The adjustment variables can be constructed from functions of the treatment assignment vector, and the researcher can use a collection of any functions correlated with the response, turning the problem of detecting interference into a feature engineering problem. The final chapter proposes new methods for designing and analyzing stochastic seeding strategies, which are an appealing way of leveraging network structure for marketing, public health, and behavioral interventions. New importance sampling estimators adapted to this setting can greatly improve precision over existing approaches.

This thesis is interdisciplinary in nature. Stein's method (Chapter 2), regression adjustments (Chapter 3), and importance sampling (Chapter 4) all command spheres of influence in certain sectors of the literature, and are here repurposed in new domains. I hope that my work shows how existing statistical technology can arise in new

arenas of application while simultaneously giving rise to new methodological questions and problems, and in this way, I hope my work is useful for both practitioners and methodologists.

Acknowledgements

First, I owe an extraordinary debt of gratitude to my advisor, Johan Ugander. His mentorship and guidance have been indispensable for bridging the gap in my education between student and researcher. His kindness, industriousness, and devotion to interdisplicinary pursuits have been an inspiration for me. I am also grateful that he has consistently allowed and encouraged me to pursue my own, independent work and carve out my own career path.

Next, I would like to thank Julia Palacios, Guenther Walther, Art Owen, and Jennifer Pan for carving time from their busy schedules to serve on my dissertation and defense committees. I thank all of the faculty, staff, and teachers at Stanford who have helped me over the last five years, as well as everyone with whom I have formally or informally collaborated. Among my collaborators I must especially thank Dean Eckles. Dean has been an excellent collaborator and mentor, and I have learned a great deal from his unique perspective on research.

I have been fortunate to complete three internships during my time at Stanford. These experiences have been invaluable for shaping my research and professional outlook. They have given me perspective on the most important and impactful aspects of research problems, imparted me with technical skills for maximizing my research efficiency, and driven me to improve my communication skills. Thanks goes to all of my industry mentors and friends from these experiences, including William Browne, Eytan Bakshy, Kostya Kashin, and April Chen.

I am grateful to have been surrounded by so many wonderful friends and classmates. This five-year journey with my PhD cohort — Yu Bai, Wenfei Du, Leying Guan, Gene Katsevich, Keli Liu, Jelena Markovic, Paulo Orenstein, Evan Patterson,

Feng Ruan, Pragya Sur, and Jeha Yang — has been filled with many adventures and milestones, all the way from our first-year courses through the present. Many friendships with other students, including Stephen Bates, Rina Friedberg, Evan Rosenman, and Andy Tsao, have been essential. My friendships with Alton Russell and Krystal Smith began at NC State and have been rejuvenated here at Stanford. I thank all of my other friends from NC State and Stanford who have provided me with so much support. I especially thank Lydia Allen for being a great friend.

My education began long before I came to Stanford. Crucial mentors during my undergraduate years include Alina Duca, Sandra Paur, and Eric Stone, and Blair Sullivan from NC State, and Gary Gordon from Lafayette College. I remain grateful to my teachers in the International Baccalaureate program at Broughton High School, who instilled in me the importance of creativity and critical thinking in my intellectual pursuits. I inherited a passion for science and mathematics from Kevin Ledger, Helen Roberts, and Dave Corsetti. I must especially thank my English and history teachers, Barbara Nichols, Richard Matkins, Lee Quinn, and Bryan Elsaesser, who taught me the essential writing skills that have still proven vital all these years later.

Finally, I thank my parents, Patrick and Haren, and my brother, Adam. Without their lifelong support and sacrifice none of this would have been possible.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis concerns causal inference under interference. It is necessarily an interdisciplinary treatment and draws together ideas from classical statistics, machine learning, econometrics, and the social sciences. In this introductory chapter I provide a brief history of the relevant areas of the causal inference, experimental design/analysis, and interference literatures.

1.1 Causal inference

There has been intellectual interest in the philosophy of causation since ancient times. Plato's version of the *axiom of causality* or *principle of universal causation* was stated as "everything that becomes or changes must do so owing to some cause; for nothing can come to be without a cause" (*Timaeus 28a*)(Zeyl, 2000). Aristotle wrote that "we do not have knowledge of a thing until we have grasped its why, that is to say, its cause," and that answers to the question "why?" come in one of four categories: matter, form, agent, and purpose (*Physics II, 3*; *Metaphysics V, 2*).

The development of mathematical treatments of causality is tightly intertwined with the origins of modern statistics. The main ideas of the Neyman-Rubin causal model or potential outcomes framework originate from the analysis of agricultural experiments as written by Neyman (1923). Indeed, Jerzy Neyman describes the concept of an "array of unknown potential yields," which is used to determine the variance

of the difference between the average of the observed yields of two varieties of crops. Despite this early introduction of a potential outcomes framework, it was not until Rubin (1974) that this concept was extended beyond the randomized experiments setting, and more recently still that this framework has been commonly viewed as an acceptable way to approach the inferences of causal effects. For a more detailed history of the Neyman-Rubin framework I find Imbens and Rubin (2015, Chapter 2) to be a helpful literature review.

Consider a finite population of n *individuals* or *units* indexed $i = 1, \ldots, n$. Each unit is subjected to a treatment condition, which is represented by the random variable $W_i \in \mathcal{W}$. In a clinical trial, W_i might represent a choice of drug; on an online video platform, W_i might represent a choice of recommendation algorithm. The set $\mathcal{W}$ is the space of treatments. In some settings, $\mathcal{W} = \mathbb{R}$, for example if W_i represents the dosage of a drug; however in this thesis we will restrict ourselves to studying the case where $\mathcal{W} = \{0, 1\}$, so that the value of the treatment is either 0 (the *control condition*) or 1 (the *treatment condition*). The distribution of the vector of treatments $\mathbf{W} = (W_1, \ldots, W_n)$ is crucial for making inferences. In experimental settings this distribution is known and controlled by the experimenter; in observational settings this distribution is unknown and may depend on hidden covariates which can lead to confounding bias.

Classically, we then assume the existence of potential outcomes $Y_i(0)$ and $Y_i(1)$, which represent the response values in the case that unit i is subject to the control and treatment condition, respectively. The potential outcomes belong to an outcome space $\mathcal{Y}$ (which is generally equal to $\mathbb{R}$). The observed response for unit i is

$$Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0),$$

that is, Y_i takes the value $Y_i(1)$ when $W_i = 1$ and the value $Y_i(0)$ when $W_i = 0$. Crucially, exactly one of the two potential outcomes is observed and the other remains unobserved.

1.1.1 What is random?

We are concerned primarily with estimating variations of the *average causal effect* or *average treatment effect* (ATE)

$$\tau = \mathbf{E}[Y_i(1) - Y_i(0)], \tag{1.1}$$

which is a difference of average outcomes between the counterfactuals of exposure to treatment and exposure to control. It is important to be clear about the source of randomness here, which in turn informs what is meant by the expectation operator in the estimand above.

Typically, the potential outcomes $Y_i(0)$ and $Y_i(1)$ are assumed to be deterministic, and no superpopulation beyond the n observed units is considered. The only randomness in the observed outcome Y_i, therefore, is induced by the distribution of the treatment assignment W_i; and one could write

$$Y_i = y_i(W_i) \tag{1.2}$$

for some deterministic function $y_i : \mathcal{W} \to \mathcal{Y}$ mapping from the treatment space to the outcome space. Then, the expectation in equation (1.1) is meant as a finite expectation over the population of n units, and equation (1.1) becomes

$$\tau = \frac{1}{n} \sum_{i=1}^{n} [y_i(1) - y_i(0)]. \tag{1.3}$$

Other times, however, one assumes that $Y_i(0)$ and $Y_i(1)$ are themselves random variables, in other words, to assume that the counterfactuals are stochastic (VanderWeele and Robins, 2012). It will be convenient for us to take this approach when working in the presence of interference; it is for this reason that I have chosen to use capital letters to introduce the potential outcomes. Then, rather than equation (1.3)

the estimand is instead interpreted as

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}[Y_i(1) - Y_i(0)]. \tag{1.4}$$

Finally, there is the question of whether one considers the population to be simply the n observed units at hand or whether one wishes to consider a larger (finite or infinite) super-population. This is partially a substantive question, as whether the researcher's primary goal is to make inferences about the unobserved counterfactuals for the units at hand or to make inferences about entire units not observed. If one wishes to conduct superpopulation inference, then the average treatment effect estimand can simply be taken to be τ defined by equation (1.1), where we interpret the expectation to be taken over repeated i.i.d. samples of variables $(Y(0), Y(1))$.

As discussed in Imbens (2004), these sample and population ATE estimands are generally exchangeable in the sense that good estimators of (1.1) are generally good estimators of (1.3) and (1.4), and vice versa. That is, unbiasedness of an estimator is generally robust across this family of estimands. Naturally, there is additional variance from sampling in the superpopulation viewpoint, though in practice, variance *estimators* are usually the same for the finite population and superpopulation views (Ding, Li, and Miratrix, 2017).

The finite population setup is quite common and arguably the more popular approach in causal inference; the texts Imbens and Rubin (2015) and Gerber and Green (2012), for example, work primarily within the finite population setting. Reasoning about inference in this arena is quite different from the usual approach in classical statistics, in which the goal is nearly always to estimate information about an unseen larger population of interest. In the randomized experiment setting, Athey and Imbens (2017) advocate using randomization-based finite population inference over sampling-based superpopulation inference. They argue that the randomization of the treatment assignment, which is under the control of the researcher via a well-understood mechanism, is a more trustworthy source of uncertainty than sampling variation from a hypothetical superpopulation, which requires additional assumptions

about units not belonging to the experiment. This stance evokes both Fisher's "reasoned basis for inference" (Fisher, 1935) and Freedman's assertion that "experiments should be analyzed as experiments, not as observational studies" (Freedman, 2006).

This discussion is even more pertinent when dealing with interference; if the researcher has access to a social network measured on the population through which interference is being transmitted, then it is not at all clear how she might conduct superpopulation inference unless she also has measurements on a larger network (a supergraph) representing the superpopulation. For this reason, it makes most sense to stick to the finite population setup when discussing interference, potentially allowing the potential outcomes to be stochastic.

1.1.2 Adjustment methodology under SUTVA

In a *randomized experiment* the researcher has control over the distribution of treatments $\mathbf{W}$; in an *observational study* the distribution of treatments is unknown and set by nature. Experiments, or randomized controlled trials, are often viewed as the "gold standard" of causal inference. Unconfounded estimation in an experimental setting is straightforward because the difference-in-means estimator (DM) is unbiased for the ATE τ, but the practice of increasing the precision of estimators by controlling for covariates is quite old, using methods such as matching (Stuart, 2010), weighting, post-stratification (Holt and Smith, 1979), and regression estimators. It is now well-established that OLS adjustment separately within treatment groups never decreases asymptotic precision with a low-dimensional number of covariates (Freedman, 2008a,b; Lin, 2013; Berk et al., 2013). Regression adjustment works when the covariates are correlated with the outcomes, and effectively uses a model imputation of treatment and control potential outcomes. Much recent work in theoretical statistics has verified that high-dimensional regression and machine learning models can then be used, provided the imputations are obtained out-of-sample via a *cross-fitting* procedure (Belloni et al., 2014; Chernozhukov et al., 2015; Bloniarz et al., 2016; Wager et al., 2016; Wu and Gagnon-Bartsch, 2017; Athey et al., 2017b; Chernozhukov et al., 2018). Similar ideas are also applied to the heterogeneous treatment effect

setting (e.g. Künzel et al., 2017; Nie and Wager, 2017).

In the observational study setting, adjustment is necessary to remove confounding bias. The goal in such settings is most often to construct estimators that transform the data into something that locally behaves like a randomized experiment. A standard requirement is an *unconfoundedness* assumption, which means that all necessary variables for adjustment are observed, or equivalently that a "randomized experiment" holds conditionally on the covariates at hand. Unconfoundedness, a term coined by (Rubin, 1990b), is also referred to as exogeneity, ignorability (Rosenbaum and Rubin, 1983), or lack of selection on observables (Barnow et al., 1980). Methodology is currently quite limited if one is unwilling to assume unconfoundedness, though one can attempt to be robust to hidden confounders via sensitivity analysis, which is an active research area (e.g. Robins et al., 2000; Ding and VanderWeele, 2016). Because the lack of hidden confounders is always difficult to verify, quasi-experimental methods such as those relying on natural experiments, instrumental variables, regression discontinuity designs may be more trustworthy in practice.

Under unconfoundedness similar regression methods as in the experimental case can be used. Chernozhukov et al. (2018) propose double machine learning, in which the analyst trains an *outcome model* of the response Y_i on the covariates X_i and a separate *propensity model* of the treatment W_i on the covariates X_i. The residuals of these fitted models are then used in a second-stage regression to obtain the final treatment effect estimate. This approach is an analog of Robinson (1988)'s residual-on-residual approach. These types of estimators are provably "doubly-robust" (Kang et al., 2007; Dudík et al., 2011, 2014a), meaning that they remain consistent if either the propensity model or the outcome model (but not both) is misspecified. For a unified, general treatment of adjustment methodology see Middleton (2018); for a survey article on ATE estimation under unconfoundedness see Imbens (2004).

One of the themes of this thesis is that interference causes randomized experiments to exhibit characteristics of both randomized experiments and observational studies. The distribution of the treatment assignment is fully known and controlled, so no propensity model needs to be estimated. However, depending on the estimand of interest, an outcome model may need to be estimated, and randomized treatment

assignments *are not enough to guarantee unconfounded estimation*. The existence of this bias is known is well-known and intuitive (e.g. Eckles et al., 2017) but has not before been explicitly framed in the language of unconfoundedness.

1.2 Interference

To my knowledge, the term *interference* is first mentioned by David Cox in the text *Planning of Experiments* (Cox, 1958), in which three pages (Section 2.4) are devoted to providing a number of examples. Cox's states his *no-interference* assumption as follows:

> The observation obtained when a particular treatment is applied to a particular experimental unit is assumed to be a quantity depending only on the particular unit, plus a quantity depending on the treatment used, and to be unaffected by the particular assignment of treatments to the other units.

More succinctly, the assumption of no interference is that unit i subjected to treatment t exhibits outcome $y_i(t)$, where $y_i(t)$ is invariant to changes in treatment of any other unit j. No interference is also one key part of the *stable unit treatment value assumption* (SUTVA). This assumption is mentioned briefly in Rubin (1980); Rubin writes

> If unit i is exposed to treatment j, the observed value of Y will be Y_{ij}; that is, there is no interference between units leading to different outcomes depending on the treatments other units received and there are no versions of treatments leading to "technical errors."

(The second part of SUTVA, no hidden versions of treatments, will not be addressed here.)

The SUTVA assumption must be invoked in order for the potential outcomes $Y_i(0)$ and $Y_i(1)$ defined in Section 1.1.1 to be well-defined. Otherwise, the treatments of unit j may induce variation in the outcomes of unit i, in which case it is no longer correct

that $Y_i = y_i(W_i)$ for a deterministic function $y_i : \mathcal{W} \to \mathcal{Y}$. Instead, the function must be defined on the n-dimensional treatment vector $y_i : \mathcal{W}^n \to \mathcal{Y}$, producing observed outcome

$$Y_i = y_i(\mathbf{W}). \tag{1.5}$$

More recently, Manski (2013) has called the SUTVA assumption assumption *individualistic treatment response* (ITR) to emphasize the fact that SUTVA is a restriction on the form of the treatment response function.

In many social, medical, and online settings, where the no-interference assumption often fails to hold, it becomes important to go beyond individualistic outcomes (equation (1.2)) and develop methods for dealing non-individualistic outcome functions (equation (1.5)). This has emerged as a popular research area over the last few years, driven both by practical necessity and by the relevant interesting methodological questions that arise (e.g. Rubin, 1990a; Rosenbaum, 2007; Hudgens and Halloran, 2008; Ogburn and VanderWeele, 2014; van der Laan, 2014; Walker and Muchnik, 2014; Aral, 2016; Choi, 2017; Jagadeesan et al., 2017). See also Halloran and Hudgens (2016); Taylor and Eckles (2017) for recent surveys.

1.2.1 Examples

Here I illustrate several examples of scenarios which may give rise to interference. These examples are broad, both with respect to the domain area of application and to the methodologies which have arisen to tackle them. Thus, research that goes beyond the "i.i.d. sampling" regime that is familiar to many statisticians is not only prudent for applications but also provides a source of many interesting problems for the methodologically- or theoretically-inclined statistician.

Example 1.1 (Repeated subjects in experimental psychology)**.** Welford, Brown, and Gabb (1950) conducted experiments on airline crew to determine the impact of flying fatigue on performance of certain tasks. They noticed that subjects who first encountered a task while tired continued to do it badly when fresh, while conversely other subjects who first encountered a task while fresh continued to do well when

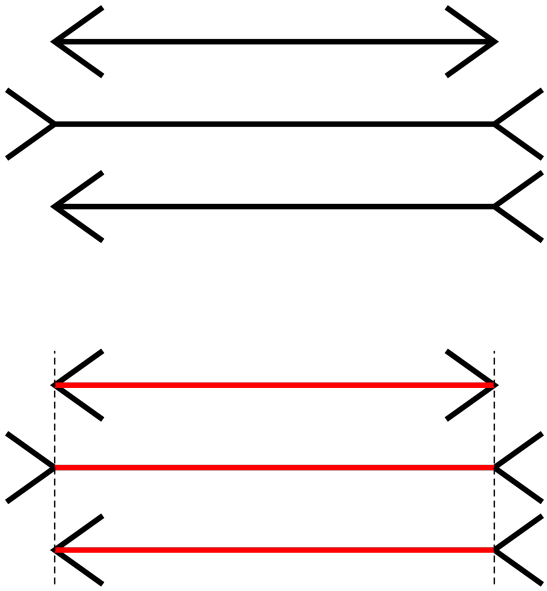

Figure 1.1: The Müller-Lyer illusion. Figure taken from Wikipedia.

tired. This phenomenon is an example of interference because individuals who repeat tasks (even days or weeks later) cannot be trusted to behave as "independent units," even after controlling for fatigue. Researchers therefore have to be cautious when determining the appropriate granularity for performing the analysis; treating each individual rather than "task instance" as a separate data point probably leads to a more accurate analysis. Cox (1958) notes that other similar effects have been reported in the literature.

For a toy example, consider the experiments conducted by Babington Smith (1951) on the Müller-Lyer illusion, which is illustrated in Figure 1.1. The perception of the lines as having different lengths is exactly due to a single individual viewing all lines in tandem.

Example 1.2 (Group interference)**.** Many data sources are naturally grouped or hierarchical in nature: students belong to classrooms which belong to schools; individuals belong to households which belong to neighborhoods or villages; online activity events can be clustered by session ID which can be clustered by user ID. In such cases it

may be reasonable to assume that interference exists *within* groups but not *across* groups; this assumption is called *partial interference* (Sobel, 2006; Hudgens and Halloran, 2008). Much early work on interference studies this grouped or blocked setting (Sobel, 2006; Rosenbaum, 2007).

Group interference is also closely tied to determining the unit of randomization upon which to conduct the experiment. For example, consider experimentation on video platforms like those found on Facebook or YouTube. In addition to randomizing entire clusters of users or individual users themselves, the analyst may choose to randomize at the *session level*, so that all videos viewed during the same session are assigned to the same treatment group but users who log on to YouTube multiple times during the course of an experiment may be assigned to different treatment groups each time. Or, the analyst may choose to randomize at the *video level*, potentially assigning individual video views to different treatment groups and disregarding session and user indicators all together. Using a more granular unit results in greater precision via larger sample sizes, yet may bias estimates in the event that any interference is present. Practically, greater precision may be worth the bias if the interference is weak enough.

More specifically, there is an extensive literature on the use of *two-stage randomized designs* or *saturation designs*, which first randomizes groups and then individuals. Using these designs a variety of direct and indirect effects and dose-response curves can be estimated (VanderWeele and Tchetgen Tchetgen, 2011; Tchetgen Tchetgen and VanderWeele, 2012; Liu and Hudgens, 2014; Baird et al., 2016; Basse et al., 2017; Basse and Feller, 2018).

Cluster-level interference is related to but not to be confused with the mechanisms that lead to the use of cluster standard errors. Interference means that outcomes are not individualistic, whereas cluster standard errors are used when outcomes are individualistic but correlated within groups. Abadie et al. (2017a) provide a careful study into how and when to appropriately use cluster standard errors (under SUTVA). One extension is to assume partial interference so that units may *interfere* within but not across groups, yet to use standard errors that protect against arbitrary *correlation* across groups.

Example 1.3 (Network spillovers and peer effects)**.** This setting is distinguished from the group interference setting not through the mechanism of interference but rather through its structure. Here we assume *general* or *arbitrary interference* in which any two units may interfere with each other. Often we assume that the researcher has access to a known, fixed network (graph) representing social connections among the units. (Of course, the measured network may not at all correspond to the pattern of interference, which can pose problems when the network is partially unmeasured/censored or when the researcher has several layers of network data available.)

Commonly, one tries to impose "more granular" potential outcomes such that "SUTVA holds conditionally" on these new potential outcomes. For example, one might define outcomes as a function of all treatments in the immediate neighborhood of unit i, rather than as a function of the treatment of unit i only. These models are known as *constant treatment response* (CTR) conditions (Manski, 2013) or *exposure conditions/models* (Aronow and Samii, 2017). The exposure model can then be coupled with a difference-in-means (DM) or inverse propensity weighted (IPW) estimator which uses only those units belonging to the contrast of interest. Exposure models are more often imposed than learned, giving rise to the problem of model misspecification. Running a clustered design, with clusters chosen according to the structure of the network (Ugander and Backstrom, 2013; Ugander et al., 2013), can be effective in increasing the number of units available for estimation (Eckles et al., 2017; Pouget-Abadie et al., 2017).

More details on the exposure modeling approach is provided in the introduction to Chapter 3 with specifics on inverse propensity weighted Hájek estimators given in Section 3.5 of that chapter.

Example 1.4 (Budget and resource constraints)**.** Interference also arises in arenas like game theory, mechanism design, and market design, whenever actors play games with a finite amount of resources. Ridesharing apps such as Lyft and Uber and marketplace apps such as AirBnB must deal with these issues carefully when structuring and analyzing A/B tests. For example, if Lyft makes a drastic change to its incentive/pricing structure for a large segment of drivers, then via perturbing the ridership

market it may also affect other drivers in ways not predicted or intended. (This is an additional concern distinct from the standard social interference induced by the bipartite driver-rider and renter-rentee networks, which is also present. For example, changing the experience of a Lyft driver will likely also change the experience of all of her riders.) Certain randomization strategies may be useful for reducing bias due to interference in large-scale advertising exchange experiments (Basse, Soufiani, and Lambert, 2016). As another example, one's success in applying for academic jobs in statistics will depend on the quantity and quality of that candidate's peers also applying for academic jobs that year, since there are so few faculty positions available.

Example 1.5 (Epidemiology, disease, and contagion)**.** Quite similar to interference for social spillovers, there is a large literature on interference in epidemiological contexts. The presence of interference in such contexts is clear: supposing that half of the individuals are vaccinated and half are not, then we would expect much more than half of the individuals to be protected from disease provided that the disease spreads socially (i.e. is contagious) and that the vaccine is effective. Perez-Heydrich et al. (2014) show that the incidence rate of cholera in a study in Bangladesh was lower among *unvaccinated* individuals living in neighborhoods with *high* vaccination rates than among *unvaccinated* individuals living in neighborhoods with *low* vaccination rates. See also Halloran and Hudgens (2016) for a methodological review of interference with a particular focus on applications to epidemiology.

Example 1.6 (Political messaging)**.** Measuring the effectiveness of political messaging or get-out-the-vote efforts is important for political campaigns deciding where to spend their money. This task is difficult because of the strong peer effects; individuals may be more easily persuaded by posts from trustworthy friends than ads/messages targeted to them directly. A distinct but related issue is social pressure in voting. Gerber et al. (2008) in the 2006 Michigan primary election in which they randomized 100,000 households to receive different mailings. One group, the Self mailer, listed the voting history of the household members along with a message exclaiming that voting records are public; a second group received the Neighbors mailer which included voting records of neighbors as well. These mailings led to increased turnout of the Self

mailer by 4.8 percentage points and the Neighbors mailer by 8.1 percentage points, which are both extroardinarily large effects. Generally treatment effects were positively correlated with greater amounts of social pressure. Numerous follow-up and replication studies as well as related experiments have been conducted (e.g. Gerber et al., 2010; Bond et al., 2012; Panagopoulos, 2013; Rogers et al., 2017)

1.2.2 Targets of estimation

The standard SUTVA intent-to-treat average treatment effect (equation (1.3) or (1.4)) can be generalized in many different ways in the presence of interference, and as a result the first challenge in conducting inference in this setting is to carefully specify and clarify the desired target estimand.

First, the analyst may want to estimate the difference in the average response when units are exposed to one policy $\mathbf{w} \in \mathcal{W}^n$ to the average response when units are exposed to another policy $\mathbf{w}' \in \mathcal{W}^n$. For example, the goal may be to compare the status quo policy to the counterfactual where all units receive the intervention or product rollout, in which case $\mathbf{w} = \mathbf{1}$, the vector of *global treatment exposure*, and $\mathbf{w}' = \mathbf{0}$, the vector of *global control exposure*. The treatment effect of interest is

$$\tau_{\text{GATE}} = \frac{1}{n} \sum_{i=1}^{n} [y_i(\mathbf{1}) - y_i(\mathbf{0})]$$

if the potential outcomes are fixed and

$$\tau_{\text{GATE}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}[Y_i(\mathbf{1}) - Y_i(\mathbf{0})]$$

if the potential outcomes are stochastic. This is called the *total average treatment effect* (TATE) (Hudgens and Halloran, 2008) or *global average treatment effect* (GATE) (Eckles et al., 2017; Chin, 2018b).

Other times, it is of scientific interest to decompose the global effect into its component parts and instead estimate version of the *direct effect* and *indirect effect*. Hudgens and Halloran (2008) define the direct effect as an individual's response to

their own treatment and the indirect effect as an individual's response to others' treatments (interference). Indirect effects are also often called spillover effects, network effects, dependent happenings (especially in epidemiology), or peer effects or social contagion (especially in the social sciences).

By partitioning the argument of the treatment response function $Y_i(\mathbf{W})$, we can more carefully understand what is meant by direct and indirect effects. Let $\mathbf{W}_{-i} = (W_1, \ldots, W_{i-1}, W_{i+1}, \ldots, W_n) \in \mathcal{W}^{n-1}$ be the vector of *indirect treatments*, consisting of all treatments except that of unit i. The *direct treatment* or *ego treatment* is W_i. Now, let us write $Y_i(W_i, \mathbf{W}_{-i})$ rather than $Y_i(\mathbf{W})$. Sävje et al. (2017) distinguishes between several careful definitions of the direct effect. The *expected average treatment effect* (EATE) is

$$\tau_{\text{EATE}} = \frac{1}{n} \sum_{i=1}^{n} \Big(\mathbf{E}[Y_i(1, \mathbf{W}_{-i})] - \mathbf{E}[Y_i(0, \mathbf{W}_{-i})] \Big)$$

and the *average distributional shift effect* (ADSE) is

$$\tau_{\text{ADSE}} = \frac{1}{n} \sum_{i=1}^{n} \Big(\mathbf{E}[Y_i(1, \mathbf{W}_{-i}) | W_i = 1] - \mathbf{E}[Y_i(0, \mathbf{W}_{-i}) | W_i = 0] \Big).$$

The EATE marginalizes over the distribution of indirect treatments $\mathbf{W}_{-i}$, and the ADSE marginalizes over the conditional distribution of $\mathbf{W}_{-i} | W_i = w$ for $w = 0, 1$. Note that this distinction is only necessary when the treatments are not assigned independently; when $W_i \perp\!\!\!\perp W_j$, then $\tau_{\text{EATE}} = \tau_{\text{ADSE}}$. Because we mostly work with independent designs in this thesis, this distinction is not consequential for our purposes, but is still important to note.

An alternative view, which I find illuminating, is to define the conditional random variables

$$Y_i(0) = Y_i | (W_i = 0), \qquad Y_i(1) = Y_i | (W_i = 1).$$

(By this, I mean that $C = A|(B = b)$ is the random variable such that $\mathbf{P}(C = c) = \mathbf{P}(A = c|B = b)$.) $Y_i(0)$ and $Y_i(1)$ are then stochastic potential outcomes which are equal to the SUTVA potential outcomes when there is no interference.

Then the ADSE above can be written

$$\tau_{\text{ADSE}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left[Y_i(1) - Y_i(0)\right].$$

This differs from equation (1.4) only in the sources of variance. This means that the variance admits an ANOVA-style decomposition into the *variance contributed by the direct effect* and the *variance contributed by the indirect effect or interference*. More details are provided in Chapter 2.

Finally, Choi (2018) proposes an estimand called the *contrast attributable to treatment* (CAT), defined as follows. Let

$$\Delta_Y = \frac{1}{N_1}\sum_{i=1}^{n} W_i Y_i - \frac{1}{N_0}\sum_{i=1}^{n}(1 - W_i)Y_i$$

be the ordinary *observed* difference-in-means and let

$$\Delta_\xi = \frac{1}{N_1}\sum_{i=1}^{n} W_i Y_i(\mathbf{0}) - \frac{1}{N_0}\sum_{i=1}^{n}(1 - W_i)Y_i(\mathbf{0})$$

be the same contrast under the counterfactual of total control. Then define

$$\tau_{\text{CAT}} = \Delta_Y - \Delta_\xi.$$

A positive value of τ_{CAT} means that the treatment caused the treated units to have higher outcomes than the control units. (Note that τ_{CAT} is an estimand whose value is conditional on the particular sampled randomization of units to treatment and control.) Because $\mathbf{E}\Delta_\xi$ is 0, τ_{CAT} can be viewed as another type of direct effect estimand, and can be estimated simply using Δ_Y. The main strength is in variance estimation; Choi (2018) establishes a variance estimator that is conservative for the true variance *no matter the structure of interference*. This means that we can always guarantee valid inference for τ_{CAT}, which is an attractive property in regimes of complicated, unknown interference, even though it is a less interpretable estimand than the SUTVA ATE.

1.2.3 Hypothesis testing

Assumption-light methodology like inference based on Choi (2018)'s τ_{CAT} is particularly valuable in social settings, in which interactions or interference make it difficult to make independence or other simplifying distributional assumptions about the behavior of individual units. In this sense, it is similar to the recent proposals for using Fisher's randomization inference (Fisher, 1925, 1935) to obtain exact tests for interference. Although Fisher's null implies Neyman's null, Fisher's null is not necessarily rejected whenever Neyman's null is, and Fisher randomization tests may often be less powerful (Ding, 2017). Athey and Imbens (2017) also recommend that people use randomization-based inference rather than sampling-based inference even under no interference. The papers (Aronow, 2012; Athey, Eckles, and Imbens, 2017a; Basse, Feller, and Toulis, 2017) extend Fisher's randomization inference to the case of non-sharp null hypotheses, which encompasses a wide range of hypotheses of interest including those for higher-order spillovers, network sparsification, and heterogeneity of interference. The basic approach is to partition the population into individuals upon which the hypothesis is being applied (the *focal units*), and individuals which are used to induce variation in the treatment (the *non-focal units*); conditional inference is then conducted.

A separate vein of literature aims to develop inferential procedures which avoid evaluation of the likelihood function. For example, even a Neyman-Pearson likelihood ratio test (Neyman and Pearson, 1933) of simple hypotheses is difficult to evaluate if the likelihoods cannot easily be calculated. Cranmer, Pavez, and Louppe (2015) show that one can instead train a calibrated discriminative classifier to distinguish between samples from the null and alternative hypotheses. It is likely that similar approaches, which rely only on samples drawn from generative models or "simulators," may be highly productive in the interference setting for generating "automated," assumption-light tests.

1.3 Contributions

The remaining contents of this thesis are divided into three sections, which are each based on a separate manuscript. Each addresses a different aspect of experimentation in the presence of interference. The first two manuscripts are entirely my own work and the third is a collaboration among myself, Dean Eckles and Johan Ugander.

Chapter 2 is based on Chin (2018a) and concerns the problem of studying the robustness of SUTVA-based estimators in the presence of interference. By this I mean the consideration of estimators that are optimal and/or typically used in the case that SUTVA holds, but studied in a regime of (usually mild or weak) interference. The problem space is thus viewed as a robustness or misspecification regime. This chapter develops methods that use Stein's method for obtaining central limit theorems of estimators in such settings. It shows that standard estimators remain asymptotically normal under mild forms of interference, and furthermore sheds light on how to adjust SUTVA-based standard errors to be robust to interference.

Chapter 3 is based on Chin (2018b) and develops new methods for estimating the global average treatment effect. Existing proposed methods for this task generally revolve around the idea of redefining potential outcomes (via an exposure model) in order to appropriately capture the pattern of interference, and then applying a standard nonparametric treatment effect estimator to those redefined counterfactuals. Such an approach has a number of practical and statistical drawbacks. Instead, my work reframes the problem as a bias reduction problem in which interference contributes bias through certain "confounders." Given access to those confounders, variations of typical regression adjustment estimators from the observational causal inference literature can then be applied.

Chapter 4 is based on Chin, Eckles, and Ugander (2018). Here we study a different flavor of experiment that is designed to empirically evaluate various *stochastic seeding strategies*, which are randomized policies about where in a social network to seed a behavior in order to maximize adoption. Interference and spillover effects, then, do not take center stage but are implicitly crucial for the underlying spreading behaviors. Analysts have heretofore designed and implemented large, expensive field

experiments for testing these strategies that have yielded quite imprecise estimates. We develop new methodology that borrows ideas from the importance sampling and counterfactual policy evaluation literatures. We show that stochastic seeding strategies can be analyzed more efficiently in such experiments, how they can be evaluated "off-policy" using existing data taken from experiments designed for other purposes, and how to design much more efficient experiments.

Chapter 2

Stein's method for interference

2.1 Introduction

In this chapter we add to a budding literature that seeks to characterize the statistical properties of certain estimators under regimes of interference. In particular, we focus on estimators that might be used *if a researcher believed that SUTVA really holds*. In other words, the goal is to work within a setting of misspecification, robustness, or sensitivity analysis.

It is not difficult to see why such tools are valuable; the existing methods discussed in the previous chapter can be expensive, in both a computational and statistical sense. In online settings, in which a standard experimental platform has been operationalized, two-stage or clustered designs may be edge use cases and so may require significant engineering effort to set up. Such experiments also sacrifice statistical power if it turns out the interference was weak or non-existent. Therefore, it is of interest to practitioners to be able to tell when controlling for interference is necessary, and when it is appropriate to use standard estimators constructed under the no-interference assumption. This is especially pertinent in the case of general interference, when there may be no clearly observable structures in the data to indicate whether interference is present.

One option is to develop hypothesis tests for testing for spillover or interference effects (Aronow, 2012; Athey et al., 2017a; Basse et al., 2017). Another study that

attempts to move the literature in this direction, and the one that is most relevant to the present work, is that of Sävje et al. (2017). In that paper, the authors develop a framework for studying the behavior of standard estimators under a weak form of interference. They characterize interference based on the notion of a *interference dependence graph*, which defines an edge between two units i and j if there exists some unit k (which is possibly i or j) whose treatment affects the responses of both i and j. They establish consistency results for various estimators and experimental designs under the restriction that the average degree of the interference dependence graph grows at rate $o(n)$.

Beyond consistency, it is desirable to know whether estimators satisfy a central limit theorem so that valid asymptotic inference can be performed. This chapter makes two contributions toward this goal. First, we demonstrate that the interference dependence graph of Sävje et al. (2017) is equivalent to the *dependency graph* introduced by Chen (1975), used in a variant of Stein's method for bounding distances between random variables. We show that in a Bernoulli randomized experiment, a central limit theorem exists for the Horvitz-Thompson version of the difference-in-means estimator if one is willing to constrain the maximal degree of the dependency graph at rate $o(n^{1/4})$, rather than the average degree at rate $o(n)$.

In practice the dependency graph may be quite dense, and may even have edges between every single pair of units in the population. As an example of how this may occur, consider the time-dynamic model studied in Eckles et al. (2017) for experiments conducted on a social network. In this model, similar in spirit to the linear-in-means model of Manski (1993) for capturing endogenous social effects, individuals observe the responses of other individuals and use that information to inform their actions in the following time period. Interference thus spreads through the network over time and, provided the network is connected, eventually creates long-range dependencies between all pairs of nodes. Therefore any local model of interference, such as the neighborhood treatment response condition, does not apply. Our second contribution is to propose a notion of *approximate local interference* that allows for long-range dependencies. We use a more general form of Stein's method to prove a central limit theorem in a setting where all units may interfere with all other units, but "strong

interference" is contained to neighborhoods of size $o(n^{1/3})$.

We find that the asymptotic variance of the difference-in-means estimator can be decomposed into two pieces: (a) the variance that results from conditioning on the standard potential outcomes $Y_i^{(0)}$ and $Y_i^{(1)}$, which would have been the true variance under SUTVA; and (b) the additional variation of $Y_i^{(0)}$ and $Y_i^{(1)}$ resulting from interference. If the additional variation due to interference is sufficiently large then standard confidence intervals may be anticonservative. The variance decomposition suggests that a conservative variance estimator may be constructed if the additional variation from interference can be estimated; in this work we show that this is indeed the case under the restricted interference conditions discussed above. Adding this additional term to a standard SUTVA variance estimator, such as the Neyman conservative variance estimator, yields confidence intervals that are robust to interference.

Our technical results rely heavily on Stein's method, a flexible family of approaches for bounding distances between functions of random variables. As such, it can be used for proving central limit theorems when it is difficult or impossible to make stronger assumptions such as independence or the existence of identically distributed random variables. Stein's method develops from the seminal paper (Stein, 1972), which provides a bound for the error in the normal approximation of a sum of random variables with a certain dependency structure. We provide a short summary of the relevant literature here, but for a longer exposition on the historical development of Stein's method we refer the reader to the surveys found in Ross (2011) and Chatterjee (2014).

The theory of dependency graphs, as a particular version of Stein's method, was developed in (Chen, 1975; Stein, 1986; Baldi and Rinott, 1989; Chen and Shao, 2004), and is used for establishing limit theorems when dependence is exactly contained within a small neighborhood of variables. The dependency graph method is also similar in spirit to the idea of m-dependence for sequences of random variables; see for example (Hoeffding and Robbins, 1948; Berk, 1973; Romano and Wolf, 2000). Section 2 of Chatterjee (2014) summarizes the main idea of the dependency graph approach.

Classical versions of Stein's method have the property that the random variables need to satisfy some "nice" condition—in the case of dependency graphs, that the degree is limited. The papers Chatterjee (2008, 2009) develop a more general version of Stein's method that Chatterjee (2014) calls the *generalized perturbative method*. This approach formalizes the idea that exact independence is really not too different from approximate independence when it comes to establishing limiting distributional results. Using this technology we are able to show that asymptotic normality still holds when there exists a weak form of long-range dependencies, even if the induced dependency graph is dense. In short, if units technically share a dependency edge but this dependence is sufficiently weak, then we may view them as essentially independent of each other.

Our results are of primary interest to two audiences. First, for practitioners, we contribute to a growing characterization in the literature of understanding when interference is a practical concern and when specialized estimators and robust confidence intervals are needed. Second, for researchers seeking to establish technical results for causal estimators under interference, our work demonstrates how Stein's method can be a useful machinery for handling the complicated dependencies that often appear among statistical objects in interference problems.

2.2 Setup

As described in Chapter 1, we work within the potential outcomes framework, or Rubin causal model (Neyman, 1923; Rubin, 1974). Consider a population of n units indexed on the set $[n] = \{1, \ldots, n\}$ and let $\mathbf{W} = (W_1, \ldots, W_n) \in \{0,1\}^n$ be a random vector of binary treatments. For every individual i and realized vector of treatments $\mathbf{w} \in \{0,1\}^n$, we posit the existence of a fixed potential outcome $Y_i(\mathbf{w})$. Note that the potential outcomes are functions of the entire treatment vector and not just the treatment of unit i. We make no parametric restrictions on the form of the potential outcomes.

It is also helpful to consider an alternative parametrization of the potential outcomes which clarifies the direct and indirect effects. Let $\mathbf{W}_{-i}$ denote the vector of

$n-1$ elements obtained by removing the i-th element from $\mathbf{W}$, and partition the vector $\mathbf{W}$ into the *direct* or *ego treatment* W_i and the *indirect treatment* $\mathbf{W}_{-i}$. Then we may index the potential outcomes by these two arguments, writing $Y_i(w_i, \mathbf{w}_{-i})$ instead of $Y_i(\mathbf{w})$.

Our estimand of interest will be a version of the direct effect. In order to formally define this estimand, we first define a random version of the SUTVA potential outcomes. Let $Y_i^{(0)}$ and $Y_i^{(1)}$ be the random variables defined as follows:

$$Y_i^{(0)} = Y_i^{(0)}(\mathbf{W}_{-i}) = Y_i(0, \mathbf{W}_{-i}) \tag{2.1}$$

$$Y_i^{(1)} = Y_i^{(1)}(\mathbf{W}_{-i}) = Y_i(1, \mathbf{W}_{-i}). \tag{2.2}$$

For $w = 0, 1$, the quantity $Y_i^{(w)}$ represents the potential outcome under the scenario in which unit i is exposed to the treatment condition $W_i = w$. We have used this notation because $Y_i^{(w)}$ are the random extension of the SUTVA potential outcomes in the presence of interference, and their values may vary depending on the treatment assignments assigned to the other units. For the rest of this chapter we will suppress the explicit dependence on $\mathbf{W}_{-i}$, but the reader should keep in mind that randomness in $Y_i^{(w)}$ results entirely from randomness in $\mathbf{W}_{-i}$. If the no-interference assumption is true, then conditioning on W_i removes all randomness in $Y_i(\mathbf{W})$, and so $Y_i^{(w)}$ are degenerate random variables and hence reduce to the standard (fixed) potential outcomes.

The conceptual advantage of viewing the potential outcomes in this way is that we can get a handle on the variation that exists before and after conditioning on the direct effect. A situation in which such conditioning removes most of the variance can be viewed as a scenario in which "SUTVA approximately holds," even if strictly speaking SUTVA is violated.

If SUTVA fails to hold, the standard average treatment effect is undefined. We follow Sävje et al. (2017) and first define the *assignment-conditional average treatment effect*

$$\tau_{\text{ACATE}}(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^{n} [Y_i(1, \mathbf{W}_{-i}) - Y_i(0, \mathbf{W}_{-i})]$$

The average effects $\tau_{\text{ACATE}}(\mathbf{W})$ are well-defined but uninterpretable and unwieldy; a seperate estimand exists for each assignment vector. Instead we focus on studying the *expected average treatment effect* (EATE)

$$\tau = \mathbf{E}[\tau_{\text{ACATE}}(\mathbf{W})],$$

where the expectation is taken over the experimental design. The EATE τ is a natural relaxation of the standard average treatment effect in the sense that they coincide whenever SUTVA holds. As noted by Sävje et al. (2017), the EATE is the expected change of changing a random unit's treatment in the current experiment.[1] It may be viewed as an expected direct effect, where the marginalization is taken over the indirect treatment assignments.

Using definitions (2.1) and (2.2), we see that τ can also be written

$$\tau = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[Y_i^{(1)} - Y_i^{(0)}], \tag{2.3}$$

In other words, we marginalize the difference of the random variables $Y_i^{(1)}$ and $Y_i^{(0)}$ both over the finite population of n units and over any randomness that is contributed by the indirect effect.

Regardless of whether or not the no-interference assumption holds, one of $Y_i^{(0)}$ and $Y_i^{(1)}$ is still unobserved. Let $Y_i = Y_i(\mathbf{W}) = W_i Y_i^{(1)} + (1-W_i)Y_i^{(0)}$ denote the observed outcome. Let $N_1 = \sum_{i=1}^{n} W_i$ and $N_0 = \sum_{i=1}^{n}(1-W_i)$ denote the within-group sample sizes. We study the behavior of the difference-in-means estimator

$$\hat{\tau} = \frac{1}{N_1}\sum_{i=1}^{n} W_i Y_i - \frac{1}{N_0}\sum_{i=1}^{n}(1-W_i)Y_i. \tag{2.4}$$

Note that this estimator is well-defined even when SUTVA is violated, as Y_i is simply

[1] Sävje et al. (2017) also consider another version of a direct effect called the *average distributional shift effect* (ADSE) In this work we only consider Bernoulli i.i.d. designs, in which case the EATE and the ADSE are the same estimand. However they are not equal to each other in general. For a further discussion of estimands under interference, the reader is directed to Section 3 of Sävje et al. (2017).

the observed outcome. Sävje et al. (2017) study a wider class of estimators, namely the design-based Horvitz-Thompson and Hájek estimators that are typically used when $\mathbf{P}(W_i = 1)$ varies with i (Horvitz and Thompson, 1952; Hájek, 1971). The difference-in-means estimator is a special case of the Hájek estimator, coinciding when the assignment probabilities are the same for every unit. For simplicity of exposition our analysis focuses on the difference-in-means estimator and an experimental design in which the assignment probabilities are constant across units. We briefly discuss generalizations at the end of this chapter.

In order to obtain asymptotic results, we follow the standard finite population regime (Freedman, 2008a,b; Lin, 2013; Abadie et al., 2017b; Sävje et al., 2017) in which we have access to a sequence of finite populations indexed by size n. Each population is comprised of its own treatments and outcomes and W_i and Y_i now represent triangular arrays of random variables. We shall largely keep the indexing on n implicit to avoid notational clutter, except when charification is helpful, such as for sequences of dependency graphs. The only randomness within each population is induced by the treatment assignment vector $\mathbf{W}$. Our goal is to study the limiting behavior of the sequence $\hat{\tau} - \tau$, subject to appropriate scaling.

Throughout this chapter we will make use of the following regularity conditions. The first two conditions, probabilistic assignment and uniformly bounded fourth moments, are standard regularity conditions for asymptotic analysis of regression estimators of treatment effects.

Assumption 2.1 (Bernoulli design and probabilistic assignment)**.** $\mathbf{P}(W_i = 1) = \pi$ *independently, where the treatment proportion* π *is bounded away from* 0 *and* 1*.*

Assumption 2.1 can be relaxed to allow different assignment probabilities per unit, $\mathbf{P}(W_i = 1) = \pi_i$, at the cost of more complicated calculations.

Assumption 2.2 (Bounded fourth moments)**.** $\mathbf{E}[|Y_i|^k]$ *are uniformly bounded by a constant for all* i, n *and all* $k \leq 4$*.*

Because we work only with Bernoulli randomized experiments as specified by Assumption 2.1, Assumption 2.2 is equivalent to a uniform bound placed on the

potential outcomes $|Y_i(\mathbf{w})^k|$ for all $i, n, \mathbf{w}$ and $k \leq 4$. This equivalence does not hold for arbitrary designs, and we state Assumption 2.2 in the manner above so as to mimic the form of Sävje et al. (2017)'s Assumption 1B.

We also assume existence of the following limits of the potential outcome moments.

Assumption 2.3 (Existence of limits)**.** *Let*

$$\bar{Y}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} Y_i^{(1)} \qquad \textit{and} \qquad \bar{Y}^{(0)} = \frac{1}{n}\sum_{i=1}^{n} Y_i^{(0)}.$$

The following limits exist:

$$\begin{aligned}
\sigma_1^2 &:= \lim_{n\to\infty} \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i^{(1)} - \bar{Y}^{(1)})^2\right] \\
\sigma_0^2 &:= \lim_{n\to\infty} \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i^{(0)} - \bar{Y}^{(0)})^2\right] \\
\sigma_{01} &:= \lim_{n\to\infty} \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i^{(1)} - \bar{Y}^{(1)})(Y_i^{(0)} - \bar{Y}^{(0)})\right] \\
\sigma_\tau^2 &:= \lim_{n\to\infty} n \operatorname{Var}\left[\bar{Y}^{(1)} - \bar{Y}^{(0)}\right].
\end{aligned} \tag{2.5}$$

The quantities inside the expectation are "population" quantities in the sense that they do not involve the treatment assignment. Because of interference they may be random, which is why the expectation is needed. A consequential implication of the assumption that σ_τ^2 exists (equation (2.5)) is that the population difference of means $\bar{Y}^{(1)} - \bar{Y}^{(0)}$ is consistent at a $n^{1/2}$ rate of convergence. It is possible that limiting results are still achievable even when the difference of means converges at a slower rate, but we do not address this case in this chapter.

2.3 A dependency graph central limit theorem

In order for central limit theorems to exist for $(\hat{\tau}_n - \tau_n)$, the observed outcomes Y_i need to be "sufficiently independent." One way to enforce this constraint is to directly

require that enough pairs of units are completely independent. This idea is formalized via the following definition.

Definition 2.1. Let $\{X_i\}_{i=1}^n$ be a collection of random variables on the nodes $[n]$ of a graph D. Then D is a *dependency graph* if for any two disjoint sets of nodes $A, B \subset [n]$ such that no edge in D crosses between A and B, the sets $\{X_i\}_{i\in A}$ and $\{X_i\}_{i\in B}$ are independent.

