ESTIMATION AND TESTING METHODS FOR CAUSAL INFERENCE WITH 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

Kevin Han

June 2023

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.

(Guido Imbens)   Principal Co-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.

(Johan Ugander)   Principal Co-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.

(Art Owen)

Approved for the Stanford University Committee on Graduate Studies

Preface

Causal inference is one of the core areas of research in modern data science that allows researchers to determine whether a specific intervention or treatment has an effect on an outcome. In its most basic form, causal inference is concerned with understanding the “cause and effect” relationship between variables. This requires going beyond correlation to understand whether changing one variable leads to a change in another. The gold standard for inferring causality is the randomized controlled trial, which randomly assigns subjects to a treatment or control group and compares outcomes. While randomized controlled trials provide us with data to do causal inference, the subsequent statistical analysis often relies on a key assumption known as the Stable Unit Treatment Value Assumption (SUTVA). This assumption states that the treatment of one unit (or individual) does not affect the outcome of another unit. However, in many real-world situations, this assumption does not hold, leading to what is called interference or a violation of SUTVA. Interference can occur in various contexts such as social networks, where the treatment of one person can influence the outcomes of others, or in marketplace, where treatment of one entity can impact other entities of same type. Understanding and handling interference is a critical and complex aspect of causal inference, and it necessitates more advanced methods to correctly estimate causal effects.

This dissertation offers new methodologies and theoretical results to address key issues in causal inference with interference. In Chapter 2, we develop inferential results for causal effect estimators in panel experiments under interference. We turn our attention to the complications brought about by network interference in Chapter 3, where we introduce novel estimation methods for causal effects. Given the difficulties interference presents to inference, the ability to detect interference becomes pivotal in determining the most suitable statistical analysis approach. To this end, Chapter 4 tackles the problem of detecting interference in online controlled experiments with increasing allocation.

Acknowledgments

The completion of this PhD thesis would not have been possible without the collective support, guidance, and encouragement of many individuals to whom I owe my deepest gratitude.

Firstly, I would like to express my deepest gratitude to my advisors, Guido Imbens and Johan Ugander. Their guidance, wisdom, and unwavering support have been instrumental in shaping my academic journey. Guido's influence in the field of causal inference is unparalleled. My understanding of causal inference has been significantly shaped by studying his textbook, reading his papers and taking his courses. He stands as a role model to me. I recall predicting to my parents three years ago that Guido would one day win the Nobel prize. It fills me with immense pride and joy to see this prediction come to reality, and I am deeply honored to be his student. On the other hand, Johan is, in my opinion, the best advisor every PhD student could ever have. Johan kindly stepped in as my co-advisor, providing me with invaluable support when Guillaume left Stanford. He is willing to devote considerable time to his students. His knowledge spans diverse areas like computational social science, social networks, causal inference, and applied mathematics, making him an invaluable resource for references and discussions. His support has been unwavering, in both my academic pursuits and personal life.

I would also like to thank my committee members, Art Owen, Jonathan Taylor, and Tatsunori Hashimoto for serving on my dissertation and defense committees. I am grateful for their invaluable input, critical feedback, and challenging questions that have significantly enriched this work. Art also carefully read the draft version of this thesis and sent me very detailed suggestions as well as a list of typos that I were not aware of beforehand.

Next I would like to express my sincere gratitude to Guillaume Basse and Iavor Bojinov, who, despite not holding official advisory roles, have significantly contributed to my academic journey. Guillaume, my co-advisor prior to his departure from Stanford, is an exceptional individual and a brilliant statistician. Although our collaboration only spanned a one and a half years, the knowledge and insights I gained from him were invaluable. Iav, my collaborator and mentor on the panel experiment project featured in this thesis, deserves special mention. During the challenging period when I was in search of another co-advisor, Iavor's support was instrumental. He also provided invaluable advice as I navigated the process of securing internships. I am deeply appreciative of his mentorship and guidance.

The decision to go to graduate school would not have been realized without the profound impact of my exceptional teachers and mentors during my college years. I am particularly indebted to Jinho Baik, Alexander Barvinok, Stephen DeBacker, Tailen Hsing, Stilian Stoev, Yuekai Sun, Roman Vershynin, Gongjun Xu and Ji Zhu. The knowledge I learned from their classes, along with their insightful guidance, were instrumental in shaping and nurturing my interests in mathematics and statistics.

I am truly fortunate to have shared this journey with an amazing cohort - Ben, Dan, Han, Hui, Kevin, Marius, Samyak, Souvik and Yuchen. The cohort dinners we had together, the classes we took, and our numerous discussions have made this journey truly memorable. To the remarkable members of U-lab, specifically Amel, Izzy, Jenny, Jordan, Martin, Samir, Serina and Yuchen, I owe a debt of gratitude for providing a supportive and stimulating environment. The past two years of attending group meetings with them have consistently been a source of joy and fulfillment.

My sincere appreciation extends to my friends at Sequoia. I am particularly thankful to Shuangning, Shuangping, Daren, Lingfu, Yuchen, Kangjie, Zhihan and Mi for providing me with steadfast support and being a constant source of joy and comfort throughout this journey. Our shared laughter and countless conversations have lightened the burden of the challenges I faced. With Shuangning, Shuangping, Daren, Lingfu, Yuchen, and Kangjie, we have spent lots of nights playing board games together, and we have also made many great memories on tennis courts. Thank you for being a part of this journey, and more importantly, for being true friends. To Xiaowei, even though he left us three years ago, his optimism and kindness have left an indelible mark on my heart. I also want to extend my thanks to my friends outside Sequoia, especially those I met at ACSSS, who have made my past three years even more special. In particular, I have been involved in the planning and production of the annual Stanford Lunar New Year Gala since 2021. This has been my most cherished memory during my time at ACSSS.

I am also grateful to the wonderful staff in our department, whose efforts make Sequoia a great place for all PhD students. I particularly want to thank Susie for her invaluable assistance and support. This amazing journey began with an offer email from Susie five years ago, an event that marked the inception of an incredible academic experience.

I owe a deep debt of gratitude to my family, whose constant love and support have been my foundation throughout this journey. To my parents and my grandparents, thank you for instilling in me the values of perseverance, curiosity, and humility. Your sacrifices and unyielding belief in me have been the driving force behind my every endeavor. Even when the path was tough, your words of encouragement and wisdom have been the compass that guided me through. You have been my first teachers and biggest cheerleaders, and for that, I am eternally grateful. My brother Han, who shares the same academic path with me since kindergarten. The bond we share as siblings and fellow PhD students has been a source of immeasurable comfort. Observing you tackle similar challenges, celebrate the same milestones, and share this unique academic journey has been an invaluable source of inspiration and mutual support. Your companionship and understanding have truly enriched my doctoral journey. I also want to express my heartfelt gratitude to my girlfriend Rita. Your love, patience, and enduring faith in me have shone a light during the most challenging times. Thank you for being my confidante, my source of joy, and my steady source of support. I am truly fortunate to have you by my side.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Causal inference is of great interest in the realm of data science as it aids researchers and practitioners in making informed decisions, designing effective policies, and unveiling the complex dynamics of various systems. In this thesis, we discuss estimation and testing methods for causal inference with interference. In this introductory chapter, we introduce the problems we will be solving in the following chapters and outline our contributions.

When researchers estimate causal effects from randomized experiments, they almost always make assumptions that restrict the number of counterfactual outcomes to simplify the subsequent inference. In standard experiments, where units are randomly assigned to either a treatment or control, researchers commonly assume that one unit's assignment does not affect another unit's response; this is usually referred to as no interference [Cox, 1958, Chapter 2]. In panel experiments, where units are exposed to different interventions over time [Bojinov et al., 2021b], in addition to no interference, researchers regularly assume that the observed outcomes were not impacted by past assignments; this is often called the no carryover assumption [Cox, 1958, Chapter 13]. Although these two assumptions are useful, there are numerous empirical examples where they are violated. This mismatch between practical applications and theoretical assumptions has catalyzed a growing amount of literature dedicated to studying relaxations of these stringent conditions for either standard or panel experiments, but not both.