The method of dependency graphs is a classical way of characterizing dependence in collections of random variables; see for example Baldi and Rinott (1989). Dependency graphs are not necessarily unique; the complete graph always satisfies Definition 2.1, for example. In this chapter we work with the dependency graph that is minimal in the sense that it has the fewest number of edges satisfying the definition.

In order to characterize interference between units, we consider dependency graphs on the collection of observed outcomes. Let D denote the dependency graph on the set of random variables $\{Y_i\}_{i=1}^n$. In this case we see that the minimal dependency graph corresponds exactly to the notion of interference dependence considered in Sävje et al. (2017), via the edge definition in Definition 5 of that paper. They define the interference dependence graph to have edges

$$D_{ij} = \begin{cases} 1 & \text{if } I_{\ell i} I_{\ell j} \text{ for some } \ell \in [n], \\ 0 & \text{otherwise,} \end{cases}$$

where

$$I_{ij} = \begin{cases} 1 & \text{if } Y_j(\mathbf{w}) \neq Y_j(\mathbf{w}') \text{ for some } \mathbf{w}, \mathbf{w}' \text{ such that } \mathbf{w}_{-i} = \mathbf{w}'_{-i}, \\ 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

In other words, because the outcomes are defined as functions of the treatment vector, units i and j are connected in this minimal dependency graph if and only if (a) the treatment of i affects the response of j, (b) the treatment of j affects the response of i, or (c) the responses of both i and j are affected by the treatment of some third

unit.

Importantly, the dependency graph is *not* the same as the social network or other network structure in which the units may be posited to interact. The dependency graph simply captures the structure of interference and does not specify the source of that interference. Furthermore, if the interference is actually induced by a social network G, then the dependency graph also depends on the process giving rise to interference. For example, if the outcome-generating process is such that only neighboring units interfere with each other, then D does have the same edge structure as G. But if interference results from a contagion process spreading over the entire network, then D may be fully-connected even if G is sparse. Throughout this work, we use D to denote the dependency graph and reserve G to denote a social network, when we need to refer to it.

Given a dependency graph defined on a collection of random variables, we can take advantage of bounds from the literature on Stein's method. Such bounds characterize the Wasserstein distance between sums of random variables and a Gaussian random variable. Recall that the Wasserstein metric between probability measures μ and ν is

$$d_{\mathcal{W}}(\mu,\nu) = \sup\left\{\left|\int h(x)d\mu(x) - \int h(x)d\nu(x)\right| : h \text{ is 1-Lipschitz}\right\},$$

where a function h is 1-Lipschitz if it satisfies $|h(x) - h(y)| \leq |x - y|$. In this chapter we are concerned only with controlling the Wasserstein distance between μ and a standard Gaussian random variable. For any random variable S, denote the distance to Gaussianity as

$$d_{\mathcal{W}}(S) = d_{\mathcal{W}}(\mu,\nu),$$

where μ is the law of S and ν is the law of a standard Gaussian random variable, having density

$$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

We rely on the following dependency graph bound, which we state as a lemma.

Lemma 2.1 (Ross (2011), Theorem 3.6)**.** *Let $X_1, \ldots, X_n$ be a collection of random variables such that $\mathbf{E}[X_i^4] < \infty$ and $\mathbf{E}[X_i] = 0$. Let $\sigma^2 = \mathrm{Var}(\sum_i X_i)$ and $S = \sum_i X_i$.*

Let d be the maximal degree of the dependency graph of $(X_1, \ldots, X_n)$. Then for constants C_1 and C_2 which do not depend on n, d or σ^2,

$$d_{\mathcal{W}}(S/\sigma) \leq C_1 \frac{d^{3/2}}{\sigma^2} \left(\sum_{i=1}^{n} \mathbf{E}[X_i^4] \right)^{1/2} + C_2 \frac{d^2}{\sigma^3} \sum_{i=1}^{n} \mathbf{E}|X_i|^3. \tag{2.6}$$

From here, we see that one can define an appropriate choice of X_i such that S is the desired treatment effect estimator, and then provide conditions so that the right-hand side converges to zero. However, a caveat is that the summand in the difference-in-means estimator $\hat{\tau}$ depends on the random sample sizes N_0 and N_1, which depend on the treatment assignments of all n units. Therefore, the dependency graph on $\{X_i\}_{i=1}^n$ unfortunately is complete (fully connected), even if the dependency graph on $\{Y_i\}_{i=1}^n$ is not, and Lemma 2.1 is not applicable.

As a result, in this section we restrict ourselves to studying a modified form of the difference-in-means estimator, defined by

$$\tilde{\tau} = \sum_{i=1}^{n} \left[\frac{W_i Y_i}{n\pi} - \frac{(1 - W_i) Y_i}{n(1-\pi)} \right]. \tag{2.7}$$

The estimator $\tilde{\tau}$ is a Horvitz-Thompson (Horvitz and Thompson, 1952) variant of the difference-in-means estimator $\hat{\tau}$, and uses the population sample sizes $n\pi$ and $n(1-\pi)$ in the denominator in place of the empirical sample sizes N_1 and N_0. Though there is little advantage to using $\tilde{\tau}$ over $\hat{\tau}$ in practice, it is still instructive for seeing how the dependency graph method works. Results are provided for the difference-in-means estimator $\hat{\tau}$ in Section 2.4.

We now define a limited interference condition that constrains the structure of the dependency graph. The metric that we use to measure the extent of interference for a collection of random variables is the maximal degree of the dependency graph. Let D_n denote the sequence of dependency graphs and d_n denote the corresponding maximal degrees. We make the following interference assumption about the limiting behavior of d_n.

Assumption 2.4 (Local interference)**.** $d_n = o(n^{1/4})$.

This assumption is a local interference assumption in the sense that it requires all interference for a given unit to come from a small number of other units. This assumption would hold, if for example, units on a social network G only interfered with neighboring units, and G itself had maximal degree $o(n^{1/4})$. For comparison, consider the restricted interference assumption (Sävje et al., 2017, Assumption 2), which requires the average degree of the dependency graph to be of order $o(n)$. Our Assumption 2.4 is stronger, but it still allows the amount of interference to grow with n. By restricting the maximal degree rather than the average degree, we can apply Lemma 2.1.

Under the notion of local dependence defined in Assumption 2.4, we obtain the following asymptotic normality result for the Horvitz-Thompson estimator.

Theorem 2.1. *Let τ and $\tilde{\tau}$ be defined as in equations* (2.3) *and* (4.10)*. Under regularity conditions (Assumptions 2.1-2.3) and the restricted dependency degree condition (Assumption 2.4), $\sqrt{n}(\tilde{\tau} - \tau)$ is asymptotically Gaussian:*

$$\sqrt{n}(\tilde{\tau} - \tau) \Rightarrow N(0, \sigma^2),$$

where

$$\sigma^2 = \lim_{n\to\infty} n \operatorname{Var}(\tilde{\tau}).$$

We arrive at Theorem 2.1 by defining an appropriate choice for X_i and evaluating the variance σ^2, which allows us to control the bound in Lemma 2.1.

Proof. First, by conditioning on the σ-field defined in equation (2.11),

$$n \operatorname{Var}(\tilde{\tau}) = n\mathbf{E}[\operatorname{Var}(\tilde{\tau}|\mathcal{F})] + \operatorname{Var}[\mathbf{E}(\tilde{\tau}|\mathcal{F})]$$

The term $\operatorname{Var}(\tilde{\tau}|\mathcal{F})$ is the variance of the Horvitz-Thompson estimator under SUTVA, which scales at rate $n^{-1/2}$. It is written in terms of second moments of the outcomes $Y_i^{(1)}$ and $Y_i^{(0)}$, so the term $n\mathbf{E}[\operatorname{Var}(\tilde{\tau}|\mathcal{F})]$ stabilizes by Assumption 2.3. The second term is equal to $\operatorname{Var}(\bar{Y}^{(1)} - \bar{Y}^{(0)}$ which also stabilizes by Assumption 2.3. Therefore the entire variance $n \operatorname{Var}(\tilde{\tau})$ converges to a non-zero limit σ^2.

Now decompose the Horvitz-Thompson estimator (4.10) as $\tilde{\tau} = \sum_{i=1}^{n} \hat{\tau}_i$ where

$$\tilde{\tau}_i = \frac{1}{n}\left[\frac{W_i Y_i^{(1)}}{\pi} - \frac{(1-W_i)Y_i^{(0)}}{1-\pi}\right].$$

Let $X_i = \sqrt{n}(\tilde{\tau}_i - \mathbf{E}[\tilde{\tau}_i])$ and denote $\sigma^2 = n\,\mathrm{Var}(\tilde{\tau})$. Then

$$S = S_n = \sum_{i=1}^{n} X_i = \sqrt{n}(\tilde{\tau} - \tau)/\sigma.$$

By the uniform moment bound (Assumption 2.2), $X_i = O_p(n^{-1/2})$, so for large enough n there exist constants C_1 and C_2 such that

$$\left(\sum_{i=1}^{n} \mathbf{E}[X_i^4]\right)^{1/2} \leq C_1 n^{-1/2}$$

and

$$\sum_{i=1}^{n} \mathbf{E}|X_i|^3 \leq C_2 n^{-1/2}.$$

We can now apply Lemma 2.1, which establishes for fixed n that

$$d_{\mathcal{W}}(S_n) \leq C_1 \frac{d_n^{3/2}}{n^{1/2}} + C_2 \frac{d_n^2}{n^{1/2}},$$

where we have ignored the σ^2 term because it stabilizes. Therefore $d_{\mathcal{W}}(S_n) \to 0$ whenever $d_n = o(n^{1/4})$, which is the constraint we have placed on the dependency graph (Assumption 2.4). Hence S_n converges to a standard Gaussian random variable. □

A curious feature of Stein's method is that it allows one to make statements about the asymptotic behavior of random objects without calculating an explicit expression for the variance. Because our primary interest is not in the Horvitz-Thompson estimator $\tilde{\tau}$, we skip calculating the limiting variance σ^2, but is not hard to express it in terms of the moments defined in Assumption 2.3. For the difference-in-means estimator in Section 2.4 we provide an explicit characterization of the limiting

variance.

2.4 A central limit theorem for approximate local interference

There are two drawbacks of relying on the dependency graph approach for studying treatment effect estimators. It does not allow us to study estimators like the difference-in-means estimator that use empirical sample sizes, and it requires exact local interference (Assumption 2.4). In this section we discuss how these issues can be overcome. Rather than require most pairs of nodes to be exactly independent, we only require approximate independence, which allows long-range interference as long as it is not too strong.

It is worth explaining how this approximate independence assumption might arise in practice. Suppose we measure a social network, G, the edges of which capture the peer interactions which we believe transmit the interference mechanism. If we believe the neighborhood exposure assumptions invoked by, e.g., Ugander et al. (2013); Forastiere et al. (2016); Sussman and Airoldi (2017); Jagadeesan et al. (2017), and others, then the dependency graph methods of Section 2.3 are sufficient. However, exact local interference is insufficient for describing more complex data generating processes. A social contagion process, such as that considered by Eckles et al. (2017), leads to a fully-connected dependency graph D even if the social graph G is sparse (but connected). This discrepancy between D and G was previously discussed in Section 2.3 and may be discouraging to practitioners.

However, if peer effects dissipate over the network, we may believe that interference from long-distance units in G may be second- or lower-order effects. One may conceptualize the existence of two different dependency graphs defined on the collection of units, one capturing strong interference and the other capturing weak interference. Here we provide a result that allows the weak interference graph to be arbitrarily dense, but requires $o(n^{1/3})$ sparsity in the strong interference graph. In this case, sparsity of the social graph G would be sufficient for the limiting results

to hold. Such sparsity has been demonstrated empirically on such social networks as Facebook (Ugander et al., 2011), MSN (Leskovec and Horvitz, 2008), and a mobile phone network (Onnela et al., 2007), and suggested theoretically by Dunbar's number, a suggested sociocognitive limit in the number of possible stable social relationships (Dunbar, 1992).

The primary assumption is provided in Assumption 2.6, but we require a technical detour to describe the main approach, developed by Chatterjee (2008). Let $\mathbf{X} = (X_1, \ldots, X_n) \in \mathcal{X}^n$ be a random vector of independent random variables on a measure space $\mathcal{X}$ and let $f : \mathcal{X}^n \to \mathbb{R}$ be a scalar-valued measurable function. The objective is to bound the distance to normality of $S := f(\mathbf{X})$. To do so, we characterize the behavior of f when $\mathbf{X}$ is "perturbed" by replacing some components of $\mathbf{X}$ by independent copies. Let $\mathbf{X}' = (X_1', \ldots, X_n')$ denote an independent copy of $\mathbf{X}$. For every $A \in [n]$ let $\mathbf{X}^A$ be the vector where the entries corresponding to A are replaced by elements of $\mathbf{X}'$, defined componentwise as

$$X_i^A = \begin{cases} X_i' & \text{if } i \in A \\ X_i & \text{if } i \notin A \end{cases}.$$

Now define

$$\begin{aligned} \Delta_i f &= f(\mathbf{X}) - f(\mathbf{X}^i), \qquad && i \in [n], \\ \Delta_i f^A &= f(\mathbf{X}^A) - f(\mathbf{X}^{A \cup i}), \qquad && A \subset [n], i \notin A, \end{aligned}$$

where we have made a notational simplification by writing $\mathbf{X}^i$ instead of $\mathbf{X}^{\{i\}}$ and $\mathbf{X}^{A \cup i}$ instead of $\mathbf{X}^{A \cup \{i\}}$. The quantities $\Delta_i f$ and $\Delta_i f^A$ can be viewed as discrete derivatives, because they measure the change in the function f in response to perturbations of $\mathbf{X}$. If perturbations in $\mathbf{X}$ act upon f mostly independently by coordinate, then we expect the resulting value $f(\mathbf{X})$ to be approximately normal. We can now state the following normal approximation theorem, which is the main result in Chatterjee (2008). We state it as a lemma.

Lemma 2.2 (Chatterjee (2008), Theorem 2.2)**.** *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector of*

independent real-valued random variables, and let $S = f(\mathbf{X})$. *Suppose* $\mathbf{E}[S] = 0$ *and* $\sigma^2 := \mathbf{E}[S^2] < \infty$. *Define*

$$T = \frac{1}{2}\sum_{i=1}^{n}\sum_{A\subset[n]\setminus\{i\}}\frac{\Delta_i f \Delta_i f^A}{n\binom{n-1}{|A|}}$$

Then $\mathbf{E}T = \sigma^2$ *and*

$$d_{\mathcal{W}}(S/\sigma) \le \frac{1}{\sigma^2}[\mathrm{Var}(\mathbf{E}(T|S))]^{1/2} + \frac{1}{2\sigma^3}\sum_{i=1}^{n}\mathbf{E}[|\Delta_i f|^3].$$

It is more convenient to study a version of Lemma 2.2 that characterizes the Wasserstein distance in terms of local dependencies. This corollary is essentially a variant of Corollary 3.2 in Chatterjee (2014).

Corollary 2.1. *Let all variables be defined as in Lemma 2.2. For every* i, j, *let* $c_{i,j}$ *be a constant such that for all* $A \in [n]\setminus\{i\}$ *and* $B \in [n]\setminus\{j\}$,

$$\mathrm{Cov}(\Delta_i f \Delta_i f^A, \Delta_j f \Delta_j f^B) \le c_{i,j}.$$

Then

$$d_{\mathcal{W}}(S/\sigma) \le \frac{1}{2\sigma^2}\left(\sum_{i,j=1}^{n} c_{i,j}\right)^{1/2} + \frac{1}{2\sigma^3}\sum_{i=1}^{n}\mathbf{E}[|\Delta_i f|^3]. \tag{2.8}$$

Proof. Notice that

$$\begin{aligned}\mathrm{Var}(\mathbf{E}(T|S)) \le \mathrm{Var}\, T &\le \frac{1}{4}\sum_{i,j=1}^{n}\sum_{\substack{A\subset[n]\setminus\{i\}\\ B\subset[n]\setminus\{j\}}}\frac{\mathrm{Cov}(\Delta_i f\Delta_i f^A, \Delta_j f\Delta_j f^B)}{n^2\binom{n}{|A|}\binom{n}{|B|}}\\ &\le \frac{1}{4}\sum_{i,j=1}^{n}\sum_{\substack{A\subset[n]\setminus\{i\}\\ B\subset[n]\setminus\{j\}}}\frac{c_{i,j}}{n^2\binom{n}{|A|}\binom{n}{|B|}} = \frac{1}{4}\sum_{i,j=1}^{n} c_{i,j}.\end{aligned}$$

Applying Lemma 2.2 completes the proof. □

One may gain intuition for Corollary 2.1 by considering the case of a dependency

graph, in which an upper bound can be provided for the number of covariance terms $c_{i,j}$ that can be nonzero. Let D be a graph with maximal degree d_n and let $\mathcal{N}_i$ denote the neighborhood of unit i. Letting $\mathbf{X}$ be a collection of independent random variables, as in Lemma 2.2, consider a function of the form $S = f(\mathbf{X}) = \sum_{i=1}^{n} g_i(X_i, \mathbf{X}_{\mathcal{N}_i})$, where $\mathbf{X}_{\mathcal{N}_i}$ are the elements of $\mathbf{X}$ restricted to $\mathcal{N}_i$. Notice now that D is a dependency graph for the collection of variables $\{g_i(X_i, \mathbf{X}_{\mathcal{N}_i})\}_{i=1}^{n}$. Then the discrete derivative has the form $\Delta_i f = \sum_{r \in \mathcal{N}_i} \Delta_i X_r$, where $\Delta_i X_r = X_r - X_r^i$ is the effect on unit r of perturbing unit i. Notice that the sum in the discrete derivative is only over the r units in $\mathcal{N}_i$, because the only arguments of g_i are elements of $\mathcal{N}_i$. Now consider the covariance between the discrete derivatives for units i and j, which can be calculated as

$$\begin{aligned}
\mathrm{Cov}(\Delta_i f, \Delta_j f) &= \sum_{r \in \mathcal{N}_i} \sum_{s \in \mathcal{N}_j} \mathrm{Cov}(\Delta_i X_r, \Delta_j X_s) \\
&= \sum_{r \in \mathcal{N}_i \cap \mathcal{N}_j} \mathrm{Cov}(\Delta_i X_r, \Delta_j X_r) \\
&\leq C d_n \mathbb{1}(|\mathcal{N}_i \cap \mathcal{N}_j| > 0),
\end{aligned}$$

where C is a constant that does not depend on n or d_n.

In other words, the covariance is always of order d_n, but is exactly zero whenever the neighborhoods of i and j do not intersect. Now, for every unit i, the number of units j such that $|\mathcal{N}_i \cap \mathcal{N}_j| > 0$ is at most d_n^2. Therefore the total number of covariances that can be nonzero is nd_n^2, and so the total magnitude of those covariances is Cnd_n^3. Assuming the variance σ^2 is of order n, and note that $\Delta_i f^A$ is of order at most d_n. Then the right-hand side of equation 2.8 is of the form

$$\frac{C_1}{n}(nd_n^3)^{1/2} + \frac{C_2}{n^{3/2}} nd_n^3 = C_1 \frac{d_n^{3/2}}{n^{1/2}} + C_2 \frac{d_n^3}{n^{1/2}}.$$

This quantity can be made small in the limit if d_n grows sufficiently slowly. The first term here and the first term in the dependency graph bound (2.6) both require a $d_n = o(n^{1/3})$ constraint; the second term here requires a $d_n = o(n^{1/6})$ constraint whereas the second term of equation (2.6) requires $d_n = o(n^{1/4})$.

We return now to the problem of obtaining a limiting result for $\hat{\tau}$. Define the

sequence of functions

$$f_n(\mathbf{W}) = \sqrt{n}(\hat{\tau} - \tau) = \sqrt{n}\sum_{i=1}^{n}\left[\frac{W_i}{N_1} - \frac{1-W_i}{N_0}\right]Y_i(\mathbf{W}) - \sqrt{n}\tau.$$

Since the treatment vector $\mathbf{W}$ is comprised of independent Bernoulli(π) random variables and is the sole source of randomness in f_n, Corollary 2.1 is applicable provided we define appropriate constraints on the behavior of f_n under perturbations of the treatment vector.

Let W_i' denote an independent copy of W_i and let $\mathbf{W}^i$ the resulting treatment vector obtained by swapping out W_i for W_i' in $\mathbf{W}$, defined componentwise as

$$W_j^i = \begin{cases} W_j & \text{if } j \neq i \\ W_j' & \text{if } j = i \end{cases}.$$

Let

$$Y_j^i = Y_r(\mathbf{W}^i)$$

denote the resulting response of unit r when i is perturbed, and define

$$\Delta_i Y_r = Y_r - Y_r^i$$

to be the change in Y_r when W_i is replaced with an independent copy. Furthermore, let

$$N_1' = N_1 + W_i' - W_i$$
$$N_0' = n - N_1'$$

denote the adjusted sample sizes.

The following lemma, which we state without proof, characterizes the discrete derivative $\Delta_i f_n$.

Lemma 2.3. *Let W_i', N_1', N_0', and Y_r^i be defined as above. Then for every $i \in [n]$,*

the discrete derivative can be written as

$$\Delta_i f_n = \sqrt{n}\left(A_i + \sum_{r \neq i} B_{i,r}\right), \tag{2.9}$$

where

$$A_i = \left[\frac{W_i}{N_1} - \frac{W_i'}{N_1'}\right] Y_i^{(1)} - \left[\frac{1-W_i}{N_0} - \frac{1-W_i'}{N_0'}\right] Y_i^{(0)}$$
$$B_{i,r} = \left[\frac{W_r}{N_1} Y_r - \frac{W_r}{N_1'} Y_r^i\right] - \left[\frac{1-W_r}{N_0} Y_r - \frac{1-W_r}{N_0'} Y_r^i\right].$$

Notice that A_i describes the change for unit i (the direct effect) and $B_{i,r}$ describes the effect that perturbing the treatment of unit i has on the response of unit r.

We now describe assumptions that constrain the behavior of $\Delta_i Y_r$, which in turn allows us to handle $\Delta_i f_n$ via the expression in Lemma 2.3. Assumption 2.5 provides a set of weak regularity conditions; the main conceptual condition of approximate local interference is provided by Assumption 2.6.

Assumption 2.5 (Covariance regularity conditions)**.** *The following global covariance constraints hold:*

(a)

$$\sum_{i=1}^{n} \sum_{j=1}^{n} |\operatorname{Cov}(Y_i, Y_j)| = o(n^2).$$

(b)

$$\sum_{i=1}^{n} \sum_{r \neq i} \sum_{j \neq i} |\operatorname{Cov}(\Delta_i Y_r, Y_j)| = o(n^2).$$

(c)

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r \neq i} \sum_{\substack{s \neq j \\ s \neq r}} |\operatorname{Cov}(\Delta_i Y_r, \Delta_j Y_s)| = o(n^2).$$

Part (a) states that the overall responses are not too dependent. Part (b) states that the effect of perturbing i on r is mostly independent of the behavior of unit j.

Part (c) states that the effect of perturbing i on r and the effect of perturbing j on s are mostly independent, when r and s are distinct from i and j and each other. To see why these assumptions may be reasonable, consider the case where SUTVA holds. Then, the expressions in part (b) and (c) are also exactly zero. Finally, $\text{Cov}(Y_i, Y_j) = 0$ whenever $i \neq j$ so that the expression in part (a) is $O(n)$.

Now we describe the approximate local interference condition. For every n, let H_n denote a undirected graph on the vertices $[n]$ which represents a dependency graph for "strong interference," with weak interference allowed outside of H_n. Let the neighborhood of unit i on H_n be denoted by $\mathcal{N}_i^{H_n} = \{j \in [n] : (H_n)_{ij} = 1\}$. Assumption 2.6 formally describes the conditions required of H_n.

Assumption 2.6 (Approximate local interference)**.** *There exists a sequence of graphs* $\{H_n\}_{n=1}^{\infty}$ *having maximal degree sequence* $h_n = o(n^{1/3})$ *such that*

$$\max \left\{ \sum_{r \notin \mathcal{N}_i^{H_n}} |\Delta_i Y_r|, \sum_{r \notin \mathcal{N}_i^{H_n}} |\Delta_r Y_i| \right\} \to 0$$

almost surely for every node $i \in [n]$.

Assumption 2.6 states that outside of the strong interference neighborhoods $\mathcal{N}_i^{H_n}$, the interference has a magnitude that is vanishing in the limit. If the quantities $\sum_{r \notin \mathcal{N}_i^{H_n}} |\Delta_i Y_r|$ and $\sum_{r \notin \mathcal{N}_i^{H_n}} |\Delta_r Y_i|$ are equal to zero exactly, then H_n becomes a sparse dependency graph of the sort discussed in Section 2.3. Notably, the neighborhood size h_n is allowed to grow at rate $o(n^{1/3})$. Note that the maximum is taken over two terms; the first describes outcomes that can be affected by the treatment of unit i and the second describes treatments that can affect the outcome of unit i.[2]

We now state the main result. In comparison to Theorem 2.1, it replaces a restriction on the dependency graph, Assumption 2.4, with the approximate local interference requirements, Assumptions 2.5 and 2.6. It also is a statement about the difference-in-means estimator $\hat{\tau}$ rather than the Horvitz-Thompson estimator $\tilde{\tau}$.

[2]It is easy to generalize H_n to directed graphs, which would capture the notion that Y_i may depend on W_j without Y_j depending on W_i and vice versa. We define H_n as undirected here only for notational and conceptual simplicity.

Theorem 2.2. *Let τ and $\hat{\tau}$ be defined as in equations* (2.3) *and* (2.4)*, and assume that the regularity conditions (Assumptions 2.1-2.3 and 2.5) hold. Assume further that the outcome functions is constrained according to Assumption 2.6. Then $\sqrt{n}(\hat{\tau} - \tau)$ is asymptotically Gaussian:*

$$\sqrt{n}(\hat{\tau} - \tau) \Rightarrow N(0, \sigma^2).$$

The limiting variance σ^2 has the form

$$\sigma^2 := \lim_{n\to\infty} n \operatorname{Var}(\hat{\tau}) = \frac{1-\pi}{\pi}\sigma_1^2 + \frac{\pi}{1-\pi}\sigma_0^2 + 2\sigma_{01} + \sigma_\tau^2, \tag{2.10}$$

where the quantities σ_1^2, σ_0^2, σ_{01}, and σ_τ^2 are defined in Assumption 2.3, and $\pi = \lim_{n\to\infty} \mathbf{P}(W_i = 1)$ is the limiting treatment proportion.

Before providing the proof, we emphasize to the reader that the asymptotic variance takes the form of a variance decomposition based on conditioning on the "potential outcomes" $Y_i^{(0)}$ and $Y_i^{(1)}$ (the random quantities defined in equations (2.1) and (2.2), not the original fixed potential outcomes $Y_i(\mathbf{w})$). To see this, denote the σ-field generated by $Y_i^{(0)}$ and $Y_i^{(1)}$ as

$$\mathcal{F} := \{Y_i^{(w)} : i \in [n], w \in \{0, 1\}\}. \tag{2.11}$$

Then by the law of total variance,

$$\operatorname{Var}(\hat{\tau}) = \mathbf{E}[\operatorname{Var}(\hat{\tau}|\mathcal{F})] + \operatorname{Var}[\mathbf{E}(\hat{\tau}|\mathcal{F})].$$

This decomposition is evident in the asymptotic variance σ^2, as

$$\lim_{n\to\infty} n\mathbf{E}[\operatorname{Var}(\hat{\tau}|\mathcal{F})] = \frac{1-\pi}{\pi}\sigma_1^2 + \frac{\pi}{1-\pi}\sigma_0^2 + 2\sigma_{01} \tag{2.12}$$

and

$$\lim_{n\to\infty} n \operatorname{Var}[\mathbf{E}(\hat{\tau}|\mathcal{F})] = \sigma_\tau^2. \tag{2.13}$$

(A proof of this decomposition is provided in the supplementary material as a part of the proof of Theorem 2.) The first three terms, (2.12), form the standard asymptotic

variance of the difference-in-means estimator under no-interference, in which $Y_i^{(0)}$ and $Y_i^{(1)}$ are fixed quantities (Freedman, 2008a,b; Lin, 2013). The last term, (2.13), captures the additional variation of the "total population" average treatment effect, which is variation that remains even if we were able to observe $Y_i^{(0)}$ and $Y_i^{(1)}$ for every unit. This decomposition implies that if σ_τ^2 stabilizes to a non-zero value, then confidence intervals constructed under the SUTVA assumption will only provide correct coverage conditional on $\mathcal{F}$, and will fail to account for the fact that $Y_i^{(0)}$ and $Y_i^{(1)}$ may exhibit additional variation under the experimental design. In Section 2.5 we provide a consistent estimator of σ_τ^2 that can be used to adjust SUTVA-based standard errors.

We now turn to the proof of Theorem 2.2. This first requires stating and proving several lemmas, which focus on getting a handle on the discrete derivative. Throughout this section, $C_1, C_2, C_3, \ldots$ indicate numerical constants that do not depend on n, and their values may change from line to line.

Lemma 2.4. *Let A_i and $B_{i,r}$ be defined as in Lemma 2.3 and assume the regularity conditions (Assumptions 2.1-2.3). For all i, j, r, and s,*

$$|\operatorname{Cov}(A_i, A_j)| \leq \frac{C_1}{n^2} |\operatorname{Cov}(Y_i, Y_j)| \tag{2.14}$$

$$|\operatorname{Cov}(B_{i,r}, A_j)| \leq \left(\frac{C_1}{n^2} + \frac{C_2}{n^3}\right) |\operatorname{Cov}(\Delta_i Y_r, Y_j)| + \frac{C_3}{n^3} |\operatorname{Cov}(Y_r, Y_j)| \tag{2.15}$$

$$\begin{aligned} |\operatorname{Cov}(B_{i,r}, B_{j,s})| \leq{}& \left(\frac{C_1}{n^2} + \frac{C_2}{n^3}\right) |\operatorname{Cov}(\Delta_i Y_r, \Delta_j Y_s)| + \left(\frac{C_3}{n^3} + \frac{C_4}{n^4}\right) |\operatorname{Cov}(\Delta_i Y_r, Y_s)| \\ &+ \left(\frac{C_5}{n^3} + \frac{C_6}{n^4}\right) |\operatorname{Cov}(Y_r, \Delta_j Y_s)| + \frac{C_7}{n^4} |\operatorname{Cov}(Y_r, Y_s)|, \end{aligned} \tag{2.16}$$

where the C_k are constants, not necessarily the same from line to line.

Proof. Note that $B_{i,r}$ can be written as

$$B_{i,r} = \left[\frac{W_r}{N_1} - \frac{1 - W_r}{N_0}\right] \Delta_i Y_r + W_r Y_r^i \frac{W_i - W_i'}{N_1 N_1'} - (1 - W_r) Y_r^i \frac{W_i - W_i'}{N_0 N_0'}.$$

Equation (2.14) follows from examining the form of A_i and noting that $N_1 = O_p(n)$

and $N_0 = O_p(n)$. For equation (2.15), note

$$\begin{aligned}|\operatorname{Cov}(B_{i,r}, A_j)| &\leq \frac{C_1}{n^2}|\operatorname{Cov}(\Delta_i Y_r, Y_j)| + \frac{C_2}{n^3}|\operatorname{Cov}(Y_r^i, Y_j)| \\ &= \frac{C_1}{n^2}|\operatorname{Cov}(\Delta_i Y_r, Y_j)| + \frac{C_2}{n^3}(|\operatorname{Cov}(Y_r, Y_j)| + |\operatorname{Cov}(Y_r^i - Y_r, Y_j)|),\end{aligned}$$

which gives equation (2.15). Similarly, we have

$$\begin{aligned}|\operatorname{Cov}(B_{i,r}, B_{j,s})| \leq& \frac{C_1}{n^2}|\operatorname{Cov}(\Delta_i Y_r, \Delta_j Y_s)| + \frac{C_2}{n^3}|\operatorname{Cov}(\Delta_i Y_r, Y_s^j)| \\ &+ \frac{C_3}{n^3}|\operatorname{Cov}(Y_r^i, \Delta_j Y_s)| + \frac{C_4}{n^4}|\operatorname{Cov}(Y_r^i, Y_s^j)|,\end{aligned}$$

and rewriting $Y_r^i = Y_r - \Delta_i Y_r$ and $Y_s^j = Y_s - \Delta_i Y_s$ gives equation (2.16). □

Lemma 2.5. *Under the regularity conditions (Assumptions 2.1-2.3), there exist constants C_1 through C_5 such that*

$$\begin{aligned}\frac{1}{n}\sum_{i,j}|\operatorname{Cov}(\Delta_i f_n, \Delta_j f_n)| \leq& \frac{C_1}{n^2}\sum_{i,j}|\operatorname{Cov}(Y_i, Y_j)| + \left(\frac{C_2}{n^2} + \frac{C_3}{n^3}\right)\sum_{i,j}\sum_{r\neq i}|\operatorname{Cov}(\Delta_i Y_r, Y_j)| \\ &+ \left(\frac{C_4}{n^2} + \frac{C_5}{n^3}\right)\sum_{i,j}\sum_{r\neq i}\sum_{s\neq j}|\operatorname{Cov}(\Delta_i Y_r, \Delta_j Y_s)|.\end{aligned}$$

Proof. By expanding the form of the discrete derivative (2.9), we have

$$\begin{aligned}\frac{1}{n}|\operatorname{Cov}(\Delta_i f_n, \Delta_j f_n)| =& |\operatorname{Cov}(A_i, A_j)| + \sum_{r\neq i}|\operatorname{Cov}(B_{i,r}, A_j)| \\ &+ \sum_{s\neq j}|\operatorname{Cov}(A_i, B_{j,s})| + \sum_{r\neq i}\sum_{s\neq j}|\operatorname{Cov}(B_{i,r}, B_{j,s})|.\end{aligned}$$

By summing over i and j substituting the bounds from Lemma 2.4, the right-hand

side above is bounded above by

$$\frac{1}{n^2}\sum_{i,j}\left[C_1|\operatorname{Cov}(Y_i,Y_j)| + \sum_{r\neq i}\left[\left(C_2+\frac{C_3}{n}\right)|\operatorname{Cov}(\Delta_i Y_r, Y_j)| + \frac{C_4}{n}|\operatorname{Cov}(Y_r,Y_j)|\right]\right.$$
$$+\sum_{s\neq j}\left[\left(C_5+\frac{C_6}{n}\right)|\operatorname{Cov}(Y_i,\Delta_j Y_s)| + \frac{C_7}{n}|\operatorname{Cov}(Y_i,Y_s)|\right]$$
$$+\sum_{r\neq i}\sum_{s\neq j}\left[\left(C_8+\frac{C_9}{n}\right)|\operatorname{Cov}(\Delta_i Y_r,\Delta_j Y_s)| + \left(\frac{C_{10}}{n}+\frac{C_{11}}{n^2}\right)|\operatorname{Cov}(\Delta_i Y_r, Y_s)|\right.$$
$$\left.\left.+\left(\frac{C_{12}}{n}+\frac{C_{13}}{n^2}\right)|\operatorname{Cov}(Y_r,\Delta_j Y_s)| + \frac{C_{14}}{n^2}|\operatorname{Cov}(Y_r,Y_s)|\right]\right].$$

We now exploit the symmetry in the summations and combine terms to give the desired result. □

Lemma 2.6 (Restricted interference)**.** *Under the regularity conditions (Assumptions 2.1-2.3 and 2.5) and approximate local interference (Assumption 2.6), for every unit i, the total amount of interference that results from perturbing the treatment of unit i satisfies*

$$\sum_{r\neq i}|\Delta_i Y_r| = O_p(n^{1/3}).$$

Proof. We divide the interference emanating from unit i into collections of "weak" and "strong" interference, this partition being specified by the neighborhood $\mathcal{N}_i^{H_n}$.

$$\sum_{r\neq i}|\Delta_i Y_r| = \sum_{r\notin\mathcal{N}_i^{H_n}}|\Delta_i Y_r| + \sum_{r\in\mathcal{N}_i^{H_n}}|\Delta_i Y_r|.$$

The first summand tends to 0 as $n\to\infty$ by Assumption 2.6. The second summand is bounded above by Ch_n because of the uniform moment bound (Assumption 2.2), and the result follows since $h_n = o(n^{1/3})$. Here C is a constant, and the quantities h_n, δ_k, H_n, and $\mathcal{N}_i^{H_n}$ are as defined in Assumption 2.6. □

Lemma 2.7. *Under the regularity conditions (Assumptions 2.1-2.3 and 2.5) and*

approximate local interference (Assumption 2.6),

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{r\neq i}|\operatorname{Cov}(\Delta_i Y_r, Y_i)| = o(1).$$

Proof. By the uniform moment bound there exists a constant C such that

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{r\neq i}|\operatorname{Cov}(\Delta_i Y_r, Y_i)| \leq \frac{C}{n^2}\sum_{i=1}^{n}\sum_{r\neq i}\mathbf{E}|\Delta_i Y_r|.$$

Now using $\sum_{r\neq i}|\Delta_i Y_r| = O_p(n^{1/3})$, as established by Lemma 2.6, we find that the right hand side of the inequality above is bounded above by

$$\frac{C}{n^2}\sum_{i=1}^{n} n^{1/3} = Cn^{-2/3} = o(1).$$

□

Lemma 2.8. *Under the regularity conditions (Assumptions 2.1-2.3 and 2.5) and approximate local interference (Assumption 2.6),*

$$\frac{1}{n^2}\sum_{i,j}\sum_{\substack{r\neq i\\ r\neq j}}|\operatorname{Cov}(\Delta_i Y_r, \Delta_j Y_r)| = o(1).$$

Proof. Denote $\Delta_r^{i,j} = \operatorname{Cov}(\Delta_i Y_r, \Delta_j Y_r)$. We proceed by partitioning the sum depending on whether r belongs to the neighborhoods of i and j as defined in Assumption 2.6.

That is, we can write

$$\frac{1}{n^2}\sum_{i,j}\sum_{\substack{r\neq i\\ r\neq j}}\Delta_r^{i,j} = \frac{1}{n^2}\sum_{i,j}\left[\sum_{\substack{r\in\mathcal{N}_i^{H_n}\\ r\in\mathcal{N}_j^{H_n}}}|\Delta_r^{i,j}| + \sum_{\substack{r\in\mathcal{N}_i^{H_n}\\ r\notin\mathcal{N}_j^{H_n}}}|\Delta_r^{i,j}| + \sum_{\substack{r\notin\mathcal{N}_i^{H_n}\\ r\in\mathcal{N}_j^{H_n}}}|\Delta_r^{i,j}| + \sum_{\substack{r\notin\mathcal{N}_i^{H_n}\\ r\notin\mathcal{N}_j^{H_n}}}|\Delta_r^{i,j}|\right]$$

$$\leq \frac{1}{n^2}\sum_{i,j}\left[\sum_{\substack{r\in\mathcal{N}_i^{H_n}\\ r\in\mathcal{N}_j^{H_n}}}|\Delta_r^{i,j}| + \sum_{r\notin\mathcal{N}_j^{H_n}}|\Delta_r^{i,j}| + \sum_{r\notin\mathcal{N}_i^{H_n}}|\Delta_r^{i,j}| + \sum_{r\notin\mathcal{N}_i^{H_n}}|\Delta_r^{i,j}|\right]$$

Now, by Assumption 2.6, each of the inner sums of the last three terms tends to zero in the limit (and the outer sums also tend to zero because there are n^2 summands offset by the n^2 in the denominator). For the first term, the sum is zero whenever the intersection of $\mathcal{N}_i^{H_n}$ and $\mathcal{N}_j^{H_n}$ is empty, and of order h_n otherwise. Therefore,

$$\begin{aligned}
\frac{1}{n^2}\sum_{i,j}\sum_{\substack{r\neq i\\ r\neq j}}\Delta_r^{i,j} &\leq \frac{C}{n^2}\sum_{i,j}h_n\mathbb{1}(|\mathcal{N}_i^{H_n}\cap\mathcal{N}_j^{H_n}|>0)+o(1)\\
&\leq \frac{Ch_n}{n^2}\sum_{i=1}^{n}\sum_{s=1}^{n}\sum_{j=1}^{n}\mathbb{1}(s\in\mathcal{N}_i^{H_n}, j\in\mathcal{N}_s^{H_n})+o(1)\\
&\leq \frac{Ch_n^3}{n}+o(1).
\end{aligned}$$

The proof is finished by noting that $h_n = o(n^{1/3})$, as specified by Assumption 2.6. □

Lemma 2.9. *In addition to the regularity conditions (Assumptions 2.1-2.3), assume that Assumption 2.6 (approximate local independence) holds. Then for all $i \in [n]$ and $A \in [n] \setminus \{i\}$,*

$$|\Delta_i f_n| = O_p(n^{-1/2}) \tag{2.17}$$

$$|\Delta_i f_n^A| = O_p(n^{-1/2}). \tag{2.18}$$

Proof. By a similar argument as in Lemma 2.5,

$$\begin{aligned}
\mathbf{E}(\Delta_i f_n)^2 \leq & n\left[\operatorname{Var}(A_i) + \sum_{r\neq i}\operatorname{Cov}(A_i, B_{i,r}) + \sum_{r\neq i}\sum_{s\neq i}\operatorname{Cov}(B_{i,r}, B_{i,s})\right] \\
\leq & \frac{C_1}{n}\operatorname{Var}(Y_i) + \left(\frac{C_2}{n} + \frac{C_3}{n^2}\right)\sum_{r\neq i}\operatorname{Cov}(Y_i, \Delta_i Y_r) \\
& + \left(\frac{C_4}{n} + \frac{C_5}{n^2}\right)\sum_{r\neq i}\sum_{s\neq i}\operatorname{Cov}(\Delta_i Y_r, \Delta_i Y_s) \\
\leq & \frac{C_1}{n} + \left(\frac{C_2}{n} + \frac{C_3}{n^2}\right)\sum_{r\neq i}\mathbf{E}|\Delta_i Y_r| + \left(\frac{C_4}{n} + \frac{C_5}{n^2}\right)\sum_{r\neq i}\sum_{s\neq i}\mathbf{E}[|\Delta_i Y_r \Delta_i Y_s|] \\
= & \frac{C_1}{n} + \left(\frac{C_2}{n} + \frac{C_3}{n^2}\right)\sum_{r\neq i}\mathbf{E}|\Delta_i Y_r| + \left(\frac{C_4}{n} + \frac{C_5}{n^2}\right)\left(\sum_{r\neq i}\mathbf{E}|\Delta_i Y_r|\right)^2.
\end{aligned}$$

The whole right-hand side is then $O(n^{-1})$ by the fact that $\sum_{r\neq i}|\Delta_i Y_r| = O_p(1)$ (Assumption 2.6). Then (2.17) follows from Markov's inequality. Equation (2.18) immediately follows because $\Delta_i f_n^A$ is equal in distribution to $\Delta_i f_n$. □

We are now ready to prove Theorem 2.2.

Proof. We first compute the limiting variance $\sigma^2 := \lim_{n\to\infty} n\operatorname{Var}(\hat{\tau})$. Let $\mathcal{F}$ be the σ-field defined by equation (2.11). By conditioning on $\mathcal{F}$ we have

$$\operatorname{Var}(\hat{\tau}) = \mathbf{E}\left[\operatorname{Var}\left[\hat{\tau}\middle|\mathcal{F}\right]\right] + \operatorname{Var}\left[\mathbf{E}\left[\hat{\tau}\middle|\mathcal{F}\right]\right].$$

Now,

$$\operatorname{Var}[\hat{\tau}|\mathcal{F}] = \operatorname{Var}\left[\sum_{i=1}^{n}\frac{W_i Y_i^{(1)}}{N_1} - \frac{(1-W_i)Y_i^{(0)}}{N_0}\middle| Y_i^{(0)}, Y_i^{(1)}\right]$$

is the usual variance of a difference-in-means estimator under SUTVA, i.e. fixed potential outcomes. This is known to be (see for example Lin, 2013)

$$\lim_{n\to\infty} n\mathbf{E}[\operatorname{Var}[\hat{\tau}|\mathcal{F}]] = \frac{1-\pi}{\pi}\sigma_1^2 + \frac{\pi}{1-\pi}\sigma_0^2 + 2\sigma_{01}.$$

For the second term, we have $\mathbf{E}[\hat{\tau}|\mathcal{F}] = \bar{Y}_n^{(1)} - \bar{Y}_n^{(0)}$, so

$$\lim_{n\to\infty} n \operatorname{Var}[\mathbf{E}[\hat{\tau}|\mathcal{F}]] = \sigma_\tau^2$$

by Assumption 2.3. This produces the variance expression (2.10).

Since the variance term σ^2 of expression (2.8) in Corollary 2.1 stabilizes, it is sufficient to show

$$\lim_{n\to\infty} \left(\sum_{i,j} c_{i,j} \right)^{1/2} = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{E}|\Delta_i f_n|^3 = 0.$$

Since $|\Delta_i f_n^A| = O_p(n^{-1/2})$ by equation (2.18) of Lemma 2.9, there exists a constant C such that

$$|\operatorname{Cov}(\Delta_i f_n \Delta_i f_n^A, \Delta_j f_n \Delta_j f_n^B)| \leq \frac{C}{n} |\operatorname{Cov}(\Delta_i f_n, \Delta_j f_n)|.$$

Then by Lemma 2.5, there exist constants $c_{i,j} \geq 0$ such that

$$|\operatorname{Cov}(\Delta_i f_n \Delta_i f_n^A, \Delta_j f_n \Delta_j f_n^B)| \leq c_{i,j}$$

and

$$\begin{aligned}\sum_{i,j} c_{i,j} \leq & \frac{C_1}{n^2} \sum_{i,j} |\operatorname{Cov}(Y_i, Y_j)| + \frac{C_2}{n^2} \left[\sum_{i=1}^{n} \sum_{r\neq i} |\operatorname{Cov}(\Delta_i Y_r, Y_i)| + \sum_{i=1}^{n} \sum_{j\neq i} \sum_{r\neq i} |\operatorname{Cov}(\Delta_i Y_r, Y_j)| \right] \\ & + \frac{C_3}{n^2} \left[\sum_{i,j} \sum_{\substack{r\neq i \\ r\neq j}} |\operatorname{Cov}(\Delta_i Y_r, \Delta_j Y_r)| + \sum_{i,j} \sum_{r\neq i} \sum_{\substack{s\neq j \\ s\neq r}} |\operatorname{Cov}(\Delta_i Y_r, \Delta_j Y_s)| \right].\end{aligned}$$

Each of the five terms in the bound captures a different relationship among the responses and discrete derivatives. The first term measures a global covariance structure which tends to zero by Assumption 2.5. The third and fifth terms concern covariances among distinct actors, which are also negligible by Assumption 2.5. The second and fourth terms are the only ones that include elements measuring strong interference.

These two terms are handled by Lemmas 2.7 and 2.8, respectively. So we conclude

$$\lim_{n\to\infty}\left(\sum_{i,j} c_{i,j}\right)^{1/2} = 0.$$

Finally, by equation (2.17) of Lemma 2.9, $\mathbf{E}|\Delta_i f_n|^3 = O(n^{-3/2})$. Hence

$$\sum_{i=1}^{n} \mathbf{E}|\Delta_i f_n|^3 = O(n^{-1/2})$$

and so tends to zero. □

2.5 Variance estimation

We can use the variance decomposition to guide the derivation of an appropriate variance estimator for $\hat{\tau}$. In order to estimate the SUTVA portion of variance, it is enough to use the standard Neyman conservative variance estimator,

$$\hat{V}^2_{\text{SUTVA}} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0},$$

where

$$S_1^2 = \frac{1}{N_1 - 1}\sum_{i=1}^{n} W_i(Y_i - \bar{Y}_1)^2, \qquad S_0^2 = \frac{1}{N_0 - 1}\sum_{i=1}^{n}(1 - W_i)(Y_i - \bar{Y}_0)^2$$

where the sample variances for units assigned to the control group and treatment group, respectively, and

$$\bar{Y}_1 = \frac{1}{N_1}\sum_{i=1}^{n} W_i Y_i, \qquad \bar{Y}_0 = \frac{1}{N_0}\sum_{i=1}^{n}(1 - W_i)Y_i$$

are the sample means. (Note that $\bar{Y}_1$ and $\bar{Y}_0$ are not the same as $\bar{Y}^{(1)}$ and $\bar{Y}^{(0)}$.) It remains to handle σ^2_τ, and under the restricted dependence conditions used in this work, it is possible to derive a consistent estimator of the variance. Denote the

(random) population average treatment effect by

$$T = \bar{Y}^{(1)} - \bar{Y}^{(0)},$$

so that $\tau = \mathbf{E}[T]$ and $\sigma_\tau^2 = \mathrm{Var}(T)/n$, and

$$\mathbf{E}[T^2] = \mathbf{E}\left[(\bar{Y}^{(1)})^2 + (\bar{Y}^{(0)})^2 - 2\bar{Y}^{(1)}\bar{Y}^{(0)}\right].$$

Then we see that the plugin estimator

$$\hat{V}_\tau^2 = \bar{Y}_1^2 + \bar{Y}_0^2 - 2\bar{Y}_1\bar{Y}_0 - \hat{\tau}^2$$

is consistent for $\mathrm{Var}(T) = \mathbf{E}[T^2] - \tau^2$ as long as $\bar{Y}_1$ and $\bar{Y}_0$ are consistent for their limiting expectations. The following proposition shows that this is indeed the case under a $o(n^{1/3})$ (approximate) maximal degree dependency structure.