In standard experiments without evoking the no interference assumption, each unit's outcome depends on the assignments received by all other experimental units. Allowing for such arbitrary population interference[1] makes causal inference challenging [Basse and Airoldi, 2018]. In practice, researchers look for an underlying structure that limits the scope of interference. For example, when studying electoral participation during a special election in 2009 in Chicago, Sinclair et al. [2012] assumed that interference occurred within-household but not across; more broadly, this type of interference has been found in many other applications, including education (Hong and Raudenbush

[1] We use the term *population interference* to emphasize that the interference occurred across units.

[2006]; Rosenbaum [2007]), economics (Sobel [2006a]; Manski [2013]) and public health (Halloran and Struchiner [1995]). Inference in this setting is challenging because interference increases the number of potential outcomes and makes observations dependent. Aronow and Samii [2017] introduce a general framework for studying causal inference with interference: they introduce the concept of exposure mapping, define useful estimands, and construct asymptotically valid confidence intervals based on the Horvitz-Thompson estimator.

The literature on panel experiments has similarly shifted towards relaxing the no carryover effects assumption that precludes outcomes from being impacted by past assignments. For example, in the most extreme case, Bojinov and Shephard [2019] allows for arbitrary carryover effects when studying whether algorithms or humans are better at executing large financial orders; more generally, these types of relaxations have also been studied in economics [Angrist and Kuersteiner, 2011, Rambachan and Shephard, 2019], epidemiology [Robins, 1986, Robins et al., 1999], public health [Boruvka et al., 2018], and political science [Blackwell and Glynn, 2018]. Similarly to relaxing the no interference assumption, removing the no carryover assumption enables researchers to develop and explore a richer class of causal estimands that capture both the contemporaneous and delayed causal effects [Bojinov and Shephard, 2019]. The latter is particularly important for technology companies seeking to understand the long-term impact of their interventions [Basse et al., 2019, Hohnhold et al., 2015]. Similarly to the population interference setting, researchers use the analogous Horvitz-Thompson type estimator estimators to analyze experiments with carryover effects.

An apparent gap in the literature is an understanding of whether the possibility of running a panel experiment alleviates the challenges posed by population interference or makes them worse. This is particularly important for researchers wishing to run field experiments for two reasons. First, it is often the case the researchers are constrained on the maximal number of experimental units they can recruit, for instance, because of costs or limits in the total population. Second, population interference often leads to increased uncertainty that reduces only by increasing the sample size. Panel experiments can alleviate these as it is often cheaper to change an experimental unit's treatment than to recruit more subjects, and uncertainty tends to decrease as the sample size and the number of time period increase. However, what happens when there is population interference and carryover effects is unclear.

We address this gap in Chapter 2 where we introduce a unifying framework for studying panel experiments with population interference. We begin by focusing on panel experiments with population interference but no carryover effects (Section 2.2). Here, we provide asymptotically valid confidence intervals for estimands defined at specific time periods and estimands that average contrasts over multiple time periods. We also introduce a novel class of assumptions enabling us to leverage past data to improve inference at a given time. Together, our results show that using panel experiments when there is population interference allows us to achieve valid inference under much weaker conditions on the population interference and even drop all restrictions for large time horizons. These results should be particularly encouraging for researchers wishing to run field experiments when the number of experimental subjects is constrained, as is often the case in Economics (for example, Andreoni and Samuelson [2006]) and Management Science (for example, Bojinov et al. [2021a]).

We then tackle the most general setting featuring both population and temporal interference (Section 2.3). Under additional assumptions, we derive a general central limit theorem, which fails to provide the same clear benefit because of the data complexity caused by carryover effects. We also state asymptotic results for a restricted type of mixed interference that generalizes the usual stratified interference to panel experiments and provides a blueprint for deriving additional results in specific contexts. Here we show a clear benefit that incorporating a temporal dimension allows us to relax the main restriction on the maximal cluster size to obtain valid inference. For researchers, these results are slightly less encouraging but, nevertheless, provide an essential next step in understanding how to leverage panel experiments in real-world settings.

Finally, Section 2.1 details our setup by introducing the potential outcomes framework, our causal estimands and corresponding estimators, and the randomization-based framework that we leverage for all our results. We conclude the chapter with simulations (Section 2.4), empirical applications (Section 2.5), and a discussion (Section 2.6). The Appendix contains a detailed discussion of inference under population interference for standard experiments, all proofs, and additional simulations.

We then shift our focus to causal effect estimation under network interference. In practice, interference typically arises from interactions among experimental units [Hong and Raudenbush, 2006, Cai et al., 2015]. It presents a significant challenge in product experiments by tech companies, especially those involving social or market interactions. Various methods have been developed to manage interference by leveraging the structure of user interactions Eckles et al. [2017], Pouget-Abadie et al. [2019a], Karrer et al. [2021]. In the no-interference literature, regression adjustment has proven to be effective in estimating the average treatment effect, both in theory Lin [2013] and practice Deng et al. [2013]. Chin Chin [2019] considers regression adjustment under interference when assuming a linear model for the outcomes with covariates derived from the social network and the assignment vector, and estimate the parameters of the model from the experimental data. This linear model assumption is not uncommon in the interference literature and has been utilized in experiment design Harshaw et al. [2022] and interference detection Pouget-Abadie et al. [2019a].

In Chapter 3, we propose a method to estimate the global average treatment effect using regression adjustment, without assuming the true set of features as per Chin's approach. We generate adjustment features in a model-free manner, based on observed experimental data. We first presents the problem setup and motivates our method through an examination of the classic linear-in-means model in econometrics. Subsequently, we outline our procedure to generate model-free covariates from observed experimental data. We then detail how to estimate and infer the global average treatment effect using these model-free covariates. Our work culminates in simulations, empirical applications, and a discussion of our findings.

The technology industry has adopted online randomized controlled experiments, also known as A/B testing, to guide product development and make business decisions [Kohavi et al., 2013, 2020]. In the past decade, firms have developed a dynamic phase release framework in which a new treatment (such as a new product feature) is gradually released to an increasing number of units in the target population through a sequence of randomized experiments [Kohavi et al., 2020]. Companies including Google, Microsoft, LinkedIn, and Meta all developed in-house platforms that implement this framework at-scale [Tang et al., 2010, Kohavi et al., 2013, Bakshy et al., 2014, Xu et al., 2015]. Contrary to the sophisticated engineering design of such platforms, the strategy to analyze A/B testing is relatively simple—often, only the most powerful experiment in the sequence is used to provide a summary of the treatment effect, using tools from classical causal inference assuming independence among test units [Imbens and Rubin, 2015].

In scenarios such as experimenting in a social network setting or in a bipartite online marketplace, interference among units may exist. Thus a natural question is whether such interference harms the validity of simple inference procedures. Specific designs have been proposed to test or correct for the interference effects in different applications [Saveski et al., 2017, Eckles et al., 2017, Ugander et al., 2013, Pouget-Abadie et al., 2019b, Johari et al., 2022]. However, these designs are limited to specific applications and often require significant engineering work to implement in parallel to the existing A/B testing infrastructure in most companies. Even when such designs are implemented, their complex nature often results in lower throughput and can slow down the decision process.

In Chapter 4, we introduce a widely applicable procedure to test for interference in generic online experiments. The proposed method utilizes data from multiple experiments in the sequence. It can be implemented on top of an existing A/B testing platform with a separate flow and does not require *a priori* the knowledge of the underlying interference mechanism. Once implemented, this test can be run as a standard screening for any A/B test running on the platform. If the test suggests that no interference exists, the experimenter can proceed with classical causal inference analysis with confidence; if the test suggests that some form of interference does exist, the experimenter may need to redesign experiments in a more delicate way. At the platform level, such screening could provide valuable and timely feedback on the choice of designs and help experimenters update development roadmaps accordingly.

Chapter 2

Population Interference in Panel Experiments

2.1 Setup

2.1.1 Assignments

Consider a randomized experiment occurring over T periods, on a finite population of n experimental units. At each time step $t \in \{1, \cdots, T\}$, unit $i \in \{1, \cdots, n\}$ can be assigned to treatment ($W_{i,t} = 1$) or control ($W_{i,t} = 0$); extensions to non-binary treatments are straightforward. We denote by $W_{i,1:t} = (W_{i,1}, W_{i,2}, \cdots, W_{i,t})$ the assignment path up to time t for unit i, $W_{1:n,t}$ the assignment vector for all n units at time step t and $W_{1:n,1:t} \in \{0,1\}^{n \times t}$ the assignment matrix. Hence, for each i and t, $W_{i,1:t}$ is a vector of length t, $W_{1:n,t}$ is a vector of length n and $W_{1:n,1:t}$ is a matrix of dimension $n \times t$.

We define an assignment mechanism (or design) to be the probability distribution of the assignment matrix $\mathbb{P}(W_{1:n,1:T})$. Following much of the literature on analyzing complex experiments, we adopt the randomization-based approach to inference, in which the assignment mechanism is the only source of randomness (see Kempthorne [1955] and Abadie et al. [2020] for extended discussions). Throughout, we use lower cases w with the appropriate subscript for realizations of the assignment matrix W.

2.1.2 Potential outcomes and exposure mappings

The goal of causal inference is to study how an intervention impacts an outcome of interest. Following the potential outcomes formulation, for panel experiments without any assumptions, each unit i at time t has 2^{nT} potential outcomes corresponding to the total number of distinct realizations of the assignment matrix, denoted by $Y_{i,t}(w_{1:n,1:T})$. For simplicity, we assume that the potential outcomes are one dimensional, although it is straightforward to relaxing this assumption.

In randomized experiments, where we control the assignment mechanism, the outcomes at time t are not impacted by future assignments that have yet to be revealed to the units [Bojinov and Shephard, 2019]. This assumption drastically reduces the total number of potential outcomes[1] and will be implicitly made throughout this chapter. Potential outcomes depend on assignments in various ways. We now introduce two concepts that we will constantly refer to hereafter.

Definition 1 (No carryover effect and population interference). We say that there is no carryover effect if and only if

$$Y_{i,t}(w_{1:n,1:t}) = Y_{i,t}(w^{'}_{1:n,1:t}) \text{ whenever } w_{1:n,t} = w^{'}_{1:n,t}.$$

And we say that there is no population interference if and only if

$$Y_{i,t}(w_{1:n,1:t}) = Y_{i,t}(w^{'}_{1:n,1:t}) \text{ whenever } w_{i,1:t} = w^{'}_{i,1:t}.$$

Unfortunately, inference is still impossible without any assumptions on the population interference structure [Basse and Airoldi, 2018]. One way forward is to assume that the outcomes of unit i depend only on the treatments assigned to a subset of the population. This intuition extends more generally to the assertion that the outcome of unit i at time t depends on a low-dimensional representation of $w_{1:n,1:t}$. Formally, for each unique i, t pair we define the exposure mapping $f_{i,t} : \{0,1\}^{n\times t} \to \Delta$, where Δ is the set of possible exposures[2] [Aronow and Samii, 2017].

Defining exposure mappings in this flexible manner allows us to unify and transparently consider restrictions on the population interference and the duration of the carryover effect. Throughout this chapter, we restrict our focus to properly specified time-invariant exposure mappings, which are formally defined below.

Assumption 2 (Properly specified time-invariant exposure mapping). The exposure mappings are properly specified if, for all pairs $i \in \{1, \cdots, n\}$ and $t \in \{1, \cdots, T\}$, and any two assignment matrices $w_{1:n,1:t}$ and $w^{'}_{1:n,1:t}$,

$$\begin{aligned} &Y_{i,t}(w_{1:n,1:t}) = Y_{i,t}(w^{'}_{1:n,1:t}) \text{ whenever} \\ &f_{i,t}(w_{1:n,1:t}) = f_{i,t}(w^{'}_{1:n,1:t}). \end{aligned}$$

For $p \in \{1, \cdots, T\}$, we say the exposure mappings are p-time-invariant if for any $t, t^{'} \in \{p, \cdots, T\}$

[1] The assumption, known as non-anticipating potential outcomes [Bojinov and Shephard, 2019], can be violated if experimental units are told what their future assignments will be and modify their present behavior as a result. For instance, this could occur for shoppers who expect to receive a considerable discount on a subsequent day and may curtail their spending until they receive the discount.

[2] To make exposure mappings useful, we assume the cardinality of Δ is (substantially) smaller than $n \times t$.

and any unit i,

$$f_{i,t}(w_{1:n,1:t}) = f_{i,t'}(w_{1:n,1:t'}) \text{ whenever}$$
$$w_{1:n,t-p+1:t} = w_{1:n,t'-p+1:t'}.$$

The exposure mappings are time-invariant if the exposure mappings are p-time-invariant for some $p \in \{1, \cdots, T\}$. We say the exposure mappings are properly specified time-invariant exposure mappings if they are both properly specified and time-invariant.

Properly specified exposure mappings can be thought of as defining "effective treatments", allowing us to write

$$Y_{i,t}(w_{1:n,1:t}) = Y_{i,t}(f_{i,t}(w_{1:n,1:t})) = Y_{i,t}(h_{i,t})$$

where $h_{i,t} = f_{i,t}(w_{1:n,1:t}) \in \Delta$. Time-invariant exposure mappings constrain the relationship between experimental units to be invariant over time. Specifically, it does not allow the exposure mappings to change across time. For example, if at time $t = 1$, the outcomes depend on the fraction of treated neighbors in the graph then it cannot be the case that at time $t = 2$ the outcomes now depend on the number of treated neighbors in the graph. We will see why such an invariance assumption is necessary in the next section when we define causal effects. Of course, the validity of Assumption 2 depends on the exact definition of the exposure mapping and should be informed by the empirical context.

Throughout this chapter, we consider a special class of exposure mappings that restrict the outcomes of unit i to depend only on the assignments of a predefined subset of units that we refer to as i's neighborhood and index by $\mathcal{N}_i$. For example, for units connected through a social network, $\mathcal{N}_i$ indexes the set of nodes connected to i by an edge; for units organized households, $\mathcal{N}_i$ indexes the set of units that live in the same household as i; and for units located in space, $\mathcal{N}_i$ indexes the set of units who are at most a certain distance away from unit i.

Definition 3 (Locally Effective Assignments (LEA))*.* We say the assignments and exposure mappings are locally effective if the exposure mappings are p-time-invariant for some $p \in \{1, \cdots, T\}$ and

$$f_{i,t}(w_{1:n,1:t}) = f_{i,t}(w_{\mathcal{N}_i,t-p+1:t}),$$

with the convention that $w_{\mathcal{N}_i,t-p+1:t} = w_{\mathcal{N}_i,1:t}$ for $t - p + 1 \leq 0$.

Although LEA imposes further structure, it still provides a great deal of flexibility as it incorporates all notions of traditional population interference and temporal carryover effects as special cases. For example, fixing $p = 1$ makes the exposure values depend only on current assignments, which is equivalent to usual population interference. On the other hand, fixing $\mathcal{N}_i = \{i\}$ is equivalent to the no interference assumption imposed on panel experiments in Bojinov et al. [2021b]. Of course, these special cases are interesting and extensively studied, but our general formulation's real benefit is to consider scenarios where there is both population interference and carryover effects.

Example 1 (Example of Locally Effective Assignments). We consider an example where the exposure values depend on past assignments. In particular, let

$$f_{i,t}(w_{1:n,1:t}) = (w_{i,t-1}, w_{i,t}, u_{i,t-1}, u_{i,t})$$

where $u_{i,t-1} = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} w_{j,t-1}$ and $u_{i,t} = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} w_{j,t}$; we use $|\mathcal{A}|$ to denote the cardinality of the set $\mathcal{A}$. Hence, one unit's assignment and the fraction of treated neighbors at previous time step matter as well. This is a special case of LEA with $p = 2$. In this example, the exposure mappings are 2-time-invariant: for $t, t^{'} \geq 2$, if $w_{1:n,(t-1):t} = w_{1:n,(t^{'}-1):t^{'}}$ then $f_{i,t}(w_{1:n,1:t}) = f_{i,t^{'}}(w_{1:n,1:t^{'}})$.