Proposition 2.1. *Under regularity conditions (Assumptions 2.1-2.3) and restricted interference (either Assumption 2.4 or Assumptions 2.5-2.6), $\hat{\sigma}_\tau^2$ is consistent for σ_τ^2.*

Proof. We wish to show that

$$\hat{V}_\tau^2 = \bar{Y}_1^2 + \bar{Y}_0^2 - 2\bar{Y}_1\bar{Y}_0 - \hat{\tau}^2$$

is consistent for $\mathrm{Var}(T) = \mathbf{E}[T^2] - \tau^2$. It is already established that $\hat{\tau} \xrightarrow{p} \tau$, so it suffices to show that $\mathrm{Var}(\bar{Y}_1^2) \to 0$ (and $\mathrm{Var}(\bar{Y}_0^2) \to 0$ is similar).

Now, the variance is decomposed as

$$\mathrm{Var}(\bar{Y}_1^2) = \mathbf{E}(\mathrm{Var}(\bar{Y}_1^2|\mathcal{F})) + \mathrm{Var}(\mathbf{E}(\bar{Y}_1^2|\mathcal{F}))$$

where $\mathcal{F}$ is the σ-field defined by equation (2.11) representing "conditioning on SUTVA." Since $\bar{Y}_1^2$ is consistent under SUTVA, we have $\mathrm{Var}(\bar{Y}_1^2|\mathcal{F}) \to 0$. Hence the first term is zero. For the second term, $\mathbf{E}(\bar{Y}_1^2|\mathcal{F}) = (\bar{Y}^{(1)})^2$, and so we require that $\mathrm{Var}((\bar{Y}^{(1)})^2) \to 0$. Notice that if Y_i have maximal dependency degree $o(n^k)$ then Y_i^2 have maximal dependency degree $o(n^{2k})$. Therefore consistency for $(\bar{Y}^{(1)})^2$ follows

from Proposition 2 of Sävje et al. (2017) whenever $k < 1/2$. (see also Assumption 2 of that paper). Hence this is satisfied for the (approximate) dependency degree restrictions used in this chapter, where $k = 1/4$ or $k = 1/3$.

□

The following corollary follows immediately, and shows that $\hat{V}_\tau^2$ can be used as an "interference adjustment" to protect the standard SUTVA variance estimator against the forms of interference discussed in this chapter.

Corollary 2.2. *The variance estimator $\hat{V} = \hat{V}_{SUTVA}^2 + \hat{V}_\tau^2$ is asymptotically conservative for the variance of $\hat{\tau}$.*

This variance estimator can then be used to construct $1 - \alpha$ confidence intervals in the usual way,

$$\hat{\tau} \pm z_{\alpha/2}\sqrt{\hat{V}},$$

where $z_{\alpha/2}$ is the appropriate Gaussian quantile.

2.6 Simulations

This section is devoted to two sets of simulations that are designed to illuminate some of the practical implications of our theoretical findings. The first simulation involves tests of Gaussianity and considers situations in which asymptotic inference may be invalid. The second simulation involves the variance decomposition provided by equations (2.12) and (2.13) and considers situations in which inference built under the SUTVA assumption may be invalid. For both simulations we use the same set of networks and generative response model.

In order to replicate as closely as possible the structural characteristics observed in real-world networks, we use five empirical networks from the `facebook100` dataset, an assortment of complete online friendship networks for one hundred colleges and universities collected from a single-day snapshot of Facebook in September 2005. A detailed analysis of the social structure of these networks was given in Traud et al. (2012). The five schools used are the California Institute of Technology, Haverford

College, Amherst College, Michigan Technological University, and Wake Forest University; the only reason for the selection of these five particular schools was so as to produce a rough stratification of population sizes. For each school, we use the largest connected component only. Table 2.1 contains basic summary statistics for the networks used.

network	nodes	edges	avg. degree	avg. pairwise dist.	diameter
Caltech	762	16651	43.70	2.34	6
Haverford	1446	59589	82.42	2.23	6
Amherst	2235	90954	81.39	2.40	7
Michigan Tech	3745	81901	43.74	2.84	7
Wake Forest	5366	279186	104.06	2.51	9

Table 2.1: Summary statistics for the five networks used in the simulation.

We use a simple response model that allows us to control the amount of dependence exhibited among the observations. For network G and nodes i and j, let $\tilde{Z}_{\rho,i,j} = 1$ if nodes i and j are exactly ρ units apart in graph G, and then define

$$Z_{\rho,i} = \left(\sum_j \tilde{Z}_{\rho,i,j} \right)^{-1} \sum_j \tilde{Z}_{\rho,i,j} W_j$$

to be the proportion of units which are exactly distance ρ from i that receive the treatment. Then we model the outcome as

$$Y_i^{(w)} = \alpha_i^{(w)} + \sum_{\rho=1}^{\rho_{\max}} \beta_\rho^{(w)} Z_{\rho,i}$$

for $w = 0, 1$. The intercept $\alpha_i = (\alpha_i^{(0)}, \alpha_i^{(1)})$ captures a heterogeneous direct effect. The maximum distance parameter $\rho_{\max}$ is an integer ranging from 0 to the diameter of the graph. By $\rho_{\max} = 0$ we mean the summation is omitted entirely, so that $Y_i^{(w)} = \alpha_i^{(w)}$, and there is no spillover effect and hence no interference. When $\rho_{\max} = 1$, each unit is subject to a direct effect and a spillover effect governed by coefficient $\beta_1^{(w)}$ and the proportion $Z_{1,i}$ of neighbors of i receiving the treatment. Analogously, higher values of $\rho_{\max}$ admit more distant sources of interference.

We model the coefficient vector as decaying exponentially in the graph distance,

$$\beta_\rho^{(1)} = 2\gamma^\rho, \qquad \beta_\rho^{(0)} = \gamma^\rho,$$

for a decay parameter $\gamma \in (0, 1)$. Therefore, each node receives a direct effect $\alpha_i^{(1)} - \alpha_i^{(0)}$ and an indirect effect

$$\sum_{\rho=1}^{\rho_{\max}} \gamma^\rho Z_{\rho,i}.$$

We control the amount of dependence by varying the parameters $\rho_{\max}$, which explicitly controls the structure of the dependency graph, and γ, which controls the rate at which spillover effects dissipate as they travel through the network.

2.6.1 Tests of normality

We first compute normality test statistics for a variety of parameter configurations. We vary the decay rate γ, and the maximum dependency distance $\rho_{\max}$. The parameter values we use are $\gamma \in \{0.5, 0.9, 0.99\}$, and $\rho_{\max} \in \{2, 6\}$. The maximum value $\rho_{\max} = 6$ was used because all networks have diameter at least 6. For every parameter configuration and each network, we generate 10 instances of the direct effect. The direct effect values $\alpha_i^{(1)}$ and $\alpha_i^{(0)}$ are sampled from independent exponential distributions with different means; the treatment group has mean 1/0.3 and the control group has mean 2.

For each of the 10 instances, we sample 500 draws of the treatment vector as independent Bernoulli(0.5) variables, and compute the outcomes and resulting difference-in-means estimate. We report the test statistic and p-value of the Shapiro-Wilk (SW) test for normality (Shapiro and Wilk, 1965). This produces 10 p-values for each network and parameter configuration, one for each instance of the direct effect. Note that we use these p-values purely for exploratory purposes and do not require nor attempt multiple comparison control.

The results are provided in Table 2.2 and displayed in Figure 2.1. Recall that p-values are uniform under the null hypothesis that the difference-in-means statistics are normally distributed, so that configurations in which most of the p-values are small

may indicate a departure from normality. The scenarios representing the greatest amount of interference are those in which the indirect effect is allowed to propagate over a long distance ($\rho_{\max} = 6$) and in which the indirect effect does not decay much ($\gamma = 0.99$) as it travels across the network. We see that the p-values are smallest for these configurations. For all networks, the value of γ needs to be quite large in order to cause serious problems; p-values appear to be roughly uniform for a dissipation rate of $\gamma = 0.5$ even when $\rho_{\max} = 6$. This supports the claim that having a sparse dependency graph is not necessary for asymptotic normality, since for these networks, the induced dependency graph when $\rho_{\max} = 6$ is either complete or nearly complete. Departures from normality also seem to be sensitive to the particular network structure; the Caltech and Michigan Tech networks seem to be quite well-behaved even under the strongest regimes of interference ($\gamma = 0.99$ and $\rho_{\max} = 6$).

2.6.2 Variance decompositions

For this simulation we explore the relationship between the strength of interference and the resulting variance components. We focus on the Caltech network, which, based on the previous simulation, appears to have a network structure such that the difference-in-means estimator for our response model appears to have a Gaussian distribution even under strong regimes of interference. We draw a single set of $\alpha_i^{(0)}$ and $\alpha_i^{(1)}$ using the same distribution as the previous simulation, with exponential distributions of mean 1/0.3 for the treatment group and mean 2 for the control group. We vary the maximum distance $\rho_{\max}$ from 0 to 5 and the decay parameter γ from 0.1 to 0.9 in increments of 0.1. For each parameter configuration, we draw 10,000 iterates of the treatment vector $\mathbf{W}$ as iid Bernoulli(0.5), and recompute the potential outcomes $Y_i^{(1)}$ and $Y_i^{(0)}$ each time. We then use the potential outcomes to compute

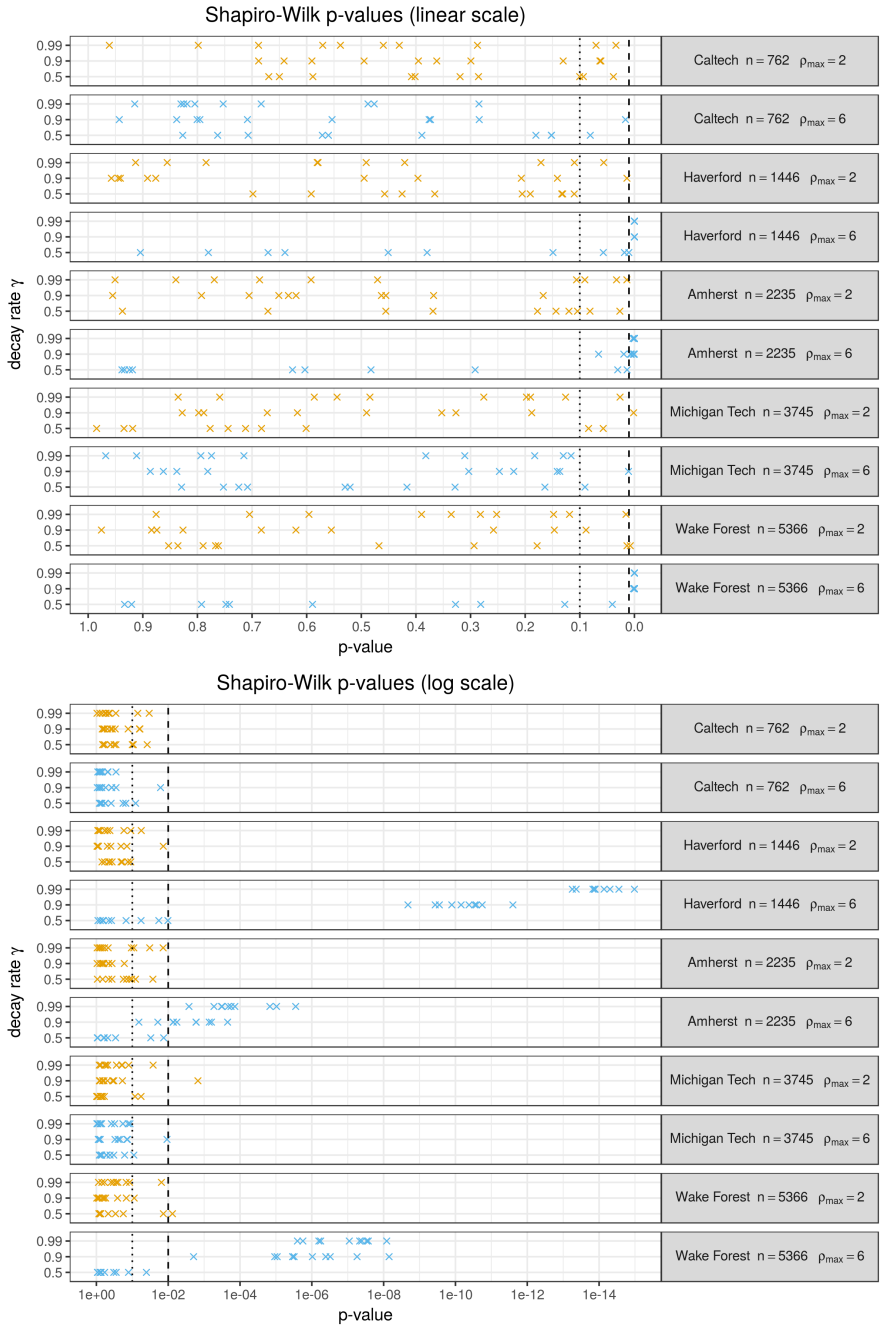

Figure 2.1: (top) Every data point is a p-value for the Shapiro-Wilk test against a Gaussian reference distribution. Each panel represents a different network and dependency distance $\rho_{\max}$ combination. The panels with orange points correspond to $\rho_{\max} = 2$ (less interference) and those with blue points correspond to $\rho_{\max} = 6$ (more interference). The vertical axis contains the three levels of the decay rate γ, ranging from $\gamma = 0.5$ (less interference) to $\gamma = 0.99$ (more interference). The nominal cutoff values 0.1 (vertical dotted line) and 0.01 (vertical dashed line) are highlighted for reference. (bottom) The same plot but using a logarithmic scale for the horizontal axis.

the variance components

$$\begin{aligned}
\sigma_1^2 &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left[(Y_i^{(1)} - \bar{Y}^{(1)})^2\right] \\
\sigma_0^2 &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left[(Y_i^{(0)} - \bar{Y}^{(0)})^2\right] \\
\sigma_{01} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left[(Y_i^{(1)} - \bar{Y}^{(1)})(Y_i^{(0)} - \bar{Y}^{(0)})\right] \\
\sigma_\tau^2 &= n \operatorname{Var}\left[\bar{Y}^{(1)} - \bar{Y}^{(0)}\right],
\end{aligned}$$

where the expectation and variance are computed as finite population moments over the 10,000 simulation replicates. We also compute the observed variance $\hat{\sigma}^2_{\mathrm{DM}}$ of the difference-in-means estimator. This is calculated as the empirical variance of the difference-in-means estimator over the 10,000 draws of the treatment vector.

The SUTVA (conditional) variance is

$$\sigma^2_{\mathrm{SUTVA}} = \sigma_1^2 + \sigma_0^2 + 2\sigma_{01}.$$

We display the ratio of the expected true variance of $\hat{\tau}$ to the conditional variance, $(\sigma^2_{\mathrm{SUTVA}} + \sigma_\tau^2)/\sigma^2_{\mathrm{SUTVA}}$, as well as the observed ratio, $\hat{\sigma}^2_{\mathrm{DM}}/\sigma^2_{\mathrm{SUTVA}}$. The resulting ratios are displayed in Figure 2.2, and the full results are given in Tables 2.3 and 2.4. The observed variances mostly track the expected variances. Under SUTVA, $Y_i^{(0)}$ and $Y_i^{(1)}$ exhibit no additional variation so the observed variance appears to match $\sigma^2_{\mathrm{SUTVA}}$. As we allow units to influence units farther away in the graph, the variance ratio grows. The discrepancy is not too large for fast decaying interference ($\gamma < 0.5$) but for γ close to 1.0 it can be drastic. When $\rho_{\max} = 5$ and $\gamma = 0.6$ the observed variance is only 7.4% larger than $\sigma^2_{\mathrm{SUTVA}}$, but for $\rho_{\max} = 5$ and $\gamma = 0.9$ the observed discrepancy is 60.1%.

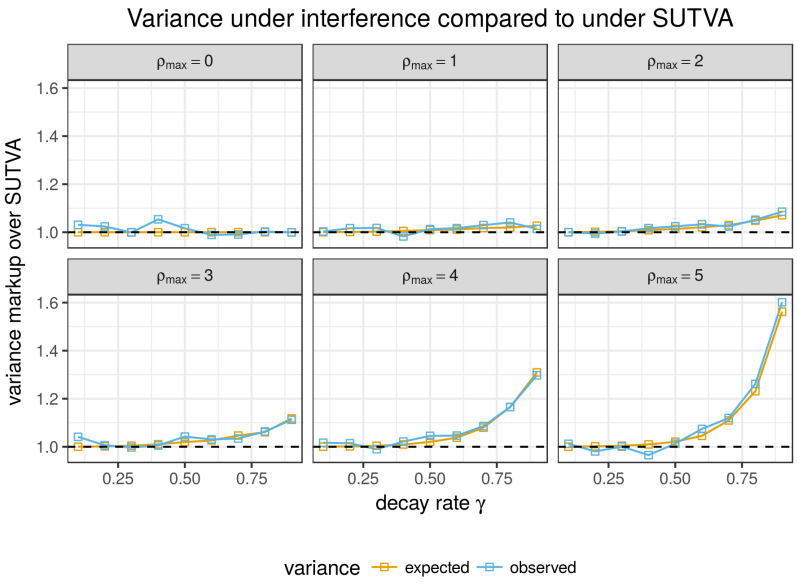

Figure 2.2: Variance ratios for the Caltech network. Each panel represents a different maximum distance $\rho_{\max}$. The horizontal axis is the decay rate γ and the vertical axis marks the variance ratios. The horizontal dotted line marks the baseline, which is a ratio of one. The orange values are the expected variance ratios $(\sigma^2_{\text{SUTVA}}+\sigma^2_\tau)/\sigma^2_{\text{SUTVA}}$, and the blue values are the observed variance ratios $\hat{\sigma}^2_{\text{DM}}/\sigma^2_{\text{SUTVA}}$. The upper-left most panel, $\rho_{\max} = 0$, is the case when SUTVA is true. The markup is greatest when $\rho_{\max}$ and γ are both large. Note that the horizontal axis starts at 100% and that the greatest observed ratio is about a 60% increase in the variance over SUTVA.

2.7 Discussion

In this work we have developed a framework for obtaining asymptotic results for causal estimators in randomized experiments in the presence of interference. We contextualize the work of Sävje et al. (2017) within Stein's method and obtain asymptotic normality results.

Our two main results—one constraining the dependency degree (Theorem 2.1) and another placing conditions on how the treatment variables interact with the response variables (Theorem 2.2)—highlight two general ways one may proceed for handling arbitrary interference. The dependency graph approach follows a motif found in the interference literature of relying on local interference assumptions, such

as the neighborhood treatment response condition. Such assumptions, which enforce a sort of "sparsity of interference," are often viewed as implausible yet necessary for tractability of results. However, we have shown that progress is possible under certain dense regimes of interference. Our result is still restrictive in the sense that it requires a condition of approximately sparse dependency. It is possible that more general advancements can be made by characterizing the behavior of the object T in the main perturbative theorem (Lemma 2.2).

As discussed in Section 3.2, the definition of σ_τ^2 (equation (2.5) in Assumption 2.3) means that we have restricted ourselves to studying rate-optimal scenarios. A useful extension would be to establish similar limiting results for the case when the difference-in-means $\bar{Y}^{(1)} - \bar{Y}^{(0)}$ converges at a slower-than-$\sqrt{n}$ rate. A related issue is efficiency in the presence of interference. The semiparametric efficiency bound (Hahn, 1998) that serves as the basis for efficient estimation of average treatment effects in observational studies is reliant on independent units. A characterization of similar semiparametric efficiency bounds for various levels of interference is a prerequisite for understanding whether efficient estimators remain optimal under interference.

We discuss how to approach statistical inference by noticing that σ_τ^2 characterizes the difference between the variance under interference and under SUTVA. It is also possible that estimates of σ_τ^2 can serve as the basis for a statistical test of whether interference exists. For a related recent idea, see Choi (2018).

Finally, our work is a novel application of Stein's method. From a technical standpoint, our results demonstrate that tools from that literature can be used for establishing theoretical results for causal estimators under interference. By overlaying the interference framework on top of Stein's method we are able to sidestep more complicated calculations or detailed assumptions about the structure of interference. We have not addressed other statistical objects such as more general estimators and designs, or different estimands including the global treatment effect that compares all units in treatment to all units in control. It is important to understand the behavior of the "direct effect" (EATE) estimand considered in Sävje et al. (2017) and this work, because it is the natural estimand of SUTVA-based estimators. However it is worth considering the utility and interpretability of such an estimand in practice. Since

interference at its core involves handling a dependent collection of random variables, we suspect that Stein's method may be useful for understanding other settings as well.

Network		Parameters		SW statistic	SW p-values		
school	nodes	$\rho_{\max}$	γ	average	avg	min	max
Caltech	762	2	0.5	0.996	0.355	0.038	0.669
Caltech	762	2	0.9	0.996	0.373	0.061	0.688
Caltech	762	2	0.99	0.997	0.484	0.034	0.961
Caltech	762	6	0.5	0.997	0.438	0.081	0.827
Caltech	762	6	0.9	0.997	0.569	0.016	0.943
Caltech	762	6	0.99	0.998	0.688	0.285	0.915
Haverford	1446	2	0.5	0.996	0.331	0.110	0.698
Haverford	1446	2	0.9	0.997	0.586	0.014	0.958
Haverford	1446	2	0.99	0.997	0.496	0.056	0.913
Haverford	1446	6	0.5	0.996	0.406	0.010	0.904
Haverford	1446	6	0.9	0.957	0.000	0.000	0.000
Haverford	1446	6	0.99	0.928	0.000	0.000	0.000
Amherst	2235	2	0.5	0.996	0.309	0.027	0.937
Amherst	2235	2	0.9	0.997	0.581	0.167	0.955
Amherst	2235	2	0.99	0.996	0.455	0.014	0.951
Amherst	2235	6	0.5	0.997	0.576	0.013	0.938
Amherst	2235	6	0.9	0.991	0.011	0.000	0.066
Amherst	2235	6	0.99	0.986	0.000	0.000	0.003
Michigan Tech	3745	2	0.5	0.997	0.649	0.057	0.985
Michigan Tech	3745	2	0.9	0.996	0.506	0.001	0.828
Michigan Tech	3745	2	0.99	0.996	0.403	0.026	0.835
Michigan Tech	3745	6	0.5	0.997	0.506	0.091	0.829
Michigan Tech	3745	6	0.9	0.996	0.443	0.011	0.886
Michigan Tech	3745	6	0.99	0.997	0.528	0.116	0.968
Wake Forest	5366	2	0.5	0.996	0.497	0.008	0.853
Wake Forest	5366	2	0.9	0.997	0.591	0.089	0.976
Wake Forest	5366	2	0.99	0.996	0.372	0.015	0.876
Wake Forest	5366	6	0.5	0.997	0.550	0.040	0.933
Wake Forest	5366	6	0.9	0.979	0.000	0.000	0.002
Wake Forest	5366	6	0.99	0.975	0.000	0.000	0.000

Table 2.2: Summary of Shapiro-Wilk p-values from Simulation 1. Average, minimum, and maximum are taken over the 10 instances of the response.

Parameters		Variances			Ratios to SUTVA	
$\rho_{\max}$	γ	SUTVA	expected	observed	expected	observed
0	0.1	14.770	14.770	15.228	1.000	1.031
0	0.2	15.205	15.205	15.570	1.000	1.024
0	0.3	14.690	14.690	14.680	1.000	0.999
0	0.4	15.382	15.382	16.208	1.000	1.054
0	0.5	14.478	14.478	14.714	1.000	1.016
0	0.6	14.321	14.321	14.164	1.000	0.989
0	0.7	16.674	16.674	16.521	1.000	0.991
0	0.8	17.574	17.574	17.623	1.000	1.003
0	0.9	16.453	16.453	16.440	1.000	0.999
1	0.1	12.717	12.722	12.758	1.000	1.003
1	0.2	14.845	14.864	15.094	1.001	1.017
1	0.3	14.694	14.736	14.954	1.003	1.018
1	0.4	14.282	14.360	14.034	1.005	0.983
1	0.5	12.739	12.856	12.906	1.009	1.013
1	0.6	16.073	16.251	16.346	1.011	1.017
1	0.7	13.262	13.497	13.655	1.018	1.030
1	0.8	15.247	15.546	15.867	1.020	1.041
1	0.9	14.324	14.713	14.528	1.027	1.014
2	0.1	16.199	16.205	16.198	1.000	1.000
2	0.2	17.088	17.113	16.990	1.001	0.994
2	0.3	15.325	15.386	15.373	1.004	1.003
2	0.4	14.046	14.168	14.283	1.009	1.017
2	0.5	15.489	15.703	15.868	1.014	1.024
2	0.6	16.697	17.040	17.247	1.021	1.033
2	0.7	17.665	18.186	18.088	1.029	1.024
2	0.8	15.419	16.159	16.219	1.048	1.052
2	0.9	14.598	15.631	15.846	1.071	1.085

Table 2.3: Table of variances for the Caltech network from Simulation 2, for $\rho = 0, 1, 2$.

Parameters		Variances			Ratios to SUTVA	
$\rho_{\max}$	γ	SUTVA	expected	observed	expected	observed
3	0.1	14.095	14.100	14.678	1.000	1.041
3	0.2	14.267	14.292	14.359	1.002	1.006
3	0.3	14.798	14.863	14.755	1.004	0.997
3	0.4	13.442	13.581	13.517	1.010	1.006
3	0.5	12.762	13.013	13.302	1.020	1.042
3	0.6	16.095	16.519	16.592	1.026	1.031
3	0.7	14.900	15.595	15.410	1.047	1.034
3	0.8	18.009	19.103	19.158	1.061	1.064
3	0.9	14.031	15.690	15.607	1.118	1.112
4	0.1	12.765	12.771	12.980	1.000	1.017
4	0.2	13.902	13.927	14.105	1.002	1.015
4	0.3	15.799	15.866	15.638	1.004	0.990
4	0.4	15.210	15.352	15.541	1.009	1.022
4	0.5	14.311	14.601	14.962	1.020	1.046
4	0.6	16.144	16.755	16.893	1.038	1.046
4	0.7	15.692	16.942	17.043	1.080	1.086
4	0.8	16.461	19.185	19.188	1.165	1.166
4	0.9	18.759	24.566	24.343	1.310	1.298
5	0.1	14.747	14.752	14.928	1.000	1.012
5	0.2	13.261	13.286	13.006	1.002	0.981
5	0.3	15.400	15.467	15.409	1.004	1.001
5	0.4	14.980	15.127	14.462	1.010	0.965
5	0.5	14.784	15.094	14.978	1.021	1.013
5	0.6	14.554	15.221	15.635	1.046	1.074
5	0.7	14.374	15.951	16.093	1.110	1.120
5	0.8	16.307	20.078	20.575	1.231	1.262
5	0.9	15.546	24.286	24.894	1.562	1.601

Table 2.4: Table of variances for the Caltech network from Simulation 2, for $\rho = 3, 4, 5$.

Chapter 3

Regression adjustments for interference

3.1 Introduction

The goal in a randomized experiment is often to estimate the *total* or *global average treatment effect* (GATE) of a binary treatment variable on a response variable. The GATE is the difference in average outcomes when all units are exposed to treatment versus when all units are exposed to control. Under the standard assumption that units do not interfere with each other (Cox, 1958), which forms a key part of the *stable unit treatment value assumption* (SUTVA) (Rubin, 1974, 1980), the global average treatment effect reduces to the standard average treatment effect.

However, in many social, medical, and online settings the no-interference assumption may fail to hold (Rosenbaum, 2007; Walker and Muchnik, 2014; Aral, 2016; Taylor and Eckles, 2017). In such settings, peer and spillover effects can bias estimates of the global average treatment effect. In the past decade, there has been a flurry of literature proposing methods for handling interference, mostly focusing on cases in which structural assumptions about the nature of interference are known. For example, if there is a natural grouping structure to the data, such as households or schools or classrooms, it may be reasonable to assume that interference exists within but not across groups. Versions of this assumption are known as *partial* or *stratified*

interference (Hudgens and Halloran, 2008). In this case two-stage randomized designs can be used to decompose direct and indirect effects, which is an approach studied by VanderWeele and Tchetgen Tchetgen (2011); Tchetgen Tchetgen and VanderWeele (2012); Liu and Hudgens (2014); Baird et al. (2016); Basse et al. (2017), among others. Baird et al. (2016) study how two-stage, random saturation designs can be used to estimate dose response curves under the stratified interference assumption. Basse and Feller (2018) study two-stage experiments in which households with multiple students are assigned to treatment or control. Other works that propose methods of handling interference include Ogburn and VanderWeele (2014), which maps out causal diagrams for interference; van der Laan (2014), which studies a targeted maximum likelihood estimator for the case where network connections and treatments possibly change over time; Choi (2017), which shows how confidence intervals can be constructed in the presence of monotone treatment effects; and Jagadeesan et al. (2017), which studies designs for estimating the direct effect that strive to balance the network degrees of treated and control units.

The *modus operandi* for general or arbitrary interference is the method of *exposure modeling*, in which the researcher defines equivalence classes of treatments that inform the interference pattern. Aronow and Samii (2017) develop a general framework for analyzing inverse propensity weighted (Horvitz-Thompson- and Hájek-style) estimators under correct specification of local exposure models. The exposure model often used is some version of an assumption that the potential outcomes of unit i are constant conditional on all treatments in a local neighborhood of i, or that the potential outcomes are a monotone function of such treatments. This assumption, known as *neighborhood treatment response* (NTR), is a generalization of partial and stratified interference to the general network setting (Manski, 2013). Methods for handling interference often rely on neighborhood treatment response as a core assumption. For example, Sussman and Airoldi (2017) develop unbiased estimators for various parametric models of interference that are all restrictions on the NTR condition, and Forastiere et al. (2016) propose propensity score estimators for observational studies using the NTR assumption.

Aronow and Samii (2017) use their methods to analyze the results of a field experiment on an anti-conflict program in middle schools in New Jersey. By defining appropriate exposure models, they are able to estimate a direct effect (the effect of receiving the anti-conflict intervention), a spillover effect (the effect of being friends with some students who received the anti-conflict intervention), and a school effect (the effect of attending a school in which some students received the anti-conflict intervention). The network structure consists of 56 disjoint social networks (schools), comprising 24,191 students in the original Paluck et al. (2016) study and a subset of 2,050 students studied in the Aronow and Samii (2017) analysis. There are a number of similar studies in which the target of scientific inquiry is the quantification of peer or spillover effects and where the dataset permits doing so by being comprised of "many sparse networks." Studies which consist of randomized experiments on such social networks include Banerjee et al. (2013), which studies a microfinance loan program in villages in India; Cai et al. (2015), which studies a weather insurance program for farmers in rural China; Kim et al. (2015), which concerns public health interventions such as water purification and microvitamin tablets in villages in Honduras; and Beaman et al. (2018), which explores social diffusion of a new agricultural technology among farmers in Malawi. (Some studies thereof do not explicitly aim to understand spillover effects—for example Kim et al. (2015) and Beaman et al. (2018) are concerned primarily with strategies for targeting influential individuals—but the presence of such effects is still crucial for their purposes.) In these settings, exposure modeling may be (and has been) a successful way of decomposing direct and spillover effects.

The difficulties of using exposure models for global effects How should one proceed if the goal is estimation of the global treatment effect rather than a decomposition into direct and spillover effects? In this setting interference is a nuisance, not an object of intrinsic scientific interest. Unbiased estimation would result from using an exposure model that accurately represents the true data-generating process. However, the complicated nature of social interactions makes it difficult to select an exposure model that is both tractable and well-specified. Eckles et al. (2017) discuss

some of the difficulties of working in this setting in the context of "implausibility of tractable treatment response assumptions":

> It is unclear how substantive judgment can directly inform the selection of an exposure model for interference in networks—at least when the vast majority of vertices are in a single connected component. Interference is often expected because of social interactions (i.e., peer effects) where vertices respond to their neighbors' behaviors: in discrete time, the behavior of a vertex at t is affected by the behavior of its neighbors at $t-1$; if this is the case, then the behavior of a vertex at t would also be affected by the behavior of its neighbors' neighbors at $t-2$, and so forth. Such a process will result in violations of the NTR assumption, and many other assumptions that would make analysis tractable.

In this setting, one primary tool that has developed in the literature is the method of *graph cluster randomization* (Ugander et al., 2013), where researchers use a clustered design in which the clusters are selected according to the structure of the graph in order to lower the variance of NTR-based inverse propensity estimators. Eckles et al. (2017) provide theoretical results and simulation experiments to show how clustered designs can reduce bias due to interference. Clusters can be obtained using algorithms developed in the graph partitioning and community detection literature (Fortunato, 2010; Ugander and Backstrom, 2013).

While the graph clustering approach can be effective at removing some bias, the structure of real-world empirical networks may make it difficult to obtain satisfactory bias reduction via clustering, which relies on having good quality graph cuts. The "six degrees of separation" phenomenon is well-documented in large social networks (Ugander et al., 2011; Backstrom et al., 2012), and the average distance between two Facebook users in February 2016 was just 3.5 (Bhagat et al., 2016). Furthermore, most users belong to one large connected component and are unlikely to separate cleanly into evenly-sized clusters. In a graph clustered experiment run at LinkedIn, the optimal clustering strategy used maintained only 35.59% of edges between nodes of the same cluster (Saveski et al., 2017, Table 1), suggesting that bias remains even

after clustering. Figure 3.1 provides an example illustration of how the structure of the network can markedly affect how much we might expect cluster randomization to help.

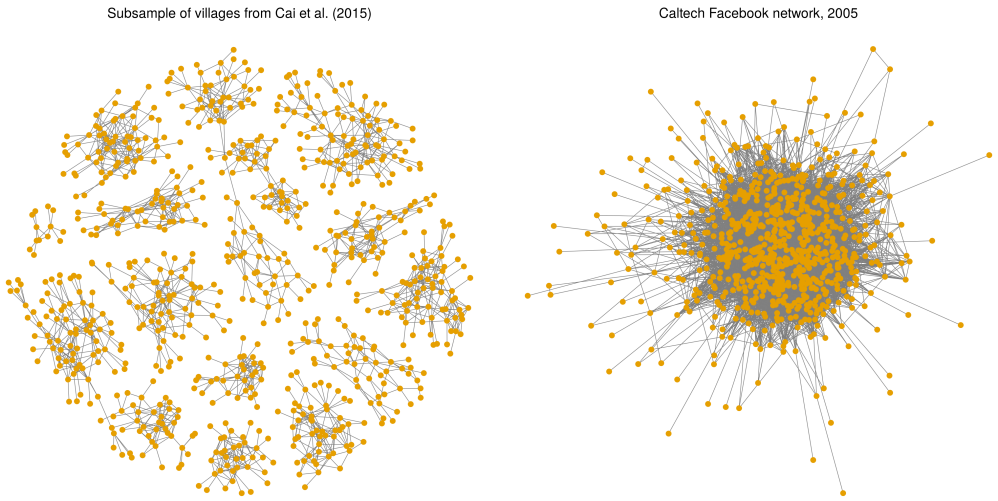

Figure 3.1: (left) A subset of 16 nearly-disjoint Chinese villages, comprising 822 nodes, from an experiment regarding weather insurance adoption conducted by Cai et al. (2015). The setup of many, sparse networks is similar to that in the anti-conflict school dataset from Paluck et al. (2016). (right) The largest connected component of the Caltech Facebook network, with 762 nodes, from a single day snapshot in September 2005, taken from the `facebook100` dataset (Traud et al., 2011, 2012). Networks were plotted with the `ggnet2` function (Tyner et al., 2017) in the `GGally` package, using the default Fruchterman-Reingold force-directed layout (Fruchterman and Reingold, 1991). We should not be surprised if methods for handling interference that might work well in the collection of networks on the left, such as exposure modeling and graph clustering, do not work so well in the network on the right.

Such experimental designs also face practical hurdles. Though cluster randomized controlled trials are commonly used in science and medicine, existing experimentation platform infrastructure in some organizations may only exist for standard (i.i.d.) randomized experiments, in which case adapting the design and analysis pipelines for graph cluster randomization would require significant ad hoc engineering effort. In regimes of only mild interference, it may simply not be worth the trouble to run a clustered or two-stage experiment, especially if there is no way to know *a priori* how much bias from interference will be present. Instead, the practitioner would prefer to

have a data-adaptive debiasing mechanism that can be applied to an experiment that has already been run. Ideally, such estimators provide robustness to deviations from SUTVA yet do not sacrifice too much in precision loss if it turns out interference was weak or non-existent.

Towards an agnostic regression approach We can take advantage of the fact that the global treatment effect estimand, as opposed to a peer or spillover effect estimand, can be defined *agnostically* without regard to any exposure model. The exposure model used, therefore, matters only insofar as it informs the corresponding estimator used. An appropriate exposure model is one that leads to estimates of the global treatment effect that are approximately unbiased, *even if it is not the exposure model corresponding to the true data-generating process.* This agnostic perspective gives us hope because of the decoupling between data generation and estimation: We can believe in a complex interference pattern without having to use the corresponding intractable exposure model for estimation.

Our approach is motivated by the rich literature on regression adjustment estimators in the non-interference setting. In randomized controlled trials, regression adjustments are used to adjust for imbalances due to randomized assignment of the empirical covariate distributions of different treatment groups, and thus improve precision of treatment effect estimators. In the observational studies setting, regression adjustments are used to adjust for inherent differences between the covariate distributions of different treatment groups. We heavily borrow tools from that literature, both in the classical regime of using low-dimensional, linear regression estimators (Freedman, 2008a,b; Lin, 2013; Berk et al., 2013) and more recent advancements that can utilize high-dimensional regression and machine learning techniques (Bloniarz et al., 2016; Wager et al., 2016; Wu and Gagnon-Bartsch, 2017; Athey et al., 2017b; Chernozhukov et al., 2018). This recent literature adopts the agnostic perspective that properties of least squares and machine learning estimators can be utilized without assuming the parametric model itself.

This chapter contains two main contributions: (a) a regression adjustment strategy for debiasing global treatment effect estimators, and (b) a class of bootstrapping

and resampling methods for constructing variance estimates of such estimators. We explore how well the analysis side of an experiment can be improved in independently-assigned (non-clustered) experiments. Our approach can be loosely motivated by the *linear-in-means* (LIM) family of models from the econometrics literature (Manski, 1993). In a simple version of this model, an individual's outcome is said to depend on the average of her peer's exogenous features. If this is true, then the peer average feature "statistic" can be adjusted for when estimating the global treatment effect, even if the linear-in-means model itself does not hold. We note that much of the linear-in-means literature focuses on the identifiability of various peer effect parameters within the LIM model (Bramoullé et al., 2009); our goal instead is estimation of the agnostic global effect.

Generally, our strategy is to learn a statistical model that captures the relationship between the outcomes and a set of unit-level statistics constructed from the treatment vector and the observed network. These statistics can be viewed as *features* or *adjustment variables*, and are to be constructed by the practitioner using domain knowledge. The model is then used to predict the unobserved potential outcomes of each unit under the counterfactual scenarios if the unit had been assigned to global treatment, and global control. The approach is thus reminiscent of regression adjustment estimators and off-policy evaluation. Figure 3.2 demonstrates how feature distributions differ between the observed design distribution and the unobserved global counterfactual distributions of interest.

In Section 3.3 we present estimators in the context of a generative linear model and in Section 3.4 we discuss the non-linear analog. Even though the results in this chapter are presented within the context of a generative model, they still make progress towards a fully agnostic solution. First, the models considered here are considerably more flexible, and more easily extended, than exposure models (which are also assumed to be generative). Second, the assumptions on the errors can be relaxed and we show via simulation experiments (Section 4.7) that these linear methods can work well in more general contexts. The non-linear context, which allows the use of arbitrary machine learning estimators, also moves closer to a fully agnostic approach by allowing a nonparametric generative model. Third, by connecting these results

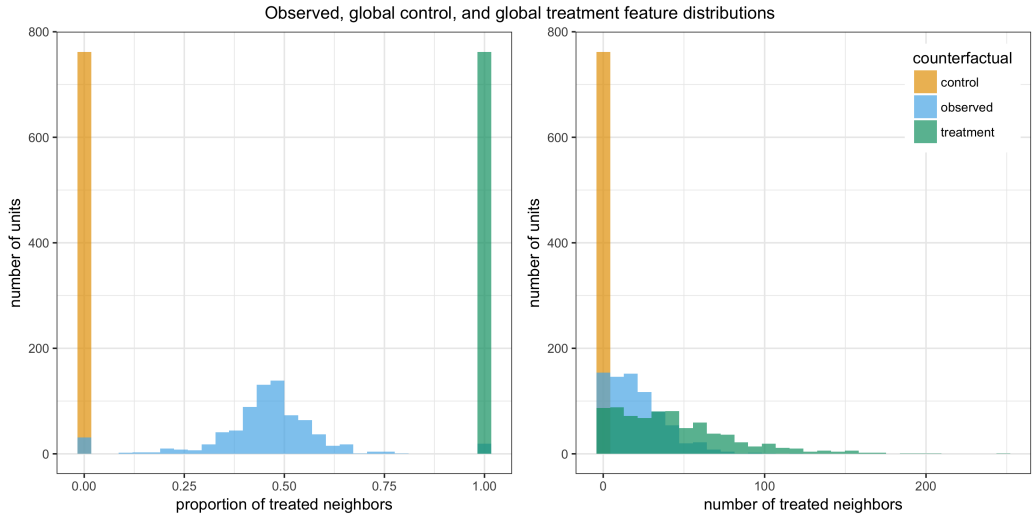

Figure 3.2: (left) Distributions for fraction of treated neighbors $d_i^{-1} \sum_{j \in \mathcal{N}_i} W_j$. (right) Distributions for number of treated neighbors $\sum_{j \in \mathcal{N}_i} W_j$. Feature distributions are under global exposure to control $\mathbf{W} = \mathbf{0}$ (orange), global exposure to treatment $\mathbf{W} = \mathbf{1}$ (green), and a single observed treatment instance from an iid Bernoulli(0.5) distribution (blue). Network is the Caltech social graph from the `facebook100` dataset (Traud et al., 2011, 2012). If the response is correlated with one or both of these features, then ideas from off-policy evaluation of the counterfactual outcomes can guide estimation of the global treatment effect. Even if the distributions are quite different, as in the left hand picture, if the response can be modeled by low dimensional model then extrapolation may not be too unreasonable.

to analogous ones in the SUTVA case we lay the foundation for how to think about an agnostic approach, which is not possible using pure exposure modeling methods. Indeed, agnostic perspectives have emerged only recently even in the SUTVA setting (Freedman, 2008a,b; Lin, 2013). This work, therefore, can be viewed as a conceptual stepping stone between existing methods that assume exposure models are generative, and future work that would establish a fully agnostic presentation.

It is shown in this chapter that an assumption of exogeneity is required, even though the treatment is randomized. Such an assumption can be likened to an *unconfoundedness*, *ignorability*, or *selection on observables* assumption. A curious feature of randomized experiments under interference, then, is that they display characteristics of observational studies as well. It is helpful to think of estimators used in

SUTVA observational studies that require the estimation of both a *propensity model* and a *response model*. (Doubly-robust estimators allow misspecification of one but not both of these models.) In a randomized experiment under interference the propensity model is fully known and does not need to be estimated; however, the response can be affected by confounding variables. In randomized experiments under interference, then, researchers must be wary of the same challenges that beset drawing causal conclusions from observational datasets, even though the treatments were assigned randomly. The exogeneity assumption is not generally verifiable from the data but is necessary in order to make any progress. Ideally, one has access to methods for conducting sensitivity analyses for interference, but such methods are in their infancy and we refrain from addressing this issue here.

Our estimators have several advantages over existing exposure modeling estimators. The correct specification of an exposure model is also a form of exogeneity assumption, yet our approach admits much more flexible forms of interference. It can handle multiple types of graph features, which do not even have to be constructed from the same network. Adjusting for interference becomes a feature engineering problem in which the practitioner is free to use his or her domain knowledge to construct appropriate features. If a feature turns out to be noninformative for interference, no additional bias is incurred (though a penalty in variance may be paid). Our adjustment framework also reduces to the standard, SUTVA regression adjustment setup in the event that static, baseline characteristics are used.

Finally, we propose methods for quantifying the variance of the proposed estimators. Variance estimation in the presence of interference is generally difficult because of the complicated dependencies created by the propagation of interference over the network structure. Confidence intervals based on asymptotic approximations may not be reliable since the dependencies can drastically reduce the effective sample size. For example, the variance of the sample mean of the features may not even scale at a n^{-1} rate, where n is the sample size. In this chapter we propose a novel way of taking advantage of the randomization distribution to produce bootstrap standard errors, assuming unconfoundedness. Since the features are constructed by the researcher from the vector of treatments, and the distribution of treatments is known

completely in a randomized experiment, we can calculate via Monte Carlo simulation the sampling distribution of any function of the design matrix under the randomization distribution. This approach ensures that we properly represent all of the dependencies exhibited empirically by the data, and can then be used to construct standard errors.

3.2 Setup and estimation in LIM models

We work within the potential outcomes framework, or Rubin causal model (Neyman, 1923; Rubin, 1974). Consider a population of n units indexed on the set $[n] = \{1, \ldots, n\}$ and let $\mathbf{W} = (W_1, \ldots, W_n) \in \mathcal{W} = \{0, 1\}^n$ be a random vector of binary treatments. We will work only with treatments assigned according to a Bernoulli randomized experimental design:

Assumption 3.1. $W_i \overset{iid}{\sim} \text{Bernoulli}(\pi)$ *for every unit* $i \in [n]$*, where* $\pi \in (0, 1)$ *is the treatment assignment probability.*

The general spirit of our approach can likely be extended to more complicated designs, but our goal in this work is to show that substantial analysis-side improvements can be made even under the simplest possible experimental design.

Suppose that each response lives in an outcome space $\mathcal{Y}$, and is determined by a mean function $\mu_i : \mathcal{W} \to \mathcal{Y}$:

$$Y_i = Y_i(\mathbf{W}) = \mu_i(\mathbf{W}) + \varepsilon_i \tag{3.1}$$

In this section we limit ourselves to an informal discussion of point estimation and defer the question of variance estimation to a future section. The only assumption we require on the residuals, therefore, is an assumption of strict exogeneity:

$$\mathbf{E}[\varepsilon_i | W_1, \ldots, W_n] = 0.$$

In particular, no independence or other assumptions about the correlational structure of the residuals are made in this section, though such assumptions will be necessary

for variance estimation, which we address in Section 3.3.

Because the units are assumed to belong to a network structure, distinguishing between finite population and infinite superpopulation setups is not so straightforward. In the SUTVA setting, good estimators for finite population estimands (or conditional average treatment effects) are usually good estimators for superpopulation estimands, and vice versa (Imbens, 2004). In order to simplify the analysis, we do not work with a fixed potential outcomes $Y_i(\mathbf{w})$ for $\mathbf{w} \in \mathcal{W}$, and allow the residuals ε_i to be random variables. We therefore consider additional variation of the potential outcomes coming from repetitions of the experiment, but we do not consider the units to be sampled from a larger population. We do this because it is easier to discuss the behavior of ε_i when they are random variables. This perspective is related to the *intrinsic nondeterminism* perspective discussed by Pearl (2009) on page 220, as well as the idea of *stochastic counterfactuals* discussed previously in the literature (Greenland, 1987; Robins and Greenland, 1989, 2000; VanderWeele and Robins, 2012).

In this chapter we focus on estimation of the *total* or *global average treatment effect* (GATE), defined by

$$\tau = \frac{1}{n} \sum_{i=1}^{n} [\mathbf{E}[Y_i(\mathbf{1})] - \mathbf{E}[Y_i(\mathbf{0})]]. \tag{3.2}$$

This parameter is called a global treatment effect because is a contrast of average outcomes between the cases when the units are globally exposed to treatment ($\mathbf{W} = \mathbf{1}$) and globally exposed to control ($\mathbf{W} = \mathbf{0}$).

Under an assumption of strict exogeneity, in which $\mathbf{E}[\varepsilon_i|\mathbf{W}] = 0$, the treatment effect is the difference of average global exposure means

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \left[\mu_i(\mathbf{1}) - \mu_i(\mathbf{0})\right],$$

In order to proceed, we must make assumptions about the structure of the mean function μ_i.

3.2.1 A simple linear-in-means model

To illustrate our approach we start with a simple model. Let G be a network with adjacency matrix A. For simplicity in this work we will mostly assume that G is simple and undirected, but one can just as easily use a weighted and directed graph. We emphasize that we assume G is completely known to the researcher. Let $\mathcal{N}_i = \{j \in [n] : A_{ij} = 1\}$ be the neighborhood of unit i and $d_i = |\mathcal{N}_i|$ be the network degree of unit i. Define

$$X_i = \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} W_j, \tag{3.3}$$

the fraction of neighbors of i that are in the treatment group. Then take the mean function μ_i in equation (3.1) to be as follows.

Model 3.1 (Exogenous LIM model)**.**

$$\mu_i(\mathbf{W}) = \alpha + \gamma W_i + \delta X_i.$$

This model is a simple version of a linear-in-means model (Manski, 1993). The model contains an intercept α as well as a direct effect γ, which captures the strength of individual i's response to changes in its own treatment assignment. Additionally, the response of unit i is correlated with mean treatment assignment of its neighbors; Manski (1993) calls δ an *exogenous social effect*, because it captures the correlation of unit i's response with the exogenous characteristics of its neighbors. The interactions are also assumed to be "anonymous" in that the unit i responds only to the mean neighborhood treatment assignment and not the identities of those treated neighbors. In this model, unit i responds to its neighbors' treatments but not to its neighbors' outcomes. Under Model 3.1, the variable X_i is the mechanism by which interference affects the outcome and thus can be viewed as playing a similar role as baseline characteristics or pretreatment covariates in an observational study. However, in this work we shall use the term *statistic* or *feature* rather than *covariate* to refer to X_i, in order to remind the reader that X_i does not represent a baseline characteristic.

Now consider the estimand (3.2) under Model 3.1. If all units are globally exposed to treatment then it is the case for all units i that $W_i = 1$ and $X_i = 1$. Therefore

$$\frac{1}{n}\sum_{i=1}^{n}\mu_i(\mathbf{1}) = \alpha + \gamma + \delta.$$

Similarly, if all units are globally exposed to control, then $W_i = 0$ and $X_i = 0$, and so

$$\frac{1}{n}\sum_{i=1}^{n}\mu_i(\mathbf{0}) = \alpha.$$

Therefore, the treatment effect under Model 3.1 is simply

$$\tau = (\alpha + \gamma + \delta) - \alpha = \gamma + \delta.$$

This parametrization suggests that if we have access to unbiased estimators $\hat{\gamma}$ and $\hat{\delta}$ for γ and δ, then an unbiased estimate for τ is given by

$$\hat{\tau} = \hat{\gamma} + \hat{\delta}.$$

In particular, one is tempted to estimate γ and δ with an OLS regression of Y_i on W_i and X_i. Of course, using $\hat{\tau}$ as an estimator for τ only makes sense if Model 3.1 accurately represents the true data generating process. We build up more flexible models in the following sections.

In contrast, we can easily see why the difference-in-means estimator, defined for sample sizes $N_1 = \sum_{i=1}^{n} W_i$ and $N_0 = \sum_{i=1}^{n}(1 - W_i)$ as

$$\hat{\tau}_{\mathrm{DM}} = \frac{1}{N_1}\sum_{i=1}^{n} W_i Y_i - \frac{1}{N_0}\sum_{i=1}^{n}(1 - W_i)Y_i, \tag{3.4}$$

is biased under Model 3.1. The mean treated response is

$$\mathbf{E}[Y_i|W_i = 1] = \alpha + \gamma + \delta\mathbf{E}[X_i|W_i = 1] = \alpha + \gamma + \delta\mathbf{E}[X_i],$$

where X_i is independent of W_i since the treatments are assigned independently and there are no self-loops in G. Similarly,

$$\mathbf{E}[Y_i|W_i = 0] = \alpha + \delta\mathbf{E}[X_i|W_i = 0] = \alpha + \delta\mathbf{E}[X_i].$$

Therefore, the difference-in-means estimator $\hat{\tau}_{\mathrm{DM}}$ has expectation γ, which need not equal $\tau = \gamma + \delta$ in general. Only if $\delta = 0$ do they coincide, in which case SUTVA holds and there is no interference. In other words, the difference-in-means estimator marginalizes out the indirect effect rather than adjusting for it; it is an unbiased estimator not for the GATE but for a version of the direct effect known as the *average distributional shift effect* (ADSE), defined as

$$\frac{1}{n}\sum_{i=1}^{n}[\mathbf{E}[Y_i|W_i = 1] - \mathbf{E}[Y_i|W_i = 0]].$$

The ADSE is a version of the direct effect that was introduced in Sävje et al. (2017) as a natural object of study for estimators which are designed for the SUTVA setting. Sävje et al. (2017); Chin (2018a) study the limiting behavior of estimators such as $\hat{\tau}_{\mathrm{DM}}$ under mild regimes of misspecification of SUTVA due to interference; this is the main focus of Chapter 2 of this thesis.

3.2.2 Linear-in-means with endogenous effects

Now we move to the more interesting version of the linear-in-means model, which contains an endogenous social effect in addition to an exogenous one. Let

$$Z_i = \frac{1}{d_i}\sum_{j\in\mathcal{N}_i} Y_j, \tag{3.5}$$

the average value of the neighboring responses. Now consider the following model:

Model 3.2(a) (LIM with endogenous social effect)**.**

$$\mu_i(\mathbf{W}) = \alpha + \beta Z_i + \gamma W_i + \delta X_i.$$

In addition to direct and exogenous spillover effects, unit i now depends on the outcomes of its neighbors through the spillover effect β. It is conventional and reasonable to assume that $|\beta| < 1$. Model 3.2(a) is often more realistic than Model 3.1; as discussed in the introduction, we often believe that interference is caused by individuals reacting to their peers' behaviors rather than to their peers' treatment assignments.

It is helpful to write Model 3.2(a) in vector-matrix form. Let $\tilde{G}$ be the weighted graph defined by degree-normalizing the adjacency matrix of G; i.e., let $\tilde{G}$ be the graph corresponding to the adjacency matrix $\tilde{A}$ with entries $\tilde{A}_{ij} = d_i^{-1} A_{ij}$. Then the matrix representation of Model 3.2(a) is

$$Y = \alpha + \beta \tilde{A} Y + \gamma W + \delta \tilde{A} W + \varepsilon, \tag{3.6}$$

where Y, W, and ε are the n-vectors of responses, treatment assignments, and residuals, respectively. Using the matrix identity $(I - \beta \tilde{A})^{-1} = \sum_{k=0}^{\infty} \beta^k \tilde{A}^k$, as in equation (6) of Bramoullé et al. (2009), one obtains the reduced form

$$Y = \frac{\alpha}{1-\beta} + \gamma W + (\gamma\beta + \delta) \sum_{k=0}^{\infty} \beta^k \tilde{A}^{k+1} W + \sum_{k=0}^{\infty} \beta^k \tilde{A}^k \varepsilon.$$

Unlike Manski (1993); Bramoullé et al. (2009) and other works in the "reflection problem" literature, we are not concerned with the identification of the social effect parameters β and δ; these are only nuisance parameters toward the end of estimating τ. We do note, however, that conditions for identifiability are generally mild enough to be satisfied by real-world networks. For example, Bramoullé et al. (2009) show that the parameters in Model 3.2(a) are identified whenever there exist a triple of individuals who are not all pairwise friends with each other; such a triple nearly certainly exists in any networks that we consider.

Now, let $X_{i,k}$ be the i-th coordinate of $\tilde{A}^k W$. That is,

$$\begin{aligned} X_{i,1} &= \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} W_j \\ X_{i,2} &= \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} \frac{1}{d_j} \sum_{k \in \mathcal{N}_j} W_k \\ X_{i,3} &= \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} \frac{1}{d_j} \sum_{k \in \mathcal{N}_j} \frac{1}{d_k} \sum_{\ell \in \mathcal{N}_k} W_\ell, \end{aligned}$$

and in general, for any $k \geq 1$,

$$X_{i,k} = \frac{1}{d_i} \sum_{j_1 \in \mathcal{N}_i} \frac{1}{d_{j_1}} \sum_{j_2 \in \mathcal{N}_{j_1}} \cdots \frac{1}{d_{j_k}} \sum_{j_{k-1} \in \mathcal{N}_{j_{k-1}}} W_{j_k}.$$

Then Model 3.2(a) is the same as

$$Y_i = \tilde{\alpha} + \tilde{\gamma} W_i + \sum_{k=0}^{\infty} \tilde{\beta}_k X_{i,k} + \tilde{\varepsilon}_i, \tag{3.7}$$

where we have reparametrized the coefficients as

$$\begin{aligned} \tilde{\alpha} &= \frac{\alpha}{1-\beta} \\ \tilde{\gamma} &= \gamma \\ \tilde{\beta}_k &= (\gamma\beta + \delta)\beta^k \\ \tilde{\varepsilon} &= \sum_{k=0}^{\infty} \beta^k A^k \varepsilon. \end{aligned}$$

Notice that equation (3.7) respects exogeneity, as

$$\mathbf{E}[\tilde{\varepsilon}|\mathbf{W}] = \sum_{k=0}^{\infty} \beta^k A^k \mathbf{E}[\varepsilon|\mathbf{W}] = 0.$$

Each feature $X_{i,k}$ represents the effect of treatments from units of graph distance k on the response of unit i. Since $|\beta| < 1$, the effects of the terms $\tilde{\beta}_k X_{i,k}$ do not contribute

much to equation (3.7) when k is large. Therefore, for any finite integer K, we may consider approximating Model 3.2(a) with a finite-dimensional model.

Model 3.2(b) (Finite linear-in-means)**.**

$$Y_i = \tilde{\alpha} + \tilde{\gamma} W_i + \sum_{k=0}^{K} \tilde{\beta}_k X_{i,k} + \tilde{\varepsilon}_i, \tag{3.8}$$

The approximation error is of order $\tilde{\beta}^{k+1} = (\gamma\beta + \delta)\beta^{K+1}$ (recall that $|\beta| < 1$). Therefore, good estimates of the coefficients in equation (3.8) should be good estimates of the coefficients in equation (3.7) as well. Unless spillover effects are extremely large, the approximation may be quite good for even small values of K. In fact, it may be reasonable to take equation (3.8) rather than equation (3.7) as the truth where K is no larger than the diameter of the network G, as spillovers for larger distances may not make sense.

As in Model 3.1, we can consider the counterfactuals of interest. If all units are globally exposed to treatment, then $W_i = 1$ and $X_{i,k} = 1$ for all i and k. Similarly, if all units are globally exposed to control, then $W_i = 0$ and $X_{i,k} = 0$ for all i and k. Therefore, by equation (3.7), the estimand τ under Model 3.2(a) is

$$\tau = \tilde{\gamma} + \sum_{k=0}^{\infty} \tilde{\beta}_k,$$

and under Model 3.2(b) it is

$$\tau = \tilde{\gamma} + \sum_{k=0}^{K} \tilde{\beta}_k.$$

Now, since Model 3.2(b) has only $K + 3$ coefficients, given $n > K + 3$ individuals one can estimate the coefficients using, say, ordinary least squares. The treatment effect estimator

$$\hat{\tau} = \hat{\gamma} + \sum_{k=1}^{K} \hat{\beta}_k$$

is then unbiased for τ under Model 3.2(b) and "approximately unbiased" for τ under Model 3.2(a). This discussion is of course quite informal, and we make more formal

arguments in Section 3.3.

One interpretation of the discussion in this section is that an endogeneous social effect in the linear-in-means model manifests as a propogation of exogenous effects through the social network, with the strength of the exogenous effect diminishing as the network distance increases. Therefore, adjusting for the exogenous features within the first few neighborhoods is nearly equivalent to adjusting for the behavior implied by the endogenous social effect.

3.2.3 Model assumptions and exposure models

The statements of the models discussed in this section couple together an interference mechanism restriction with a functional form assumption. It is worth disentangling these assumptions and discussing why it may be sometimes advantageous for the analyst to consider them jointly. First consider Model 3.1. It implies that the interference mechanism is restricted to influence from units only one step away in the graph, and furthermore, that this one-step influence is transmitted only through the statistic X_i. This interference mechanism restriction can be framed in the language of *constant treatment response* (CTR) mappings (Manski, 2013):

$$\mu_i(\mathbf{w}) = \mu_i(\mathbf{w}') \text{ for all } \mathbf{w}, \mathbf{w}' \in \mathcal{W} \text{ such that } w_i = w_i', x_i = x_i'. \tag{3.9}$$

The CTR statement (3.9) is equivalent to specifying an exposure model that the potential outcomes depend only on $\mathbf{W}$ through W_i and X_i. However, it makes no assumptions about the functional form of $\mu_i(\cdot)$, yet Model 3.1 goes further and makes a strong parametric functional form assumption about the response. It is conceptually useful to recognize the different meanings and implications of these assumptions.

One tempting approach, then, might be for an analyst to first consider verifying whether the exposure model holds, using domain knowledge or otherwise. The analyst then separately proceeds to consider appropriate functional forms (and perhaps only if nonparametric estimators exhibit low power). This logic may succeed for simple exposures of the form implied by Model 3.1 but can lead to issues for more complex data-generating processes likely to be encountered in the real world.

This is made clear by the discussion of Models 3.2(a) and 3.2(b). By rewriting Model 3.2(a) as the infinite series given by equation (3.7), we find that there is no data-reducing exposure model or CTR assumption that can handle such endogenous social effects! This is discouraging unless the analyst jointly considers the parametric implications of an endogenous effect $|\beta| < 1$, which suggests a way forward via the finite approximation Model 3.2(b). Even if the linearity in equation (3.8) is too strong, a natural relaxation might be a kind of generalized additive model of the form

$$Y_i = \alpha + \gamma W_i + \sum_{k=0}^{K} f(k) g(X_{i,k}) + \varepsilon_i,$$

where g is arbitrary but f is restricted to be decreasing in k in order to ensure that spillovers decrease in graph distance.

Furthermore, statisticians are well-versed in distinguishing and handling modeling violations of the mean function (here, corresponding to the interference function form) and the covariance function (corresponding to the interference restriction assumption), whereas statements like (3.9) may be a bit more abstruse for the practicing statistician. The implications here are further clarified by the discussions in the following sections as well as the simulation examples provided in Section 4.7.

3.3 Interference features and the general linear model

In Section 3.2, we showed that the mean function in the linear-in-means model is comprised of a linear combination of statistics $X_{i,k}$ which are constructed as functions of the treatment vector. This fact suggests extending our approach to a linear model containing other functions of the treatment vector that are correlated with Y_i, not just the ones implied by the linear-in-means model. We now formulate the general linear model. We suppose that each unit i is associated with a p-dimensional vector of *interference features* or *interference statistics* $X_i \in \mathbb{R}^p$ that inform the pattern of interference for unit i. We assume that the X_i are low-dimensional ($p \ll n$). Because X_i is to be used for adjustment, the main requirement is that it not be a "post-treatment variable"; that is, that it not be correlated with the treatment W_i.

Therefore, we require the following assumption:

Assumption 3.2. *$X_i \perp\!\!\!\perp W_i$ for all $i \in [n]$.*

Let $\mathbf{W}_{-i}$ denote the vector of *indirect treatments*, which is the $n-1$ vector of all treatments except for W_i. The key feature of our approach is that even though X_i must be independent of W_i, *it is not necessary that X_i be independent of the vector of indirect treatments $\mathbf{W}_{-i}$.* In fact, in order for X_i to be useful for adjusting for interference, we expect that X_i will be correlated with some entries of $\mathbf{W}_{-i}$. In particular, X_i may be a deterministic function $x_i(\cdot)$ of the indirect treatments,

$$X_i = x_i(\mathbf{W}_{-i}). \tag{3.10}$$

Adjusting for such a variable X_i will not cause post-treatment adjustment bias as long as the entries of $\mathbf{W}$ are independent of each other. This holds automatically in a Bernoulli randomized design (Assumption 3.1).

The features X_i may depend on static structural information about the units such as network information provided by G, though since G is static we supress this dependence in the notation. For example, X_i defined as in equation (3.3), which represents the proportion of treated neighbors, captures a particular form of exogenous social influence. Provided there are no self-loops in G so that $A_{ii} = 0$, W_i does not appear on the right-hand side of equation (3.3) and so X_i and W_i are independent.

We assume that we can easily sample from the distribution of X_i. In particular, if $X_i = x_i(\mathbf{W}_{-i})$, then the distribution of X_i can be constructed by Monte Carlo sampling from the randomization distribution of the treatment $\mathbf{W}$. In this work and in all the examples we use, we assume that X_i is a function of $\mathbf{W}_{-i}$ as in equation (3.10), so that conditioning on $\mathbf{W}_{-i}$ removes all randomness in X_i. But the generalization is easily handled.

In this section we assume that the response is linear in X_i; we address nonparametric response surfaces in Section 3.4.

Model 3.3 (Linear model)**.** *Given X_i, let the response Y_i follow*

$$Y_i = W_i\mu^{(1)}(X_i) + (1 - W_i)\mu^{(0)}(X_i) + \varepsilon_i,$$

where the conditional response surfaces

$$\mu^{(0)}(x) = \mathbf{E}[Y_i^{(0)}|X = x], \qquad \mu^{(1)}(x) = \mathbf{E}[Y_i^{(1)}|X = x]$$

satisfy

$$\mu^{(0)}(x) = \beta_0^\top x, \qquad \mu^{(1)}(x) = \beta_1^\top x$$

for $x \in \mathbb{R}^p$ and $\beta_0, \beta_1 \in \mathbb{R}^p$. That is, they follow a "separate slopes" linear model in X_i. We assume $p < n$.

In the above parametrization, we assume that the first coordinate of each X_i is set to 1, so that the vectors β_0 and β_1 contain coefficients corresponding to the intercept as in the classical OLS formulation.

3.3.1 Feature engineering

Before considering assumptions on the residuals ε_i, we pause here to emphasize the flexibility provided by modeling the interference pattern as in Model 3.3. In this framework, the researcher can use domain knowledge to construct graph features that are expected to contribute to interference. In essence, we have transformed the problem of determining the structure of the interference pattern into a feature engineering problem, which is perhaps a more intuitive and accessible task for the practitioner.

To elaborate, consider the problem of selecting an exposure model. Ugander et al. (2013) propose and study a number of different exposure models for targeting the global treatment effect, including *fractional exposure* (based on the fraction of treated neighbors), *absolute exposure* (based on the raw number of treated neighbors), and extensions based on the k-core structure of the network. In reality, it may be the case that fractional and absolute exposure both contribute partial effects of interference, so ideally one wishes to avoid having to choose between one of the two exposure models. On the other hand, both features are easily included in Model 3.3 by encoding both the fraction and raw number of treated neighbors in X_i. (Including

both features only makes sense when working with a complex network. If the interference structure is comprised of large, disjoint, and equally-sized clusters, as in partial interference, then the fraction and number of treated neighbors encode roughly the same information and one obtains a collinearity scenario that violates the full-rank assumption of Proposition 3.1. The methods in this work are primarily motivated by the complex network setting.)

In a similar manner, the researcher may wish to handle longer-range interference, such as that coming from two-step or greater neighborhoods. It is possible to handle two-step information by working with the graph corresponding to the adjacency matrix A^2, but this approach is unsatisfactory because presumably one-step interference is stronger than two-step interference, and this distinction is lost by using A^2. On the other hand, if one-step and two-step network information are encoded as separate features, both effects are included and the magnitudes of their coefficients will reflect the strength of the corresponding interference contributed by each feature.

Furthermore, nothing in our framework requires the variables to be constructed from a single network. Often, the researcher has access to multiple networks defined on the same vertex set—i.e., a *multilayer network* (Kivelä et al., 2014)—representing different types of interactions among the units. For example, social networking sites such as Facebook and Twitter contain multiple friendship or follower networks based on the strength and type of interpersonal relationship (e.g. family, colleagues, and acquaintances), as well as activity-based networks constructed from event data such as posts, tweets, likes, or comments. Often these networks are also dynamic in time. Given the sociological phenomenon that the strength of a tie is an indicator of its capacity for social influence (Granovetter, 1973) and that people use different mediums differently when communicating online (Haythornthwaite and Wellman, 1998), any or all of these network layers can conceivably be a medium for interference in varying amounts depending on the treatment variable and outcome metric in question. In our framework graph features from different network layers are easily included in the model.

3.3.2 Exogeneity assumptions

Consider the following assumptions on the residuals.

Assumption 3.3. *(a) The errors are strictly exogenous:* $\mathbf{E}[\varepsilon_i | X_1, \ldots, X_n] = 0$ *for all* $i \in [n]$.

(b) The errors are independent.

(c) The errors are homoscedastic: $\mathrm{Var}(\varepsilon_i | X_1, \ldots, X_n) = \sigma^2$ *for all* $i \in [n]$.

Assumption 3.3(a) captures the requirement that the features contain all of the information needed to adjust for the bias contributed by interference, and thus is similar to an unconfoundedness or ignorability assumption often invoked in observational studies. Point estimates can be constructed based only on Assumption 3(a), but variance estimation requires Assumption 3.3(b) so that each data point contributes additional independent information. Note that in the SUTVA case Assumption 3.3(a) is all that is needed for valid inference. However under interference, it is possible that conditioning on the features removes all bias but interference is still present in the errors, in which case i.i.d.-based standard errors would be incorrect. These assumptions cannot be verified from the data, and so this setup borrows all of the problems that come with selecting an exposure model or being able to verify unconfoundedness. However, our setup is slightly different because of the flexibility afforded by the features. Compared to what we envision as the usual observational studies setting, our features are constructed from the treatment vector and social network rather than being collected in the wild, and so they are quite cheap to construct via feature engineering. That said, more work for conducting sensitivity analysis for interference or spillover effects is certainly needed.

Assumption 3.3(c) is the easiest to deal with if violated. One may use a heteroscedisticity-consistent estimate of the covariance matrix, also known as the sandwich estimator or the Eicker-Huber-White estimator (Eicker, 1967; Huber, 1967; White, 1980). In this chapter we invoke Assumption 3.3(c) mainly to simplify notation, but heteroscedasticity-robust extensions are straightforward.

For X_i following equation (3.10), denote

$$X_i^{(\mathbf{0})} = x_i(\mathbf{W}_{-i} = \mathbf{0}), \qquad X_i^{(\mathbf{1})} = x_i(\mathbf{W}_{-i} = \mathbf{1}).$$

The variable $X_i^{(\mathbf{0})}$ represents the context for unit i under the counterfactual scenario that i is exposed to global control, and the variable $X_i^{(\mathbf{1})}$ represents the context for unit i under the counterfactual scenario that i exposed to global treatment. Both of these values are non-deterministic.[1] For example, if X_i be the "mean treated" statistic as defined in equation (3.3), then $X_i^{(\mathbf{0})} = 0$ and $X_i^{(\mathbf{1})} = 1$ for every unit $i \in [n]$.

We now consider the estimand under Model 3.3 and Assumption 3.3. The GATE for Model 3.3 is

$$\begin{aligned}
\tau &= \frac{1}{n}\sum_{i=1}^{n} \left[\mathbf{E}[Y_i|\mathbf{W} = \mathbf{1}] - \mathbf{E}[Y_i|\mathbf{W} = \mathbf{0}]\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \left[\mu^{(1)}(X_i^{(\mathbf{1})}) - \mu^{(0)}(X_i^{(\mathbf{0})})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \left[(X_i^{(\mathbf{1})})^\top \beta_1 - (X_i^{(\mathbf{0})})^\top \beta_0\right],
\end{aligned}$$

where the second equality is by Assumption 3.3(a). Now introduce the quantities

$$\omega_0 = \frac{1}{n}\sum_{i=1}^{n} X_i^{(\mathbf{0})}, \qquad \omega_1 = \frac{1}{n}\sum_{i=1}^{n} X_i^{(\mathbf{1})},$$

which are the mean counterfactual feature values for global control and global treatment, averaged over the population. We emphasize that ω_0 and ω_1 are non-deterministic and known, because the distribution of X_i is assumed to be known. We then have

$$\tau = \omega_1^\top \beta_1 - \omega_0^\top \beta_0. \tag{3.11}$$

[1] In the event X_i are not defined through a function $x_i(\cdot)$, one may work with $X_i^{(\mathbf{0})} = X_i|(\mathbf{W}_{-i} = \mathbf{0})$ and $X_i^{(\mathbf{1})} = X_i|(\mathbf{W}_{-i} = \mathbf{1})$, where this notation means that $X_i^{(\mathbf{0})}$ follows the conditional distribution of X_i, conditionally on the event that $\mathbf{W}_{-i} = \mathbf{0}$, and similarly for $X_i^{(\mathbf{1})}$. In this case $X_i^{(\mathbf{0})}$ and $X_i^{(\mathbf{1})}$ may be random, and estimands can be defined using $\mathbf{E}[X_i^{(\mathbf{0})}]$ and $\mathbf{E}[X_i^{(\mathbf{1})}]$ instead.

Such an estimand, which focuses on the statistics of the finite population at hand, is natural in the network setting where there is no clear superpopulation or larger network of interest.

We now construct an estimator by estimating the regression coefficients with ordinary least squares. For $w = 0, 1$, let X_w be the $N_w \times p$ design matrix corresponding to features belonging to treatment group w, where the first column of X_w is a column of ones. Let y_w be the N_w-vector of observed responses Y_i for treatment group w. Then we use the standard OLS estimator

$$\hat{\beta}_w = (X_w^\top X_w)^{-1} X_w^\top y_w. \tag{3.12}$$

The estimate of the treatment effect is taken to be the difference in mean predicted outcomes under the global treatment and control counterfactual distributions,

$$\hat{\tau} = \omega_1^\top \hat{\beta}_1 - \omega_0^\top \hat{\beta}_0. \tag{3.13}$$

Assuming Model 3.3 holds, $\hat{\tau}$ is an unbiased estimate of τ, which follows from unbiasedness of the OLS coefficients.

Proposition 3.1. *Suppose Model 3.3 and Assumptions 3.1, 3.2, and 3.3(a) hold. Let τ and $\hat{\tau}$ be defined as in equations* (3.11) *and* (3.13)*, and let $\hat{\beta}_w$ for $w = 0, 1$ be OLS estimators as defined in equation* (3.12)*. Then conditionally on X_w being full (column) rank,*[2] *$\hat{\beta}_w$ is an unbiased estimator of β_w and $\hat{\tau}$ is an unbiased estimator of τ.*

Proof. Let ε_w be the N_w vector of w-group residuals. As $y_w = X_w\beta_w + \varepsilon_w$, for $w = 0, 1$,

[2] Since X_w is random and depends on $\mathbf{W}$, conditioning on X_w having full column rank is necessary, even though this condition may not be fulfilled for all realizations of the treatment vector. For example, if X_w contains a column for the fraction of neighbors treated, then it is possible though highly unlikely for all units to be assigned to treatment, in which case this column is collinear with the intercept and X_w is not full rank. We shall, for the most part, ignore this technicality and assume that the features are chosen so that the event that $X_w^\top X_w$ is singular doesn't happen very often, and is in fact negligible asymptotically. Understanding combinations of network structures and interference mechanisms that give rise to singular X_w is of interest to practitioners but outside the scope of our study here.

conditionally on X_w being full rank we have

$$\begin{aligned}\mathbf{E}[\hat{\beta}_w] &= \mathbf{E}[(X_w^\top X_w)^{-1} X_w^\top y_w] \\ &= \mathbf{E}[(X_w^\top X_w)^{-1} X_w^\top (X_w \beta_w + \varepsilon_w)] \\ &= \beta_w + \mathbf{E}[(X_w^\top X_w)^{-1} X_w^\top \varepsilon_w].\end{aligned}$$

Assumption 3.3(a) ensures that the second term is zero, and thus $\hat{\beta}_w$ is unbiased for β_w.

Unbiasedness of $\hat{\tau}$ then follows by linearity of expectation. □

Notice that the treatment group predicted mean is

$$\omega_1^\top \hat{\beta}_1 = \omega_1^\top (X_1^\top X_1)^{-1} X_1^\top y_1$$

and the control group predicted mean is

$$\omega_0^\top \hat{\beta}_0 = \omega_0^\top (X_0^\top X_0)^{-1} X_0^\top y_0.$$

Therefore $\hat{\tau}$ is linear in the observed response vector y. That is, $\tau = a_0^\top y_0 + a_1^\top y_1$ where the weight vectors $a_0 \in \mathbb{R}^{N_0}$ and $a_1 \in \mathbb{R}^{N_1}$ are given by

$$a_0^\top = \omega_0^\top (X_0^\top X_0)^{-1} X_0^\top \tag{3.14}$$

$$a_1^\top = \omega_1^\top (X_1^\top X_1)^{-1} X_1^\top. \tag{3.15}$$

These weights allow us to compare the reweighting strategy with that of other linear estimators, such as the Hájek estimator, which is a particular weighted mean of y. More details are provided in Section 3.5, with an example given in Section 3.6.2.

3.3.3 Inference

Now we provide variance expressions under the assumption that the errors are exogenous, independent, and homoscedastic, as in Assumption 3.3.

Theorem 3.1. *Suppose Model 3.3 and Assumptions 3.1, 3.2, and 3.3 hold. Then*

$$\mathrm{Var}(\hat{\tau}) = \sigma^2(\|\omega_0\|^2_{\Gamma_0} + \|\omega_1\|^2_{\Gamma_1}), \tag{3.16}$$

where $\|v\|^2_M = v^\top M v$*, and* $\Gamma_w = \mathbf{E}[(X_w^\top X_w)^{-1}]$*, and* ω_w *is the mean of the counterfactual feature distribution (including an intercept) for* $w = 0, 1$*.*

Proof. We first calculate the variance of $\hat{\beta}_w$. By the law of total variance, we have

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_w) &= \mathrm{Var}[(X_w^\top X_w)^{-1} X_w^\top y_w] \\
&= \mathrm{Var}[(X_w^\top X_w)^{-1} X_w^\top \varepsilon_w] \\
&= \mathbf{E}[(X_w^\top X_w)^{-1} X_w^\top \mathrm{Var}(\varepsilon_w | X_w) X_w (X_w^\top X_w)^{-1}] + \mathrm{Var}[(X_w^\top X_w)^{-1} X_w^\top \mathbf{E}(\varepsilon_w | X_w)].
\end{aligned}$$

The second term is equal to zero by Assumption 3.3(a), and so by Assumption 3.3(b) and (c),

$$\mathrm{Var}(\hat{\beta}_w) = \sigma^2 \mathbf{E}[(X_w^\top X_w)^{-1}].$$

The coefficient estimates of the two groups are uncorrelated because the residuals are uncorrelated. That is,

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= \mathbf{E}[\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1 | X)] + \mathrm{Cov}(\mathbf{E}[\hat{\beta}_0 | X], \mathbf{E}[\hat{\beta}_1 | X]) \\
&= \mathbf{E}[\mathrm{Cov}((X_0^\top X_0)^{-1} X_0^\top \varepsilon_0, (X_1^\top X_1)^{-1} X_1^\top \varepsilon_1)] + 0 \\
&= 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Var}(\hat{\tau}) &= \mathrm{Var}((\omega_1)^\top \hat{\beta}_1 - (\omega_0)^\top \hat{\beta}_0) \\
&= \sigma^2 \left((\omega_0)^\top \mathbf{E}[(X_0^\top X_0)^{-1}] \omega_0 + (\omega_1)^\top \mathbf{E}[(X_1^\top X_1)^{-1}] \omega_1 \right),
\end{aligned}$$

which produces the variance expression in equation (3.16). □

Algorithm 1 Estimating Γ_0 and Γ_1 by Monte Carlo

for b = 1:B **do**

Sample treatment $\mathbf{W}_b \in \mathcal{W}$ and compute corresponding features $X_{b,i}$ and sample sizes $N_{b,0}$ and $N_{b,1}$

Calculate sample covariances

$$(\tilde{X}_0^\top \tilde{X}_0)_b \leftarrow \frac{1}{N_{b,0}} \sum_{i=1}^{n} (1 - W_{b,i}) X_{b,i} X_{b,i}^\top$$

$$(\tilde{X}_1^\top \tilde{X}_1)_b \leftarrow \frac{1}{N_{b,1}} \sum_{i=1}^{n} W_{b,i} X_{b,i} X_{b,i}^\top$$

end for

return Moment estimates

$$\hat{\Gamma}_w \leftarrow \widehat{\mathbf{E}}[(\tilde{X}_w^\top \tilde{X}_w)^{-1}] = \frac{1}{B} \sum_{b=1}^{B} (\tilde{X}_w^\top \tilde{X}_w)_b^{-1}$$

for $w = 0, 1$.

Variance estimation

In order to estimate the variance (3.16), we must estimate the quantities $\Gamma_0 = \mathbf{E}[(X_0^\top X_0)^{-1}]$ and $\Gamma_1 = \mathbf{E}[(X_1^\top X_1)^{-1}]$, which are the expected inverse sample covariance matrices. Of course, $(X_0^\top X_0)^{-1}$ and $(X_1^\top X_1)^{-1}$ are observed and unbiased estimators. However, unlike standard baseline characteristics collected in the wild, we envision that the X_i are constructed from the graph G and the treatment vector $\mathbf{W}$, and so we can take advantage of the fact that the distribution of X_i is completely known to the researcher. It is thus possible to compute Γ_0 and Γ_1 up to arbitrary precision by repeated Monte Carlo sampling from the randomization distribution of $\mathbf{W}$. For clarity, this estimation procedure is illustrated in Algorithm 1.

Finally, we can estimate σ^2 in the usual way, with the residual mean squared error

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - W_i(\hat{\beta}_1^\top X_i) - (1 - W_i)(\hat{\beta}_0^\top X_i) \right)^2.$$

Equipped with $\hat{\sigma}^2$ and Monte Carlo estimates $\hat{\Gamma}_w$, we can use the variance estimate

$$\widehat{\mathrm{Var}}(\hat{\tau}) = \hat{\sigma}^2 \left(\|\omega_0\|^2_{\hat{\Gamma}_0} + \|\omega_1\|^2_{\hat{\Gamma}_1} \right). \tag{3.17}$$

3.3.4 Asymptotic results

Proposition 3.1 and Theorem 1 characterize the finite n expectation and variance of the treatment effect estimator under Model 4. Establishing an asymptotic result is more nuanced, as because of the dependence among units implied by interference, the quantities $\mathbf{E}[(X_0^\top X_0)^{-1}]$ and $\mathbf{E}[(X_1^\top X_1)^{-1}]$ may not be $O(n^{-1})$ in which case $\hat{\tau}$ would not converge at a $\sqrt{n}$ rate. This is a problem with dealing with interference in general, making comparisons to the semiparametric efficiency bound (Hahn, 1998), a standard benchmark in the SUTVA case, difficult in this setting. However we can state a $\sqrt{n}$ central limit theorem in the event that the sample mean and covariance do scale and converge appropriately. To do so, we implicitly assume existence of a sequence of populations indexed by their size n, and that the parameters associated with each population setup, such as β_0, β_1, π, and σ^2, converge to appropriate limits. Such an asymptotic regime is the standard for results of this sort (cf. Freedman, 2008a,b; Lin, 2013; Abadie et al., 2017a,b; Sävje et al., 2017; Chin, 2018a). We suppress the index on n to avoid notational clutter.

So that we can compare to previous works, it is helpful to reparametrize the linear regression setup so that the intercept and slope coefficients are written separately. That is, let X_i and ω_w be redefined to exclude the intercept, and let $\beta_w = (\alpha_w, \eta_w)$ so that the mean functions are written $\mu^{(w)}(x) = \alpha_w + \eta_w^\top x$, where α_w is the intercept parameter and η_w is the vector of slope coefficients. Then the GATE is

$$\tau = (\alpha_1 + \omega_1^\top \eta_1) - (\alpha_0 + \omega_0^\top \eta_0).$$

Denote the within-group sample averages by

$$\bar{y}_1 = \frac{1}{N_1} \sum_{i=1}^{n} W_i Y_i, \qquad \bar{y}_0 = \frac{1}{N_0} \sum_{i=1}^{n} (1 - W_i) Y_i$$

and

$$\bar{X}_0 = \frac{1}{N_1}\sum_{i=1}^{n} W_i X_i, \qquad \bar{X}_0 = \frac{1}{N_0}\sum_{i=1}^{n}(1-W_i)X_i.$$

Since the intercept is determined by $\hat{\alpha}_w = \bar{y}_w - \bar{X}_w^\top \hat{\eta}_w$, the estimator $\hat{\tau}$, equation (3.13), is written as

$$\begin{aligned}\hat{\tau} &= (\hat{\alpha}_1 + \omega_1^\top \hat{\eta}_1) - (\hat{\alpha}_0 + \omega_0^\top \hat{\eta}_0). \\ &= \bar{y}_1 - \bar{y}_0 + (\omega_1 - \bar{X}_1)^\top \hat{\eta}_1 - (\omega_0 - \bar{X}_0)^\top \hat{\eta}_0. \end{aligned} \tag{3.18}$$

Now, $\hat{\tau}$ is seen to be an adjustment of the difference-in-means estimator $\bar{y}_1 - \bar{y}_0$. The adjustment depends on both the estimated strength of interference, $\hat{\eta}_w$, and the discrepancy between the means of the observed distribution and the reference or target distribution, $\bar{X}_w - \omega_w$. This linear shift is a motif in the regression adjustment literature, and is reminiscent of, e.g., equation (16) of Aronow and Middleton (2013).

We now state a central limit theorem for $\hat{\tau}$.

Theorem 3.2. *Assume the setup of Theorem 3.1. Assume further that the sample moments converge in probability:*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu_X,$$

$$S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^\top (X_i - \bar{X}) \xrightarrow{p} \Sigma_X,$$

where Σ_X is positive definite, and that all fourth moments are bounded. Then $\sqrt{n}(\hat{\tau} - \tau) \Rightarrow N(0, V)$, where

$$V = \sigma^2 \left(\frac{1}{\pi(1-\pi)} + \frac{\|\omega_0 - \mu_X\|^2_{\Sigma_X^{-1}}}{1-\pi} + \frac{\|\omega_1 - \mu_X\|^2_{\Sigma_X^{-1}}}{\pi} \right). \tag{3.19}$$

The terms in expression (3.19) are unpacked versions of the terms in expression (3.16), and can be stated in this way since the feature moments converge at the appropriate rate.

In order to prove the above statement, we first establish the following lemma, which provides some basic convergence results.

Lemma 3.1. *Let $\bar{X}_w$ and S_w denote the within-group sample means and covariances. Under Assumptions 3.1, 3.2, and the assumptions in the statement of Theorem 3.2, the following statements hold for $w = 0, 1$.*

(a) $\bar{X}_w \xrightarrow{p} \mu_X$.

(b) $S_w \xrightarrow{p} \Sigma_X$.

(c) $\hat{\eta}_w \xrightarrow{p} \eta_w$.

(d) $\sqrt{n\pi}(\bar{X}_1 - \mu_X) \Rightarrow N(0, \Sigma_X)$ *and* $\sqrt{n(1-\pi)}(\bar{X}_0 - \mu_X) \Rightarrow N(0, \Sigma_X)$.

(e) $\sqrt{n\pi}(\hat{\eta}_1 - \eta_1) \Rightarrow N(0, \sigma^2\Sigma_X^{-1})$ *and* $\sqrt{n(1-\pi)}(\hat{\eta}_0 - \eta_0) \Rightarrow N(0, \sigma^2\Sigma_X^{-1})$.

(f) $\sqrt{n}(\bar{\varepsilon}_1 - \bar{\varepsilon}_0) \Rightarrow N\left(0, \frac{\sigma^2}{\pi(1-\pi)}\right)$.

Proof. (a) Because of Bernoulli random sampling it holds that

$$\lim_{n\to\infty} \mathbf{E}[\bar{X}_1] = \lim_{n\to\infty} \mathbf{E}\left[\frac{1}{N_1}\sum_{i=1}^{n} W_i X_i\right] = \mu_X.$$

By conditioning on X we have

$$\mathrm{Var}(\bar{X}_1) = \mathbf{E}[\mathrm{Var}(\bar{X}_1|X)] + \mathrm{Var}[\mathbf{E}(\bar{X}_1|X)].$$

For the first term, we have

$$\mathbf{E}[\mathrm{Var}(\bar{X}_1|X)] = \mathbf{E}\left[\mathrm{Var}\left(\frac{1}{n\pi}\sum_{i=1}^{n} W_i X_i + r_n\right)\right],$$

where

$$\mathrm{Var}\left(\frac{1}{n\pi}\sum_{i=1}^{n} W_i X_i\right) = \frac{1-\pi}{n^2\pi}\sum_{i=1}^{n} X_i^2 = O_p(n^{-1})$$

and

$$r_n = \left(\frac{1}{N_1} - \frac{1}{np}\right) \sum_{i=1}^{n} W_i X_i = O_p(n^{-1})$$

since $N_1/n \to \pi$ in probability. For the second term, we have

$$\text{Var}[\mathbf{E}(\bar{X}_1|X] = \text{Var}(\bar{X}) \to 0$$

since $\bar{X} - \mu_X = o_p(1)$. Therefore, we conclude $\text{Var}(\bar{X}_1) \to 0$, and so consistency follows from Chebychev's inequality.

The result similarly holds for $\bar{X}_0$.

(b) This result is established in a similar manner to part (a), using the fact that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^\top (X_i - \bar{X}) \xrightarrow{p} \Sigma_X,$$

and the fact that fourth moments are bounded.

(c) The convergence of $\hat{\eta}_w$ to η_w follows conditionally on X from standard OLS theory. Then, letting

$$S_w = \frac{1}{n}(X_w - \bar{X}_w)^\top (X_w - \bar{X}_w)$$

denote the sample covariance matrix, we find

$$\begin{aligned}\text{Var}(\hat{\eta}) &= \text{Var}[\mathbf{E}[\hat{\eta}_w|X]] + \mathbf{E}[\text{Var}[\hat{\eta}_w|X]] \\ &= \text{Var}[\eta_w] + \frac{\sigma^2}{n}\mathbf{E}[S_w^{-1}] \to 0.\end{aligned}$$

Convergence in probability follows from Chebychev's inequality.

(d) This result follows from Bernoulli sampling and the convergence of the finite population means, $\bar{X} \xrightarrow{p} \mu_X$.

(e) As in the proof of part (c), we write

$$\hat{\eta}_w = \frac{1}{n} S_w^{-1}(X_w - \bar{X}_w)^\top (y_w - \bar{y}_w).$$

Since $y_w = X_w\eta_w + \varepsilon_w$, we can write

$$\begin{aligned}\sqrt{n}(\hat{\eta}_w - \eta_w) &= \sqrt{n}\left[\frac{1}{n}S_w^{-1}(X_w - \bar{X}_w)^\top(y_w - \bar{y}_w) - \eta_w\right] \\ &= \frac{1}{\sqrt{n}}S_w^{-1}(X_w - \bar{X}_w)^\top(\varepsilon_w - \bar{\varepsilon}_w) \\ &= \frac{1}{\sqrt{n}}\Sigma_X^{-1}(X_w - \bar{X}_w)^\top(\varepsilon_w - \bar{\varepsilon}_w) + R,\end{aligned}$$

where the remainder is

$$R = \frac{1}{\sqrt{n}}(S_w^{-1} - \Sigma_X^{-1})(X_w - \bar{X}_w)^\top(\varepsilon_w - \bar{\varepsilon}_w)$$

Since $S_w^{-1} - \Sigma_X^{-1} = o_p(1)$ is implied by $S_w \xrightarrow{p} \Sigma_X$, and $\sqrt{n}(X_w - \bar{X}_w)^\top = O_p(1)$ and $\sqrt{n}(\varepsilon_w - \bar{\varepsilon}_w) = O_p(1)$, the remainder satisfies $R = o_p(1)$.

Then $\sqrt{n}(\hat{\eta}_w - \eta_w)$ is asymptotically Gaussian with mean zero and variance

$$\lim_{n\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{n}}\Sigma_X^{-1}(X_w - \bar{X}_w)^\top(\varepsilon_w - \bar{\varepsilon}_w)\right) = \sigma^2\Sigma_X^{-1}\lim_{n\to\infty}\mathrm{Var}(X_w)\Sigma_X^{-1}.$$

Using the result of part (d), this variance equals $\frac{\sigma^2}{\pi}\Sigma_X^{-1}\Sigma_X\Sigma_X^{-1} = \frac{\sigma^2}{\pi}\Sigma_X^{-1}$ when $w = 1$ and $\frac{\sigma^2}{1-\pi}\Sigma_X^{-1}$ when $w = 0$.

(f) From Assumption 3.3, $\bar{\varepsilon}_1$ is independent of $\bar{\varepsilon}_0$ with variances $\sigma^2/(n\pi)$ and $\sigma^2/(n(1-\pi))$, respectively. A standard central limit theorem shows that $\sqrt{n}(\bar{\varepsilon}_1 - \bar{\varepsilon}_0)$ is asymptotically Gaussian with mean 0 and variance

$$\frac{\sigma^2}{\pi} + \frac{\sigma^2}{1-\pi} = \frac{\sigma^2}{\pi(1-\pi)}.$$

□

Equipped with these preliminary results we now proceed to the proof of Theorem 3.2.

Proof. We characterize the treatment effect estimator as

$$\begin{aligned}\hat{\tau} - \tau &= \bar{y}_1 - \bar{y}_0 + (\omega_1 - \bar{X}_1)^\top \hat{\eta}_1 - (\omega_0 - \bar{X}_0)^\top \hat{\eta}_0 - (\alpha_1 - \alpha_0) - (\omega_1^\top \eta_1 - \omega_0^\top \eta_0) \\ &= \bar{\varepsilon}_1 - \bar{\varepsilon}_0 + (\omega_1 - \bar{X}_1)^\top (\hat{\eta}_1 - \eta_1) - (\omega_0 - \bar{X}_0)^\top (\hat{\eta}_0 - \eta_0),\end{aligned}$$

which implies that

$$\sqrt{n}(\hat{\tau} - \tau) = \sqrt{n}(\bar{\varepsilon}_1 - \bar{\varepsilon}_0) + \sqrt{n}(\omega_1 - \bar{X}_1)^\top (\hat{\eta}_1 - \eta_1) - \sqrt{n}(\omega_0 - \bar{X}_0)^\top (\hat{\eta}_0 - \eta_0).$$

Now,

$$\sqrt{n}(\omega_w - \bar{X}_w)^\top (\hat{\eta}_w - \eta_w) = \sqrt{n}(\omega_w - \mu_w)^\top (\hat{\eta}_w - \eta_w) + \sqrt{n}(\mu_w - \bar{X}_w)^\top (\hat{\eta}_w - \eta_w),$$

for $w = 0, 1$, where the second term is $o_p(1)$ since $\bar{X}_w \xrightarrow{p} \mu_X$ and $\hat{\eta}_w \xrightarrow{p} \eta_w$ following from parts (a) and (c) of Lemma 3.1. Therefore,

$$\sqrt{n}(\hat{\tau} - \tau) = \sqrt{n}(\bar{\varepsilon}_1 - \bar{\varepsilon}_0) + \sqrt{n}(\omega_1 - \mu_X)^\top (\hat{\eta}_1 - \eta_1) - \sqrt{n}(\omega_0 - \mu_X)^\top (\hat{\eta}_0 - \eta_0) + o_p(1).$$

The three terms are uncorrelated, with

$$\begin{aligned}\sqrt{n}(\bar{\varepsilon}_1 - \bar{\varepsilon}_0) &\Rightarrow N\left(0, \frac{\sigma^2}{\pi(1-\pi)}\right) \\ \sqrt{n}(\omega_1 - \mu_X)^\top (\hat{\eta}_1 - \eta_1) &\Rightarrow N\left(0, \frac{\sigma^2}{\pi}\|\omega_1 - \mu\|^2_{\Sigma_X^{-1}}\right) \\ \sqrt{n}(\omega_0 - \mu_X)^\top (\hat{\eta}_0 - \eta_0) &\Rightarrow N\left(0, \frac{\sigma^2}{1-\pi}\|\omega_0 - \mu\|^2_{\Sigma_X^{-1}}\right),\end{aligned}$$

established in parts (e) and (f) of Lemma 3.1. Combining the terms produces the variance expression in equation (3.19), and completes the proof. □

3.3.5 Relationship with standard regression adjustments

The practitioner may also wish to perform standard regression adjustments to adjust for static, contextual node-level variables such as age, gender, and other demographic

variables. This fits easily into the framework of Model 4, as any such static variable X_i can be viewed as simply a constant function of the indirect treatment vector $\mathbf{W}_{-i}$. Then the adjustment is not used to remove bias but simply to reduce variance by balancing the feature distributions. In this case the counterfactual (global exposure) distribution is the same as the observed distribution, and in particular, $\omega_0 = \mu_X$ and $\omega_1 = \mu_X$. Hence we see that Theorem 3.2 reduces to the standard asymptotic result for regression adjustments using OLS.

Corollary 3.1. *Assume the setup of Theorem 3.2. Suppose X_i is independent of $\mathbf{W}_{-i}$. Then $\omega_0 = \mu_X$ and $\omega_1 = \mu_X$ and*

$$\sqrt{n}(\hat{\tau} - \tau) \Rightarrow N\left(0, \frac{\sigma^2}{\pi(1-\pi)}\right),$$

Proof. If X_i is independent of $\mathbf{W}_{-i}$, then

$$\omega_0 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[X_i | \mathbf{W}_{-i} = \mathbf{0}] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[X_i],$$

and so is equal to μ_X in the limit (with the understanding that ω_0 is actually a sequence associated with each finite population). The same holds true for ω_1. Then the result follows immediately from equation (3.19), as the second and third terms are equal to zero. □

This variance in Corollary 3.1 is the same asymptotic variance as in the standard regression adjustment setup (cf. Wager et al., 2016, Theorem 2). In practice, if some components of X_i are static covariates and some are interference variables, then the resulting variance will be decomposed into the components stated in Theorem 3.2 and Corollary 3.1. Conditioning on both baseline covariates and interference features may in fact be necessary to ensure that Assumption 3.3(a) holds. For example if X_i is the number of treated neighbors it may be believed that the potential outcomes depend on node degree (in the graph G) as well.