We conclude this section by pointing out that though the LEA(p) assumption does not assume any model of the outcomes, it does have one limitation. Namely, it rules out long-range dependency of past assignments and also population interference beyond one's neighborhood. The long-range dependency in time is not uncommon in econometrics literature [Judson and Owen, 1999, Wooldridge, 2010]. For example, if we consider the following parametric model [Wooldridge, 2010]:

$$Y_{i,t} = \rho Y_{i,t-1} + \beta W_{i,t} + \epsilon_{i,t}, \text{ where } \rho \in (-1, 1) \text{ and } \mathbb{E}[\epsilon_{i,t} | Y_{i,1}, \cdots, Y_{i,t}, W_{i,1}, \cdots, W_{i,t}] = 0.$$

Such a specification leads to infinite-long dependency of past assignments on the current outcomes. Population interference beyond local interference has also been studied a lot in econometrics literature [Manski, 1993, Bramoullé et al., 2009, Leung, 2022]. In summary, LEA(p) brings in the flexibility of doing inference without making modeling assumption but loses the flexibility of accounting for long-range time and population dependency.

2.1.3 Causal effects

Causal effects, within the potential outcomes framework, are defined as contrasts of each unit's potential outcomes under alternate assignments [Imbens and Rubin, 2015]. As the number of possible contrasts grows exponentially with the number of distinct potential outcomes, we focus on two important special cases.

The first—which is well-defined regardless of the interference structure—compares the difference in the potential outcomes across two extreme scenarios: assigning every unit to treatment, $W_{1:n,1:t} = 1_{1:n,1:t}$, as opposed to control, $W_{1:n,1:t} = 0_{1:n,1:t}$.

Definition 4 (Total effect at time t). The total effect at time t is

$$\tau_t^{TE} = \frac{1}{n} \sum_{i=1}^{n} Y_{i,t}(1_{1:n,1:t}) - \frac{1}{n} \sum_{i=1}^{n} Y_{i,t}(0_{1:n,1:t}).$$

Our total effect at time t corresponds to the Global Average Treatment Effect sometimes used in single time experiments [Ugander and Yin, 2020]. In the absence of interference and carryover effects, the total effect at time t reduces to the usual average treatment effect at time t.

The second—which requires Assumption 2—provides a much richer class of causal effects with important practical applications. The TEC estimand is the generalization of the usual exposure contrast estimands [Aronow and Samii, 2017] to the panel experiment setting. Hereafter, the letter k will always represent values in Δ.

Definition 5 (Temporal exposure contrast (TEC)). For any time step t and exposure values $k, k^{'} \in \Delta$, we define the temporal exposure contrast between k and $k^{'}$ to be

$$\tau_t^{k,k^{'}} = \frac{1}{n}\sum_{i=1}^{n} Y_{i,t}(k) - \frac{1}{n}\sum_{i=1}^{n} Y_{i,t}(k^{'})$$

Notice that if there exist carryover effects then TEC may not be well-defined for the first few time steps. In this case, we may assume that all units are in the control group prior to the first time step in the panel experiment.

In panel experiments, researchers are often less interested in the idiosyncratic effects at each point in time and instead focus on the temporal average causal effect that captures the intervention's average impact across both time and units [Bojinov and Shephard, 2019, Bojinov et al., 2021b, 2022]. For example, Bojinov and Shephard [2019] are not interested in the relative difference between an algorithm or a human executing a large financial order on an arbitrary day of the experiment but are instead interested in the average difference across multiple trades on the same market. Technology companies like DoorDash [DoorDash, 2018] use switchback design for panel experiments and consider average effect across time to make product decision.

Definition 6 (Average total effect). The average total effect is

$$\overline{\tau^{TE}} = \frac{1}{T}\sum_{t=1}^{T} \tau_t^{TE}.$$

Similar to the total effect, in many applications, we are interested in the TEC's temporal average.

Definition 7 (Average temporal exposure contrast (ATEC)). For any exposure values $k, k^{'} \in \Delta$, we define the average temporal exposure contrast between k and $k^{'}$ to be

$$\bar{\tau}^{k,k^{'}} = \frac{1}{T}\sum_{t=1}^{T} \tau_t^{k,k^{'}}$$

Remark 1. Without assuming that the exposure mappings are time-invariant, the definition of the ATEC becomes more cumbersome as an exposure $k \in \Delta$ may be in the image of $f_{i,t}$ for some t, but not in the image of $f_{i,t'}$. That is, $Y_{i,t}(k)$ might be well-defined while $Y_{i,t'}(k)$ is not, which makes taking temporal averages difficult.

To conclude this section, we note by passing that there are certainly many other causal estimands of interest. For example, a vast literature in econometrics and statistics studies estimation and inference of spillover effects under either different designs or different model assumptions [Leung, 2020, Bramoullé et al., 2020, Vazquez-Bare, 2022].

2.1.4 Estimation and inference

The observed data

For any choice of exposure mappings $\{f_{i,t}\}$, the observed assignment path $W_{1:n,1:t}$ induces the exposure $H_{i,t} = f_{i,t}(W_{1:n,1:t})$ for each i and t; in particular, the assignment mechanism $\mathbb{P}(W_{1:n,1:t})$ induces a distribution for the exposures $\mathbb{P}(H_{i,t})$ for each i and t. Under Assumption 2[3], the observed outcomes $Y_{i,t}$ for unit i at time t can therefore be written:

$$Y_{i,t} = \sum_{k \in \Delta} \mathbf{1}(H_{i,t} = k) Y_{i,t}(k), \quad \forall i \in 1, \cdots, n, \forall t \in 1, \cdots, T,$$

We will use the observed data to estimate the causal effects defined in 2.1.3.

Estimation

For the different interference structures studied in the following sections, we will rely on Horvitz-Thompson estimators [Horvitz and Thompson, 1952], or variations of it; e.g., to estimate $\tau_t^{k,k^{'}}$, we will use:

$$\hat{\tau}_t^{k,k^{'}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}(H_{i,t} = k)}{\mathbb{P}(H_{i,t} = k)} Y_{i,t} - \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}(H_{i,t} = k^{'})}{\mathbb{P}(H_{i,t} = k^{'})} Y_{i,t}. \tag{2.1}$$

Taking the temporal average of (2.1) provides a natural estimator of $\bar{\tau}^{k,k^{'}}$,

$$\hat{\bar{\tau}}^{k,k^{'}} = \frac{1}{T} \sum_{t=1}^{T} \hat{\tau}_t^{k,k^{'}}. \tag{2.2}$$

Similarly, if we let $f_{i,t}(1_{1:n,1:t}) = h_{i,t}^1$ and $f_{i,t}(0_{1:n,1:t}) = h_{i,t}^0$, then we can estimate total effect at time t (c.f. Definition 4) by the following estimator

$$\hat{\tau}_t^{TE} = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}(H_{i,t} = h_{i,t}^1)}{\mathbb{P}(H_{i,t} = h_{i,t}^1)} Y_{i,t} - \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}(H_{i,t} = h_{i,t}^0)}{\mathbb{P}(H_{i,t} = h_{i,t}^0)} Y_{i,t}. \tag{2.3}$$

[3] We additionally assume that each unit fully complies with the assignment, leaving the relaxation of this assumption as future work.

Again, we have a natural estimator of average total effect induced by the above estimator

$$\hat{\bar{\tau}}^{TE} = \frac{1}{T}\sum_{t=1}^{T} \hat{\tau}_t^{TE}. \tag{2.4}$$

The properties of these estimators are discussed in details in the rest of this manuscript.

Randomization-based inference

Throughout this chapter, we adopt the randomization-based framework— that is, we consider the potential outcomes as fixed, with the assignment being the only source of randomness. This framework has seen a recent uptake in causal inference [Lin, 2013, Li and Ding, 2017, Li et al., 2019] and has become the standard for analyzing experiments with population interference [Aronow and Samii, 2017, Basse and Feller, 2018, Chin, 2018, Sävje et al., 2021] and unbounded carryover effects [Bojinov and Shephard, 2019, Rambachan and Shephard, 2019, Bojinov et al., 2021b, 2022].

There are two dominant inferential strategies within the randomization framework. The first is to use Fisher (conditional) randomization tests for sharp null hypotheses of no exposure effects, or for pairwise null hypotheses contrasting two exposures. While these tests deliver p-values that are exact and non-asymptotic, they are challenging to run with complex exposure mappings [Athey et al., 2018, Basse et al., 2019, Puelz et al., 2022].