3.4 Nonparametric adjustments

In this section we relax the linear model, Model 3.3:

Model 3.4 (Non-linear response surface)**.** *Let Y_i follow*

$$Y_i = W_i\mu^{(1)}(X_i) + (1 - W_i)\mu^{(0)}(X_i) + \varepsilon_i,$$

with conditional mean response surfaces

$$\mu^{(0)}(x) = \mathbf{E}[Y_i^{(0)}|X = x], \qquad \mu^{(1)}(x) = \mathbf{E}[Y_i^{(1)}|X = x].$$

We make no parametric assumptions on the form of $\mu^{(0)}(x)$ and $\mu^{(1)}(x)$.

We maintain Assumption 3.3, namely that SUTVA holds conditionally on $X_1, \ldots, X_n$.

In the SUTVA setting, adjustment with OLS works best when the adjustment variables are highly correlated with the potential outcomes; that is, the precision improvement largely depends on the prediction accuracy. This fact suggests that predicted outcomes obtained from an arbitrary machine learning model can be used for adjustment, an idea formalized by Wager et al. (2016); Wu and Gagnon-Bartsch (2017). Based on ideas from Aronow and Middleton (2013), these papers propose using the estimator

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{-i}^{(1)}(X_i) - \hat{\mu}_{-i}^{(0)}(X_i)\right) + \frac{1}{N_1}\sum_{i=1}^{n} W_i\left(Y_i - \hat{\mu}_{-i}^{(1)}(X_i)\right) \\ - \frac{1}{N_0}\sum_{i=1}^{n}(1 - W_i)\left(Y_i - \hat{\mu}_{-i}^{(0)}(X_i)\right),\end{aligned} \tag{3.20}$$

where $\hat{\mu}_{-i}^{(0)}$ and $\hat{\mu}_{-i}^{(1)}$ are predictions of the potential outcomes obtained without using the i-th observation. This doubly-robust style approach is called *cross-estimation* by Wager et al. (2016) and the *leave-one-out potential outcomes* (LOOP) estimator by Wu and Gagnon-Bartsch (2017) who focus on imputing the outcomes using a version of leave-one-out cross validation. This estimator is also reminiscent of

the *double machine learning* (DML) cross-fitting estimators developed for the observational study setting (Chernozhukov et al., 2018), which consists of the following two-stage procedure: (a) train predictive machine learning models $\hat{e}(\cdot)$ of X_i on W_i (the propensity model) and $\hat{m}(\cdot)$ of X_i on Y_i (the response model), and then (b) use the out-of-sample residuals $W_i - \hat{e}(X_i)$ and $Y_i - \hat{m}(X_i)$ in a final stage regression. The difference in the experimental setting is that the propensity scores are known and so no propensity model is needed. Wu and Gagnon-Bartsch (2017) study the behavior of (3.20) in the finite population setting where the only randomization comes from the treatment assignment, and Wager et al. (2016) provide asymptotic results for estimating the population average treatment effect. As long as the predicted value $\hat{\mu}_{-i}^{(w)}$ does not use the i-th observation, estimator (3.20) allows us to obtain asymptotically unbiased adjustments and valid inference using machine learning algorithms such as random forests or neural networks. In practice, such predictions are obtained by a cross validation-style procedure in which the data are split into K folds, and the predictions for each fold k are obtained using a model fitted on data from the other $K-1$ folds. (Cross validation on graphs is in general difficult (Chen and Lei, 2018; Li et al., 2018), but our procedure is unrelated to that problem because the features are constructed from the entire graph and fixed beforehand.)

In this section we apply insights from the above works to the interference setting. Under Model 3.4, the global average treatment effect has the form

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \left[\mu^{(1)}(X_i^{(\mathbf{1})}) - \mu^{(0)}(X_i^{(\mathbf{0})}) \right].$$

To develop an estimator of τ, consider the form of the OLS estimator given by equation (3.18), which can be rewritten as

$$\begin{aligned}\hat{\tau} =& \bar{y}_1 - \bar{y}_0 + (\omega_1 - \bar{X}_1)^\top \hat{\eta}_1 - (\omega_0 - \bar{X}_0)^\top \hat{\eta}_0 \\ =& \omega_1^\top \hat{\eta}_1 - \omega_0^\top \hat{\eta}_0 + (\bar{y}_1 - \bar{X}_1^\top \hat{\eta}_1) - (\bar{y}_0 - \bar{X}_0^\top \hat{\eta}_0) \\ =& \frac{1}{n}\sum_{i=1}^{n}\left((X_i^{(\mathbf{1})})^\top \hat{\eta}_1 - (X_i^{(\mathbf{0})})^\top \hat{\eta}_0\right) + \frac{1}{N_1}\sum_{i=1}^{n} W_i\left(Y_i - X_i^\top \hat{\eta}_1\right) \\ & - \frac{1}{N_0}\sum_{i=1}^{n}(1-W_i)\left(Y_i - X_i^\top \hat{\eta}_0\right). \end{aligned} \tag{3.21}$$

Now, by analog, we define the estimator for the nonparametric setting as

$$\begin{aligned}\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{-i}^{(1)}(X_i^{(\mathbf{1})}) - \hat{\mu}_{-i}^{(0)}(X_i^{(\mathbf{0})})\right) + \frac{1}{N_1}\sum_{i=1}^{n} W_i\left(Y_i - \hat{\mu}_{-i}^{(1)}(X_i)\right) \\ - \frac{1}{N_0}\sum_{i=1}^{n}(1-W_i)\left(Y_i - \hat{\mu}_{-i}^{(0)}(X_i)\right). \end{aligned} \tag{3.22}$$

One sees that equations (3.21) and (3.22) agree whenever $\hat{\mu}^{(w)}(x) = \hat{\alpha}_w + x^\top \hat{\eta}_w$. Furthermore, equation (3.22) is equal to its SUTVA version, equation (3.20), whenever $X_i^{(\mathbf{0})} = X_i^{(\mathbf{1})} = X_i$.

Because the units can be arbitrarily connected, the cross-fitting component partitions are not immediately guaranteed to be exactly independent, and so any theoretical guarantees must assume some form of approximate independence of the out-of-sample predictions. In this work we leave such theoretical results open for future work; our primary contribution is the proposal of estimator (3.22) and a bootstrap variance estimation method that respects the empirical structure of interference.

3.4.1 Bootstrap variance estimation

Here we discuss a method for placing error bars on the estimate $\hat{\tau}$ defined in equation (3.22). We propose using a bootstrap estimator to estimate the sampling variance. Under exogeneity (Assumption 3.3), the features and residuals contribute orthogonally to the total variance, and so the model and residuals can be resampled separately.

Instead of using the fixed, observed $X_1, \dots X_n$ as in a standard residual bootstrap, we propose capturing the entire variance induced by the feature distribution by sampling a new X_i from its population distribution for each bootstrap replicate. That is, for each of B bootstrap repetitions, we sample a new treatment vector $\mathbf{W}^b$ and compute bootstrapped features $X_i^b = x_i(\mathbf{W}_{-i}^b)$. The means are then computed using the fitted function as $\hat{\mu}_{-i}^{(0)}(X_i^b)$ and $\hat{\mu}_{-i}^{(1)}(X_i^b)$. Provided that the adjustments are consistent in the sup norm sense, that is, that

$$\sup_x |\hat{\mu}^{(0)}(x) - \mu^{(0)}(x)| \xrightarrow{p} 0, \qquad \sup_x |\hat{\mu}^{(1)}(x) - \mu^{(1)}(x)| \xrightarrow{p} 0,$$

then $\hat{\mu}^{(0)}(\cdot)$, $\hat{\mu}^{(1)}(\cdot)$ serve as appropriate stand-ins for $\mu^{(0)}(\cdot)$, $\mu^{(1)}(\cdot)$ in large samples.

For the residual portion, we take the initial fitting functions $\hat{\mu}_{-i}^{(0)}(\cdot)$ and $\hat{\mu}_{-i}^{(1)}(\cdot)$ and compute the residuals

$$\hat{\varepsilon}_i = Y_i - W_i \hat{\mu}_{-i}^{(1)}(X_i) - (1 - W_i)\hat{\mu}_{-i}^{(0)}(X_i).$$

Under an assumption of independent errors, it is appropriate to compute bootstrap residuals $\varepsilon_1^b, \dots, \varepsilon_n^b$ by sampling with replacement from the observed residuals $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$. We can then construct an artificial bootstrap response

$$Y_i^b = W_i^b \hat{\mu}_{-i}^{(1)}(X_i^b) + (1 - W_i^b)\hat{\mu}_{-i}^{(0)}(X_i^b) + \varepsilon_i^b.$$

We then compute $\hat{\tau}^b$ using data (Y_i^b, X_i^b, W_i^b), and then take the bootstrap distribution $\{\hat{\tau}^b\}_{b=1}^B$ as an approximation to the true distribution of $\hat{\tau}$. To construct a

$1-\alpha$ confidence interval, one can calculate the endpoints using approximate Gaussian quantiles,

$$\hat{\tau} \pm z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\tau}_b)}.$$

Alternatively, one may use the $\alpha/2$ and $1-\alpha/2$ quantiles of the empirical bootstrap distribution (a percentile bootstrap), which is preferable if the distribution of $\hat{\tau}$ is skewed.

We wish to emphasize that the main insight here is that exogeneity allows the feature and residual variances to be handled separately, and that the feature variance can be computed from the design, however complicated the structure of X_i itself may be. The bootstrap residuals ε_i^b as described above rely on independent errors, but in fact the practitioner is free to utilize the entirety of the rich bootstrap literature stemming from Efron (1979) in the event that this independence assumption is violated. For example, one may use versions of the block bootstrap (Künsch, 1989) to try and protect against correlated errors. One can use more complicated bootstrap methods to be more faithful to the empirical distribution, such as incorporating higher-order features of the distribution via bias-corrected and accelerated (BCa) intervals (Efron, 1987), or handling heteroscedasticity via the wild bootstrap (Wu, 1986).

3.5 Exposure modeling

One of the main alternative approaches is to use a version of an inverse propensity weighted estimator derived from a local neighborhood exposure model, so it is important to provide detail into that approach now. We briefly describe the exposure model-based estimators framed in the language of *constant treatment response* assumptions (Manski, 2013). For $Y_i(\mathbf{w}) = \mu_i(\mathbf{w}) + \varepsilon_i$, this approach partitions the space of treatments $\mathcal{W}$ into classes of treatments that map to the same mean response $\mu_i(\cdot)$ for unit i. The partition function is assumed known, and is called an *exposure function*. The no-interference portion of SUTVA can be specified as an exposure model, since no-interference is equivalent to the requirement that $\mu_i(\mathbf{w}_1) = \mu_i(\mathbf{w}_2)$ for any two treatment vectors $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$ in which the i-th components of $\mathbf{w}_1$ and $\mathbf{w}_2$

agree. Manski (2013) refers to this formulation as *individualistic treatment response* (ITR).

The exposure model most commonly used for local interference is a version of the *neighborhood treatment response* (NTR) assumption, which given a graph G, posits that $\mu_i(\mathbf{w}_1) = \mu_i(\mathbf{w}_2)$ whenever $\mathbf{w}_1$ and $\mathbf{w}_2$ agree in all components j such that $j \in \mathcal{N}_i \cup \{i\}$. In other words, NTR assumes that Y_i depends on unit i's own treatment and possibly any other unit in its neighborhood $\mathcal{N}_i$, but that it does not respond to changes in the treatments of any units outside of its immediate neighborhood.

Several versions of one-step neighborhood interference have been used which all roughly refer to the same idea; in addition to the NTR assumption, Sussman and Airoldi (2017) propose several variants of a *neighborhood interference assumption* (NIA) and Forastiere, Airoldi, and Mealli (2016) study the *stable unit treatment on neighborhood value assumption* (SUTNVA). For the purposes of estimating the global treatment effect, one may use *fractional q-NTR* (Ugander et al., 2013), where given a threshold parameter $q \in (0.5, 1]$, q-NTR assumes that a unit is effectively in global treatment if at least a fraction q of its neighbors are assigned to treatment, and similarly for global control. NTR is thus a graph analog of partial interference for groups and q-NTR is a corresponding version of stratified interference. The threshold q is a tuning parameter; larger values of q result in less bias due to interference, but greater variance because there are fewer units available for estimation. Eckles et al. (2017) provide some theoretical results for characterizing the amount of bias reduction. There is not much guidance for selecting q to manage this bias-variance tradeoff; Eckles et al. (2017) uses $q = 0.75$.

3.5.1 IPW estimators

Aronow and Samii (2017) study the behavior of inverse propensity weighted (IPW)

estimators based on a well-specified exposure model. Toward this end, let

$$E_i^{(\mathbf{1})} = \mathbb{1}\left\{\frac{1}{d_i}\sum_{j\in\mathcal{N}_i} W_j \geq q\right\}$$
$$E_i^{(\mathbf{0})} = \mathbb{1}\left\{\frac{1}{d_i}\sum_{j\in\mathcal{N}_i} W_j \leq 1-q\right\}$$

be the events that unit i is q-NTR exposed to global treatment and q-NTR exposed to global control, respectively. Let their expectations be denoted by

$$\pi_i^{(\mathbf{1})} = \mathbf{E}(E_i^{(\mathbf{1})}), \qquad \pi_i^{(\mathbf{0})} = \mathbf{E}(E_i^{(\mathbf{0})}),$$

which represent the *propensity scores* for unit i being exposed to the global potential outcome conditions. Then the inverse propensity weighted estimators under consideration are defined as

$$\hat{\tau}_{\text{HT}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{E_i^{(\mathbf{1})}Y_i}{\pi_i^{(\mathbf{1})}} - \frac{E_i^{(\mathbf{0})}Y_i}{\pi_i^{(\mathbf{0})}}\right]$$
$$\hat{\tau}_{\text{Hájek}} = \left(\sum_{i=1}^{n}\frac{E_i^{(\mathbf{1})}}{\pi_i^{(\mathbf{1})}}\right)^{-1}\sum_{i=1}^{n}\frac{E_i^{(\mathbf{1})}Y_i}{\pi_i^{(\mathbf{1})}} - \left(\sum_{i=1}^{n}\frac{E_i^{(\mathbf{0})}}{\pi_i^{(\mathbf{0})}}\right)^{-1}\sum_{i=1}^{n}\frac{E_i^{(\mathbf{0})}Y_i}{\pi_i^{(\mathbf{0})}} \tag{3.23}$$

The estimator $\hat{\tau}_{\text{HT}}$ is the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), and $\hat{\tau}_{\text{Hájek}}$ is the Hájek estimator (Hájek, 1971); these names stem from the survey sampling literature and are commonly used in the interference literature. In the importance sampling and off-policy evaluation literatures, analogs of $\hat{\tau}_{\text{HT}}$ and $\hat{\tau}_{\text{Hájek}}$ are known as unnormalized and self-normalized importance sampling estimators, respectively. In the finite potential outcomes framework the Horvitz-Thompson estimator is unbiased under the experimental design distribution, but suffers from excessive variance when the probabilities of global exposure are small, as is usually the case. The Hájek estimator, which forces the weights to sum to one and is thus interpretable as a difference of weighted within-group means, incurs a small amount of finite sample bias but is asymptotically unbiased, and is nearly always preferable

to $\hat{\tau}_{\text{HT}}$. For our simulations we will therefore avoid using $\hat{\tau}_{\text{HT}}$.

One of the main insights in the exposure modeling framework developed by Aronow and Samii (2017) is that even if the initial treatment assignment probability π is constant across units, the global treatment propensity scores need not be; indeed, $\pi_i^{(\mathbf{1})}$ and $\pi_i^{(\mathbf{0})}$ depend on the network structure and choice of exposure model. Therefore inverse propensity weighting is needed to produce unbiased (or consistent) estimators for contrasts between exposures even in a Bernoulli randomized design.

Given a design and a (simple enough) exposure model, the propensities can be calculated exactly. If the treatments are assigned according to independent Bernoulli coin flips, the exact exposure probabilities are expressed straightforwardly using the binomial distribution function. That is, for treatment probability $\pi = \mathbf{P}(W_i = 1)$ and degree d_i, the probability of unit i being q-NTR exposed to global treatment is

$$\pi_i^{(\mathbf{1})} = \pi(1 - F_{d_i,\pi}(\lfloor d_i q \rfloor)), \tag{3.24}$$

where

$$F_{n,p}(k) = \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j}$$

is the distribution function of a Binomial(n, p) random variable. Similarly, the probability that unit i is q-NTR exposed to global control is

$$\pi_i^{(\mathbf{0})} = (1 - \pi) F_{d_i,\pi}(\lfloor d_i (1 - q) \rfloor). \tag{3.25}$$

In a cluster randomized design, exposure probabilities for fractional neighborhood exposure can be computed using a dynamic program (Ugander et al., 2013).

A further comment on the propensity scores $\pi_i^{(\mathbf{1})}$ and $\pi_i^{(\mathbf{0})}$ is necessary. Importantly, these propensity scores are exact only to the extent to which the exposure model is correct. Thus, when the exposure model is unknown, these propensities scores should be viewed as *estimated* propensities, in which case even small estimation errors in the propensities can lead to large estimation errors in their inverses. It is therefore the case that $\hat{\tau}_{\text{HT}}$ and $\hat{\tau}_{\text{Hájek}}$ can suffer from the same high-variance problems as IPW estimators based on a fitted propensity model used in observational

studies, even if the exposure model is only mildly misspecified.

In our simulations we use the Hájek estimator, $\hat{\tau}_{\text{Hájek}}$, defined by equation (4.11) and the q-NTR exposure probabilities (3.24) and (3.25). We fix $q = 0.75$, which is the same threshold used in Eckles et al. (2017). For the other values of q that we tried, performance was roughly on par with or worse than $q = 0.75$.

3.5.2 Conservative variance estimation for the Hájek estimator

Aronow and Samii (2017) use the following approach for conservative variance estimation, which we explain below for pedagogical and comparison purposes. Denote

$$\hat{\mu}^{(\mathbf{1})} = \frac{1}{n}\sum_{i=1}^{n}\frac{E_i^{(\mathbf{1})}Y_i}{\pi_i^{(\mathbf{1})}}, \qquad \hat{\mu}^{(\mathbf{0})} = \frac{1}{n}\sum_{i=1}^{n}\frac{E_i^{(\mathbf{0})}Y_i}{\pi_i^{(\mathbf{0})}}$$

so that

$$\hat{\tau}_{\text{HT}} = \hat{\mu}^{(\mathbf{1})} - \hat{\mu}^{(\mathbf{0})}.$$

Define

$$\pi_{i,j}^{(\mathbf{0},\mathbf{1})} = \mathbf{E}[E_i^{(\mathbf{0})}E_j^{(\mathbf{1})}]$$

as the joint probability that unit i is globally exposed to control and unit j is globally exposed to treatment. Let $d_{ij} = \{k : k \in \mathcal{N}_i \cap \mathcal{N}_j\}$ be the set of units adjacency to both i and j. Then the joint propensity is calculated under the Bernoulli design as

$$\pi_{i,j}^{(\mathbf{0},\mathbf{1})} = \sum_{k=0}^{d_{ij}} \underbrace{\binom{d_{ij}}{k}\pi^k(1-\pi)^{d_{ij}-k}}_{k \text{ shared neighbors treated}} \times \underbrace{\pi(1 - F_{d_i,\pi}(\lfloor d_i q - k \rfloor))}_{\text{remaining } i \text{ neighbors in treatment}} \tag{3.26}$$

$$\times \underbrace{(1-\pi)F_{d_j,\pi}(\lfloor d_j(1-q) + (d_{ij}-k)\rfloor)}_{\text{remaining } j \text{ neighbors in control}}. \tag{3.27}$$

Note that $\pi_{i,j}^{(\mathbf{0},\mathbf{1})} = 0$ for some pairs i and j.

Then an estimator for the Horvitz-Thompson variance is given by

$$\widehat{\mathrm{Var}}(\hat{\tau}_{\mathrm{Hájek}}) = \widehat{\mathrm{Var}}(\hat{\mu}^{(\mathbf{1})}) + \widehat{\mathrm{Var}}(\hat{\mu}^{(\mathbf{0})}) - 2\widehat{\mathrm{Cov}}(\hat{\mu}^{(\mathbf{0})}, \hat{\mu}^{(\mathbf{1})}),$$

where

$$\begin{aligned}
\widehat{\mathrm{Var}}(\hat{\mu}^{(\mathbf{1})}) &= \sum_{i=1}^{n} \frac{1-\pi_i^{(\mathbf{1})}}{(\pi_i^{(\mathbf{1})})^2} E_i^{(\mathbf{1})} Y_i^2 \\
\widehat{\mathrm{Var}}(\hat{\mu}^{(\mathbf{0})}) &= \sum_{i=1}^{n} \frac{1-\pi_i^{(\mathbf{0})}}{(\pi_i^{(\mathbf{0})})^2} E_i^{(\mathbf{0})} Y_i^2 \\
\widehat{\mathrm{Cov}}(\hat{\mu}^{(\mathbf{0})}, \hat{\mu}^{(\mathbf{1})}) &= \sum_{i=1}^{n} \sum_{\substack{j:j\neq i \\ \pi_{i,j}^{(\mathbf{0},\mathbf{1})}>0}} \frac{E_i^{(\mathbf{0})} E_j^{(\mathbf{1})}}{\pi_{i,j}^{(\mathbf{0},\mathbf{1})}} \frac{Y_i Y_j}{\pi_i^{(\mathbf{0})} \pi_j^{(\mathbf{1})}} (\pi_{i,j}^{(\mathbf{0},\mathbf{1})} - \pi_i^{(\mathbf{0})} \pi_j^{(\mathbf{1})}) \\
&\quad - \sum_{i} \sum_{\substack{j:j\neq i \\ \pi_{i,j}^{(\mathbf{0},\mathbf{1})}=0}} \left[\frac{E_i^{(\mathbf{0})} Y_i^2}{2\pi_i^{(\mathbf{0})}} + \frac{E_j^{(\mathbf{1})} Y_j^2}{2\pi_j^{(\mathbf{1})}} \right]
\end{aligned}$$

This quantity is a conservative estimate for the true Horvitz-Thompson variance.

For the Hájek estimator, one can use the above variance estimator except that in place of $Y_i(\mathbf{1})$ and $Y_i(\mathbf{0})$, one substitutes in the residuals

$$Y_i(\mathbf{1}) - \left(\sum_{i=1}^{n} \frac{E_i^{(\mathbf{1})}}{\pi_i^{(\mathbf{1})}} \right)^{-1} \sum_{i=1}^{n} \frac{E_i^{(\mathbf{1})} Y_i(\mathbf{1})}{\pi_i^{(\mathbf{1})}}$$

and

$$Y_i(\mathbf{0}) - \left(\sum_{i=1}^{n} \frac{E_i^{(\mathbf{0})}}{\pi_i^{(\mathbf{0})}} \right)^{-1} \sum_{i=1}^{n} \frac{E_i^{(\mathbf{0})} Y_i(\mathbf{0})}{\pi_i^{(\mathbf{0})}},$$

respectively. In this case the resulting intervals can be shown to be asymptotically correct via the delta method. For further details the reader is referred to (Sections 5, 7.2, Aronow and Samii, 2017) as well as the earlier results in Aronow and Samii (2012).

3.6 Simulations

This section is devoted to running a number of simulation experiments. Our goals in these simulations are to (a) verify that our adjustment estimators and variance estimates are behaving as intended, (b) compare the performance of our proposed estimators to that of existing inverse propensity weighted estimators based on exposure models, and (c) empirically explore the behavior of our estimators in regimes of mild model misspecification.

For the network G we use a subset of empirical social networks from the `facebook100` dataset, an assortment of complete online friendship networks for one hundred colleges and universities collected from a single-day snapshot of Facebook in September 2005. A detailed analysis of the social structure of these networks was given in Traud et al. (2011, 2012). We use an empirical network rather than an instance of a random graph model in order to replicate as closely as possible the structural characteristics observed in real-world networks. We use the largest connected components of the Caltech and Stanford networks. Some summary statistics for the networks are given in Table 3.1.

network	Caltech	Stanford
number of nodes	762	11586
number of edges	16651	568309
diameter	6	9
average pairwise distance	2.33	2.82

Table 3.1: Summary statistics for the `facebook100` networks.

In all simulation regimes we compare our regression estimators to two other estimators, which we describe now. As a baseline we use the SUTVA difference-in-means estimator

$$\hat{\tau}_{\mathrm{DM}} = \frac{1}{N_1}\sum_{i=1}^{n} W_i Y_i - \frac{1}{N_0}\sum_{i=1}^{n}(1-W_i)Y_i.$$

3.6.1 Variance estimates in a linear model

We first run a basic simulation in which we compute estimates, variances and variance estimates in an ordinary linear model. We consider two features,

$$X_{1,i} = \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} W_j,$$

the proportion of treated neighbors, and

$$X_{2,i} = \sum_{j \in \mathcal{N}_i} W_j,$$

the number of treated neighbors. It is conceivable that Y_i may depend on both of these features. Let the data-generating process for Y_i be as in Model 3.3; that is, the mean function for Y_i is linear in $X_i = (X_{1,i}, X_{2,i})$, given parameters $\alpha_w \in \mathbb{R}$ and $\beta_w = (\beta_{w,1}, \beta_{w,2}) \in \mathbb{R}^2$ for $w = 0, 1$. We simulate $\varepsilon_i \sim N(0, \sigma^2)$.

Let $\bar{d} = n^{-1} \sum_{i=1}^n d_i$ be the average degree of G. Then the true global treatment effect is

$$\alpha_1 - \alpha_0 + \beta_{1,1} + \bar{d}\beta_{1,2}.$$

We fix $\alpha_1 = 1$ and $\alpha_0 = 0$, so that the direct effect is 1. We fix the noise variance at $\sigma^2 = 1$. We vary the "proportion" coordinate of β_0 in $\{0, 0.1\}$, the "number" coordinate of β_0 in $\{0, 0.01\}$, the "proportion" coordinate of β_1 in $\{0, 0.2\}$, and the "number" coordinate of β_1 in $\{0, 0.05\}$, giving 16 total parameter configurations. SUTVA holds when $\beta_0 = \beta_1 = 0$.

We use equation (3.17) to estimate the variance of the adjusted estimator, using 200 bootstrap samples from the feature distribution to calculate the inverse covariance matrices. We also compute the difference-in-means (DM) estimator for comparison purposes, for which we use the standard Neyman conservative variance estimate

$$\frac{S_0^2}{N_0} + \frac{S_1^2}{N_1},$$

where S_0^2 and S_1^2 are the within-group sample variances. We compute confidence

intervals based on Gaussian quantiles for a 90% nominal coverage rate.

We then run 1000 simulated experiments, sampling a new treatment vector $\mathbf{W}$ and computing the two estimators each time. The results are shown in Table 3.2. The bias of the DM estimator increases with greater departures from SUTVA, and confidence intervals for that estimator are only theoretically valid under SUTVA (the first row in Table 3.2). Otherwise, the confidence intervals are anticonservative, both due to bias of the DM estimator and due to invalidity of the Neyman variance estimate, which assumes fixed potential outcomes. On the other hand, the adjustment estimator is unbiased and has valid coverage for all parameter configurations.

Parameters			Bias		SE		SE Ratio		Coverage rate	
β_0	β_1	τ	DM	adj	DM	adj	DM	adj	DM	adj
(0, 0)	(0, 0)	1	0.007	-0.013	0.074	1.149	0.982	1.031	**0.891**	**0.913**
(0, 0.01)	(0, 0)	1	-0.006	-0.028	0.072	1.189	1.004	0.996	**0.899**	**0.901**
(0.1, 0)	(0, 0)	1	-0.053	0.027	0.072	1.151	1.004	1.029	0.808	**0.919**
(0.1, 0.01)	(0, 0)	1	-0.052	0.017	0.074	1.155	0.973	1.025	0.801	**0.906**
(0, 0)	(0, 0.05)	1.05	-0.025	0.005	0.073	1.211	0.990	0.976	0.866	**0.882**
(0, 0.01)	(0, 0.05)	1.05	-0.026	0.005	0.070	1.173	1.036	1.010	**0.894**	**0.909**
(0.1, 0)	(0, 0.05)	1.05	-0.075	0.058	0.075	1.232	0.960	0.961	0.707	**0.884**
(0.1, 0.01)	(0, 0.05)	1.05	-0.078	-0.019	0.073	1.151	0.996	1.031	0.699	**0.912**
(0, 0)	(0.2, 0)	1.2	-0.097	0.058	0.073	1.168	0.993	1.014	0.615	**0.910**
(0, 0.01)	(0.2, 0)	1.2	-0.104	0.007	0.073	1.142	0.999	1.039	0.577	**0.916**
(0.1, 0)	(0.2, 0)	1.2	-0.151	0.002	0.073	1.197	0.991	0.988	0.334	**0.892**
(0.1, 0.01)	(0.2, 0)	1.2	-0.152	0.044	0.072	1.208	1.014	0.981	0.315	**0.894**
(0, 0)	(0.2, 0.05)	1.25	-0.125	-0.054	0.072	1.154	1.003	1.025	0.476	**0.908**
(0, 0.01)	(0.2, 0.05)	1.25	-0.130	-0.014	0.070	1.149	1.029	1.031	0.446	**0.920**
(0.1, 0)	(0.2, 0.05)	1.25	-0.174	0.016	0.074	1.206	0.983	0.982	0.232	**0.894**
(0.1, 0.01)	(0.2, 0.05)	1.25	-0.182	-0.014	0.074	1.172	0.985	1.012	0.194	**0.903**

Table 3.2: Results of the basic simulation setup from Section 3.6.1, showing bias, true standard error, ratio of estimated standard error to true standard error, and coverage rate of 90% nominal Gaussian confidence interval. Coverage rates which fall within a 99% one-sided interval of the nominal coverage rate (that is, coverage rates above $0.9 - 2.326\sqrt{0.9 \times 0.1/1000} \approx 0.878$) are **bolded**.

3.6.2 Estimator weights

Both the OLS adjustment estimator and the Hájek estimator are linear reweighting estimators. The OLS weights are given by equations (3.14) and (3.15), and the Hájek

weights are implied by the definition of the Hájek estimator in equation (4.11). Both depend on only the network structure, treatment assignment, and exposure model or choice of features, but not on the realized outcome variable. The weights for a single Bernoulli(0.5) draw of the treatment vector $\mathbf{W}$ for the Caltech graph are displayed in Figure 3.3, assuming that the Hájek estimator is to be constructed under the q-NTR exposure condition for $q = 0.75$, and the OLS estimator uses the fraction of treated neighbors as the only adjustment variable. We see that the Hájek estimator trusts a few select observations to be representative of the global exposure conditions. A graph cluster randomized design would increase the number of units used in the Hájek estimator. The OLS estimator, on the other hand, gives all units non-zero weight. Some units that are in the treatment group but are surrounded by control individuals are treated as diagnostic for the control mean and vice versa, which is a reasonable thing to do if the linear model is true.

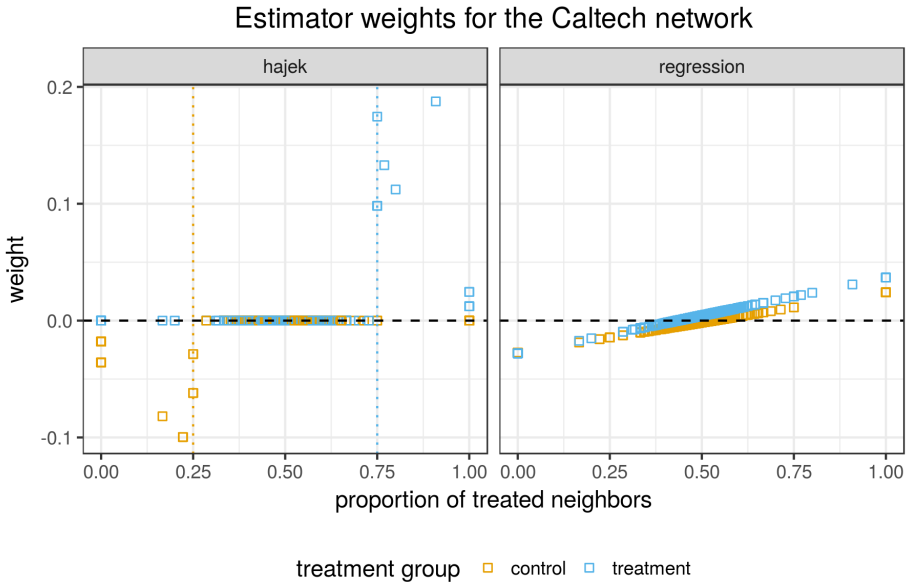

Figure 3.3: Estimator weights for the case where the only feature is the proportion of treated neighbors. (left) The Hájek estimator selects a few individuals from treatment and control and takes a weighted average of those individuals with weights determined by exposure probabilities. Vertical dotted lines are the thresholds used for selecting observations. (right) The regression estimator takes a more democratic approach, giving all units non-zero weight.

3.6.3 Dynamic linear-in-means

Here we replicate portions of the simulation experiments conducted by Eckles et al. (2017). That paper uses a discrete-time dynamic model, which can be viewed as a noisy best-response model (Blume, 1995), in which individuals observe and respond to the behaviors of their peers, using that information to guide their actions in the following time period. Given responses $Y_{i,t-1}$ for time period $t-1$, let

$$Z_{i,t-1} = \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} Y_{i,t-1},$$

a time-varying version of Z_i defined in equation (3.5), which represents the average behavior of unit i's neighbors at time $t-1$. Then we model

$$Y_{i,t} = \alpha + \beta W_i + \gamma Z_{i,t-1} + \varepsilon_{i,t}. \tag{3.28}$$

The noise is taken to be $\varepsilon_{i,t} \sim N(0, \sigma^2)$, which is independent and homoscedastic across time and individuals. Eckles et al. (2017) add an additional thresholding step that transforms equation (3.28) into a probit model and Y into a binary outcome variable, but here we study the non-thresholded case which is closer to the original linear-in-means model specified by Manski (1993). Starting from initial values $Y_{i,0} = 0$, the process is run up to a maximum time T and then the final outcomes are taken to be $Y_i = Y_{i,T}$. The choice of T, along with the strength of the spillover effect γ, governs the amount of interference. If T is larger than the diameter of the graph, then the interference pattern is fully dense, and no exposure model holds.

We construct two different adjustment variables. First, let

$$X_{1,i} = \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} W_j,$$

the proportion of treated neighbors. Now let

$$\mathcal{N}_i^{(2)} = \{k \in [n] \setminus \{i\} : \text{there exists } j \text{ such that } A_{ij} A_{jk} = 1\}$$

be the two-step neighborhood of unit i. Then define

$$X_{2,i} = \frac{1}{|\mathcal{N}_i^{(2)}|} \sum_{k \in \mathcal{N}_i^{(2)}} A_{ij} A_{jk} W_k,$$

the proportion of individuals belonging to $\mathcal{N}_i^{(2)}$ who are treated. (Note that unit i itself does *not* belong to its own two-step neighborhood.)

We use a small-world network (Watts and Strogatz, 1998), which is the random graph model used in the simulations by Eckles et al. (2017), with $n = 1000$ vertices, initial neighborhood size 10, and rewiring probability 0.1. We also run our simulation on the empirical Caltech network.

As in Eckles et al. (2017), we compute the "true" global treatment effects by Monte Carlo simulation. For every parameter configuration we sample 5000 instances of the response vector under global exposure to treatment $\mathbf{W} = \mathbf{1}$, and 5000 instances of the response vector under global exposure to control $\mathbf{W} = \mathbf{0}$, and then average the resulting difference in response means. For the response model, we fix the intercept at $\alpha = 0$ and the direct effect at $\beta = 1$. We vary the spillover effect $\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$ and the maximum number of time steps $T \in \{2, 4\}$. Larger values of γ and T indicate more interference. We also use two different levels for the noise standard deviation, $\sigma \in \{1, 3\}$.

We consider two versions of the linear adjustment estimator defined in equation (3.13), one that adjusts for $X_{1,i}$ only, and one that adjusts for both $X_{1,i}$ and $X_{2,i}$. The first model adjusts for one-step neighborhood information, whereas the second model adjusts for both one- and two-step neighborhood information. We compare to the difference-in-means estimator and the Hájek estimator with $q = 0.75$ fractional NTR exposure.

We emphasize that the all of the estimators that we consider are misspecified under the data generating process that we use in this simulation. For $T \geq 2$, local neighborhood exposure fails, so the propensity scores used in the Hájek estimator do not align with the true propensity scores. Our adjustment estimators are also misspecified for $T \geq 2$; not only is the linear model misspecified, but the residuals

are neither independent nor exogenous, violating Assumption 3.3.

The results are displayed in Figure 3.4. We see that the two OLS adjustment estimators are uniformly better at bias reduction than the Hájek estimator. The two-step adjustment is nearly unbiased even though it is misspecified, even in the presence of strong spillover effects. This is because interference is dissipating exponentially, so that units don't really respond to the behavior of individuals that are distance 3 or 4 away. The two-step adjustment has higher variance than the one-step adjustment because it involves fitting a more complex model. Furthermore, estimators appear to have more trouble handling the real-world network structure of the Caltech network, compared to the artificial small-world network.

The difference-in-means estimator outperforms the adjustment estimators in regimes of weak interference, which is expected since difference-in-means is the best that can be done under correct specification of SUTVA. In terms of RMSE the Hájek estimator sometimes outperforms the two-step adjustment estimator because of large variance. However, if the main goal is robustness to interference, then unconfounded estimation coupled with valid confidence intervals is likely the priority over optimizing an error metric such as RMSE. In this case, since the Hájek estimator neither achieve sufficient bias reduction nor provide correct coverage, it has no real advantage over the adjustment estimators.

Figure 3.5 displays the coverage rates obtain from variance estimates using equation (3.17) under the dynamic treatment response setup. The coverage is not always correct due to misspecification, especially for `adj1`. We see that coverage rates for `adj2` are often conservative even though it too is misspecified. We note that standard variance estimators for the difference-in-means estimator and those derived in Aronow and Samii (2017) for the Hájek estimator also would fail here because they rely on correct specification of SUTVA and an exposure model, respectively. In short, we are plagued with the same difficulties that beset attempting to do valid inference in observational studies when we do not know whether unconfoundedness holds.

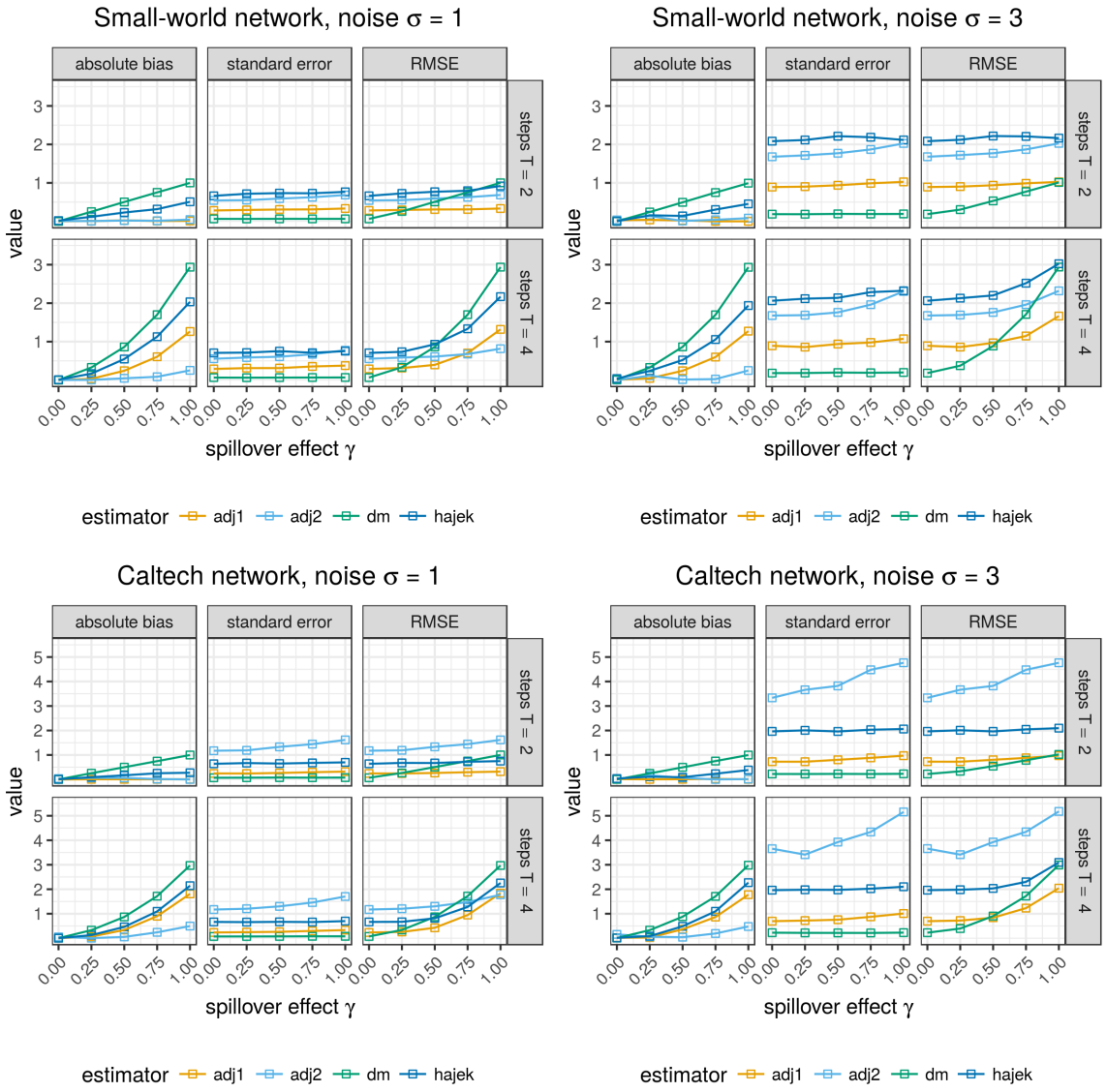

Figure 3.4: Results for linear-in-means simulation. `dm` is the difference-in-means estimator, `hajek` is the Hájek estimator, `adj1` is adjustment based on a one-step neighborhood, and `adj2` is adjustment based on a two-step neighborhood.

3.6.4 Average + aggregate peer effects

In this example we consider a response model in which individuals respond partially to the *average* behavior of their peers and partially to the *aggregate* behavior of their

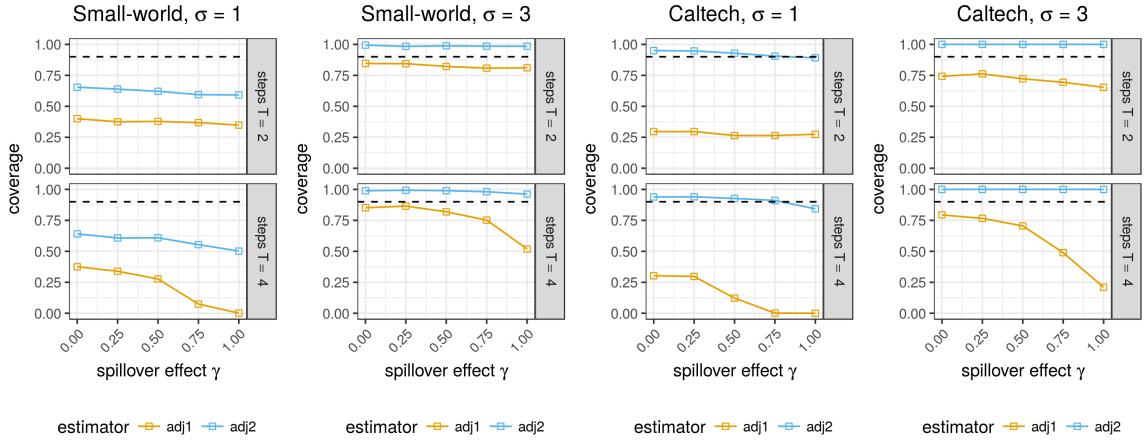

Figure 3.5: Coverage rates for 90% nominal interval.

peers. Let

$$X_i^{\text{frac}} = \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} W_j$$

be the fraction of treated neighbors and

$$X_i^{\text{num}} = \sum_{j \in \mathcal{N}_i} W_j$$

be the number of treated neighbors. X_i^{frac} captures a notion of fractional neighborhood exposure and X_i^{num} captures a notion of absolute neighborhood exposure. It seems reasonable that both of these features may contribute interference. In order to use an exposure model estimator one would need to focus on either fractional exposure or absolute exposure, or otherwise define a more complicated exposure model, but our adjustments easily handle both features.

We consider the following response function:

$$Y_i = -5 + 2(2 + E_i)W_i + 0.03X_i^{\text{frac}} + \frac{1}{1 + 0.001e^{-0.03(X_i^{\text{num}} - 300)}} + \frac{10}{3 + e^{-8(X_i^{\text{frac}} - 0.4)}} + \varepsilon_i,$$

where $E_i \sim N(0, 2)$ introduces heterogeneity into the direct effect and $\varepsilon_i \sim N(0, 1)$ is homoscedastic noise. This function captures a possible way in which individuals could respond nonlinearly to their peer exposures through X_i^{frac} and X_i^{num}. Figure 3.6

plots a single draw of this response on individuals from the Stanford network. The continuous response exhibits a logistic dependence on both features. We see that individuals with less than half of their neighbors exposed to the treatment condition experience a steadily increasing peer effect as the proportion of treated neighbors increases. For individuals with more than half of their neighbors exposed to the treatment condition, the effect is nearly constant across values of X_i^{frac}, capturing the idea that after a certain threshold observing additional peer exposures doesn't add much. For X_i^{num}, we see that a small number of treated neighbors essentially contributes no interference, but once a large number of neighbors are exposed to treatment this has a measurable impact on the response. We also see that there is a noticeable bump around $X_i^{\text{frac}} = 0.5$; this is because individuals with peers nearly equally assigned to the two groups are more likely to have high degree. The model extends the idea of neighborhood exposure to capture the intuition that having a high proportion of treated neighbors is evidence for being subject to interference, but such evidence is stronger when the individual in question has many friends and not just one or two friends. The true treatment effect is $\tau = 6.336$, which was computed using 2000 Monte Carlo draws each of global treatment and global control.

In our experience larger populations seem to be needed for fitting the more complex, nonlinear functions, so we work with the Stanford network which has 11586 nodes. We predict the response surfaces using a generalized additive model (GAM) (Hastie and Tibshirani, 1986), which is easy and fast to fit in `R`, but other methods such as local regression or random forests could of course be used instead. We split the dataset into $K = 2$ folds, and within each fold, train a GAM separately in the treatment and control groups for a total of 4 fitted models. The models are then used to obtain predicted responses on the held-out fold. Standard errors were computed via the bootstrap as described in Section 3.4.1, using 50 bootstrap replications.

We compare to the difference-in-means estimator, the Hájek estimator using a threshold of $q = 0.75$ on the X_i^{frac} variable, and the OLS adjustment. The results are displayed in Table 3.3. The DM estimator exhibits the most bias, as it does not adjust for any sort of interference. The Hájek estimator removes some bias, but

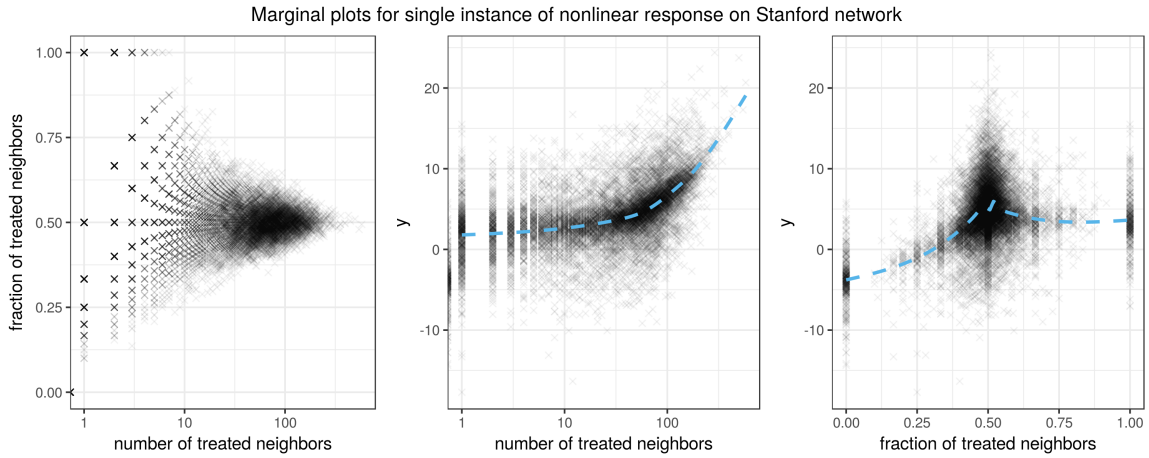

Figure 3.6: One draw of the features and response for the nonlinear setup. The left panel shows the relationship between the two features, and the right two panels show the relationship of the response with each covariate. The horizontal axis for "number of treated neighbors" (X_i^{num}) is on a logarithmic scale. A local linear regression, for exploratory purposes, is plotted in blue.

because it is based on a fractional exposure model it is unable to respond to the effect of having a high treated degree. Both the OLS and GAM estimators remove about 95% of the bias. The GAM adjustment does only slightly better than OLS; for this setup what matters most is adjusting for both axes of the interference statistic, and the added flexibility provided by the GAM does not seem to be crucial. We note also that average bootstrapped standard error is 1.076 times greater than the true standard error, suggesting that confidence intervals built on this standard error will have the approximately correct length.