The second, which we focus on in this chapter, is to construct confidence intervals based on the asymptotic distribution of our estimators. Intuitively, the asymptotic distribution represents a sequence of hypothetical randomized experiments in which either the number of units increases, the number of time steps increases, or both [Li and Ding, 2017, Bojinov et al., 2021b]. Within each step, we apply the analogous assignment mechanism, obtain the observed data, and compute our proposed estimand to estimate the causal effect of interest [Aronow and Samii, 2017, Chin, 2018].

Under the randomization framework, it is easy to show that the Horvitz-Thompson estimators $\hat{\tau}_t^{k,k'}$, $\hat{\bar{\tau}}^{k,k'}$, $\hat{\tau}_t^{TE}$ and $\hat{\bar{\tau}}^{TE}$ are unbiased for $\tau_t^{k,k'}$, $\bar{\tau}^{k,k'}$, τ_t^{TE} and $\bar{\tau}^{TE}$, respectively[4]; obtaining central limit theorems in this setting, however, is notoriously challenging. In the next two sections, we develop such results for the above four estimators under different experiment assumptions.

2.2 Panel experiments with population interference and no carryover effects

Depending on their structure and on the researcher's goals, panel experiments with multiple treatment periods may be a blessing or a curse. Suppose the temporal dimension does not interact with the interference mechanism, which occurs when there is only purely population interference.

[4] For example, see Bojinov and Shephard [2019] and Aronow and Samii [2017] for explicit proof.

In that case, the inference is equivalent to the standard experimental setup (Appendix 2.7.1), or it may benefit from the additional information if we are willing to consider a different estimand (Section 2.2.1) or additional assumptions (Section 2.2.2). In contrast, the presence of population spillovers in addition to the carryover effects from the previous section — a setting we call "mixed interference" — significantly compromises our ability to draw inference (see Section 2.3). In this setting, however, if the researcher has control over the temporal dimension, it may be possible to reduce the carryover effects by increasing the physical time between randomization points or adding a "cool-off" period Bojinov et al. [2022].

We work exclusively with temporally independent assignment mechanisms in this section, i.e., $W_{1:n,t}$ and $W_{1:n,t'}$ are independent for any t and $t^{'}$. We assume the same set of assumptions for each time step t as in Appendix 2.7.1.

2.2.1 Average temporal exposure contrast

In Section 2.7.1, we showed that under pure population interference, inference on the TEC at time t could strictly be reduced to the cross-sectional setting, where the only relevant asymptotic regime takes $n \to \infty$. When considering the ATEC and its natural estimator $\hat{\bar{\tau}}^{k,k'}$, however, the asymptotic picture changes and we may now consider, broadly speaking, three regimes: (1) T fixed and $n \to \infty$; (2) $T \to \infty$ and $n \to \infty$; (3) $T \to \infty$ and n fixed. An important insight that we will emphasize in this section is that inference in these three regimes requires different constraints on the population interference mechanism. Roughly speaking, the larger T is relative to n, the more interference we can tolerate.

To make this more formal, denote by $d_n^{(t)}$ the maximal number of dependent exposure values for any unit i at time step t. Let $d_n = \limsup_{t\to\infty} d_n^{(t)}$ with the convention that for fixed T, $d_n = \max\{d_n^{(1)}, \cdots, d_n^{(T)}\}$. Hence, d_n in this section is different from d_n in the previous section in the sense that it is a bound on all time steps. Our first result establishes a central limit theorem in the fixed T regime.

Theorem 8. *Suppose we have pure population interference, then for any T, under Assumption 2 and Assumptions 16-18 in Appendix 2.7.1 and the condition that $d_n = o(n^{1/4})$, we have*

$$\frac{\sqrt{nT}(\hat{\bar{\tau}}^{k,k^{'}} - \bar{\tau}^{k,k^{'}})}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\sigma_{n,t}^2}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$, where $\sigma_{n,t}^2 = Var(\sqrt{n}\hat{\tau}_t^{k,k^{'}})$.

This first theorem states a central limit theorem for the regime where T is fixed and $n \to \infty$, making it relevant for applications where N is much larger than T. Notice that, like Theorem 19, it requires $d_n = o(n^{1/4})$. Intuitively, this is because this asymptotic regime is closest to that of the previous section: any finite number of time periods T is negligible compared with infinitely many observations n.

At the other extreme, we consider the regime where $T \to \infty$ and n is fixed:

Theorem 9. *Suppose we have pure population interference and Assumptions 2, 16, 17 are satisfied. Let $\sigma_{n,t}^2 = Var(\sqrt{n}\hat{\tau}_t^{k,k'})$, we further assume that $\frac{1}{T}\sum_{t=1}^T \sigma_{n,t}^2$ is bounded away from 0 for any T. We then have*

$$\frac{\sqrt{nT}(\hat{\bar{\tau}}^{k,k'} - \bar{\tau}^{k,k'})}{\sqrt{\frac{1}{T}\sum_{t=1}^T \sigma_{n,t}^2}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $T \to \infty$.

This central limit theorem makes no assumption whatsoever on the interference mechanism, beyond assuming that there are no carryover effects: in particular, we allow a unit's outcome to depend on any other unit's assignment. This perhaps surprising fact sheds some light into the nature of inference for the ATEC, and how it differs from the TEC. Intuitively, a central limit theorem requires enough "nearly independent" observations: this means that even if at any time step t, the observations are all correlated, we can still have infinitely many independent observations if: (1) observations are uncorrelated across time and (2) we observe infinitely many time periods.

The next theorem formalizes this intuition, by making the trade-off between the growth rates of T and d_n explicit:

Theorem 10. *Suppose we have pure population interference and Assumption 2, Assumptions 16-18 are satisfied, then for $T = T(n)$ such that either*

$$\frac{n}{T} \to 0 \tag{2.5}$$

or

$$\frac{\min\{d_n^2, n\}}{\sqrt{nT}} \to 0 \tag{2.6}$$

holds, we have

$$\frac{\sqrt{nT}(\hat{\bar{\tau}}^{k,k'} - \bar{\tau}^{k,k'})}{\sqrt{\frac{1}{T}\sum_{t=1}^T \sigma_{n,t}^2}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$, where $\sigma_{n,t}^2 = Var(\sqrt{n}\hat{\tau}_t^{k,k'})$.

Condition (2.5) is actually a special case of condition (2.6): if we do not impose any assumptions on the interference, $\min\{d_n^2, n\}$ is just n, so we need $\frac{n}{\sqrt{nT}} \to 0$, which is equivalent to $\frac{n}{T} \to 0$. Condition 2.6 gives us more subtle control over the rate of growth required of T for any given level of interference. For instance, while for finite T we would require $d_n = o(n^{1/4})$, if T grows as $T(n) = \sqrt{n}$ we only require $d_n = o(n^{1/2})$. As with the previous theorem, the intuition behind this result is that as d_n becomes larger, the number of "nearly independent" observations at each time point shrinks — this must be counterbalanced by an increase in the the number of temporal observation, i.e, an increase in the rate of $T = T(n)$.