3.7 Reanalysis of a farmer's insurance experiment

In this section we apply our methods to a field experiment conducted on individuals in 185 villages in rural China (Cai et al., 2015). The purpose of the study was to quantify the network (spillover) effects of certain information sessions for a farmer's weather insurance product on the eventual adoption of that product. Though they do not frame their approach explicitly in the language of exposure models as in (Aronow

estimator	estimate	absolute bias (%)	SE (ratio)
DM	-0.002	6.339 (100%)	0.077 (—)
Hájek	2.653	3.683 (58.1%)	1.601 (—)
OLS	6.683	0.347 (5.5%)	0.252 (0.942)
GAM	6.655	0.319 (5.0%)	0.246 (1.076)

Table 3.3: Nonlinear simulation results. The bias column displays the absolute and relative bias from the truth $\tau = 6.336$. The SE column displays the true standard error over 200 simulation replications, and for the adjustment estimators we display in parentheses the ratio of the estimated standard error to the true standard error.

and Samii, 2017), the estimands that are implied by the regression coefficients in the models that they use in that paper can be thought of as contrasts between exposures in an appropriately-defined exposure model. The authors did not consider estimating the global treatment effect; our proposed methods essentially allow us to perform an off-policy analysis of that estimand.

In the original field experiment, the researchers consider four treatment groups obtained by assigning villagers to either a simple or intensive information session in one of two rounds that were held three days apart. Here, for simplicity, we ignore the temporal distinction between the two rounds and consider a villager to be treated if they were exposed to either of the two intensive sessions.[3] The outcome variable is a binary indicator for whether the villager decided to purchase weather insurance.

We drop all villagers that were missing information about the treatment or the response, as well as villages lacking network information. Though the study was conducted in separate villages (for the purpose of administering the insurance information sessions), we combine all of the villagers into one large graph G. The network has 4,382 nodes and 17,069 edges. Because some social connections exist across villages, the villages do not partition exactly into separate connected components; our graph G has 36 connected components. The summary statistics for the processed dataset are given in Table 3.4.

[3]According to Cai et al. (2015) the treatment groups in the study are stratified by household size and farm size, but it is not clear from the data if and how exactly this was done, so for simplicity we analyze the experiment as if it were an unstratified, Bernoulli randomized experiment.

number of nodes	4832
number of edges	17069
number (%) treated	2406 (49.8%)
average takeup (mean response)	44.6%

Table 3.4: Summary statistics for the Cai et al. (2015) dataset.

Now let $\mathcal{N}_i$ and $\mathcal{N}_i^{(2)}$ be the one- and two-step neighborhoods for unit i, as we have denoted previously. We construct four variables from the graph: the fraction of units in $\mathcal{N}_i$ who are treated (`frac1`), the fraction of units in $\mathcal{N}_i^{(2)}$ who are treated (`frac2`), the number of units in $\mathcal{N}_i$ who are treated (`num1`), and the number of units in $\mathcal{N}_i^{(2)}$ who are treated (`num2`). Figure 3.7 displays the scatterplot matrix for these four variables as well as the response. As might be expected, these four variables are positively correlated with each other, and each is (weakly) positively correlated with the response variable. This correlation with the response suggests that these variables may be useful for adjustment.

We compute the OLS adjusted estimator as well as an adjustment estimator that used predictions from a logistic regression with $K = 5$ folds. We construct standard errors using the variance estimator given by equation (3.17) in the OLS case, and the parametric bootstrap variance estimator described in Section 3.4.1 with 200 bootstrap replications for the logistic regression case. We compare to the difference-in-means estimator and Hájek estimators based on thresholding on the `frac1` and `frac2` variables with $q = 0.75$.

The estimates are displayed in Table 4.4. Considering the strong positive spillover effects discovered by Cai et al. (2015), the difference-in-means estimate of 0.0774 is likely to be an underestimate of the true global treatment effect. The Hájek estimators produce estimates of 0.1630 (one-step fractional NTR) and 0.1672 (two-step fractional NTR). Though we do not know the truth, it may make us nervous that these estimates are more than twice the magnitude of the difference-in-means estimator, which if true would suggest that magnitude of the spillover effect is larger than the magnitude of the direct effect. The true treatment effect likely falls in between the estimates produced by difference-in-means and Hájek (though we have no way of knowing for sure). The

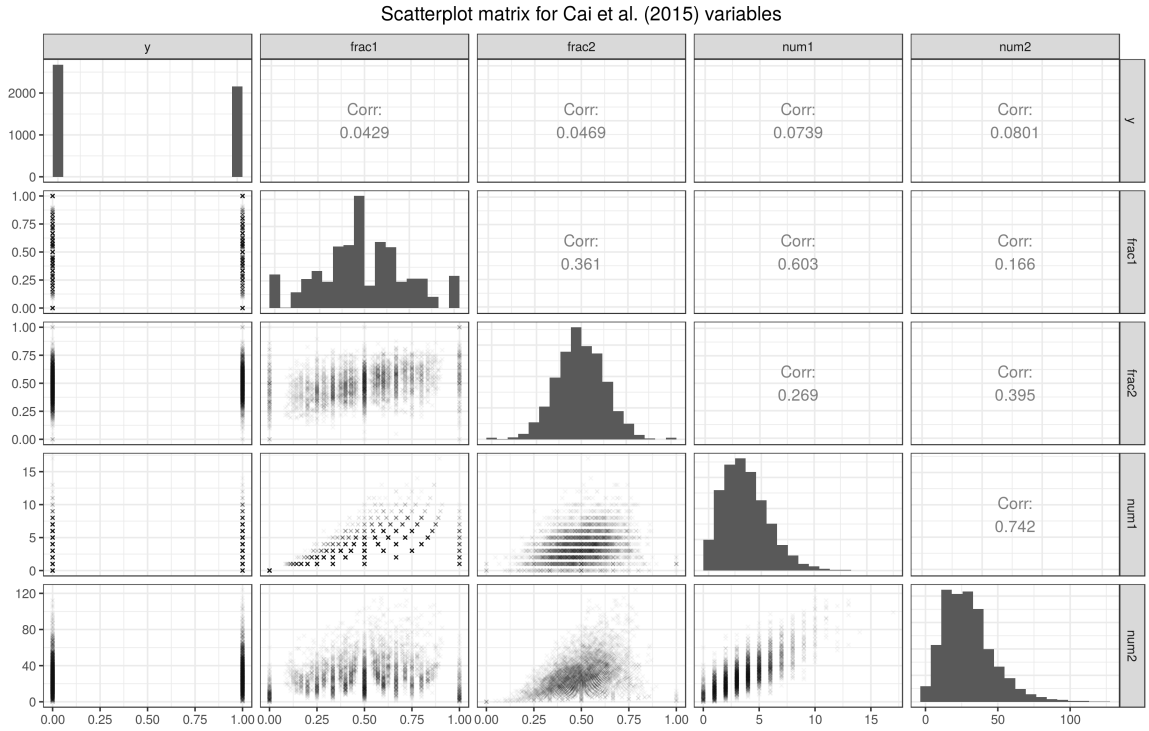

Figure 3.7: Scatterplot matrix for the variables used in the Cai et al. (2015) analysis.

OLS (0.1218) and logistic regression estimates (0.1197) are similar to each other and both within this range; an advantage they have over the Hájek estimators is that they incorporate information about the raw number of treated neighbors. The standard error estimates of 0.0561 (linear regression adjustment) and 0.0559 (logistic regression adjustment) are quite wide, suggesting some caution when interpreting this result.

Note that we have omitted computation of standard error estimates for the difference-in-means and Hájek estimators for several reasons. SUTVA and the neighborhood exposure conditions both likely fail to hold, so it is unclear how we should interpret such standard errors. Secondly, the conservative variance estimators proposed for the Hájek estimator (cf. Sections 5, 7.2, Aronow and Samii, 2017) are themselves inverse propensity estimators relying on small propensities, and consequently we found them to be quite unstable. For example, the variance estimate was often much greater than 1, which is the maximum possible variance of a $[-1, 1]$-valued random variable. Of course, we also do not know if the exogeneity assumptions hold or if other variables

should be included. In the regression analyses conducted by Cai et al. (2015), they also consider some other social network measures including indicator variables for varying numbers of friends and differentiation between strong and weak ties; a more sophisticated analysis here could include these features as well.

estimator	estimate	standard error
DM	0.0774	—
Hájek 1 ($q = 0.75$)	0.1630	—
Hájek 2 ($q = 0.75$)	0.1672	—
Linear	0.1218	0.0561
Logistic (5-fold)	0.1197	0.0559

Table 3.5: Estimates and standard errors for estimating the global treatment effect of intensive session on insurance adoption.

3.8 Discussion

We propose regression adjustments for interference in randomized experiments, which opens the world of the rich regression adjustment literature to the interference setting. We show in simulation experiments that the adjustments can do well, and we show how to do inference under exogeneity/unconfoundedness assumptions. Our reanalysis of the Cai et al. (2015) study shows that our approach can produce sensible estimates of the global treatment effect on real data.

There is much work to do to ensure that this approach can be reliably used in practical settings. First, we would like to extend the methods to handle more complicated designs. In reality a combination of design-side methods (graph clustering) and analysis-side methods (adjustment) could be the most effective approach. It would also be useful to have a thorough understanding of the combinations of network structures and experimental designs that correspond to the mathematical assumptions (exogeneity, full-rank design) listed in this work.

Secondly, it is necessary to formalize the placement of the methods discussed here within the agnostic perspective to treatment effect estimation. This would clarify the exogeneity/unconfoundedness requirement and better elucidate how interference

causes a randomized experiment to behave in some ways like an observational study. However, such assumptions are not new, and also needed to employ both standard estimators for observational studies in the SUTVA setting and exposure modeling estimators in the interference setting.

This issue simply highlights the need for better methods that can detect interference; there are several budding possibilities here. First, several works have proposed ways of doing sensitivity analysis for interference. VanderWeele et al. (2014) extend Robins et al. (2000)-style sensitivity analysis to cover some of the interference estimators studied in Hudgens and Halloran (2008), and Egami (2017) propose using an auxiliary network to perform sensitivity analysis on estimates obtained using the primary network. But clearly more work in this area is needed. Second, hypothesis tests for network or spillover effects of the type developed in (Aronow, 2012; Athey et al., 2017a; Basse et al., 2017), could be informative if applied to the residuals of a fitted interference model. Finally, one can always use more robust standard error constructions such as Eicker-Huber-White (Eicker, 1967; Huber, 1967; White, 1980) standard errors for heteroscedasticity or cluster bootstrap methods for dependence, though if the network structure is such that graph cluster randomization is unlikely to work well, then clustered bootstrap probably won't work well either. It is also possible that work based on dependency central limit theorems like the ones considered in Chin (2018a) could be used to develop more robust variance calculations. Broadly, any of the above methods ideas can be applied to the residuals of an interference model. If the bulk of interference can be captured in the mean function, then it is perhaps easier to deal with the remaining interference in the residuals.

Chapter 4

Better seed targeting experiments

4.1 Introduction

Interventions are often targeted to individuals based on their observed characteristics (Manski, 2004). Given the large theoretical and empirical literature showing the prevalence of peer effects in adoption processes, it is reasonable to assume that targeting strategies that incorporate network characteristics may be particularly successful. For example, a simple strategy might be to target an intervention at a small number of "seed individuals" who are well-connected and centrally located in the social network of the target population. Such a strategy aims to go beyond direct targeting and identify individuals who are expected to have an outsized impact on total adoption (Kempe et al., 2003; Hinz et al., 2011; Zubcsek and Sarvary, 2011; Libai et al., 2013).[1]

In this chapter we show how to efficiently evaluate such targeting strategies from field experiments with sufficiently randomized intervention policies. Crucial to our approach is the insight that many strategies of interest are *stochastic seeding strategies*, meaning that the targeted individuals are not determined deterministically but rather selected from a probability distribution of eligible seed sets. This insight allows

[1]We generally focus on maximizing total adoption. Naturally, one might intervene to reduce an outcome, such as with vaccinations (Cohen et al., 2003) or other preventative measures (Paluck et al., 2016). Similar ideas also apply to allocating a budget for measurement in a network, rather than intervention, as with selecting nodes to serve as sensors for outbreak detection (Leskovec et al., 2007; Christakis and Fowler, 2010).

us to construct reweighting estimators using ideas adapted from the counterfactual policy evaluation and importance sampling literatures. In our analysis these estimators sometimes provide more than a four-fold boost in effective sample size over simple difference-in-means estimators. Given these estimators, we also show how to gain further precision via optimized experimental designs, and how to conduct off-policy evaluation of field experiments that were not explicitly designed to evaluate targeting strategies.

Our analysis can be used to compare arbitrary stochastic seeding strategies considered over the same eligible seed sets, but for concreteness much of our discussion and analysis focuses on a particular strategy that we call *one-hop targeting*.[2] In one-hop targeting a seed set of k nodes is assembled by first randomly selecting k nodes and then randomly selecting one of their network neighbors as a seed. This strategy seeks to take advantage of the *friendship paradox*, Feld's (1991) observation that the average node has fewer connections than the average neighbor.[3] Because friends of randomly-selected individuals are likely to have more connections than the randomly-selected individuals themselves, this one-hop strategy ostensibly increases the chances of selecting influential seeds. See Figure 4.1 for an illustration of the probabilities of different seed sets on a small network. This strategy, which is *local* in the sense that it does not require observation of the entire network, is further motivated by the fact that observing or measuring the entire social network for a population can be impractical or so expensive that it would be better to spend that budget simply treating a few more nodes (cf. Akbarpour et al., 2017).

While one-hop targeting has received substantial attention and advocacy, evidence that it leads to greater adoption rates remains limited. In a field experiment on a study population of 5,773 individuals in 32 villages in rural Honduras, Kim et al. (2015) compare one-hop targeting to *random targeting*, a baseline strategy where seeds are

[2]Kim et al. (2015) call this "nomination" targeting. However, note that it does not involve individuals nominating a particular peer to be selected; rather they specify a set of people as friends, kin, etc., and then one peer is selected at random.

[3]We note that one-hop targeting is different from edge sampling, the original mechanism identified by Feld (1991) as driving the friendship paradox. See Lattanzi and Singer (2015) and Kumar et al. (2018) for further discussion of these differences.

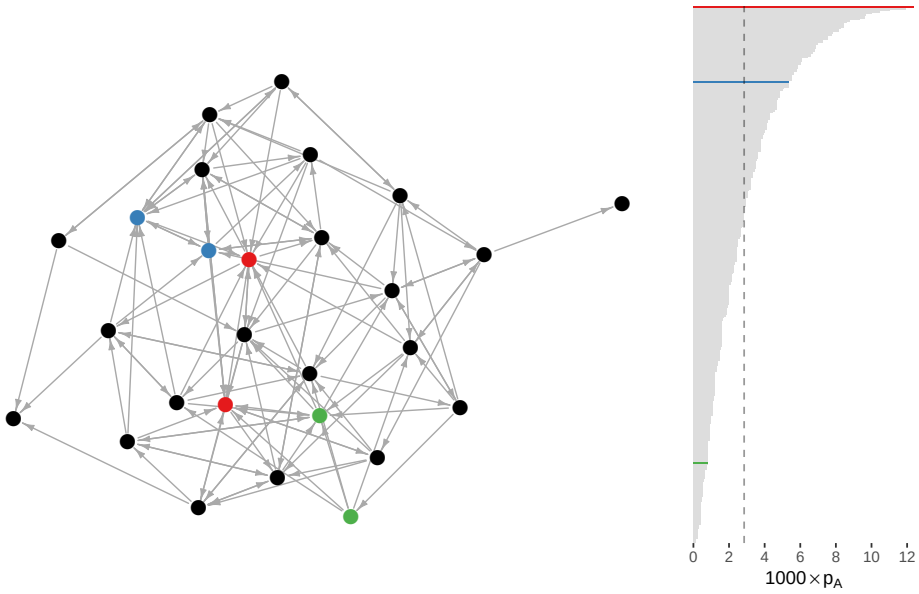

Figure 4.1: Illustration of the probability of a seed set being selected under the one-hop seeding strategy with seed sets of size $k = 2$. (left) Network for one village in Cai et al. (2015), with three possible seed sets highlighted: the seed sets with maximum probability under one-hop targeting (red) and two other example seed sets (blue, green). (right) Probability under the one-hop strategy (p_A) of all seed sets, with the highlighted seed sets (red, blue, and green) corresponding to those in the network on the left. The dashed line represents the uniform probability of each seed set under random seeding.

selected uniformly at random without using network information.[4] The goal was to measure the effectiveness of these strategies for two public health interventions: chlorine for water purification and multivitamins for micronutrient deficiencies. Selected seed individuals were provided information and coupons for these interventions, and the overall coupon adoption/redemption rate was used as the village-level outcome. The authors reported no statistically significant results for the redemption of chlorine coupons but a 95% confidence interval of a 6.9% to 17.9% increase in the redemption of multivitamin coupons, suggesting that one-hop targeting significantly increased the adoption rate for that intervention. This inference relies on strong parametric

[4]The authors use a third strategy where they target in-degree nodes, which we do not address here.

and within-village independence assumptions that would seem to be violated by the influence processes posited to cause the observed differences. Reflecting optimism about these targeting methods, but also the need for further empirical research, that team is conducting a larger, follow-up field experiment (Shakya et al., 2017).

The Kim et al. (2015) study demonstrates that even for quite large field experiments aimed at studying one-hop targeting, statistical power and nonparametric inference remain a challenge. Though the study contained 5,773 individuals, the nature of the intervention and village-level outcome means that the true effective sample size may be as small as 32 villages. Since field experiments remain expensive and imprecise, better methods are needed. To see how our insights into stochastic strategies apply here, consider a seed set selected by a uniformly at random strategy. If the observed seed set has positive probability under one-hop targeting (which it will, assuming all nodes in the seed set have positive in-degree) then it also provides information about the outcome under the one-hop targeting distribution, regardless of the fact that it was selected under the uniformly at random strategy. In fact, because the strategies are stochastic, a seed set selected uniformly at random will sometimes even have higher degrees than a set selected by one-hop targeting. This plays out in the Kim et al. (2015) study; of the eight villages that were assigned to one-hop and random targeting for the two different products, in three villages (37.5%) the random seed set had a higher mean in-degree than the one-hop seed set (see Figure 4.2, which is based on data presented in Table S3 of Kim et al. (2015)). Using these villages in a simple difference-in-means comparison of the sort used in the analysis by Kim et al. (2015) then becomes statistically counterproductive. In this work we show how we can instead use potentially *all* villages to learn about *both* targeting strategies, and conduct valid nonparametric asymptotic inference, as well as Fisherian randomization inference for small samples.

Our methods are not just restricted to the analysis of experiments that were designed specifically for the purpose of comparing seeding strategies, such as the Kim et al. (2015) study. In fact, our estimators can be applied to *any* experiment that randomizes node treatment in a collection of networks. This opens the possibility of

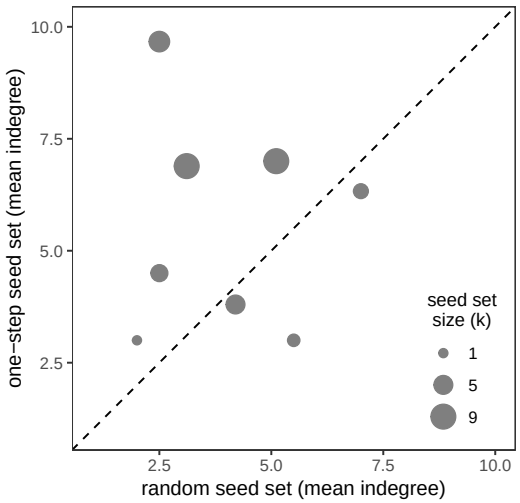

Figure 4.2: Mean indegree of the seed sets selected in Kim et al. (2015) for those villages assigned to one-hop targeting for one product (multivitamins or chlorine) and random targeting for the other.

analyzing a much larger collection of studies, as a substantial number of recent targeting experiments assign interventions using random targeting without the stated goal of evaluating the efficacy of other strategies. In this chapter we provide new empirical results pertaining to targeting strategies based on two such studies, reanalyzing the data from the farmer's insurance experiment in China conducted by Cai et al. (2015) and the peer conflict study in U.S. schools conducted by Paluck et al. (2016). We also present ways to optimize the design of experiments for distinguishing two targeting strategies. The optimization is based on the idea that some seed sets will be much more informative for comparing strategies than others, and such seed sets can be identified before having run the experiment.

Finally, as stated before, the framework we provide does not just apply to one-hop and random targeting but rather can be applied to evaluate the difference between any two stochastic seeding strategies with the same support. There are many natural stochastic strategies based on both local or non-local information. One-hop and random targeting both assume no more than local knowledge of the network; other local-information stochastic strategies include multi-hop targeting, targeting random

neighbors of a set with particular demographic traits, a mixed strategy that targets either a random node or a one-hop neighbor, or perhaps targeting both a random set *and* one of their random friends. The last example would be expected to behave favorably in domains governed by complex contagion (Centola and Macy, 2007). In settings where the full network is available, a wide range of stochastic strategies admit themselves: one could target nodes randomly proportional to degree (which is roughly achieved by one-hop targeting), randomly proportional to some more sophisticated centrality measure, or using some randomized algorithm to, e.g., greedily construct independent sets (Luby, 1986). Relatively little empirical attention has been devoted to some of the strategies mentioned here, but we hope that our proposed methods help to stimulate further research in this area and lower the barrier to entry for running more powerful field experiments.

4.1.1 Related work

The present work is related to both prior work on targeting in social networks and work in statistical methodology.

The literature on peer effects or "social contagion" is large, including purchasing behavior for products with direct (e.g., Tucker, 2008; Eckles et al., 2016) and indirect (e.g., Nair et al., 2004) network effects, and even those plausibly absent network effects (e.g., Aral et al., 2009). Social network information may be used for targeting in at least two ways. First, individuals may be indirectly affected by interventions received by others. Second, due to (dis)assortativity, social networks encode otherwise latent information about individuals (e.g., McPherson et al., 2001; Hill et al., 2006; Currarini et al., 2010; Altenburger and Ugander, 2018), so characteristics of network neighbors can be used for targeting even if there are no expectations of peer effects. Furthermore, there can be substantial heterogeneity in peer effects (Bakshy et al., 2012; Aral and Walker, 2012, 2014) and, in aggregate, in combination with network structure, large differences in the total influence of an individual's adoption (Bakshy et al., 2011; Iyengar et al., 2011; Yoganarasimhan, 2012; Banerjee et al., 2017; Galeotti et al., 2017).

There is a substantial literature on optimal seeding and approximations thereof given a known network and a model of how individuals are affected by the intervention and others' adoptions. In this literature on "influence maximization", computer scientists have developed algorithms for the problem of finding a set of k seeds so as to maximize expected adoption under various models (Domingos and Richardson, 2001). This influence maximization problem is NP-hard (Kempe et al., 2003), so much of that work (e.g., Chen et al., 2009) is concerned with efficient algorithms for tractable approximations under various models of social influence. Empirical evaluation of these proposed methods through field experiments is needed, as the efficacy of seeding strategies can be quite sensitive to deviations from the simple theoretical models sometimes used. For example, the costs associated with causing an individual to adopt can vary (cf. Bakshy et al., 2011) and influence and susceptibility may be correlated in the network (Aral et al., 2013; Aral and Dhillon, 2018); thus, the most influential individuals may be the most difficult to induce to adopt. Researchers in many disciplines have conducted empirical studies of seeding strategies making use of network information (e.g., Hinz et al., 2011; Kim et al., 2015; Beaman and Dillon, 2018; **?**; Banerjee et al., 2017). For example, **?** select seeds deterministically for some villages based on an optimization that involves an exhaustive search over all possible seed sets; this remains tractable because they select only $k = 2$ seeds per village. Much of the prior research has examined how one-hop targeting shifts the distribution of degree and centrality measures of seeds (e.g., Lattanzi and Singer, 2015; Kumar et al., 2018), compared with random targeting. Some work has analyzed resulting outcomes under assumed models of contagion (Kumar and Sudhir, 2018; Chami et al., 2017). Ideas based on the friendship paradox (which motivates one-hop targeting) have also been applied beyond seeding to the related problems of outbreak detection (Christakis and Fowler, 2010) and immunization (Cohen et al., 2003; Gallos et al., 2007; Chami et al., 2017).

Our proposed estimators are adaptations of estimators familiar from the literatures on importance sampling, counterfactual policy evaluation in bandit and reinforcement learning (e.g., Dudík et al., 2014b; Li et al., 2011; Precup et al., 2000; Li et al., 2014; Swaminathan and Joachims, 2015; Swaminathan et al., 2017), treatment rules (e.g.,

Hirano and Porter, 2009; Manski, 2004, 2007), and dynamic treatment regimes (e.g., Robins, 1986; Murphy et al., 2001; Murphy, 2003; Hernán et al., 2006). Dudík et al. (2014b) and Athey and Wager (2017) include reviews that span these multiple literatures. Much of this related methodological work focuses on evaluating and learning deterministic policies, though some work considers extensions where stochastic policies are addressed. For example, Murphy (2003) consider policies that simply assign treatment with some non-degenerate probability; Muñoz and van der Laan (2012) consider shifts to the distribution of a continuous or many-valued treatment, such as the number of hours spent in vigorous exercise or time from onset of symptoms to treatment. Stochastic policies are sometimes considered not because they are of interest *per se*, but because they solve positivity problems that may exist for evaluating deterministic interventions (Muñoz and van der Laan, 2012; Kennedy, 2018). In other cases, these policies may be of interest because policy-makers have limited control over treatment; this latter case is more similar to the present setting.

4.2 Problem formulation

Suppose we are interested in estimating the difference in adoption between one-hop and random targeting. We have data from an experiment in which each of n villages (or, e.g., schools, firms), labeled $i = 1, \ldots, n$, was randomly assigned to either random targeting or one-hop targeting. For each village we also have access to a graph $G_i = (V_i, E_i)$ that records the social connections among the $m_i = |V_i|$ residents of village i.

We focus on cases where the targeting strategies in the experiment only targeted seed sets of size k. Let $\mathcal{S}_i = \{s \subseteq V_i : |s| = k\}$ denote the collection of all such seed sets, which comprise the *eligible seed sets* in village i. Let S_i be a random variable that represents the seed set selected in village i, sampled from $\mathcal{S}_i$. Formally, each stochastic targeting strategy for village i imposes a non-degenerate probability distribution for the random variable S_i. Let p_i^A denote the seed set probability distribution for one-hop targeting in village i, and let p_i^B denote the seed set probability for random targeting in village i. Throughout will use A to represent one-hop targeting and B to

represent random targeting but, more generally, much of our approach can be used to compare arbitrary stochastic strategies A and B as well.

We first discuss the random targeting probabilities because they are more straightforward. In random targeting, all eligible seed sets are equally likely, so p_i^B is characterized by the uniform probabilities

$$\mathbf{P}_i^B(S_i = s) = \binom{m_i}{k}^{-1}, \qquad \text{for all } s \in \mathcal{S}_i,\ i = 1, \ldots, n, \tag{4.1}$$

where $\mathbf{P}_i^B$ denotes probability with respect to p_i^B. Notice that this is independent of the graph structure of G_i (the network structure of village i), depending only on the number of residents m_i.

The one-hop targeting distributions p_i^A, meanwhile, depend on the structure of G_i. There are multiple variations on one-hop targeting that we describe here. One form of the one-hop targeting strategy proceeds in sequential fashion. A "nominator" individual is randomly selected, and then a seed individual is then selected randomly from among the neighbors of the nominator. This seed selection process is repeated until the desired seed set size k is reached, while requiring that no individual is used as a nominator more than once and that all selected seeds are unique. Because nominators are used only once, the effective graph is modified every time a node is removed. For an arbitrary network there is no simple analytic form for p_i^A under this version of one-hop targeting.[5] Alternatively, we consider a simpler form of the one-hop targeting strategy which assumes independence among draws of the seed individuals, meaning that the nominators are drawn with replacement. As a result, seeds can be thought of as drawn independently (using the fixed network G_i) rather than sequentially.

To compute seed set probabilities under this one-hop targeting strategy we first consider the case in which seeds themselves are also drawn with replacement (meaning seed sets are in fact multisets, possibly containing multiple copies of a single seed).

[5]For fixed k the probabilities can be calculated exactly in $O(m_i^k)$ time or approximated via Monte Carlo sampling from the targeting procedure, though even this quickly becomes intractable for many realistic settings.

The probability of selecting an individual node v is simply

$$\frac{1}{m_i} \sum_{u \in \mathcal{N}_{\text{in}}(v)} \frac{1}{d_u^{\text{out}}}, \tag{4.2}$$

where $\mathcal{N}_{\text{in}}(v)$ denotes the set of in-neighbors of v, and d_u^{out} denotes the out-degree of node u. Then $\mathbf{P}_i^{A,\text{repl}}$, the probability with respect to p_i^A with replacement, is then:

$$\mathbf{P}_i^{A,\text{repl}}(S_i = s) = k! \prod_{v \in s} \frac{1}{m_i} \sum_{u \in \mathcal{N}_{\text{in}}(v)} \frac{1}{d_u^{\text{out}}}. \tag{4.3}$$

The $k!$ in the above expression comes from the fact that we seek the probabilities for unordered sets. For a given seed set s this probability is straight-forward to compute.

With the above probabilities in hand, we still need to translate each probability with replacement to one without replacement (where the seeds are unique). We can do this translation if we know the overall probability of selecting a set of size k that is unique. In G_i there are $\binom{m_i}{k}$ unique seed sets of size k, and when this quantity is manageable we can simply compute the total probability

$$\pi_i = \sum_{s \in \mathcal{S}_i : s \text{ unique}, |s|=k} \mathbf{P}_i^{A,\text{repl}}(S_i = s), \tag{4.4}$$

which lets us use a simple normalization to compute the one-hop targeting probabilities without replacement:

$$\mathbf{P}_i^{A}(S_i = s) = \frac{1}{\pi_i} \mathbf{P}_i^{A,\text{repl}}(S_i = s). \tag{4.5}$$

As noted above, this computation is only manageable when $\binom{m_i}{k}$, the number of unique seed sets of size k, is a manageable quantity. One village in the Cai et al. (2015) data we analyze has 49 nodes and 13 seeds, meaning that there are $\binom{49}{13} \approx 262$ billion unique seed sets. In such settings (or in larger networks, such as those in the Kim et al. (2015) experiment), we can still follow the above approach if we have a suitable estimator of π_i, the probability of selecting a unique seed set. We discuss such a suitable estimator $\hat{\pi}_i$ in Section 4.6, which we employ in our reanalysis of Cai

et al. (2015) in Section 4.8.

For every village the experiment produces a village-level response Y_i. We follow the potential outcomes framework (Neyman, 1923; Rubin, 1974) and assume that the potential outcomes are fixed at the seed set level, meaning that $Y_i = y_i(S_i)$, where $y_i : \mathcal{S}_i \to \mathcal{Y}$ is a function mapping from the space of seed sets to the outcome space $\mathcal{Y}$. In most of the cases that we study, Y_i is a count or fraction of adopters of the village and so $\mathcal{Y} = \mathbb{Z}$ or $\mathcal{Y} = [0, 1]$, but this fact is not important for any of our results.

Given the above notation, we can now state the estimand for the experiment in both a finite population and superpopulation framework. The goal of the experiment is to estimate the difference in expected outcomes for each of the two targeting strategies:

$$\tau_{\text{fp}} = \frac{1}{n} \sum_{i=1}^{n} \left[\mathbf{E}_i^A[y_i(S_i)] - \mathbf{E}_i^B[y_i(S_i)] \right], \tag{4.6}$$

where $\mathbf{E}_i^A$ and $\mathbf{E}_i^B$ denote expectation over $S_i \sim p_i^A$ and $S_i \sim p_i^B$, respectively. Here τ_{fp} is a finite population estimand, which considers the villages to be fixed.

We also consider estimation of the superpopulation estimand, in which the villages (and the corresponding networks G_i) are viewed as an i.i.d. sample from an infinite superpopulation. In this case, we may consider a single *design distribution* p_Δ, from which seed sets are sampled i.i.d. The goal is to study the difference between the superpopulation one-hop targeting distribution (denoted p_A) and the superpopulation random targeting distribution (denoted p_B). The finite population distributions discussed above, p_i^A and p_i^B, result from conditioning p_A and p_B on the observed network G_i for each sampled village i.

In the superpopulation framework seed sets are i.i.d. draws of a random variable S. We observe i.i.d. realizations of a response variable $Y = y(S)$, and the superpopulation estimand is then

$$\tau_{\text{sp}} = \mathbf{E}_A[y(S)] - \mathbf{E}_B[y(S)], \tag{4.7}$$

where $\mathbf{E}_A$ and $\mathbf{E}_B$ denote expectation with respect to the superpopulation targeting distributions.

Notice that τ_{sp} can be also be written as

$$\tau_{\text{sp}} = \mathbf{E}_{\Delta}\left[\frac{p_A(S) - p_B(S)}{p_{\Delta}(S)} y(S)\right].$$

4.3 Estimators

We now derive estimators for the above estimands. Let Z_i be the village-level treatment indicator, where $Z_i = 1$ if village i is assigned to one-hop targeting and $Z_i = 0$ if village i is assigned to random targeting. The observed data consist of the realized treatment assignments $z_1, \ldots, z_n$ and outcomes $y_1, \ldots, y_n$. We use lowercase letters to indicate that these are observed values.

The simplest estimator is the difference-in-means estimator, which is simply the difference in sample means for villages assigned to each targeting strategy:

$$\hat{\tau}_{\text{DM}} = \frac{1}{\sum_{i=1}^{n} z_i} \sum_{i=1}^{n} z_i y_i - \frac{1}{\sum_{i=1}^{n} (1 - z_i)} \sum_{i=1}^{n} (1 - z_i) y_i.$$

The difference-in-means estimator is unbiased for both τ_{fp} and τ_{sp}, but we can increase precision (Särndal, 1976) by noting that each observation is potentially informative about multiple targeting strategies: if an observed seed set has positive probability under a particular targeting strategy, then it provides information about that strategy.

We assume that the experiment selects $Z_i \sim \text{Bernoulli}(\rho)$, where $\rho \in (0, 1)$ is the treatment assignment probability. Each seed set S_i is then sampled from the mixture distribution

$$S_i \sim \rho p_i^A + (1 - \rho) p_i^B.$$

We refer to this distribution as the *design distribution* for village i, denoted by p_i^{Δ}, with probabilities given by

$$\mathbf{P}_i^{\Delta}(S_i = s) = \rho \mathbf{P}_i^A(S_i = s) + (1 - \rho)\mathbf{P}_i^B(S_i = s). \tag{4.8}$$

Since p_i^A, p_i^B, and p_i^{Δ} are all completely known, we in particular know the exact

probabilities corresponding to the observed seed sets. Let s_i be the observed seed set for village i and let

$$a_i = \mathbf{P}_i^A(S_i = s_i),$$
$$b_i = \mathbf{P}_i^B(S_i = s_i),$$
$$d_i = \mathbf{P}_i^\Delta(S_i = s_i),$$

denote the corresponding observed probabilities. (We use lowercase, Latin letters to emphasize that a_i, b_i, and d_i are known and observed.) We can then compute reweighting estimators by defining the weights

$$w_i^A = a_i/d_i, \qquad w_i^B = b_i/d_i. \tag{4.9}$$

Then the Horvitz–Thompson estimator (Horvitz and Thompson, 1952) is defined by

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}(w_i^A - w_i^B)y_i. \tag{4.10}$$

By defining the normalized weights

$$\tilde{w}_i^A = \frac{w_i^A}{n^{-1}\sum_j w_j^A}, \qquad \tilde{w}_i^B = \frac{w_i^B}{n^{-1}\sum_j w_j^B},$$

we obtain the Hájek estimator

$$\tilde{\tau} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{w}_i^A - \tilde{w}_i^B)y_i. \tag{4.11}$$

These estimators are familiar in the importance sampling literature, where the Horvitz–Thompson estimator is known as an *unnormalized importance sampling estimator* and the Hájek estimator is known as an *self-normalized importance sampling estimator* (see, for example, Owen (2013)). In that context, data is provided by means of an *importance distribution* but the desired quantities are population moments of a different *reference distribution*. For our application, the design distribution, p_i^Δ,

serves as the importance distribution and the two targeting distributions, p_i^A and p_i^B, are the reference distributions.

In the importance sampling literature there are competing arguments for whether to use the unnormalized estimator $\hat{\tau}$ or the normalized estimator $\tilde{\tau}$, and the optimal choice will depend on the particular application. See, for example, the discussion in Owen (2013, Ch. 9). The Horvitz–Thompson estimator is unbiased, as the unnormalized weights w_i^A and w_i^B have mean one. However, if they have excessive variance then the resulting Horvitz–Thompson estimator will be quite unstable. The Hájek estimator remedies this problem by forcing the mean of the weights $\tilde{w}_i^A$ and $\tilde{w}_i^B$ to be exactly equal to one. The Hájek estimator is biased, but this bias is negligible in large sample sizes. In our case, since the seed set probabilities are usually extremely small, we generally expect the self-normalized estimator to be more precise.

This approach can also be used for off-policy evaluation. Such off-policy evaluation requires positivity:

Assumption 4.1 (Positivity)**.** *For every village i and $s_i \in \mathcal{S}_i$, if $p_i^A(s_i) > 0$ or $p_i^B(s_i) > 0$, then $p_i^\Delta(s_i) > 0$.*

Positivity is satisfied for the mixture distribution considered above and for other cases. For example, suppose a seeding experiment on a collection of networks was designed for another purpose and used completely random assignment to treatment. In this case the design distribution is simply the random targeting distribution, $p_\Delta = p_B$, rather than the mixture distribution given by equation (4.8). Then $w_i^B = b_i/b_i = 1$, so of course the random targeting mean is estimated using the standard sample mean of the observations. The off-policy estimate for one-hop targeting is obtained by using the weights $w_i^A = a_i/b_i$.

Remark 4.1. In some common cases, Assumption 4.1 may not be strictly satisfied for the stochastic targeting strategies we have considered so far. For example, an experiment may have used a design that blocked (i.e., pre-stratified) on observables (e.g., household income), such that is was impossible for all k seeds to be, e.g., households in the highest income category. One can then consider variations on the

stochastic seeding strategies that condition on, e.g., the relevant balance between seeds and non-seeds.

Remark 4.2. Experimental designs may often be mixtures of stochastic and deterministic strategies (e.g., Kim et al. (2015) use a mixture of random seeding, one-hop seeding, and selecting the maximum in-degree nodes). The unconditional design distribution p_i^{Δ} may still satisfy Assumption 4.1 even if not all, or even none, of the component distributions do so individually.

4.4 Inference

In the previous section we described how counterfactual evaluation of village-level outcomes is possible for non-deterministic targeting strategies. The nature of our problem makes standard Neyman-style variance estimates (cf. Aronow and Middleton, 2013; Imbens and Rubin, 2015) for the finite population treatment effect τ_{fp} problematic. This is because the observations are drawn from seed sets of arbitrarily different sample spaces $\mathcal{S}_i$. As a result, observations from village i provide no information about village j, even if i and j were exposed to the same treatment strategy. For example, consider the Horvitz–Thompson estimator $\hat{\tau}$ defined in equation (4.10). Since villages are independent, $\hat{\tau}$ has variance

$$\text{Var}(\hat{\tau}) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}[(w_i^A - w_i^B) y_i].$$

The village-level variances must be estimated separately, and because we observe only a single seed set in each village i, estimating any of the terms $\text{Var}[(w_i^A - w_i^B) y_i]$ is impossible. Therefore, we focus on inference for the superpopulation average treatment effect τ_{sp}, so that information can be combined across different villages.

4.4.1 Asymptotic inference

In the standard importance sampling problem, the goal is to estimate a single population mean. Our problem differs slightly in that the importance sampled data are

repurposed to estimate a difference of two population means, rather than a single population mean. These estimates are correlated, so the standard importance sampling variance expressions and variance estimates do not directly apply. In what follows we compute novel expressions for the variances and variance estimates.

Proposition 4.1. *Let $S \sim p_\Delta$ be a random seed set and let $P_A = p_A(S)$, $P_B = p_B(S)$, and $P_\Delta = p_\Delta(S)$ be random variables representing the probabilities corresponding to seed set S. Let $Y = y(S)$, $W_A = P_A/P_\Delta$, and $W_B = P_B/P_\Delta$. Then the Horvitz–Thompson estimator $\hat{\tau}$, defined in equation* (4.10)*, has expectation $\mathbf{E}[\hat{\tau}] = \tau_{sp}$ and $\mathrm{Var}(\hat{\tau}) = V_{\hat{\tau}}/n$, where*

$$V_{\hat{\tau}} = \mathbf{E}\left[\frac{1}{P_\Delta^2}\left((P_A - P_B)Y - \tau_{sp}P_\Delta\right)^2\right] = \mathbf{E}[((W_A - W_B)Y - \tau_{sp})^2]. \tag{4.12}$$

Proof. To show unbiasedness, it suffices consider a single seed set $S \sim p_\Delta$, since the seed sets are sampled iid. For a single seed set S, the Horvitz–Thompson estimator (4.10) is

$$\hat{\tau} = (W_A - W_B)Y = \frac{(p_A(S) - p_B(S))y(S)}{p_\Delta(S)}.$$

This quantity has expectation

$$\mathbf{E}[\hat{\tau}] = \mathbf{E}_\Delta\left[\frac{p_A(S)y(S)}{p_\Delta(S)}\right] - \mathbf{E}_\Delta\left[\frac{p_B(S)y(S)}{p_\Delta(S)}\right] = \mathbf{E}_A[y(S)] - \mathbf{E}_B[y(S)] = \tau_{\mathrm{sp}}.$$

The variance of $\hat{\tau}$, for a sample of size n, is calculated as

$$\begin{aligned}\mathrm{Var}(\hat{\tau}) &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}(w_i^A - w_i^B)y_i\right] = \frac{1}{n}\mathbf{E}\left[\frac{(P_A - P_B)^2Y^2}{P_\Delta^2}\right] - \tau_{\mathrm{sp}}^2 \\ &= \frac{1}{n}\mathbf{E}\left[\frac{1}{P_\Delta^2}\left((P_A - P_B)Y - \tau_{\mathrm{sp}}P_\Delta\right)^2\right] = \frac{1}{n}\mathbf{E}[((W_A - W_B)Y - \tau_{\mathrm{sp}})^2] = V_{\hat{\tau}}/n,\end{aligned}$$

where $V_{\hat{\tau}}$ is defined in equation (4.12). □

It is easy to construct an unbiased estimate of the variance expression in Proposition 4.1 by substituting in observed sample quantities, producing the variance estimate

$$\hat{V}_{\hat{\tau}} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{d_i^2}((a_i - b_i)y_i - \hat{\tau} d_i)^2 = \frac{1}{n}\sum_{i=1}^{n}((w_i^A - w_i^B)y_i - \hat{\tau})^2. \tag{4.13}$$

The Hájek estimator is not unbiased, but it is correct in large samples in the sense that it is consistent and asymptotically normal.

Proposition 4.2. *Let S, P_A, P_B, P_Δ, and Y be as in Proposition 4.1. Let $\mu_A = \mathbf{E}_A[Y]$ and $\mu_B = \mathbf{E}_B[Y]$. Then the Hájek estimator $\tilde{\tau}$, defined in (4.11), satisfies $\tilde{\tau} \to \tau_{sp}$ as $n \to \infty$ and*

$$\sqrt{n}(\tilde{\tau} - \tau_{sp}) \Rightarrow N(0, V_{\tilde{\tau}}),$$

where

$$V_{\tilde{\tau}} = \mathbf{E}\left[\frac{1}{P_\Delta^2}\Big(\mu_A P_A - \mu_B P_B - Y(P_A - P_B)\Big)^2\right]. \tag{4.14}$$

Proof. Denote the sample averages

$$\bar{w}_A = \frac{1}{n}\sum_{i=1}^{n} w_i^A \qquad \bar{w}_B = \frac{1}{n}\sum_{i=1}^{n} w_i^B$$

and

$$\overline{wy}_A = \frac{1}{n}\sum_{i=1}^{n} w_i^A y_i \qquad \overline{wy}_B = \frac{1}{n}\sum_{i=1}^{n} w_i^B y_i$$

so that

$$\tilde{\tau} = \frac{\overline{wy}_A}{\bar{w}_A} - \frac{\overline{wy}_B}{\bar{w}_B}.$$

Consistency follows from the consistency of these sample averages and the continuous mapping theorem.

The normality result is a straightforward but slightly tedious application of the delta method, and follows the standard approach for characterizing the limiting behavior of ratio estimators (see, for example, Särndal et al. (1992, Section 5.6) or Owen (2013, Ch. 9)). Let $\hat{\beta} = (\bar{w}_A, \bar{w}_B, \overline{wy}_A, \overline{wy}_B)^\top$, and notice that $\beta := \mathbf{E}[\hat{\beta}] =$

$(1, 1, \mu_A, \mu_B)$. Furthermore $\sqrt{n}(\hat{\beta}-\beta) \Rightarrow N(0, \Sigma)$, where the entries of the asymptotic variance Σ are defined for $\Omega = A, B$, by

$$\begin{aligned}
\sqrt{n}\operatorname{Var}(\bar{w}_\Omega) &= \mathbf{E}\left[\frac{P_\Omega^2}{P_\Delta^2}\right] - 1 \\
\sqrt{n}\operatorname{Var}(\overline{wy}_\Omega) &= \mathbf{E}\left[\frac{P_\Omega^2 Y^2}{P_\Delta^2}\right] - \mu_\Omega^2 \\
\sqrt{n}\operatorname{Cov}(\bar{w}_\Omega, \overline{wy}_\Omega) &= \mathbf{E}\left[\frac{P_\Omega^2 Y}{P_\Delta^2}\right] - \mu_\Omega \\
\sqrt{n}\operatorname{Cov}(\bar{w}_A, \bar{w}_B) &= \mathbf{E}\left[\frac{P_A P_B}{P_\Delta^2}\right] - 1 \\
\sqrt{n}\operatorname{Cov}(\overline{wy}_A, \overline{wy}_B) &= \mathbf{E}\left[\frac{P_A P_B Y^2}{P_\Delta^2}\right] - \mu_A\mu_B \\
\sqrt{n}\operatorname{Cov}(\bar{w}_A, \overline{wy}_B) &= \mathbf{E}\left[\frac{P_A P_B Y}{P_\Delta^2}\right] - \mu_B \\
\sqrt{n}\operatorname{Cov}(\bar{w}_B, \overline{wy}_A) &= \mathbf{E}\left[\frac{P_B P_A Y}{P_\Delta^2}\right] - \mu_A
\end{aligned}$$

where $S \sim P_\Delta$. Now $\tilde{\tau} = f(\hat{\beta})$, where $f(a, b, c, d) = c/a - d/b$ and the gradient evaluated at β is $\nabla_\beta f = (-\mu_A, \mu_B, 1, -1)$. Then by the delta method,

$$\begin{aligned}
V_{\tilde{\tau}} = \nabla_\beta f^\top \Sigma \nabla_\beta f =& \mu_A^2\left(\mathbf{E}\left[\frac{P_A^2}{P_\Delta^2}\right] - 1\right) + \mu_B^2\left(\mathbf{E}\left[\frac{P_B^2}{P_\Delta^2}\right] - 1\right) \\
&+ \mathbf{E}\left[\frac{P_A^2 Y^2}{P_\Delta^2}\right] - \mu_A^2 + \mathbf{E}\left[\frac{P_B^2 Y^2}{P_\Delta^2}\right] - \mu_B^2 \\
&- 2\mu_A\left(\mathbf{E}\left[\frac{P_A^2 Y}{P_\Delta^2}\right] - \mu_A\right) - 2\mu_B\left(\mathbf{E}\left[\frac{P_B^2 Y}{P_\Delta^2}\right] - \mu_B\right) \\
&- 2\mu_A\mu_B\left(\mathbf{E}\left[\frac{P_A P_B}{P_\Delta^2}\right] - 1\right) - 2\left(\mathbf{E}\left[\frac{P_A P_B Y^2}{P_\Delta^2}\right] - \mu_A\mu_B\right) \\
&+ 2\mu_A\left(\mathbf{E}\left[\frac{P_A P_B Y}{P_\Delta^2}\right] - \mu_B\right) - 2\mu_B\left(\mathbf{E}\left[\frac{P_B P_A Y}{P_\Delta^2}\right] - \mu_A\right). \\
=& \mathbf{E}\left[\frac{1}{P_\Delta^2}\left(\mu_A^2 P_A^2 + \mu_B^2 P_B^2 + 2\mu_A P_A Y(P_B - P_A) - 2\mu_B P_B Y(P_A + P_B)\right.\right. \\
&\left.\left. + P_A^2 Y^2 + P_B^2 Y^2 - 2P_A P_B Y^2 - 2\mu_A\mu_B P_A P_B\right)\right].
\end{aligned}$$

Rearranging terms produces the expression in equation (4.14). □

We can estimate the Hájek variance with

$$\hat{V}_{\tilde{\tau}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i^2}(\tilde{\mu}_A a_i - \tilde{\mu}_B b_i - y_i(a_i - b_i))^2, \tag{4.15}$$

where $\tilde{\mu}_A = n^{-1}\sum_{i=1}^{n}\tilde{w}_i^A y_i$ and $\tilde{\mu}_B = n^{-1}\sum_{i=1}^{n}\tilde{w}_i^B y_i$ are the Hájek plug-in estimates of the population means.

These variance estimators can be used to construct confidence intervals and conduct hypothesis tests using normal theory.

4.4.2 Effective sample size diagnostics

A measure of effective sample size, obtained by comparing the variance to standard unweighted averages, can be a useful diagnostic for determining whether the importance distribution carries enough information to estimate the mean of the target distribution and may inform whether or not asymptotic approximations for inference are appropriate (Owen, 2013, Section 9.3). We focus on the self-normalized (Hájek) estimator in this section, as it is the weighted equivalent of an unweighted sample average.

Off-policy population mean

Consider the off-policy setting where we have observed outcomes under an importance distribution p_B and wish to estimate the population mean under a target distribution p_A. This is the case, for example, when seed sets are assigned according to random targeting (B) but we wish to make inferences about one-hop targeting (A). Given weights $w_i^A = a_i/b_i$ as in equation (4.9), that mean μ_A is estimated using the weighted average

$$\hat{\mu}_A = \frac{\sum_{i=1}^{n} w_i^A y_i}{\sum_{i=1}^{n} w_i^A}$$

Now, let $\sigma^2 = \mathrm{Var}(y_i)$. Then, conditionally on the observed weights,

$$\mathrm{Var}(\hat{\mu}_A) = \sigma^2 \frac{\sum_{i=1}^n (w_i^A)^2}{(\sum_{i=1}^n w_i^A)^2} = \sigma^2 \frac{\overline{w_A^2}}{n\bar{w}_A^2},$$

where

$$\bar{w}_A = \frac{1}{n}\sum_{i=1}^n w_i^A, \qquad \overline{w_A^2} = \frac{1}{n}\sum_{i=1}^n (w_i^A)^2.$$

In contrast, an unweighted sample average of n_{eff} independent observations has variance σ^2/n_{eff}, so $\hat{\mu}_A$ has the same variance as an unweighted average of

$$n_{\text{eff}} = \frac{(\sum_{i=1}^n w_i^A)^2}{\sum_{i=1}^n (w_i^A)^2} = \frac{n\bar{w}_A^2}{\overline{w_A^2}} \tag{4.16}$$

observations. If n_{eff} is much smaller than n, then it may be the case that p_A is too different from p_B to be able to estimate μ_A using observations from p_B. This notion of defining an effective sample size for weighted averages is quite old and is also known as a *design effect* in the survey sampling literature (see, for example, Kish, 1965).

The w_i^A are i.i.d. observations of a random variable $W_A = w_A(S)$, having population moments $\mathbf{E}_A[\bar{w}_A] = 1$ and $\mathbf{E}_A[\overline{w_A^2}] = \mathbf{E}_A W_A^2 = n\mathbf{E}_B W_A$. So a population version of n_{eff} can be written as

$$n_{\text{eff}}^* = \frac{n(\mathbf{E}_A[\bar{w}_A])^2}{\mathbf{E}_A[\overline{w_A^2}]} = \frac{n}{\mathbf{E}_A W_A^2} = \frac{n}{\mathbf{E}_B W_A}. \tag{4.17}$$

If the seed set distributions are known in advance then n_{eff}^* can be computed prior to launching a field experiment, which can give some indication of the informativeness of the experiment, say, with respect to different counterfactual policies. This may be useful when the entire social network is observed and when the seed sets are of small enough size to permit computation of the expectation specified in the expression for n_{eff}^*. Otherwise a Monte Carlo estimate of n_{eff}^* can easily be constructed by sampling seed sets, or the sample version n_{eff} can be used instead.

Average treatment effect

Now consider an experiment designed to compare strategies A and B, with observations assigned to both strategies, as in the field experiment conducted by Kim et al. (2015). Let n_A be the number of observations assigned to A and $n_B = n - n_A$ be the number of observations assigned to B. Consider the Hájek estimator $\tilde{\tau}$ of the average treatment effect

$$\tilde{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{w_i^A}{\bar{w}_A} - \frac{w_i^B}{\bar{w}_B} \right) y_i.$$

As in the off-policy case, if we consider the weights as fixed, then the variance is the sum of squares of the weights,

$$V_{\tilde{\tau}} = \frac{\sigma^2}{n^2} \sum_{i=1}^{n} \left(\frac{w_i^A}{\bar{w}_A} - \frac{w_i^B}{\bar{w}_B} \right)^2 = \frac{\frac{1}{n} \sum_{i=1}^{n} (\bar{w}_B w_i^A - \bar{w}_A w_i^B)^2}{n \bar{w}_A^2 \bar{w}_B^2}.$$

In contrast to the off-policy case, there is no standard notion of effective sample size for this two-sample scenario. The appropriate point of comparison is less clear because we are estimating a difference between two means. One possibility is to use as a comparison an ordinary two-sample equal-variance difference-in-means estimator.

In a hypothetical setup where we observe two independent samples of sizes n_A and $n_B = n - n_A$, the weights correspond to

$$w_i^A = \begin{cases} 1 & \text{if } i = 1, \ldots, n_A \\ 0 & \text{if } i = n_A + 1, \ldots, n \end{cases}$$
$$w_i^B = 1 - w_i^A.$$

Then the Hájek estimator reduces to the difference-in-means estimator and the variance is

$$V_{\hat{\tau}_{\mathrm{DM}}} = \frac{\sigma^2}{n^2} \sum_{i=1}^{n} \left(\frac{w_i^A}{\bar{w}_A} - \frac{w_i^B}{\bar{w}_B} \right)^2 = \frac{\sigma^2}{n^2} \left(\frac{n_A}{\bar{w}_A^2} + \frac{n_B}{\bar{w}_B^2} \right) = \sigma^2 \left(\frac{1}{n_A} + \frac{1}{n_B} \right),$$

the variance of an ordinary two-sample difference-in-means. For example, if our experiment consists of data from 100 villages where 50 villlages are assigned to random targeting and 50 villages are assigned to nomination targeting, then the difference-in-means (DM) estimator has variance $\sigma^2 \times (1/50 + 1/50) = \sigma^2/25$.

One reasonable definition of effective sample size, then, is to scale the original sample size n by the ratio of these two variances,

$$n_{\text{eff}} = \underbrace{n}_{\text{original sample size}} \times \underbrace{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}_{\text{DM scaling}} \times \underbrace{\frac{n\bar{w}_A^2\bar{w}_B^2}{\frac{1}{n}\sum_{i=1}^{n}(\bar{w}_B w_i^A - \bar{w}_A w_i^B)^2}}_{\text{inverse Hájek variance}}. \tag{4.18}$$

Of course, this is the same as

$$n_{\text{eff}} = \frac{1}{\rho(1-\rho)} \times \frac{n\bar{w}_A^2\bar{w}_B^2}{\frac{1}{n}\sum_{i=1}^{n}(\bar{w}_B w_i^A - \bar{w}_A w_i^B)^2},$$

where ρ is the proportion of units assigned to group A (one-hop targeting). In essence, n_{eff} is defined exactly so that the difference-in-means estimator in a Bernoulli(ρ) experiment has $n_{\text{eff}} = n$ (and should thus be viewed conditionally on the proportion parameter ρ).

As $\mathbf{E}[\bar{w}_A] = \mathbf{E}[\bar{w}_B] = 1$, the denominator on the right-hand side of equation (4.18) is a plug-in estimator for $\mathbf{E}[(W_A - W_B)^2]$. Therefore, a population version of n_{eff} can be stated as

$$n_{\text{eff}}^* = \frac{1}{\rho(1-\rho)} \times \frac{n}{\mathbf{E}[(W_A - W_B)^2]} = \frac{1}{\rho(1-\rho)} \times \frac{n}{\mathbf{E}\left[\left(\frac{P_A - P_B}{P_\Delta}\right)^2\right]}. \tag{4.19}$$

Notice that the effective sample size depends on the targeting distributions only through $\mathbf{E}\left[\left(\frac{P_A - P_B}{P_\Delta}\right)^2\right]$, capturing the intuition that the n_{eff}^* is lowered if discordant seed sets (those with very different probabilities between the two targeting strategies) are not accounted for by the design. Evaluating n_{eff}^* can be a useful indicator for how powerful hypothesis tests for comparing strategies A and B may be, and can be done before running any experiments provided the network structures are known.

4.4.3 Exact inference in finite samples

In order to use the preceding analysis of variance and associated variance estimators for inference, one can use asymptotic approximations. Measures of effective sample size, or other diagnostics, may caution against relying exclusively on such approximations. We may instead wish to conduct exact finite-sample inference without relying on parametric assumptions. We thus briefly consider Fisherian randomization inference (Fisher, 1925, 1935) for τ. In particular, we consider tests of the hypothesis $H_0 : \tau = 0$ and a sharp null hypothesis that outcomes are not affected by the seed set:

$$H_0^{\text{sharp}} : y_i(s) = y_i(s') \text{ for all } s, s' \in \mathcal{S}_i, i \in \{1, ..., n\}.$$

We can conduct an exact test of H_0^{sharp} via Fisherian randomization inference by drawing counterfactual seed sets according to the design and computing a test statistic with these counterfactual seed sets and the observed outcomes, as the outcomes are unchanged under H_0^{sharp}. While such a test is valid (i.e., results in nominal Type I error rates under H_0^{sharp}) with any choice of test statistic, it is common to use a "Studentized" test statistic based on the expectation that this will result in tests that are also asymptotically valid under the non-sharp null H_0 (Chung et al., 2013); for the Hájek estimator, this Studentized test statistic is $\tilde{\tau}/\sqrt{V_{\tilde{\tau}}}$.

4.5 Optimizing the experimental design

A benefit of our approach is that we can use the variance expressions, equations (4.12) and (4.14), as guidance for designing experiments targeted to maximize power for testing the difference between targeting strategies when the experiment will use a measured network, as in Kim et al. (2015).

First consider the Horvitz–Thompson estimator. Following from equation (4.12), a good choice of P_Δ is one that minimizes

$$\sigma_\Delta^2 := \mathbf{E}\left[\frac{(P_A - P_B)^2 Y^2}{P_\Delta^2}\right] - \tau_{\text{sp}}^2 = \mathbf{E}\left[\frac{(p_A(S) - p_B(S))^2 y(S)^2}{p_\Delta(S)^2}\right] - \tau_{\text{sp}}^2.$$

Then the choice

$$p_{\Delta^*} \propto \frac{|(p_A(S) - p_B(S))y(S)|}{\tau_{\text{sp}}}$$

is optimal. To see this, let p_Δ be any other design. Then

$$\begin{aligned}
\sigma^2_{\Delta^*} + \tau^2_{\text{sp}} &= \mathbf{E}_{\Delta^*}\left[\frac{(p_A(S) - p_B(S))^2 y(S)^2}{p_{\Delta^*}(S)^2}\right] = \tau^2_{\text{sp}} \\
&= \mathbf{E}_{\Delta}\left[\frac{(p_A(S) - p_B(S))y(S)}{p_{\Delta}(S)}\right]^2 \\
&\leq \mathbf{E}_{\Delta}\left[\frac{(p_A(S) - p_B(S))^2 y(S)^2}{p_{\Delta}(S)^2}\right] = \sigma^2_{\Delta} + \tau^2_{\text{sp}},
\end{aligned}$$

using the Cauchy-Schwarz inequality. Of course, p_{Δ^*} is not actually a useful distribution since it depends on the unknown true treatment effect, but it can provide hints on how to proceed. In particular, it suggests that seed sets for which the difference in probabilities $|P_A - P_B|$ and the magnitude of the response $|Y|$ are both large provide the most information about the hypothesis, and the distribution of the experimental design should thus place more weight on these seed sets.

Since the target distributions P_A and P_B are known, we must proceed by making assumptions about the response Y. The simplest such assumption, which we explore in Section 4.7.2, is to assume that Y is constant and take P_Δ to be proportional to $|P_A - P_B|$. This approach would maximize power in the event that the true data generating process in independent of the seeding strategies under consideration.

Under this optimized design, the design probability of a seed set will be zero whenever the probability of the set under one-hop targeting equals the probability under uniform random targeting. This design does not satisfy the positivity assumption (Assumption 4.1) as we've stated it in this work, but this optimized design is in fact unproblematic. As noted by Owen (2013, Section 9.1), it is enough to have $P_\Delta > 0$ whenever $|P_A - P_B| > 0$.

A more sophisticated approach would model the response using domain knowledge, perhaps via a social influence model such as the independent cascade model or the linear threshold model (Kempe et al., 2003). The most reliable approach is to use the results of a previous, pilot experiment to inform the design of the next experiment.

Such a bootstrapping procedure is a form of adaptive importance sampling (Owen, 2013, Section 10.5). Note that all of the analysis in this section relies on the fact that Y is non-negative; otherwise, one may treat the positive and negative parts of Y separately.

Designing an experiment for the Hájek estimator is similar. Examining the variance expression in equation (4.14), the optimal design is given by

$$p_{\Delta^*} \propto |\mu_A p_A(S) - \mu_B p_B(S) - y(S)(p_A(S) - p_B(S))|.$$

Again, p_{Δ^*} relies on the unknown quantities Y, μ_A and μ_B, which must be estimated in some way using historical data or domain knowledge.

4.6 Estimating one-hop targeting probabilities

As discussed in Section 4.2, the seed set probabilities under one-hop targeting strategy with replacement (of seeds) are given by

$$\mathbf{P}_i^{A,\text{repl}}(S_i = s) = k! \prod_{v \in s} \frac{1}{m_i} \sum_{u \in \mathcal{N}_{\text{in}}(v)} \frac{1}{d_u^{\text{out}}},$$

where $\mathbf{P}_i^{A,\text{repl}}$ denotes probability with respect to p_i^A with replacement, $\mathcal{N}_{\text{in}}(v)$ denotes the set of in-neighbors of v, and d_u^{out} denotes the out-degree of node u. For a given seed set s this probability is straight-forward to compute.

To translate probabilities with replacement to probabilities without, we compute the total probability

$$\pi_i = \sum_{s \in \mathcal{S}_i : s \text{ unique}, |s|=k} \mathbf{P}_i^{A,\text{repl}}(S_i = s), \tag{4.20}$$

which lets us use a simple normalization to compute the one-hop targeting probabilities without replacement:

$$\mathbf{P}_i^{A}(S_i = s) = \frac{1}{\pi_i} \mathbf{P}_i^{A,\text{repl}}(S_i = s).$$

In this section we discuss an estimator $\hat{\pi}_i$ for π_i, which is suitable when $|\mathcal{S}_i| = \binom{m_i}{k}$, the number of seed sets of size k (the number of terms in the sum (4.20) to compute π_i), is large.

For notational convenience, in the following presentation we will suppress village subscripts ($m_i = m$, $S_i = S$, $\pi_i = \pi$, etc.). Let $N = \binom{m}{k}$ and let

$$p_j = \mathbf{P}^{A,\text{repl}}(S = s_j) = k! \prod_{v \in s_j} \frac{1}{m} \sum_{u \in \mathcal{N}_{\text{in}}(v)} \frac{1}{d_u^{\text{out}}}$$

be the probability, with replacement, for each of the $j = 1, .., N$ seed sets of size k with unique elements. Note that $\pi = \sum_{j=1}^{N} p_j$.

4.6.1 The estimator $\hat{\pi}$

Our general approach will be to estimate π using a modest sample of R probabilities—probabilities with replacement—for seed sets that we can sample uniformly from sets of unique elements (sets *without* replacement). Individual probabilities with replacement are not hard to compute, we can compute them exactly using (4.20). The difficult in computing π is that there are so many sets. Producing a uniform sample can be done easily in $O(km)$ time, and each sample's probability under one-hop targeting can be computed in $O(km)$ time, so this computation that follows is a modest $O(kmR)$.

Let $X_1, ..., X_R$ be the i.i.d. random variables representing that probability computed for each sample. Then $\mathbf{P}(X_\ell = p_j) = 1/N$, for each X_ℓ, each j.

Now consider the estimator

$$\hat{\pi} = \frac{N}{R} \sum_{i=1}^{R} X_i.$$

It is clear that $\mathbb{E}[\hat{\pi}] = \frac{N}{R} R \frac{1}{N} \sum_{j=1}^{N} p_j = \pi$.

Since the estimator π relies in R samples from the uniform distribution over sets of size k with unique elements, a natural question is if we can improve the efficiency of the estimator, specifically by using stratified sampling of the seed sets. We briefly

give a simple stratification technique (that also has computational advantages) that we successfully employ in this work. At the same time we note that in the later analysis of the variance of $\hat{\pi}$ in this section, we assume all seed sets are sampled i.i.d.

Rather than sample R sets each of size k, a simple way to sample sets is to take a uniform shuffle of the node order and split the set sequentially into sets of size k, producing r=$\lceil m/k \rceil$ sets from a single shuffle. These r sets all constitute unbiased, albeit dependent, samples from the uniform distribution. Their dependence is a useful one, effectively serving as a stratified sample, requiring each node to appear at least once in each set of samples. In practice we find that this stratification reduces the variance of our estimates.

We have not thoroughly explored the possibilities for stratified sampling strategies for the estimator $\hat{\pi}$, nor have we considered other possible ways to estimate π that may be more efficient. The estimators (basic and stratified) discussed here are adequate for the scale of data we handle in this work.

4.6.2 Variance estimation and bound

The variance of $\hat{\pi}$ is given by:

$$\begin{aligned} \mathrm{Var}[\hat{\pi}] &= \frac{N^2}{R^2} \sum_{i=1}^{R} \mathrm{Var}[X_i] \\ &= \frac{N^2}{R} \frac{1}{N} \sum_{j=1}^{N} (p_j - \bar{p})^2, \end{aligned} \tag{4.21}$$

where $\bar{p} = \frac{1}{N} \sum_{j=1}^{N} p_j$. Notice $N\bar{p} = \pi$, so if we knew $\bar{p}$ then we wouldn't need $\hat{\pi}$.

Notice that if the probabilities p_j are uniformly $1/N$, as they are for any in- and out-regular graph, then the variance of $\hat{\pi}$ is zero. More interestingly, we can bound this variance in non-vacuous ways, without estimating it, using efficiently computable properties of the graph.

Variance estimate. Let S_p^2 be the unbiased sample variance of $X_1, \ldots, X_R$, the

R probabilities corresponding to uniform samples,

$$S_p^2 = \frac{1}{R-1}\sum_{i=1}^{R}(X_i - \bar{X})^2. \tag{4.22}$$

This sample variance is an unbiased estimate of the variance of each of the i.i.d. X_i's. As such, we can estimate the variance of $\hat{\pi}$ as:

$$\hat{V}_{\hat{\pi}} = \widehat{\mathrm{Var}}[\hat{\pi}] = \frac{N^2}{R}S_p^2. \tag{4.23}$$

Variance bound. We will give two upper bounds: one bound that depends on efficiently computable properties of the graph and one that depends merely on the maximum and minimum in- and out-degrees.

We begin by noting that $\frac{1}{N}\sum_{j=1}^{N}(p_j - \bar{p})^2$ in (4.21) is the variance of a discrete random variable that has compact support on $[0, 1]$, and we can tighten this support further by deriving maximum and minimum probabilities $p_{\max}$ and $p_{\min}$. For a discrete random variable with compact support on $[p_{\min}, p_{\max}]$, we can bound (4.21) by:

$$\begin{aligned}
\mathrm{Var}[\hat{\pi}] &= \frac{N^2}{R}\frac{1}{N}\sum_{j=1}^{N}(p_j - \bar{p})^2 \\
&\leq \frac{N^2}{R}\frac{1}{2}\left[\left(p_{\max} - \frac{p_{\max}+p_{\min}}{2}\right)^2 + \left(p_{\min} - \frac{p_{\max}+p_{\min}}{2}\right)^2\right] \\
&= \frac{N^2}{4R}(p_{\max} - p_{\min})^2.
\end{aligned} \tag{4.24}$$

Recall the probability (with replacement) of each seed set s:

$$p_j = \frac{k!}{m^k}\prod_{v\in s}\underbrace{\sum_{u\in\mathcal{N}_{\mathrm{in}}(v)}\frac{1}{d_u^{\mathrm{out}}}}_{c_v},$$

where we've isolated the terms c_v as those that depend on the nodes v in the set. We can compute $\{c_v\}_{v\in V}$ from the graph for all nodes in $O(n)$ time. Let $c_1 \leq c_2 \leq \cdots \leq c_{n-1} \leq c_n$ be the individual terms in sorted order. Given these quantities in sorted

order, we then know that $p_{\min}$ and $p_{\max}$ are simply the products of the smallest and largest k elements, respectively:

$$p_{\min} = \min_{s_j \in \mathcal{S}, s_j \text{ unique}} \mathbf{P}^{A,\text{repl}}(S = s_j) = \frac{k!}{m^k} \prod_{i=1}^{k} c_i,$$
$$p_{\max} = \max_{s_j \in \mathcal{S}, s_j \text{ unique}} \mathbf{P}^{A,\text{repl}}(S = s_j) = \frac{k!}{m^k} \prod_{i=n-k+1}^{n} c_i.$$

Returning to (4.24), we have the computable bound:

$$\begin{aligned}\text{Var}[\hat{\pi}] &\leq \frac{N^2}{4R} \left(\frac{k!}{m^k} \prod_{i=n-k+1}^{n} c_i - \frac{k!}{m^k} \prod_{i=1}^{k} c_i \right)^2 \\ &= \frac{1}{4R} \left(\frac{k!\binom{m}{k}}{m^k} \right)^2 \left(\prod_{i=n-k+1}^{n} c_i - \prod_{i=1}^{k} c_i \right)^2. \end{aligned} \tag{4.25}$$

In the last step we've re-introduced $N = \binom{m}{k}$ to highlight the fact that the ratio $k!\binom{m}{k}/m^k \leq 1$, for all m, k. For the largest village in the Cai et al. (2015) dataset $m = 49$ and $k = 13$, so $(k!\binom{m}{k}/m^k)^2 \approx 0.174^2 \approx 0.0303$. The quantity (4.25) is efficiently computable for any graph.

We can also furnish a simpler (but almost always looser) bound using the maximum and minimum in-/out-degree of the graph. Let $d_{\text{in-max}} = \max_{v \in V} d_v^{\text{in}}$, $d_{\text{in-min}} = \min_{v \in V} d_v^{\text{in}}$, $d_{\text{out-max}} = \max_{v \in V} d_v^{\text{out}}$, and $d_{\text{out-min}} = \min_{v \in V} d_v^{\text{out}}$. Then

$$\prod_{i=n-k+1}^{n} c_i \leq \left(\frac{d_{\text{in-max}}}{d_{\text{out-min}}} \right)^k, \qquad \prod_{i=1}^{k} c_i \geq \left(\frac{d_{\text{in-min}}}{d_{\text{out-max}}} \right)^k.$$

The bound (4.25) then reduces to

$$\text{Var}[\hat{\pi}] \leq \frac{1}{4R} \left(\frac{k!\binom{m}{k}}{m^k} \right)^2 \left[\left(\frac{d_{\text{in-max}}}{d_{\text{out-min}}} \right)^k - \left(\frac{d_{\text{in-min}}}{d_{\text{out-max}}} \right)^k \right]^2. \tag{4.26}$$

For a graph that is in- and out-regular (all in- and out- degrees equal), this upper bound is zero, indicating that this bound is not always loose. While (4.26) is extremely

simple to compute, we derive it mostly to guide intuition. It can be much looser than (4.25) in practice, as is discussed below. The bound (4.25) is entirely straight-forward and always preferred.

Thus, given a desired precision for our estimate of π, the bound (4.25) can be used to select a sufficiently large R.

4.6.3 Estimator evaluation

We briefly evaluate the above estimator of the one-hop targeting probabilities in isolation from the many moving parts of our broader analysis of treatment effects. In Figure 4.3 we analyze the root mean squared error (RMSE) of the estimator across the 150 villages in the Cai et al. (2015) dataset, varying the number of samples R and size of seed sets k.

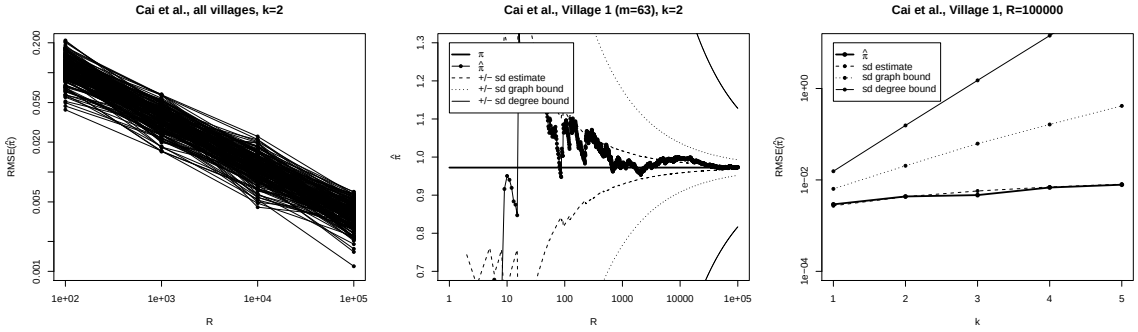

Figure 4.3: (left) The RMSE of $\hat{\pi}$ as a function of the number of samples R for each of the 150 villages in the Cai et al. (2015) dataset. (center) A convergence plot for Village 1 of that dataset, showing the estimate $\hat{\pi}$ explicitly as a function of R alongside estimates and bounds of the standard deviation. (right) The RMSE as a function of k, the size of the seed sets. The actual RMSE increases much slower than the graph- and degree-based bounds.

4.7 Simulations

In order to study our estimators in a setting where we can observe counterfactual outcomes, in this section we run simulations of behavior spreading on village networks

according to a known model. In order to accurately capture the network structure and heterogeneity exhibited among villages, we use the networks from Cai et al. (2015), the same networks we study in the empirical analysis of actual insurance decisions in Section 4.8. See the empirical analysis for a discussion of our pre-processing steps for that data, which are less relevant to the present simulations.

We study the accuracy of the variance estimates and resulting coverage rates, the feasibility of off-policy evaluations, and a comparison of a commonly used experimental design for comparing seeding strategies vs. an optimal design.

4.7.1 Performance in simple designs

In this simulation, villages are assigned to one-hop or random targeting using Bernoulli(0.5) coin flips. This is similar to the design in Kim et al. (2015), but without blocking by village characteristics for simplicity. We fix the seed sets for all interventions to $k = 2$.

To generate outcomes, we use a model with endogenous social interactions such that latent utilities are linear-in-means. Our model is a dynamic model similar to that used in Eckles et al. (2017), which can be regarded as a noisy myopic best response model in a semi-anonymous game with strategic complements. Let S_{ij} be an indicator for whether individual j in village i is selected as a seed individual. Let $Y_{ij,t} \in \{0, 1\}$ denote the adoption state of individual j in village i at time t. We set the initial set of adopters to be the seed sets, $Y_{ij,0} = \mathbb{1}\{S_{ij} = 1\}$. Then we define the t-th time step response using the probit model

$$\begin{aligned} Y^*_{ij,t} &= \alpha + \beta Z_{ij,t} + \gamma X_i + \varepsilon_{ij,t}, \\ Y_{ij,t} &= \max\{Y_{ij,t-1}, \mathbb{1}(Y^*_{ij,t} > 0)\}. \end{aligned} \tag{4.27}$$

The intercept α captures a baseline threshold for adoption. Let G_{ijk} denote the jk-entry of the adjacency matrix for the network of village i. Let $d^-_{ij} = \sum_k G_{ijk}$ and $d^+_{ij} \sum_k G_{ikj}$ be in the out- and in- degrees of individual j in village i. Define $\tilde{G}_{ijk} = G_{ijk}/d^+_{ij}$ as that entry in the row-normalized adjacency matrix if $d^+_{ij} > 0$ and zero otherwise.

Then for time step t, we let

$$Z_{ij,t} = \sum_k \tilde{G}_{ijk} Y_{ik,t-1}$$

be the mean of neighboring responses of the previous time step. The parameter β thus captures the endogenous social effect portion of the utility linear-in-means model. We also include a term γ for a static, village-level variable

$$X_i = \sum_j S_{ij} d_{ij},$$

the sum of degrees of seed set individuals. Including this feature allows us to further vary the treatment effect between one-hop and random targeting in our simulations because one-hop targeting generates high degree seeds more often than random targeting. The term could, for example, capture social contagion that occurs outside of the observed network.

We use independent $\varepsilon_{ij,t} \sim N(0,1)$ noise, which is homoscedastic across time and individuals. The linear response $Y^*_{ij,t}$ is then thresholded at zero. We also require that $Y_{ij,t} = 1$ if $Y_{ij,t-1} = 1$, which enforces the constraint that adopters cannot revert to a state of non-adoption. The village-level response Y_i is the fraction of adopters after the maximum number of time steps T have been completed, $Y_i = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij,T}$. We set $T = 3$, noting that the average pairwise distance for most villages is less than 3.

For parameter values, we vary $\alpha \in \{-3, -2, -1, 0\}$, $\beta \in \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$, and $\gamma \in \{0, 0.1\}$. For each of 1,000 Monte Carlo replicates, we consider the following simulation procedure. We then conduct a simulated experiment by assigning each village to either one-hop targeting or random targeting via a Bernoulli(0.5) random variable and compute difference-in-means, Hájek, and Horvitz–Thompson estimators as well as the corresponding variance estimates described in Section 4.4. For the difference-in-means estimator we use the standard Neyman conservative variance estimator

$$\hat{V}_{\hat{\tau}_{DM}} = \frac{S_1^2}{N_1} + \frac{S_0^1}{N_0},$$

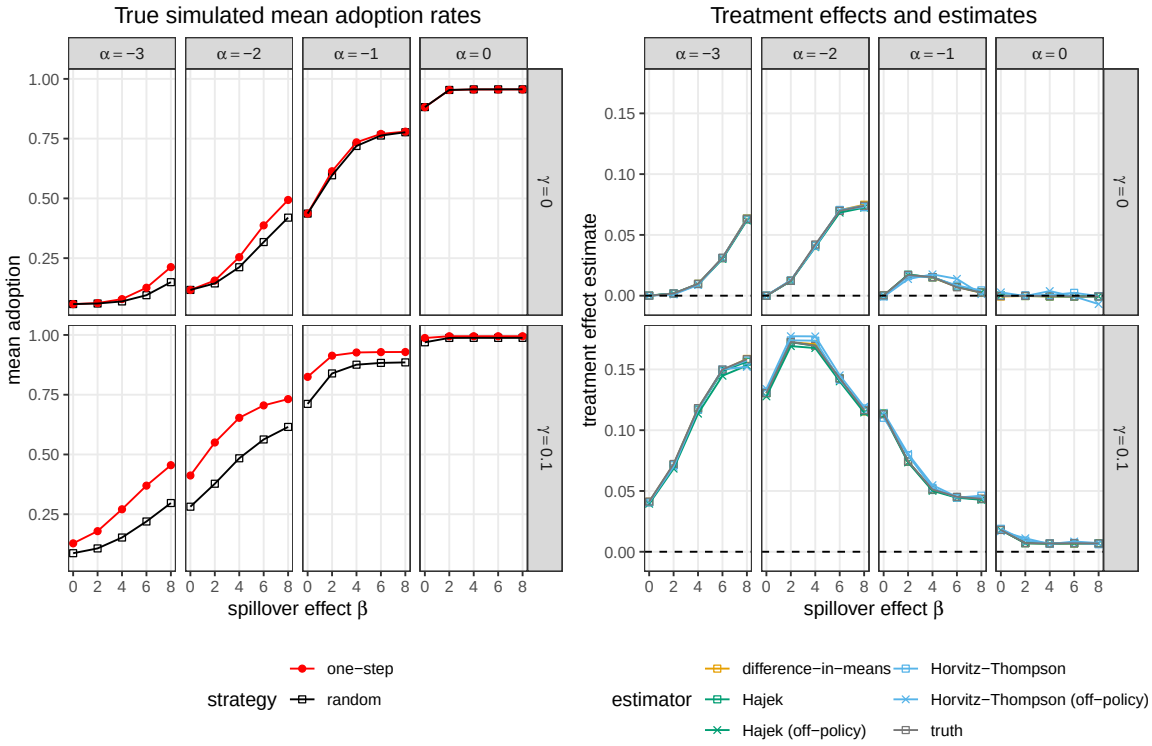

Figure 4.4: (left) True mean adoption rates for random and one-hop targeting. (right) True treatment effect and estimated values from the difference-in-means, Horvitz–Thompson, and Hájek estimators for the simulation setup described in Section 5.1. Each panel represents a pair (α, γ) of parameters from the model defined by equation (4.27); columns vary the intercept α and rows vary the degree effect γ. The horizontal axis varies the spillover effect β.

where S_1^2 and S_0^2 are the within-group sample variances and N_1 and N_0 are the group sample sizes. We also include off-policy Hájek and Horvitz–Thompson estimators that only use the data from villages assigned to random targeting; these estimators are thus handicapped by a smaller sample size and by reduced relevance of the sampled seed sets.

Results

Figure 4.4 (left) displays the true mean adoption rates under random and one-hop targeting. The parameter values used result in substantial variation in adopted rates

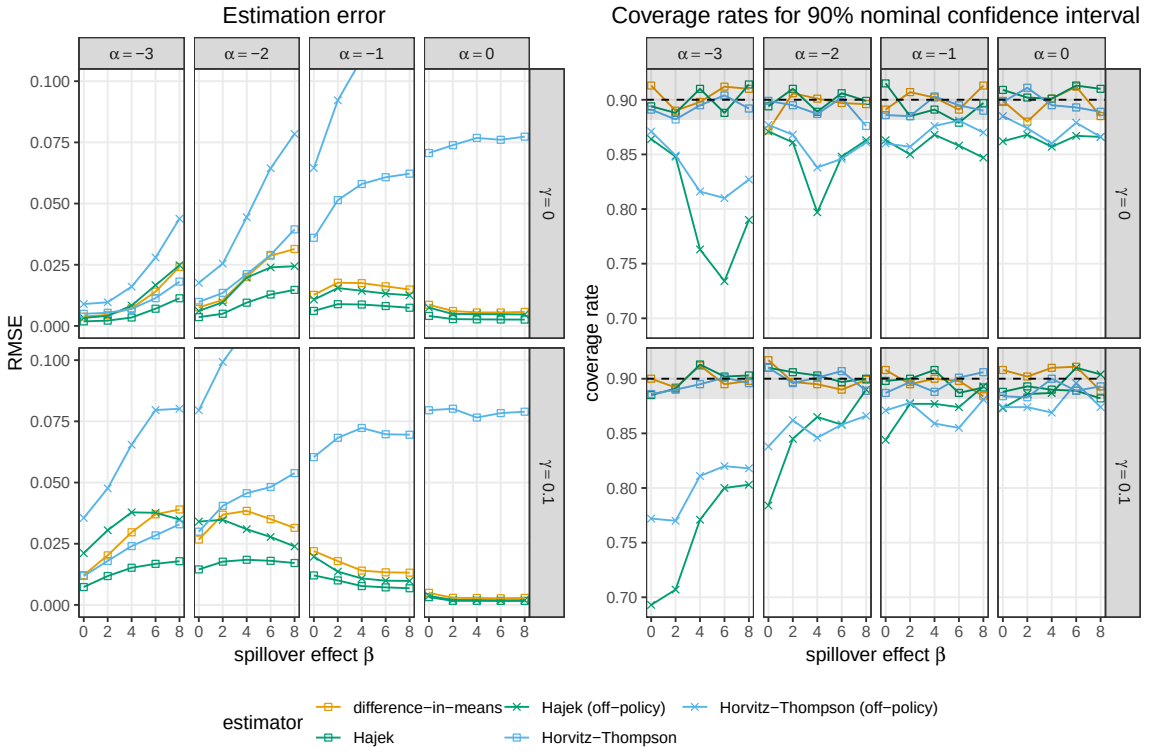

Figure 4.5: (left) Root mean-squared-error of estimators. Error increases with β as it produces more within-village dependence. The off-policy Horvitz–Thompson estimator has high variance and so is not visible in some panels. (right) Coverage rates for a 90% nominal confidence interval; shaded area is the 95% acceptance region ($p > 0.05$) for coverage being at least the nominal rate. All estimators have approximately nominal coverage, with the exception of the off-policy estimators, which are working with a much smaller sample size and effective sample size.

and treatment effects. Figure 4.4 (right) displays the average estimates along with the true treatment effect. These estimators are evaluated first according to their total error (Figure 4.5, left), as all are approximately unbiased. As expected, the Horvitz–Thompson estimators suffer from imprecision. On the other hand, the Hájek estimator generally has lower error than the difference-in-means. Notably, the off-policy Hájek estimator, making use of only half of the data and without any randomization to the one-hop strategy, sometimes outperforms the difference-in-means estimator with respect to RMSE. We next evaluate the variance estimators via the coverage of the resulting confidence intervals (Figure 4.5, right). The coverage is

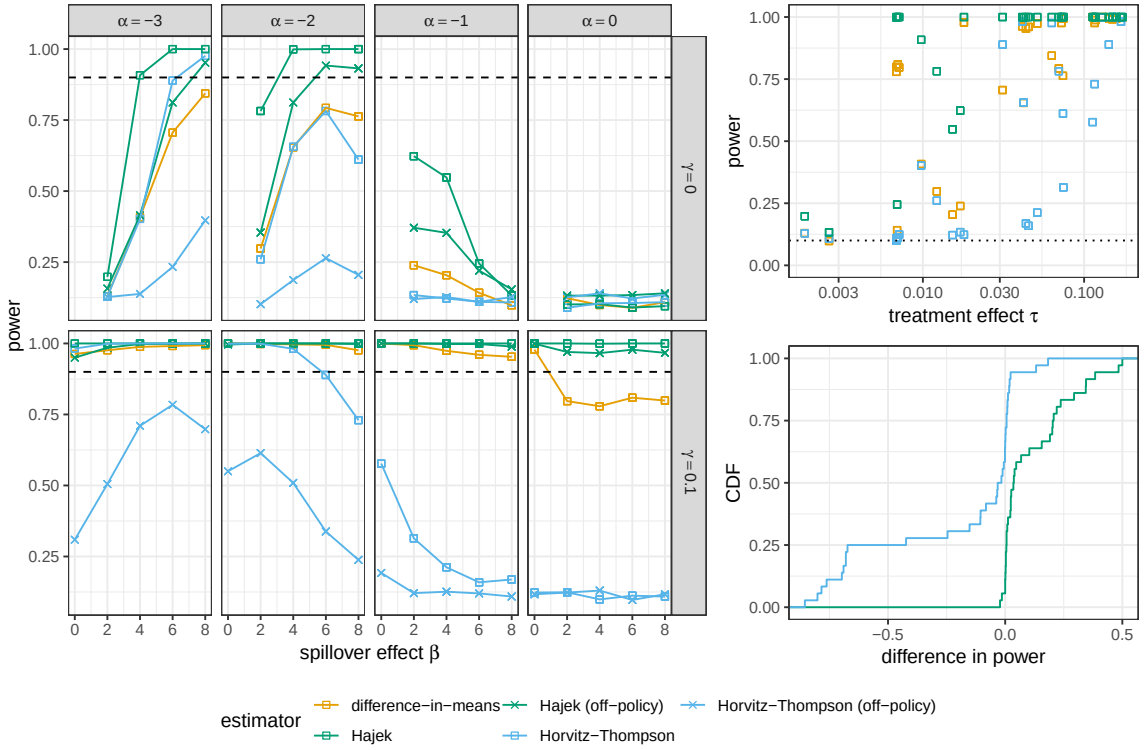

Figure 4.6: (left) Power of estimators. Estimators with below nominal coverage in a setting are not shown. (top right) Power for as a function of the true treatment effect. Horizontal axis is on a logarithmic scale. Off-policy estimators are not shown. (bottom right) Distribution of differences in power compared with difference-in-means across all settings. Off-policy estimators are not shown.

generally approximately at the nominal rate. The notable exception is for off-policy inference in settings with non-zero effects.

Given that the statistical inference is valid, we can ask how much the reduced error apparent in Figure 4.5 (left) translates into increased statistical power for the proposed estimators. Across all settings, the Hájek estimator has a true standard error that is 36% to 55% smaller; this corresponds to the gains from collecting data from 143% to 398% more villages. Figure 4.6 plots the power of the experiment (the fraction of experiments in which the null was rejected) as a function of both the simulation parameters (left) and the true treatment effect (top right). For the response model and parameter distribution used in our simulations, the Hájek estimator generally has

substantially more power than the difference-in-means estimator, while the Horvitz–Thompson estimator is underpowered due to excessive variance (Figure 4.6, bottom right).

4.7.2 Design and effective sample size

We now examine how the effective sample size, defined in Section 4.4.2, of the Hájek estimator varies for different designs and network structures. We analyze effective sample size for four additional collections of networks: the social networks of middle school students in New Jersey (Paluck et al., 2016), the social networks of students in the AddHealth study in the United States (Resnick et al., 1997), the social networks of villages from a study of the diffusion of microfinance in India (Banerjee et al., 2013), and the friendship networks of villages from a study of community health in Uganda (Chami et al., 2017). For each of these collections we compute the population effective sample size n^*_{eff} for the Hájek estimator, equation (4.19), with seed sets of size $k = 2$ under three different designs (Bernoulli, optimal under the null, and only random targeting). Calculating n^*_{eff} requires only knowledge of the network structure and no actual experimental outcome data.

The results are shown in Table 4.1. The difference-in-means estimator with a Bernoulli design would have effective sample size exactly equal to n. The results show that using the Hájek estimator brings substantial increases in precision. For example, for the Cai et al. (2015) dataset, using the Hájek estimator is equivalent to having run an experiment on 631 villages rather than the original 150, a fourfold increase. Precision is boosted further by using the optimal design described in Section 4.5. Note also that for all but one of the data sets, the effective sample size for off-policy estimation is greater than that for naïve estimation with the Bernoulli design; that is, the proposed estimators yield greater precision from an experiment not designed for the purpose of comparing one-hop and random targeting ($\tilde{\tau}$, off-policy from random targeting) than a difference-in-means estimator for a field experiment conducted for that purpose ($\hat{\tau}_{DM}$, Bernoulli(0.5)).

The calculations in Table 4.1 fix the seed set size at $k = 2$. In Table 4.2 we

Dataset	n	Population effective sample size n^*_{eff} $\tilde{\tau}$, Bernoulli(0.5)	$\tilde{\tau}$, optimal	$\tilde{\tau}$, off-policy
Cai et al.	150	631.72	871.16	233.36
Paluck et al.	56	351.60	539.04	128.34
AddHealth	85	319.36	448.80	131.04
Banerjee et al.	75	214.32	274.08	80.24
Chami et al.	17	37.84	47.92	9.32

Table 4.1: The population effective sample size n^*_{eff}, calculated using equations (4.17) and (4.19), for the Hájek estimator $\tilde{\tau}$ of the average treatment effect τ (between random and one-hop targeting) on five datasets, all collections of networks, under different designs all targeting $k = 2$ seed nodes. The off-policy evaluation is for estimating one-hop targeting from random targeting data. These n^*_{eff} can be interpreted as the number of villages needed for a difference-in-means estimator in a Bernoulli(0.5) experiment to have the same precision. The Hájek estimator always increases the effective sample size (in expectatation) over difference-in-means in a Bernoulli design, sometimes drastically. For all networks except Chami et al. (2017), the off-policy estimator has greater power even than an experiment designed explicitly for the purpose of comparing strategies.

study how the size of the seed sets impacts the effective sample size. We limit this investigation to $k = 1, \ldots, 4$, because for larger k the population effective sample size n^*_{eff} requires taking an expectation over many sets and becomes prohibitively expensive to compute. As seed sets become larger the selection probabilities under one-hop and random targeting diverge, reducing the benefits of both our Hájek estimator and optimized design. This reduction in power as k grows suggests that designs involving smaller seed sets are better for testing hypotheses about differences between one-hop and random targeting; seeding with a small k is also typically how these problems are posed. That said, these effective sample size calculations do not take into consideration the possible social influence mechanisms that may underly an outcome (Aral et al., 2013; Aral and Dhillon, 2018).

k	Population effective sample size n^*_{eff}		
	$\tilde{\tau}$, Bernoulli(0.5)	$\tilde{\tau}$, optimal	$\tilde{\tau}$, off-policy
1	1112.24	1585.48	370.28
2	631.72	871.16	233.36
3	466.68	619.40	149.92
4	383.72	490.76	98.08

Table 4.2: The population effective sample size n^*_{eff} for the Cai et al. (2015) dataset of 150 villages, for varying seed set sizes k. The values for $k = 2$ are the same as the values for Cai et al. (2015) in Table 4.1. The off-policy evaluation is for estimating one-hop targeting from random targeting data. The effective sample size decreases with k because the support of the distribution (number of seed sets) grows in k.

4.8 Empirical applications

The proposed estimators can be applied to existing field experiments. First, they can be used to increase the precision of estimation in experiments that do directly compare one-hop and random targeting. Second, they can be used for off-policy estimation of contrasts between one-hop and random targeting, even when, e.g., only random targeting was conducted, which we apply here.

4.8.1 Farmer's insurance experiment

We use our method to provide a new analysis of the data studied in Cai et al. (2015). The authors conducted a field experiment in villages in rural China to study peer effects in adoption of farmer's insurance. Villagers were assigned to one of four groups that varied the timing and intensiveness of the marketing intervention, and the presentation of information about village-wide uptake in the case of later sessions. We take the seed set to be the set of villagers assigned to the "intensive" session at the first period.[6] The response variable is the proportion of villagers who chose to purchase the farmer's insurance product.

[6]Cai et al. (2015) describes the experiment was stratified on median household size and rice area. Exploratory analyses seem inconsistent with the most natural interpretation of this description. Thus, for now and for simplicity, we analyze the experiment as if the design were an unstratified, completely randomized experiment.

	mean	st. dev.	min	max
Edges	93.0	37.4	17.0	188.0
Nodes m_i	27.6	9.4	8.0	49.0
Treated %	22.9	7.0	10.0	50.0
Treated count k_i	6.2	2.3	1.0	13.0
$\log_{10}\binom{m_i}{k_i}$	12.7	4.9	2.3	26.3
Takeup %	44.9	19.5	5.1	95.8

Table 4.3: Summary statistics for the 150 villages from Cai et al. (2015) analyzed here.

There are a small number of edges between residents of different villages; we drop these edges and consider the villages to be entirely disjoint. Not all villages had households assigned to treatments that varied within the village, and a few villages had insufficient network information; we drop all villages containing fewer than 25 edges. After this preprocessing, we are left with 150 villages, which contrasts with the 185 villages originally analyzed by Cai et al. (2015). Summary statistics for these 150 villages are given in Table 4.3.

For each village, we compute the random and one-hop targeting probabilities of the observed seed set, conditional on the observed seed set size. The random targeting probability is uniform across all seed sets of the same size, as in equation (4.1). For the one-hop targeting probability in equation (4.5), all of which involve large seed sets, we estimate the probabilities as discussed extensively in Section 4.6. Many (67) observed seed sets are not possible under one-hop targeting because they include nodes with zero in-degree. Aside from these cases the order of magnitude of the probabilities for both strategies are mostly determined by the size of the seed set, but there is enough discrepancy between the probabilities to facilitate off-policy estimation.

We compute the weights of the importance sampling estimators, w_i^A and w_i^B. Since the seed sets were assigned according to random targeting, the random targeting weights are constant, $w_i^B = b_i/b_i = 1$. The one-hop targeting weights are the ratio of one-hop to random targeting probabilities, $w_i^B = a_i/b_i$. The probabilities and normalized weights $(\tilde{w}_i^A, \tilde{w}_i^B)$ are displayed in Figure 4.7.