Unfortunately, as is typical in finite population causal inference, $\mathrm{Var}(\hat{\tau}_t^{k,k'})$ contains terms that are products of potential outcomes that can never be simultaneously observed from a single experiment, making it non-identifiable [Basse and Bojinov, 2020]. Instead, researchers derive an upper bound to the variance and compute unbiased estimates for this bound, allowing them to conduct conservative inference (i.e., derive confidence intervals with higher coverage than the nominal level). Without making assumptions on the assignment mechanism, we can obtain a simple bound by replacing all non-observable products of potential outcomes with the sum of their squares [Aronow and Samii, 2017], we denote the estimate of the bound by $\widehat{\mathrm{Var}}(\sqrt{n}\hat{\tau}_t^{k,k'})$. The specific expression can be found in the following proposition:

Proposition 1. *(Estimator of variance) We let*

$$\begin{aligned}
\widehat{Var}(\sqrt{n}\hat{\tau}^{k,k'}) = \frac{1}{n}\Bigg\{ &\sum_{i=1}^{n} \mathbf{1}(H_i = k)(1-\pi_i(k))\left[\frac{Y_i}{\pi_i(k)}\right]^2 + \sum_{i=1}^{n}\sum_{j\neq i, \pi_{ij}(k)=0}\left[\frac{\mathbf{1}(H_i = k)Y_i^2}{2\pi_i(k)} + \frac{\mathbf{1}(H_j = k)Y_j^2}{2\pi_j(k)}\right] \\
&+ \sum_{i=1}^{n}\sum_{j\neq i, \pi_{ij}(k)>0} \mathbf{1}(H_i = k)\mathbf{1}(H_j = k) \times \frac{\pi_{ij}(k) - \pi_i(k)\pi_j(k)}{\pi_{ij}(k)}\frac{Y_i}{\pi_i(k)}\frac{Y_j}{\pi_j(k)} \\
&+ \sum_{i=1}^{n} \mathbf{1}(H_i = k')(1-\pi_i(k'))\left[\frac{Y_i}{\pi_i(k')}\right]^2 + \sum_{i=1}^{n}\sum_{j\neq i, \pi_{ij}(k')=0}\left[\frac{\mathbf{1}(H_i = k')Y_i^2}{2\pi_i(k')} + \frac{\mathbf{1}(H_j = k')Y_j^2}{2\pi_j(k')}\right] \\
&+ \sum_{i=1}^{n}\sum_{j\neq i, \pi_{ij}(k')>0} \mathbf{1}(H_i = k')\mathbf{1}(H_j = k') \times \frac{\pi_{ij}(k') - \pi_i(k')\pi_j(k')}{\pi_{ij}(k')}\frac{Y_i}{\pi_i(k')}\frac{Y_j}{\pi_j(k')} \\
&- 2\sum_{i=1}^{n}\sum_{j\neq i, \pi_{ij}(k,k')>0} \Big(\pi_{ij}(k,k') - \pi_i(k)\pi_j(k')\Big) \times \frac{\mathbf{1}(H_i = k)\mathbf{1}(H_j = k')}{\pi_{ij}(k,k')}\frac{Y_i}{\pi_i(k)}\frac{Y_j}{\pi_j(k')} \\
&+ 2\sum_{i=1}^{n}\sum_{j\neq i, \pi_{ij}(k,k')=0}\left[\frac{\mathbf{1}(H_i = k)Y_i^2}{2\pi_i(k)} + \frac{\mathbf{1}(H_j = k')Y_j^2}{2\pi_j(k')}\right]\Bigg\},
\end{aligned}$$

then $\mathbb{E}\left[\widehat{Var}(\sqrt{n}\hat{\tau}^{k,k'})\right] \geq Var(\sqrt{n}\hat{\tau}^{k,k'})$

We omit the proof here as it can be found in Aronow and Samii [2017]. We drop the subscript t to ease notations. With the above proposition and central limit theorems, inference proceeds as follows:

Proposition 2. *Suppose Theorem 8 or 10 holds, then for any* $\delta > 0$*,*

$$\mathbb{P}\left(\bar{\tau}^{k,k'} \in \left[\hat{\bar{\tau}}^{k,k'} - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{1-\delta}}\sqrt{\frac{1}{T^2}\sum_{t=1}^{T}\widehat{Var}(\hat{\tau}_t^{k,k'})}, \hat{\bar{\tau}}^{k,k'} + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{1-\delta}}\sqrt{\frac{1}{T^2}\sum_{t=1}^{T}\widehat{Var}(\hat{\tau}_t^{k,k'})}\right]\right) \geq 1-\alpha$$

for large n. Moreover, suppose Theorem 9 holds, then for any $\delta > 0$,

$$\mathbb{P}\left(\bar{\tau}^{k,k'} \in \left[\hat{\bar{\tau}}^{k,k'} - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{1-\delta}}\sqrt{\frac{1}{T^2}\sum_{t=1}^{T}\widehat{Var}(\hat{\tau}_t^{k,k'})}, \hat{\tau}^{k,k'} + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{1-\delta}}\sqrt{\frac{1}{T^2}\sum_{t=1}^{T}\widehat{Var}(\hat{\tau}_t^{k,k'})}\right]\right) \geq 1-\alpha$$

for large T.

The proof of the above proposition builds on the proof of Proposition 9 in Appedix 2.7.1.

2.2.2 Shrinkage estimator under stability assumption

In Section 2.7.1 and Section 2.2.1, we considered inference on the TEC and ATEC under population interference. In this section, we focus on the following question: can we leverage the temporal information to improve inference on the TEC? Our results here are weaker than in the previous section — indeed, we provide neither central limit theorem nor asymptotic confidence interval — but we believe they are an exciting avenue for future work.

So far, we have considered only Horvitz-Thompson estimators which, while analytically tractable, are known to have large variance when exposure probabilities are small. The key idea of this section is that if the potential outcomes do not vary too much across time, then estimates of the TEC at time $t' < t$ can be used to improve our estimate of the TEC at time t. This assumption can be formalized as follows:

Assumption 11 (Weak stability of potential outcomes)*.* We say the potential outcome matrix $Y_{i,t}, i = 1, \cdots, N,\ t = 1, \cdots, T$ is ϵ-weakly stable if for each i and exposure value k, we have $|Y_{i,t}(k) - Y_{i,t+1}(k)| \leq \epsilon, \forall t \in \{1, \cdots, T-1\}$. If we further assume that $\epsilon = 0$, we then say that the potential outcome matrix is strongly stable.

Our results can be easily generalized to the case where the uniform bound ϵ in the definition is replaced by a time dependent bound ϵ_t. Throughout, we focus on the estimation of the total effect at time t as an example to illustrate how we can leverage temporal information under weak stability.

Under pure population interference and time-invariant exposure mappings,

$$\tau_t^{TE} = \frac{1}{n}\sum_{i=1}^{n} Y_{i,t}(h_i^1) - \frac{1}{n}\sum_{i=1}^{n} Y_{i,t}(h_i^0), \tag{2.7}$$

where $h_i^1 = f_i(1_{t,1:n})$ and $h_i^0 = f_i(0_{t,1:n})$.

To build some intuition we first investigate how to leverage just a single past time period, $t' = t-1$ to improve estimation at time t. The idea is that by considering a convex combination $\hat{\tau}_t^c = \alpha\hat{\tau}_t^{TE} + (1-\alpha)\hat{\tau}_{t-1}^{TE}$, for some $\alpha \in [0,1]$ as an estimator of $\hat{\tau}_t^{TE}$, we introduce some bias but reduce the variance — the hope being that under weak stability, the bias introduced will be modest compared to the reduction in variance. This is formalized in the following proposition.

Proposition 3 (Bound on the bias of $\hat{\tau}_t^c$)**.**

$$|\mathbb{E}[\hat{\tau}_t^c] - \tau_t^{TE}| \leq 2(1-\alpha)\epsilon \tag{2.8}$$

As we can see, the absolute bias of $\hat{\tau}_t^c$ is bounded by a quantity that grows linearly with ϵ: if ϵ is very small, then so will the maximum bias. In particular, $\hat{\tau}_t^c$ is unbiased for τ_t^{TE} if $\epsilon = 0$ — which corresponds to the assumption that the potential outcomes do not vary across time. Under some conditions, it can be guaranteed that the gain in bias is more than counterbalanced by a reduction in variance — making it a worthwhile trade-off, in terms of the mean squared error (MSE).

Proposition 4. *Suppose $Var(\hat{\tau}_t^{TE}) > Cov(\hat{\tau}_t^{TE}, \hat{\tau}_{t-1}^{TE})$, then there exists some $\alpha \in (0,1)$ such that $\hat{\tau}_t^c = \alpha\hat{\tau}_t^{TE} + (1-\alpha)\hat{\tau}_{t-1}^{TE}$ has lower MSE than $\hat{\tau}_t^{TE}$. Moreover, if we have $Var(\hat{\tau}_t^{TE}) - Var(\hat{\tau}_{t-1}^{TE}) > 4\epsilon^2$ then we know that $\hat{\tau}_t^c = \frac{1}{2}\hat{\tau}_t^{TE} + \frac{1}{2}\hat{\tau}_{t-1}^{TE}$ has lower MSE than $\hat{\tau}_t^{TE}$.*

By Cauchy-Schwartz inequality,

$$\mathrm{Cov}(\hat{\tau}_t^{TE}, \hat{\tau}_{t-1}^{TE}) \leq \sqrt{\mathrm{Var}(\hat{\tau}_t^{TE})}\sqrt{\mathrm{Var}(\hat{\tau}_{t-1}^{TE})},$$

hence if $\mathrm{Var}(\hat{\tau}_t^{TE}) > \mathrm{Var}(\hat{\tau}_{t-1}^{TE})$, then

$$Var(\hat{\tau}_t^{TE}) > Cov(\hat{\tau}_t^{TE}, \hat{\tau}_{t-1}^{TE}).$$

Therefore, as long as the current variance is larger, by choosing some α, the convex combination type estimator would give us a better estimator in terms of MSE. Moreover, as the proposition suggests, if we know the difference is bigger than $4\epsilon^2$, we know that $\alpha = \frac{1}{2}$ is sufficient.