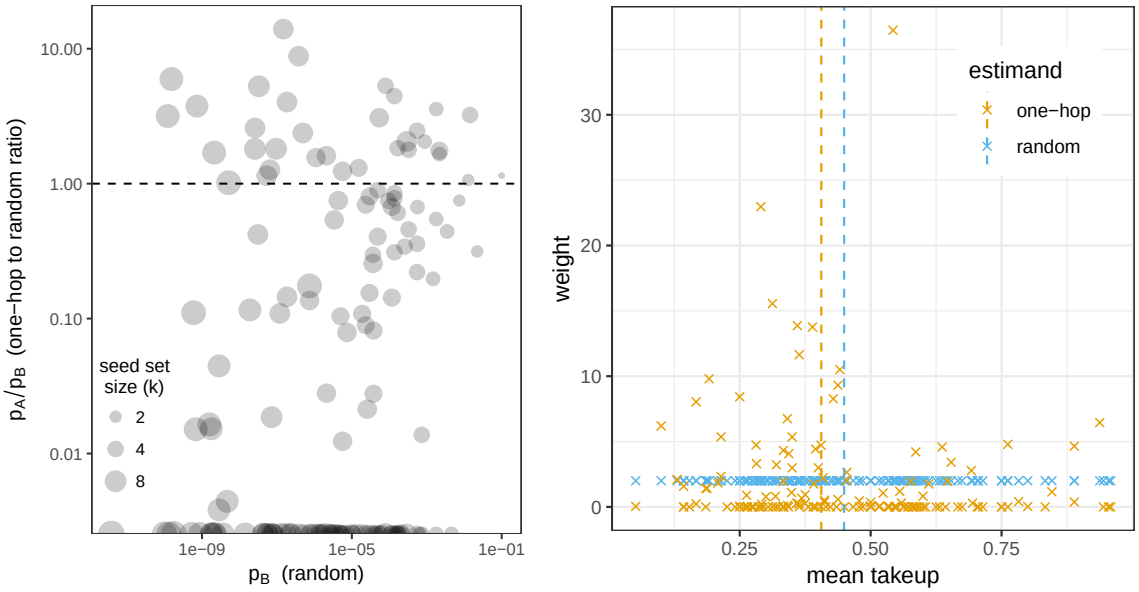

Figure 4.7: (left) The ratio of one-hop and random targeting probabilities for the 150 villages analyzed from the Cai et al. (2015) study. Absolute probabilities are correlated with seed set size, but there is considerable variation in the ratio. Seed sets with $p_B = 0$ are plotted at the bottom of the y-axis. (right) Weights as a function of the response and mean insurance takeup. Since the seed sets in the study were assigned via random targeting, the estimate for that strategy is an unweighted sample mean, whereas the one-hop targeting estimate applies reweighting. The vertical dashed lines are the (Hájek) estimated means.

Table 4.4 shows the Hájek estimate and associated inference. Asymptotic inference would lead to the conclusion that the one-hop strategy would *reduce* takeup of insurance by 1 to 7 percentage points. Bootstrap inference leads to more conservative but still suggestive conclusions. Given the simulation results in Section 4.7, in which we observed undercoverage in some settings for off-policy estimation, we should be cautious in relying on this statistical inference without further analysis. First, the one-hop targeting estimator has an effective sample size of $n_{\text{eff}} = 28.5$ using the off-policy effective sample size expression given in equation (4.16). The random targeting estimator of course has $n_{\text{eff}} = n = 150$. This suggests a great deal of caution in using estimated variance to conduct inference (e.g., to construct confidence intervals) based on normal theory as we have done here. Thus, we also conduct a hypothesis

	mean	st. dev.	min	max
Edges	3157.5	1395.3	1046.0	6561.0
Nodes m_i	426.4	172.9	138.0	835.0
Treated %	6.8	2.3	3.8	14.5
Treated count k_i	26.0	4.6	20.0	32.0
$\log_{10} \binom{m_i}{k_i}$	94.5	23.1	54.8	133.1
Peer conflict rate ($\times 100$)	15.2	13.4	0.0	49.1

Table 4.5: Summary statistics for the 28 treatment schools from Paluck et al. (2016) analyzed here.

test using Fisherian randomization inference as discussed in Section 4.4.3. This test also provides some evidence against the null of no effects of choice of seeds.

This reanalysis both demonstrates how our proposed estimators can be used off-policy and provides some cautionary results compared with previous evidence about the one-hop strategy. In particular, these results suggest that one-hop seeding may in some cases perform worse than simple random seeding.

estimate (one-hop − rand)	-0.0436
SE (analytic)	0.0209
SE (bootstrap)	0.0257
95% CI (analytic)	[-0.0846, -0.0027]
95% CI (bootstrap)	[-0.0909, 0.0088]
p-value (analytic)	0.0367
p-value (Fisherian)	0.0974

Table 4.4: Hájek estimate and inference for the difference in insurance takeup rates between one-hop and random seeding for Cai et al. (2015), which provide some evidence that one-hop seeding would have *reduced* adoption of insurance.

4.8.2 School conflict experiment

Paluck et al. (2016) conducted a field experiment in 56 middle schools in New Jersey, in which they randomly assigned an intervention designed to reduce bullying and other

peer conflict. We analyze data from the 28 schools assigned to treatment. A within-school randomization then assigned some students to be seeds: these students were invited to participate in a program that encouraged them to take a public stance against conflict among their peers at school. Paluck et al. (2016) measure several outcome variables of interest; here we focus on the number of peer conflict events per student as measured by administrative reports. For peer conflict events, lower values of the outcome are viewed as desirable. Summary statistics for these schools are given in Table 4.5.

Like our analysis of Cai et al. (2015), this analysis is also an off-policy evaluation, but Paluck et al. (2016) treat a smaller fraction of nodes, making it perhaps more typical of seeding with a limited budget. This intervention was also hypothesized to be more effective if more central individuals were seeds. In fact, Paluck et al. (2016) find that treatment reduces peer conflict more when a larger fraction of the seeds that are "social referents," defined as being in the top decile of in-degree for that school. Thus, one might expect that one-hop seeding would be effective in this setting.

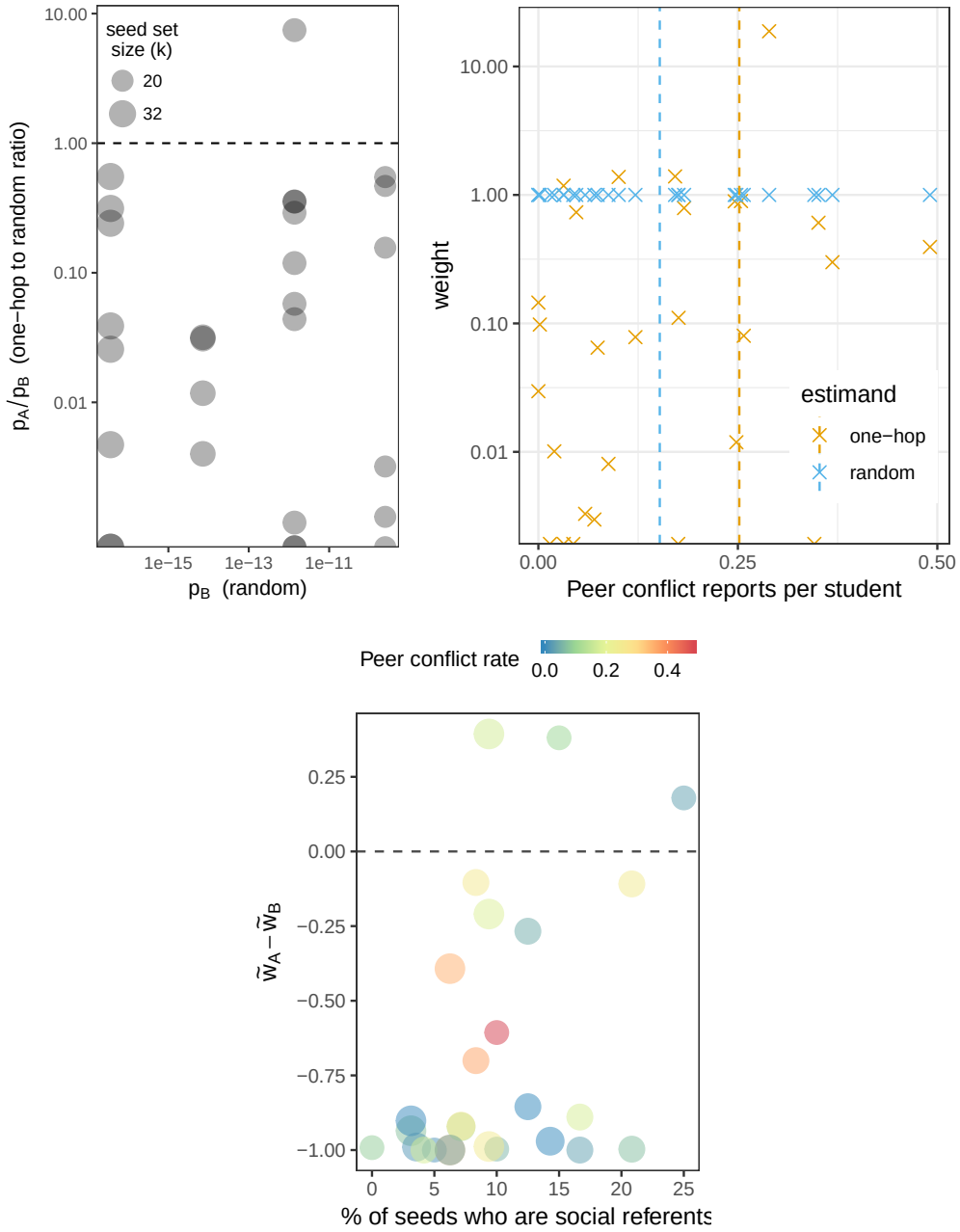

Figure 4.8: (left) The ratio of one-hop and random targeting probabilities for the 28 treated schools from the Paluck et al. (2016) study. One school has a seed with higher probability under one-hop than random seeding. (right) Weights as a function of a primary outcome, number of administrative peer conflict reports per student. The vertical dashed lines are the (Hájek) estimated means. (bottom) Relationship between measure of seed centrality used by Paluck et al. (2016) and the difference in weights used by our estimator. The school with very large positive $\tilde{w}_i^A - \tilde{w}_i^B$ (17.8) is not shown; 17% of its seeds were social referents.

Paluck et al. (2016) use a blocked (stratified) randomization to balance selected seed sets by four blocks formed by grade and gender. We thus consider a variation on one-hop seeding that conditions on selecting the observed number of seeds k_{ib} in each block b of school i; we could think of this as reflecting a desire to reach a select set of seeds. As with the previous analysis, Figure 4.8 (left) shows the ratio of probabilities p_A/p_B of the observed seed sets, illustrating that some (five) schools have seed sets that are impossible under one-hop seeding. Only one observed seed set is more probable under one-hop than random targeting.

estimate (one-hop − rand)	0.0997
SE (analytic)	0.0231
SE (bootstrap)	0.0432
95% CI (analytic)	[0.0543, 0.1451]
95% CI (bootstrap)	[0.0098, 0.1542]
p-value (analytic)	1.7e-05
p-value (Fisherian)	0.0846

Table 4.6: Hájek estimate and inference for the difference in peer conflict per student one-hop and random seeding for Paluck et al. (2016), which provide some evidence that one-hop seeding would have *increased* peer conflict (i.e., an undesirable outcome).

Table 4.6 shows the Hájek estimate of the average treatment effect on peer conflict of one-hop vs. random targeting and associated inference. Asymptotic and bootstrap inference would lead to the conclusion that the one-hop strategy would *increase* rates of peer conflict, as measured by administrative reports, by 0.01 to 0.16 incidents per student. Given the small number of schools, we conduct Fisherian randomization inference, which also provides some evidence against the null of no effect of choice of seed set. These results again suggest that one-hop seeding may in some cases perform worse than simple random seeding.

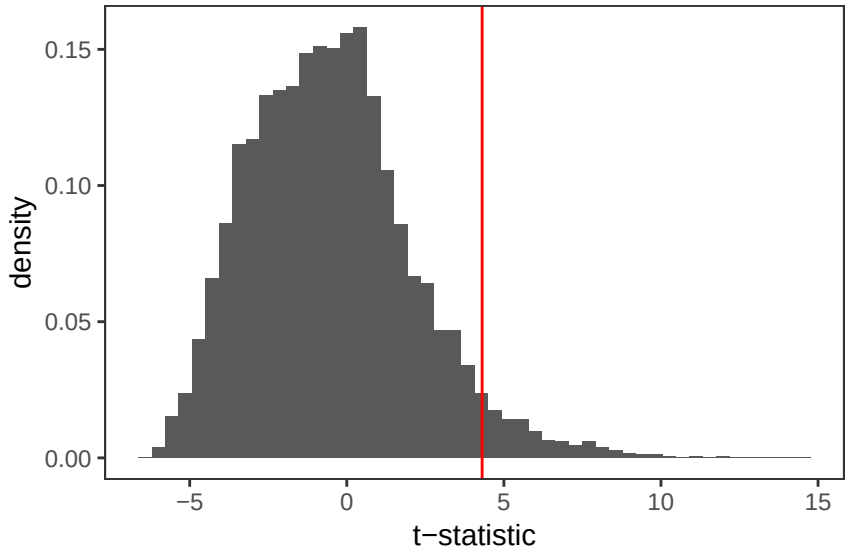

Figure 4.9: Fisherian randomization inference for peer conflict per student using the distribution of Studentized Hàjek estimator (i.e., t-statistic) under a sharp null. This figure elaborates on Table 4.6. Observed statistic shown by red line.

Here we provide results for two additional outcomes: the self-reported rate of wearing an anticonflict wristband and the self-reported number of friends talking about peer conflict. An increase in both of these outcomes is viewed as a desirable result. These results, like those in Paluck et al. (2016) for the fraction of social referents, do not provide much evidence against the null, at least when using Fisherian randomization inference.

estimate (one-hop − rand)	0.0544
SE (analytic)	0.0141
SE (bootstrap)	0.0260
95% CI (analytic)	[0.0269, 0.0820]
95% CI (bootstrap)	[-0.0075, 0.0886]
p-value (analytic)	0.00011
p-value (Fisherian)	0.1376

Table 4.7: Hájek estimate and inference for the difference in self-reported wristband-wearing one-hop and random targeting for Paluck et al. (2016).

estimate (one-hop − rand)	0.0600
SE (analytic)	0.0190
SE (bootstrap)	0.0256
95% CI (analytic)	[0.0228, 0.0973]
95% CI (bootstrap)	[0.0111, 0.1153]
p-value (analytic)	0.0016
p-value (Fisherian)	0.2706

Table 4.8: Hájek estimate and inference for the difference in self-reported friends talking about peer conflict one-hop and random targeting for Paluck et al. (2016).

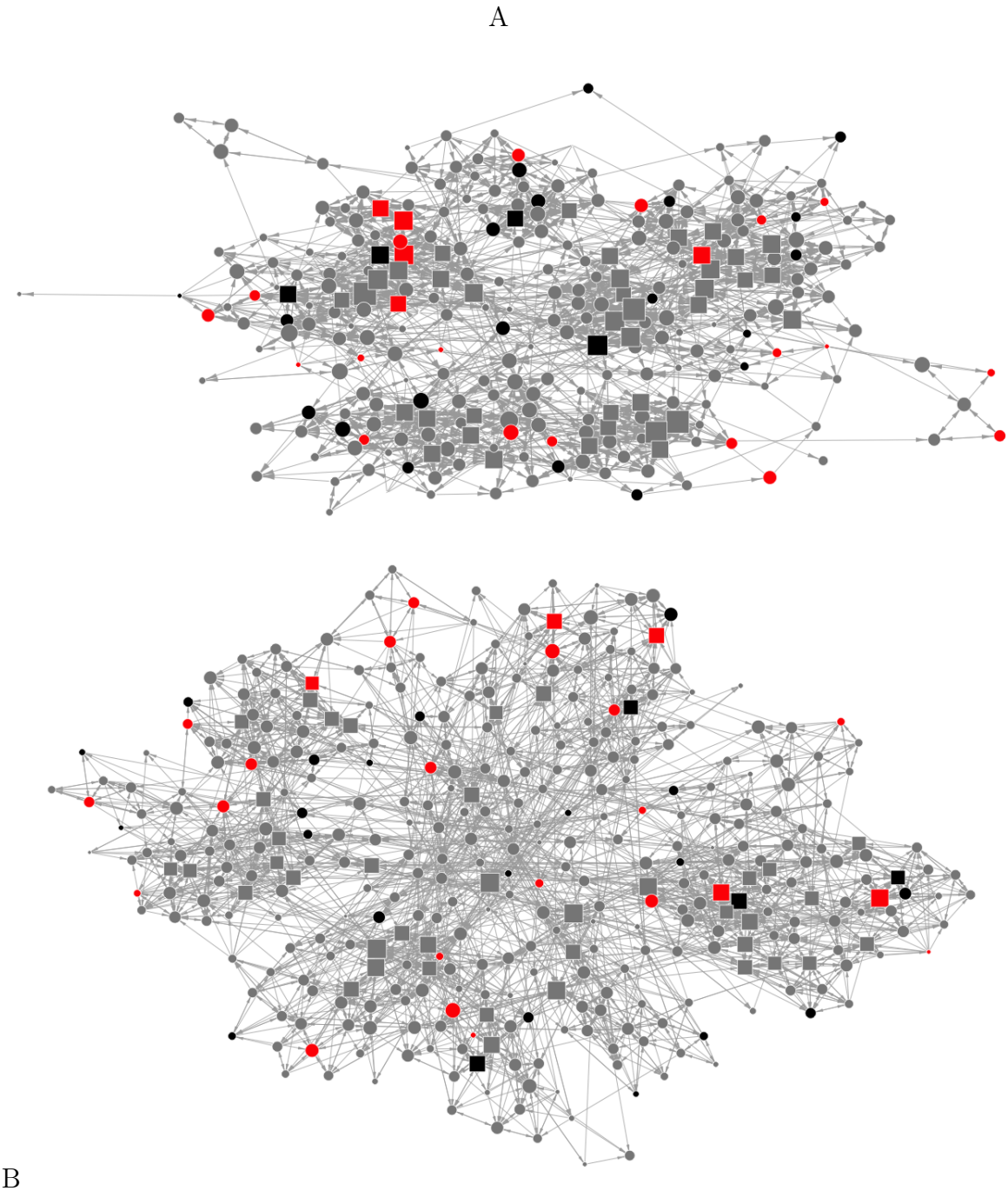

Figure 4.10: Social networks for two schools in Paluck et al. (2016) showing social referents (squares) and students eligible to be selected (black) and students selected as seeds (red). Each node v is sized proportional to $\mathbf{P}_i^{A,\text{repl}}(S_i = v)$ (i.e., row-normalized in-degree), not accounting for being eligible for treatment. Both have a somewhat similar fraction of the seed set who are social referents, with A a bit larger than B (A: 0.208, B: 0.167). But this is notably reversed for w_A (A: 1.1e-4, B: 0.00395).

4.9 Discussion

Stochastic seeding strategies are an appealing way of leveraging network structure for marketing, public health, and behavior change interventions when faced with limited network information and a limited budget. One-hop seeding, like other (mostly deterministic) strategies, is typically motivated by appeal to theory or simulations with an assumed model of social contagion. The basic theoretical premise of one-hop targeting is that by seeding an intervention at nodes with higher than average (normalized) in-degrees, more nodes in the network will have a social connection to the intervention and hopefully adopt themselves. If one had access to a full social network survey at the onset of the intervention then one could choose to specifically target the highest in-degree individuals. But besides being conveniently feasible without a full social network survey, one-step targeting has a specific potential advantage over targeting the maximum in-degree individuals: the maximum in-degree individuals are often tightly clustered in the network—as is common in networks with core-periphery structure (Borgatti and Everett, 2000; Rombach et al., 2017)—while the high in-degree individuals selected by one-hop targeting will be relatively spread out due to the initially random "nominators". Thus, the neighbors of a seed set selected by one-hop targeting is likely to be less redundant and perhaps more influential as a whole (cf. Kim et al., 2015).[7]

In this work we have developed methods for using network data for empirically estimating the effects of employing strategies such as one-hop seeding, even when the data arise from, e.g., unconditional random assignment. When an experiment has been conducted that varies the seeding strategy used, our proposed estimators offer potentially large increases in statistical precision and power. A much larger follow-up experiment to Kim et al. (2015), registered in Shakya et al. (2017), employs the same basic design as its predecessor. Given our simulations here, we expect the preregistered analyses will have lower power than achievable through better design

[7] Avoiding redundancy is key to maximizing influence in the widely studied independent cascade and linear threshold models (Kempe et al., 2003); if individual decision processes resemble a complex contagion process (Centola and Macy, 2007) then such spreading of seeds would in fact be undesirable.

and analysis. Our hope is that this perspective on these seeding strategies will in turn inform the design and analysis of future studies of seeding as well as the practice of seeding in marketing, public health, education, and development economics.

We conducted reanalyses of two field experiments using the proposed estimators. In each case, characteristics of the setting and the original results might suggest that one-hop seeding would be a promising way to increase the desired outcomes. However, in both cases, we found some evidence that one-hop seeding would in fact have backfired compared with random seeding: lowering insurance rates and increasing peer conflict compared to uniform random targeting. This emphasizes the importance of credible empirical evaluation of these strategies.

Widely accepted theoretical reasoning motivates why one-hop targeting should lead to a higher adoption rate than random targeting (for interventions seeking to maximize adoption). Why, then, do our results suggest that one-hop targeting is no more effective (and possibly less effective) than random targeting? Here we offer are a number of possible explanations, none of which are definitive and all of which suggest important follow-up work.

First, it is possible that the social networks collected from the surveys in these studies are not the networks that matter in terms of influence processes guiding the relevant adoption decisions. The study of name generators (Campbell and Lee, 1991; Perkins et al., 2015) in sociology has long established that different questions lead to different networks, e.g. "Who are your friends?" vs. "With whom do you discuss important matters?" (Bearman and Parigi, 2004). Some name generators have a tendency to elicit strong ties while others elicit weak ties (Momeni and Rabbat, 2017). It is well known that strong and weak ties figure differently in information diffusion and social decision making (Rapoport and Horvath, 1961; Granovetter, 1973; McAdam, 1986). If trying to maximize adoption, it is natural to then ask what name generator leads to the greatest adoption under one-hop targeting, and also quite natural that one-hop targeting paired with some name generators would lead to less adoption than random seeding. In this vein, Chami et al. (2017) asked both about close friends and about trusted sources of health advice. Banerjee et al. (2013) collected responses to a total of twelve different name generators, although most

analyses of that study (including our use of these networks in Section 4.7.2) analyze only the flattened network of "all relationships."

Questions about name generators raise an important dimension in which one-hop targeting can be refined when it is actually being deployed as a seeding strategy in the absence of a network survey (meaning the experimenter actually asks subjects to "nominate a random friend" as opposed to taking a survey of "all their friends" and randomly selecting a seed from the friends). As a small change to the protocol, don't ask subjects to nominate a random friend but instead ask them to nominate "the farmer they most respect" (for the weather insurance experiment) or "the student most people look up to" (for the anti-bullying intervention). This strategy would still count as a stochastic seeding strategy (when the nominators are a random set of individuals), and many of the ideas here could be applied to evaluate such strategies. This can be regarded as an explicitly stochastic variation on name generators used for identifying "opinion leaders" (Flodgren et al., 2011).

Even if the seeds selected by one-hop targeting may be more influential, they may not be susceptible to the initial intervention. Aral and Walker (2012) present some evidence of (negative) correlations between susceptibility and influence; this can have substantial consequences for approximately optimal seeding (Aral and Dhillon, 2018). Similarly, Bakshy et al. (2011) highlight that if more influential individuals require larger inducements to adopt or promote a behavior, then targeting them may be a poor use of a limited budget.

As a more speculative possible explanation of our findings, it could be that the behavior in our interventions spread via a "push" mechanism (the seed needs to tell people about the intervention for it to spread), as opposed to a "pull" mechanism (the friends of the seed observe them). The sharply different dynamics of diffusion processes under push and pull mechanisms have been widely studied in the computer science literature (Demers et al., 1987; Chierichetti et al., 2011). A behavior that spread via a push mechanism would benefit from being seeded at nodes with high out-degree, as opposed to high in-degree for a pull mechanism. While the in-degree and out-degree of nodes are often correlated, it is generically possible that a seeding strategy that climbs in average in-degree could decline in average out-degree. Again,

this highlights the importance of the specific name generator and survey technique used, as in Paluck et al. (2016) out-degree was effectively capped as students were asked for up to ten students they chose to spend time with; over 40% of students named exactly ten such students.

As a final variation on concerns about the surveyed network possibly being the "wrong" network, it is possible that the actual social network describing influential relationships is regular (everyone has the same number of influential relations) or nearly regular. In a recent study of the friendship paradox and contact strength, Bagrow et al. (2017) analyzed contact frequencies (as a proxy for tie strength) on Twitter and in cellular phone networks and found that networks of frequent ties are nearly regular, leading to a tempered friendship paradox: "your closest friends have [only] slightly more friends than you do." A finding of little or no difference in adoption rates between random and one-step targeting would be consistent with the adoption decisions in the two field experiments we analyzed here relying on social networks that are nearly regular.

It is important to stress that our empirical findings are not inconsistent with established findings that one-hop targeting can be successfully leveraged to design efficient sensing strategies. In prior work on epidemiological outbreak detection using "one-hop measurement" by Christakis and Fowler (2010), instances of the flu occurred earlier in a population selected by a one-hop strategy than a random population. But in the language of this discussion section, that finding only means that the one-hop strategy was successful in reaching a population that was more (epidemiologically) susceptible, but not necessarily (epidemiologically) influential.

Bibliography

A. Abadie, S. Athey, G. W. Imbens, and J. Wooldridge. When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research, 2017a.

A. Abadie, S. Athey, G. W. Imbens, and J. M. Wooldridge. Sampling-based vs. design-based uncertainty in regression analysis. *arXiv preprint arXiv:1706.01778*, 2017b.

M. Akbarpour, S. Malladi, and A. Saberi. Just a few seeds more: Value of targeting for diffusion in networks. Working paper, Stanford Graduate School of Business, 2017.

K. M. Altenburger and J. Ugander. Monophily in social networks introduces similarity among friends-of-friends. *Nature Human Behaviour*, 2(4):284, 2018.

S. Aral. *Networked experiments*. Oxford, UK: Oxford University Press, 2016.

S. Aral and P. S. Dhillon. Social influence maximization under empirical influence models. *Nature Human Behaviour*, 2018. Forthcoming.

S. Aral and D. Walker. Identifying influential and susceptible members of social networks. *Science*, page 1215842, 2012.

S. Aral and D. Walker. Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Management Science*, 60(6):1352–1370, 2014.

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

S. Aral, L. Muchnik, and A. Sundararajan. Engineering social contagions: Optimal network seeding in the presence of homophily. *Network Science*, 1(2):125–153, 2013.

Aristotle. Metaphysics V, 2. a.

Aristotle. Physics II, 3. b.

P. M. Aronow. A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, 41(1):3–16, 2012.

P. M. Aronow and J. A. Middleton. A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, 1(1): 135–154, 2013.

P. M. Aronow and C. Samii. Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities. *Survey Methodology*, 39(1):231–241, 2012.

P. M. Aronow and C. Samii. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947, 2017.

S. Athey and G. W. Imbens. The econometrics of randomized experimentsa. In *Handbook of Economic Field Experiments*, volume 1, pages 73–140. Elsevier, 2017.

S. Athey and S. Wager. Efficient policy learning. Working paper, Stanford Graduate School of Business. https://arxiv.org/abs/1702.02896, 2017.

S. Athey, D. Eckles, and G. W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, pages 1–11, 2017a.

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. 2017b. URL: `https://arxiv.org/pdf/1604.07125.pdf`.

B. Babington Smith. On some difficulties encountered in the use of factorial designs and analysis of variance with psychological experiments. *British Journal of Psychology. General Section*, 42(3):250–268, 1951.

L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 33–42. ACM, 2012.

J. P. Bagrow, C. M. Danforth, and L. Mitchell. Which friends are more popular than you?: Contact strength and the friendship paradox in social networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 103–108. ACM, 2017.

S. Baird, J. A. Bohren, C. McIntosh, and B. Özler. Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*, (0), 2016.

E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 65–74. ACM, 2011.

E. Bakshy, D. Eckles, R. Yan, and I. Rosenn. Social influence in social advertising: Evidence from field experiments. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 146–161. ACM, 2012.

P. Baldi and Y. Rinott. On normal approximations of distributions in terms of dependency graphs. *The Annals of Probability*, pages 1646–1650, 1989.

A. Banerjee, A. G. Chandrasekhar, E. Duflo, and M. O. Jackson. The diffusion of microfinance. *Science*, 341(6144):1236498, 2013.

A. Banerjee, A. G. Chandrasekhar, E. Duflo, and M. O. Jackson. Using gossips to spread information: theory and evidence from a randomized controlled trial. *arXiv preprint arXiv:1406.2293*, 2017.

B. S. Barnow et al. Issues in the analysis of selectivity bias. discussion papers. revised. 1980.

G. Basse and A. Feller. Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, 113(521):41–55, 2018.

G. Basse, A. Feller, and P. Toulis. Exact tests for two-stage randomized designs in the presence of interference. *arXiv preprint arXiv:1709.08036*, 2017.

G. W. Basse, H. A. Soufiani, and D. Lambert. Randomization and the pernicious effects of limited budgets on auction experiments. In *Artificial Intelligence and Statistics*, pages 1412–1420, 2016.

L. Beaman and A. Dillon. Diffusion of agricultural information within social networks: Evidence on gender inequalities from mali. *Journal of Development Economics*, 133: 147–161, 2018.

L. Beaman, A. BenYishay, J. Magruder, and A. M. Mobarak. Can network theory-based targeting increase technology adoption? Technical report, 2018. Working paper, National Bureau of Economic Research.

P. Bearman and P. Parigi. Cloning headless frogs and other important matters: Conversation topics and network structure. *Social Forces*, 83(2):535–557, 2004.

A. Belloni, V. Chernozhukov, and C. Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2): 29–50, 2014.

K. N. Berk. A central limit theorem for m-dependent random variables with unbounded m. *The Annals of Probability*, pages 352–354, 1973.

R. Berk, E. Pitkin, L. Brown, A. Buja, E. George, and L. Zhao. Covariance adjustments for the analysis of randomized field experiments. *Evaluation Review*, 37(3-4): 170–196, 2013.

S. Bhagat, M. Burke, C. Diuk, I. O. Filiz, and S. Edunov. Three and a half degrees of separation. *Facebook research note*, 2016. URL: `https://research.fb.com/three-and-a-half-degrees-of-separation/`.

A. Bloniarz, H. Liu, C.-H. Zhang, J. S. Sekhon, and B. Yu. Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390, 2016.

L. E. Blume. The statistical mechanics of best-response strategy revision. *Games and Economic Behavior*, 11(2):111–145, 1995.

R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295, 2012.

S. P. Borgatti and M. G. Everett. Models of core/periphery structures. *Social networks*, 21(4):375–395, 2000.

Y. Bramoullé, H. Djebbari, and B. Fortin. Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55, 2009.

J. Cai, A. De Janvry, and E. Sadoulet. Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108, 2015.

K. E. Campbell and B. A. Lee. Name generators in surveys of personal networks. *Social networks*, 13(3):203–221, 1991.

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

G. F. Chami, S. E. Ahnert, N. B. Kabatereine, and E. M. Tukahebwa. Social network fragmentation and community health. *Proceedings of the National Academy of Sciences*, 114(36):E7425–E7431, 2017.

S. Chatterjee. A new method of normal approximation. *The Annals of Probability*, 36(4):1584–1610, 2008.

S. Chatterjee. Fluctuations of eigenvalues and second order poincaré inequalities. *Probability Theory and Related Fields*, 143(1-2):1–40, 2009.

S. Chatterjee. A short survey of stein's method. *arXiv preprint arXiv:1404.1392*, 2014.

K. Chen and J. Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113(521): 241–251, 2018.

L. H. Chen. Poisson approximation for dependent trials. *The Annals of Probability*, pages 534–545, 1975.

L. H. Chen and Q.-M. Shao. Normal approximation under local dependence. *The Annals of Probability*, 32(3):1985–2028, 2004.

W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM, 2009.

V. Chernozhukov, C. Hansen, and M. Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *Annu. Rev. Econ.*, 7 (1):649–688, 2015.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

F. Chierichetti, S. Lattanzi, and A. Panconesi. Rumor spreading in social networks. *Theoretical Computer Science*, 412(24):2602–2610, 2011.

A. Chin. Central limit theorems via Stein's method for randomized experiments under interference. *arXiv preprint arXiv:1808.08683*, 2018a.

A. Chin. Regression adjustments for estimating the global treatment effect in experiments with interference. *arXiv preprint arXiv:1804.03105*, 2018b.

A. Chin, D. Eckles, and J. Ugander. Stochastic seeding strategies in networks: Estimation, inference, and experimental design. *arXiv preprint*, 2018.

D. Choi. Estimation of monotone treatment effects in network experiments. *Journal of the American Statistical Association*, 112(519):1147–1155, 2017.

D. Choi. Using exposure mappings as side information in experiments with interference. *arXiv preprint arXiv:1806.11219*, 2018.

N. A. Christakis and J. H. Fowler. Social network sensors for early detection of contagious outbreaks. *PloS One*, 5(9):e12948, 2010.

E. Chung, J. P. Romano, et al. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507, 2013.

R. Cohen, S. Havlin, and D. Ben-Avraham. Efficient immunization strategies for computer networks and populations. *Physical Review Letters*, 91(24):247901, 2003.

D. R. Cox. Planning of experiments. 1958.

K. Cranmer, J. Pavez, and G. Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*, 2015.

S. Currarini, M. O. Jackson, and P. Pin. Identifying the roles of race-based choice and chance in high school friendship network formation. *Proceedings of the National Academy of Sciences*, 107(11):4857–4861, 2010.

A. Demers, D. Greene, C. Hauser, W. Irish, J. Larson, S. Shenker, H. Sturgis, D. Swinehart, and D. Terry. Epidemic algorithms for replicated database maintenance. In *Proceedings of the sixth annual ACM Symposium on Principles of distributed computing*, pages 1–12. ACM, 1987.

P. Ding. A paradox from randomization-based causal inference. *Statistical Science*, 32(3):331–345, 2017.

P. Ding and T. J. VanderWeele. Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27(3):368, 2016.

P. Ding, X. Li, and L. W. Miratrix. Bridging finite and super population causal inference. *Journal of Causal Inference*, 5(2), 2017.

P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 57–66. ACM, 2001.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

M. Dudík, D. Erhan, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *Statistical Science*, pages 485–511, 2014a.

M. Dudík, D. Erhan, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014b.

R. I. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of human evolution*, 22(6):469–493, 1992.

D. Eckles, R. F. Kizilcec, and E. Bakshy. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, 113 (27):7316–7322, 2016.

D. Eckles, B. Karrer, and J. Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1), 2017.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

B. Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.

N. Egami. Unbiased estimation and sensitivity analysis for network-specific spillover effects: Application to an online network experiment. *arXiv preprint arXiv:1708.08171*, 2017.

F. Eicker. Limit theorems for regressions with unequal and dependent errors. 1967.

S. L. Feld. Why your friends have more friends than you do. *American Journal of Sociology*, 96(6):1464–1477, 1991.

R. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, London, 1925.

R. Fisher. *Design of Experiments*. Oliver and Boyd, London, 1935.

G. Flodgren, E. Parmelli, G. Doumit, M. Gattellari, M. A. O'Brien, J. Grimshaw, and M. P. Eccles. Local opinion leaders: Effects on professional practice and health care outcomes. *The Cochrane Database of Systematic Reviews*, (8):CD000125, 2011.

L. Forastiere, E. M. Airoldi, and F. Mealli. Identification and estimation of treatment and interference effects in observational studies on networks. *arXiv preprint arXiv:1609.06245*, 2016.

S. Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

D. A. Freedman. Statistical models for causation: what inferential leverage do they provide? *Evaluation review*, 30(6):691–713, 2006.

D. A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008a.

D. A. Freedman. On regression adjustments in experiments with several treatments. *The annals of applied statistics*, 2(1):176–196, 2008b.

T. M. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.

A. Galeotti, B. Golub, and S. Goyal. Targeting interventions in networks. *arXiv preprint arXiv:1710.06026*, 2017.

L. K. Gallos, F. Liljeros, P. Argyrakis, A. Bunde, and S. Havlin. Improving immunization strategies. *Physical Review E*, 75(4):045104, 2007.

A. S. Gerber and D. P. Green. *Field experiments: Design, analysis, and interpretation*. WW Norton, 2012.

A. S. Gerber, D. P. Green, and C. W. Larimer. Social pressure and voter turnout: Evidence from a large-scale field experiment. *American political Science review*, 102(1):33–48, 2008.

A. S. Gerber, D. P. Green, and C. W. Larimer. An experiment testing the relative effectiveness of encouraging voter participation by inducing feelings of pride or shame. *Political Behavior*, 32(3):409–422, 2010.

M. S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6): 1360–1380, 1973.

S. Greenland. Interpretation and choice of effect measures in epidemiologic analyses. *American journal of epidemiology*, 125(5):761–768, 1987.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

J. Hájek. Comment on 'An essay on the logical foundations of survey sampling, part 1' by D. Basu. In V. Godambe and D. A. Sprott, editors, *Foundations of Statistical Inference*, page 236, Toronto, 1971. Holt, Rinehart and Winston.

M. E. Halloran and M. G. Hudgens. Dependent happenings: a recent methodological review. *Current epidemiology reports*, 3(4):297–305, 2016.

T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3): 297–310, 1986.

C. Haythornthwaite and B. Wellman. Work, friendship, and media use for information exchange in a networked organization. *Journal of the American society for information science*, 49(12):1101–1114, 1998.

M. A. Hernán, E. Lanoy, D. Costagliola, and J. M. Robins. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & Clinical Pharmacology & Toxicology*, 98(3):237–242, 2006.

S. Hill, F. Provost, and C. Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, pages 256–276, 2006.

O. Hinz, B. Skiera, C. Barrot, and J. U. Becker. Seeding strategies for viral marketing: An empirical comparison. *Journal of Marketing*, 75(6):55–71, 2011.

K. Hirano and J. R. Porter. Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701, 2009.

W. Hoeffding and H. Robbins. The central limit theorem for dependent random variables. *Duke Mathematical Journal*, 15(3):773–780, 1948.

D. Holt and T. F. Smith. Post stratification. *Journal of the Royal Statistical Society. Series A (General)*, pages 33–46, 1979.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260): 663–685, 1952.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. 1967.

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

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

R. Iyengar, C. Van den Bulte, and T. W. Valente. Opinion leadership and social contagion in new product diffusion. *Marketing Science*, 30(2):195–212, 2011.

R. Jagadeesan, N. Pillai, and A. Volfovsky. Designs for estimating the treatment effect in networks with interference. *arXiv preprint arXiv:1705.08524*, 2017.

J. D. Kang, J. L. Schafer, et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146. ACM, 2003.

E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 2018. Forthcoming.

D. A. Kim, A. R. Hwong, D. Stafford, D. A. Hughes, A. J. O'Malley, J. H. Fowler, and N. A. Christakis. Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet*, 386(9989):145–153, 2015.

L. Kish. Survey sampling. 1965.

M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

V. Kumar and K. Sudhir. Can friends seed more buzz and adoption? Working Paper, Yale University, 2018.

V. Kumar, D. Krackhardt, and S. Feld. Network interventions based on inversity: Leveraging the friendship paradox in unknown network structures. Working Paper, Yale University, 2018.

H. R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, pages 1217–1241, 1989.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv preprint arXiv:1706.03461*, 2017.

S. Lattanzi and Y. Singer. The power of random neighbors in social networks. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 77–86. ACM, 2015.

J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915–924. ACM, 2008.

J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 420–429. ACM, 2007.

L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 297–306. ACM, 2011.

L. Li, R. Munos, and C. Szepesvári. On minimax optimal offline policy evaluation. Working paper. https://arxiv.org/abs/1409.3653, 2014.

T. Li, E. Levina, and J. Zhu. Network cross-validation by edge sampling. *arXiv preprint arXiv:1612.04717*, 2018.

B. Libai, E. Muller, and R. Peres. Decomposing the value of word-of-mouth seeding programs: Acceleration versus expansion. *Journal of Marketing Research*, 50(2): 161–176, 2013.

W. Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

L. Liu and M. G. Hudgens. Large sample randomization inference of causal effects in the presence of interference. *Journal of the American Statistical Association*, 109 (505):288–301, 2014.

M. Luby. A simple parallel algorithm for the maximal independent set problem. *SIAM journal on computing*, 15(4):1036–1053, 1986.

C. F. Manski. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 1993.

C. F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246, 2004.

C. F. Manski. *Identification for Prediction and Decision*. Harvard University Press, 2007. ISBN 0674026535.

C. F. Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1), 2013.

D. McAdam. Recruitment to high-risk activism: The case of freedom summer. *American journal of sociology*, 92(1):64–90, 1986.

M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

J. A. Middleton. A unified theory of regression adjustment for design-based inference. *arXiv preprint arXiv:1803.06011*, 2018.

N. Momeni and M. G. Rabbat. Inferring structural characteristics of networks with strong and weak ties from fixed-choice surveys. *IEEE Transactions on Signal and Information Processing over Networks*, 3(3):513–525, 2017.

I. D. Muñoz and M. van der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

S. A. Murphy, M. J. van der Laan, J. M. Robins, and C. P. P. R. Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

H. Nair, P. Chintagunta, and J.-P. Dubé. Empirical analysis of indirect network effects in the market for personal digital assistants. *Quantitative Marketing and Economics*, 2(1):23–58, 2004.

J. Neyman. On the application of probability theory to agricultural experiments. Essay on Principles. section 9. (translated and edited by D. M. Dabrowska and T. P. Speed, *Statistical Science* (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51, 1923.

J. Neyman and E. S. Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A*, 231(694-706):289–337, 1933.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

E. L. Ogburn and T. J. VanderWeele. Causal diagrams for interference. *Statistical Science*, 29(4):559–578, 2014.

J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences*, 104(18):7332–7336, 2007.

A. B. Owen. *Monte Carlo Theory, Methods and Examples.* 2013. URL `http://statweb.stanford.edu/~owen/mc/`.

E. L. Paluck, H. Shepherd, and P. M. Aronow. Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, 113(3):566–571, 2016.

C. Panagopoulos. Positive social pressure and prosocial motivation: Evidence from a large-scale field experiment on voter mobilization. *Political Psychology*, 34(2): 265–275, 2013.

J. Pearl. *Causality.* Cambridge University Press, 2009.

C. Perez-Heydrich, M. G. Hudgens, M. E. Halloran, J. D. Clemens, M. Ali, and M. E. Emch. Assessing effects of cholera vaccination in the presence of interference. *Biometrics*, 70(3):731–741, 2014.

J. M. Perkins, S. Subramanian, and N. A. Christakis. Social networks and health: a systematic review of sociocentric network studies in low-and middle-income countries. *Social science & medicine*, 125:60–78, 2015.

Plato. Timaeus 28a.

J. Pouget-Abadie, M. Saveski, G. Saint-Jacques, W. Duan, Y. Xu, S. Ghosh, and E. M. Airoldi. Testing for arbitrary interference on experimentation platforms. *arXiv preprint arXiv:1704.01190*, 2017.

D. Precup, R. S. Sutton, and S. P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 759–766, 2000.

A. Rapoport and W. J. Horvath. A study of a large sociogram. *Behavioral Science*, 6(4):279–291, 1961.

M. D. Resnick, P. S. Bearman, R. W. Blum, K. E. Bauman, K. M. Harris, J. Jones, J. Tabor, T. Beuhring, R. E. Sieving, M. Shew, et al. Protecting adolescents from harm: findings from the national longitudinal study on adolescent health. *Journal of the American Medical Association*, 278(10):823–832, 1997.

J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

J. Robins and S. Greenland. The probability of causation under a stochastic model for individual risk. *Biometrics*, pages 1125–1138, 1989.

J. M. Robins and S. Greenland. Causal inference without counterfactuals: comment. *Journal of the American Statistical Association*, 95(450):431–435, 2000.

J. M. Robins, A. Rotnitzky, and D. O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 1–94. Springer, 2000.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

T. Rogers, D. P. Green, J. Ternovski, and C. F. Young. Social pressure and voting: a field experiment conducted in a high-salience election. *Electoral Studies*, 46:87–100, 2017.

J. P. Romano and M. Wolf. A more general central limit theorem for m-dependent random variables with unbounded m. *Statistics & probability letters*, 47(2):115–124, 2000.

P. Rombach, M. A. Porter, J. H. Fowler, and P. J. Mucha. Core-periphery structure in networks (revisited). *SIAM Review*, 59(3):619–646, 2017.

P. R. Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200, 2007.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

D. B. Rubin. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 1980.

D. B. Rubin. [on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies. *Statist. Sci.*, 5(4):472–480, 11 1990a. doi: 10.1214/ss/1177012032. URL `https://doi.org/10.1214/ss/1177012032`.

D. B. Rubin. Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference*, 25(3):279–292, 1990b.

C. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer-Verlag, 1992.

C. E. Särndal. On uniformly minimum variance estimation in finite populations. *The Annals of Statistics*, pages 993–997, 1976.

M. Saveski, J. Pouget-Abadie, G. Saint-Jacques, W. Duan, S. Ghosh, Y. Xu, and E. M. Airoldi. Detecting network effects: Randomizing over randomized experiments. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1027–1035. ACM, 2017.

F. Sävje, P. M. Aronow, and M. G. Hudgens. Average treatment effects in the presence of unknown interference. *arXiv preprint arXiv:1711.06399*, 2017.

H. B. Shakya, D. Stafford, D. A. Hughes, T. Keegan, R. Negron, J. Broome, M. McKnight, L. Nicoll, J. Nelson, E. Iriarte, et al. Exploiting social influence to magnify population-level behaviour change in maternal and child health: Study protocol for a randomised controlled trial of network targeting algorithms in rural Honduras. *BMJ Open*, 7(3):e012996, 2017.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

M. E. Sobel. What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. II: Probability theory*, pages 583–602, 1972.

C. Stein. Approximate computation of expectations. *Lecture Notes-Monograph Series*, 7:i–164, 1986.

E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25 (1):1, 2010.

D. L. Sussman and E. M. Airoldi. Elements of estimation theory for causal effects in the presence of network interference. *arXiv preprint arXiv:1702.03578*, 2017.

A. Swaminathan and T. Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015.

A. Swaminathan, A. Krishnamurthy, A. Agarwal, M. Dudik, J. Langford, D. Jose, and I. Zitouni. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems*, pages 3632–3642, 2017.

S. J. Taylor and D. Eckles. Randomized experiments to detect and estimate social influence in networks. *arXiv preprint arXiv:1709.09636*, 2017.

E. J. Tchetgen Tchetgen and T. J. VanderWeele. On causal inference in the presence of interference. *Statistical methods in medical research*, 21(1):55–75, 2012.

A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53 (3):526–543, 2011.

A. L. Traud, P. J. Mucha, and M. A. Porter. Social structure of Facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

C. Tucker. Identifying formal and informal influence in technology adoption with network externalities. *Management Science*, 54(12):2024–2038, 2008.

S. Tyner, F. Briatte, and H. Hofmann. Network visualization with ggplot2. *The R Journal*, 2017.

J. Ugander and L. Backstrom. Balanced label propagation for partitioning massive graphs. In *Proceedings of the sixth ACM International Conference on Web Search and Data Mining*, pages 507–516. ACM, 2013.

J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the Facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

J. Ugander, B. Karrer, L. Backstrom, and J. 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. ACM, 2013.

M. J. van der Laan. Causal inference for a population of causally connected units. *Journal of Causal Inference*, 2(1):13–74, 2014.

T. J. VanderWeele and J. M. Robins. Stochastic counterfactuals and stochastic sufficient causes. *Statistica Sinica*, 22(1):379, 2012.

T. J. VanderWeele and E. J. Tchetgen Tchetgen. Effect partitioning under interference in two-stage randomized vaccine trials. *Statistics & Probability Letters*, 81(7):861–869, 2011.

T. J. VanderWeele, E. J. Tchetgen Tchetgen, and M. E. Halloran. Interference and sensitivity analysis. *Statistical Science: A review journal of the Institute of Mathematical Statistics*, 29(4):687, 2014.

S. Wager, W. Du, J. Taylor, and R. J. Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678, 2016.

D. Walker and L. Muchnik. Design of randomized experiments in networks. *Proceedings of the IEEE*, 102(12):1940–1951, 2014.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440, 1998.

A. T. Welford, R. A. Brown, and J. Gabb. Two experiments on fatigue as affecting skilled performance in civilian air crew. *British Journal of Psychology. General Section*, 40(4):195–211, 1950.

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838, 1980.

C.-F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, pages 1261–1295, 1986.

E. Wu and J. Gagnon-Bartsch. The LOOP estimator: Adjusting for covariates in randomized experiments. *arXiv preprint arXiv:1708.01229*, 2017.

H. Yoganarasimhan. Impact of social network structure on content propagation: A study using YouTube data. *Quantitative Marketing and Economics*, 10(1):111–150, 2012.

D. Zeyl. Plato, Timaeus. Translated, with Introduction, 2000.

P. P. Zubcsek and M. Sarvary. Advertising to a social network. *Quantitative Marketing and Economics*, 9(1):71–107, 2011.