Note that if we further assume that assignments are also independent across time, then $\mathrm{Cov}(\hat{\tau}_t^{TE}, \hat{\tau}_{t-1}^{TE}) = 0$, hence we have the Proposition 5.

Proposition 5. *Suppose that the assignments are independent across time, then there exists some $\alpha \in (0,1)$ such that $\hat{\tau}_t^c = \alpha\hat{\tau}_t^{TE} + (1-\alpha)\hat{\tau}_{t-1}^{TE}$ has lower MSE than $\hat{\tau}_t^{TE}$. The optimal α is given by $\alpha = 1 - \frac{Var(\hat{\tau}_t^{TE})}{4\epsilon^2 + Var(\hat{\tau}_t^{TE}) + Var(\hat{\tau}_{t-1}^{TE})}$.*

We show in the simulation section that the reduction in mean squared error is significant when n is small.

Under the $\epsilon-$stability assumption, Algorithm 1 provides a data dependent approach to estimate ϵ and allows us to obtain estimate $\hat{\alpha}$ of the weight parameter α,

$$\hat{\alpha} = 1 - \frac{\widehat{\mathrm{Var}}(\hat{\tau}_t^{TE})}{\widehat{\mathrm{Var}}(\hat{\tau}_t^{TE}) + \widehat{\mathrm{Var}}(\hat{\tau}_{t'}^{TE}) + 4(t-t')^2\hat{\epsilon}^2},$$

Algorithm 1 Algorithm to estimate ϵ

```
Initialize $\hat{\epsilon} = 0$.
For $t = 1$ to $T - 1$:
    • For $i = 1, 2, \cdots, n$ compute $h_{i,t}$ and $h_{i,t+1}$.
    • If $h_{i,t} = h_{i,t+1}$, i.e., the exposure values at two adjacent time points are the same, compute
      $\epsilon_{i,t} = |y_{i,t} - y_{i,t+1}|$.
    • If $\epsilon_{i,t} > \hat{\epsilon}$, set $\hat{\epsilon} = \epsilon_{i,t}$.
Output $\hat{\epsilon}$.
```

where $\widehat{\text{Var}}(\hat{\tau}_t^{TE})$ can be any estimator of the variance $\text{Var}(\hat{\tau}_t^{TE})$: we discuss a few options in Proposition 11 of Appendix 2.7.2. In addition, under pure population interference and temporally independent assignments,

$$\begin{aligned}\text{Var}(\hat{\tau}_t^c) &= \text{Var}\bigg(\alpha\hat{\tau}_t^{TE} + (1-\alpha)\hat{\tau}_{t-1}^{TE}\bigg) \\ &= \alpha^2\text{Var}(\hat{\tau}_t^{TE}) + (1-\alpha)^2\text{Var}(\hat{\tau}_{t-1}^{TE}),\end{aligned}$$

which suggests the following plug-in estimator of the variance:

$$\widehat{\text{Var}}(\hat{\tau}_t^c) = \hat{\alpha}^2\widehat{\text{Var}}(\hat{\tau}_t^{TE}) + (1-\hat{\alpha})^2\widehat{\text{Var}}(\hat{\tau}_{t-1}^{TE}).$$

We also give the expression of $\text{Cov}(\hat{\tau}_t^{TE}, \hat{\tau}_{t-1}^{TE})$ and an estimator for it in Proposition 10 and Proposition 12 of Appendix 2.7.2 respectively. Equipped with the variance and the covariance estimators, we can directly check the condition in Proposition 12. The optimal α is given in the proof and can be estimated in the similar way as in the independent assignment case.

Up to now, the type of the estimator we consider is a convex combination of $\hat{\tau}_t^{TE}$ and $\hat{\tau}_{t-1}^{TE}$. We now consider a general version of this estimator such that we combine $\hat{\tau}_t^{TE}$ and $\hat{\tau}_{t'}^{TE}$ for arbitrary $t^{'} < t$. We now give the analogous results.

Proposition 6. *Suppose $Var(\hat{\tau}_t^{TE}) > Cov(\hat{\tau}_t^{TE}, \hat{\tau}_{t'}^{TE})$, then there exists some $\alpha \in (0,1)$ such that $\hat{\tau}_t^c = \alpha\hat{\tau}_t^{TE} + (1-\alpha)\hat{\tau}_{t'}^{TE}$ has lower MSE than $\hat{\tau}_t^{TE}$. Moreover, if we have $Var(\hat{\tau}_t^{TE}) - Var(\hat{\tau}_{t'}^{TE}) > 4(t-t^{'})^2\epsilon^2$, then $\hat{\tau}_t^c = \frac{1}{2}\hat{\tau}_t^{TE} + \frac{1}{2}\hat{\tau}_{t'}^{TE}$ has lower MSE than $\hat{\tau}_t^{TE}$.*

Proposition 7. *Suppose that the assignments are independent across time, then there exists some $\alpha \in (0,1)$ such that $\hat{\tau}_t^c = \alpha\hat{\tau}_t^{TE} + (1-\alpha)\hat{\tau}_{t'}^{TE}$ has lower MSE than $\hat{\tau}_t^{TE}$. The optimal α is given by $\alpha = 1 - \frac{Var(\hat{\tau}_t^{TE})}{4(t-t^{'})^2\epsilon^2 + Var(\hat{\tau}_t^{TE}) + Var(\hat{\tau}_{t'}^{TE})}$.*

As mentioned in the introduction to this section, we do not have formal inferential results at the moment — this is an open area for future work. However, based on the variance estimator above, we do have two crude ways to construct confidence intervals. The first one ignores the bias of $\hat{\tau}_t^c$

and uses Gaussian confidence interval. The second one takes advantage of Chebyshev's inequality. Specifically, note that

$$\mathbb{P}(|\hat{\tau}_t^c - (\mathbb{E}[\hat{\tau}_t^c] - \tau_t^{TE}) - \tau_t^{TE}| \geq \epsilon) \leq \frac{\text{Var}(\hat{\tau}_t^c)}{\epsilon^2},$$

hence $\forall \delta > 0$,

$$\mathbb{P}\left(\tau_t^{TE} \in \left[\hat{\tau}_t^c - (\mathbb{E}[\hat{\tau}_t^c] - \tau_t^{TE}) - \epsilon, \hat{\tau}_t^c - (\mathbb{E}[\hat{\tau}_t^c] - \tau_t^{TE}) + \epsilon\right]\right) \geq 1 - \delta$$

for $\epsilon = \sqrt{\frac{\text{Var}(\hat{\tau}_t^c)}{\delta}}$. Let $b(\hat{\tau}_t^c) = \mathbb{E}[\hat{\tau}_t^c] - \tau_t^{TE} = (1-\alpha)(\tau_{t-1}^{TE} - \tau_t^{TE})$ be the bias of our convex combination estimator. If we estimate $b(\hat{\tau}_t^c)$ by $\hat{b}(\hat{\tau}_t^c) = (1-\hat{\alpha})(\hat{\tau}_{t-1}^{TE} - \hat{\tau}_t^{TE})$, then we can use the following interval as an approximate $(1-\delta)$-level confidence interval of τ_t^{TE}:

$$\left[\hat{\tau}_t^c - \hat{b}(\hat{\tau}_t^c) - \sqrt{\frac{\widehat{\text{Var}}(\hat{\tau}_t^c)}{\delta}}, \hat{\tau}_t^c - \hat{b}(\hat{\tau}_t^c) + \sqrt{\frac{\widehat{\text{Var}}(\hat{\tau}_t^c)}{\delta}}\right].$$

We explore empirically the coverage of the above approximate confidence intervals with a simulation study in Section 2.4.

The approach we have described in this section naturally extends to using the $k-1$ previous time steps, yielding the weighted combination estimator:

$$\hat{\tau}_t^c = \alpha_1 \hat{\tau}_{t-k+1}^{TE} + \cdots + \alpha_k \hat{\tau}_t^{TE},$$

where $\alpha_1, \ldots, \alpha_k$ can be estimated by solving a slightly more involved convex optimization problem. We describe this approach in full details in Appendix 2.7.3.

2.3 Panel experiments with population interference and carryover effects

Section 2.2.1 shows that adding a temporal dimension does not hurt inference and may even help if interference remains confined to the population dimension. Mixed interference, in contrast, affects our ability to draw inference both for the TEC and ATEC, albeit in different ways. For temporal exposure contrasts (TEC), the same theorem as in Section 2.7.1 holds (recall that d_n is the maximal degree of the dependency graph of $H_1, \cdots, H_n$):

Theorem 12. *Under Assumption 2, Assumptions 16-18 and the condition that $d_n = o(n^{1/4})$, we have*

$$\frac{\sqrt{n}(\hat{\tau}_t^{k,k'} - \tau_t^{k,k'})}{Var(\sqrt{n}\hat{\tau}_t^{k,k'})^{1/2}} \xrightarrow{d} \mathcal{N}(0,1)$$

The difference with the pure population setting is not mathematical but conceptual: in the mixed setting, the exposures involve the assignments over previous time steps. Consequently, there are generally many more exposures than in the pure population setting, and each unit has a lower probability of receiving each. This leads to Horvitz-Thompson estimators with a much larger variance.

For the average temporal exposure contrast, the difference between population interference and mixed interference is starker. The main difficulty is that mixed interference breaks the temporal independence that powered the results of section 2.2.1. We first establish a general theorem and then give a specific setting under which we have a concrete result. Throughout, we assume that all the potential outcomes are uniformly bounded and the overlap assumption is satisfied for every time step t and every unit i. As in the cross-sectional setting, we impose a condition that controls the rate at which the variance shrinks:

Assumption 13. Assume that

$$\liminf_{n\to\infty} \mathrm{Var}(\sqrt{nT}\hat{\bar{\tau}}^{k,k'}) \geq \epsilon > 0$$

for some ϵ.

This technical assumption (we are generally more worried about the variance not shrinking fast enough) rules out the pathological case that the variance vanishes as $n \to \infty$.

Theorem 14. *Under Assumption 2 and Assumption 13, suppose $\{H_{i,t}\}_{i=1}^n$ is an s-dependent sequence of random variables for a fixed t and LEA(p) assumption is satisfied with some finite p. If s, n, T are such that $s^5T^4 = o(n^{1-\alpha})$ for some $0 < \alpha < 1$, then we have that*

$$\frac{\sqrt{nT}(\hat{\bar{\tau}}^{k,k'} - \bar{\tau}^{k,k'})}{\sqrt{Var(\sqrt{nT}\hat{\bar{\tau}}^{k,k'})}} \xrightarrow{d} \mathcal{N}(0,1)$$

as $n \to \infty$.

$\{H_{i,t}\}_{i=1}^n$ is an s-dependent sequence of random variables if and only if for any index set $I, J \subseteq [n]$, $\{H_{i,t}\}_{i=1}^n$ and $\{H_{i,t}\}_{i=1}^n$ are independent so long as $\min_{j\in J} j - \max_{i\in I} i > s$. The above theorem requires general assumptions on exposure values as well as the asymptotic variance though independent assignments are not required. We now focus on a specific setting, to illustrate the type of results that can be derived under mixed interference.

Consider the following natural temporal extension of the stratified interference setting (Hudgens and Halloran [2008]; Basse and Feller [2018]):

$$f_{i,t}(w_{1:n,1:t}) = f(w_{i,t-1}, w_{i,t}, \{w_{j,t-1}\}_{j\in\mathcal{N}_i, j\neq i}, \{w_{j,t}\}_{j\in\mathcal{N}_i, j\neq i})$$

where $\mathcal{N}_i$ is the group to which unit i belongs. For convenience, we fix each group to be of size r [5].

[5] this also ensures that each unit is associated with exactly the same set of exposure values so that the exposure contrast between two arbitrary exposure values is well-defined.

Theorem 15. *With the above setting and temporally independent assignments, under Assumption 13, suppose n, r, T are such that $r = o((nT)^{\frac{1}{4}})$, then*

$$\frac{\sqrt{nrT}(\hat{\bar{\tau}}^{k,k'} - \bar{\tau}^{k,k'})}{\sqrt{Var(\sqrt{nrT}\hat{\bar{\tau}}^{k,k'})}} \xrightarrow{d} \mathcal{N}(0,1)$$

as $n \to \infty$.

The theorem holds for heterogeneous group sizes as long as $\max_i r_i = o((nT)^{\frac{1}{4}})$ where $r_i = |\mathcal{N}_i|$ is the size of the group unit i belongs to. To do inference, we consider a specific example of stratified interference:

$$f_{i,t}(w_{1:n,1:t}) = \left(w_{i,t-1}, w_{i,t}, \sum_{j \in \mathcal{N}_i, j \neq i} w_{j,t-1}, \sum_{j \in \mathcal{N}_i, j \neq i} w_{j,t} \right).$$

We focus on the Bernoulli design where each unit is independently assigned to treatment with probability $\frac{1}{2}$. We consider the exposures $k = (1, 1, r-1, r-1)$ and $k' = (0, 0, 0, 0)$. Such exposure contrast is exactly the same as the total effect since essentially we are comparing the world of everyone getting treatment to the world of everyone getting control. Notice that in this case, r cannot be infinite, otherwise the overlap assumption would be violated. To ease notations, we index each unit i by a tuple (l, q), meaning that unit i is the q-th unit in the l-th group[6].

Proposition 8. *Assuming the above setup, we can estimate the asymptotic variance by*

$$\widehat{B_n}^2 = \sum_{t=1}^{T} \widehat{Var}(X_{n,t}) + 2 \sum_{t=1}^{T-1} \widehat{Cov}(X_{n,t}, X_{n,t+1}),$$

[6] If we use a tuple (l, q) to represent the $q-$th unit in the $l-$th household, then we note by passing that $0 < C_1 \leq Y_{(l,q),t}(k) \leq C_2$ for all l, q, t, k for some C_1, C_2 is sufficient for Assumption 13.

where

$$\widehat{Var}(X_{n,t}) = \frac{1}{nrT}\Bigg[\sum_{l=1}^{n}\sum_{q=1}^{r}(2^{2r}-1)\frac{\mathbf{1}(H_{(l,q),t}=k)Y^2_{(l,q),t}}{\mathbb{P}(H_{(l,q),t}=k)} + \sum_{l=1}^{n}\sum_{q=1}^{r}(2^{2r}-1)\frac{\mathbf{1}(H_{(l,q),t}=k^{'})Y^2_{(l,q),t}}{\mathbb{P}(H_{(l,q),t}=k')}$$
$$+\sum_{l=1}^{n}\sum_{q=1}^{r}\left(\frac{\mathbf{1}(H_{(l,q),t}=k)Y^2_{(l,q),t}}{\mathbb{P}(H_{(l,q),t}=k)} + \frac{\mathbf{1}(H_{(l,q),t}=k^{'})Y^2_{(l,q),t}}{\mathbb{P}(H_{(l,q),t}=k')}\right)$$
$$+\sum_{l=1}^{n}\sum_{q_1=1}^{r}\sum_{q_2\neq q_1}\Bigg((2^{2r}-1)\frac{\mathbf{1}(H_{(l,q_1),t}=k, H_{(l,q_2),t)}=k)Y_{(l,q_1),t}Y_{(l,q_2),t}}{\mathbb{P}(H_{(l,q_1),t}=k, H_{(l,q_2),t)}=k)}+$$
$$+\ (2^{2r}-1)\frac{\mathbf{1}(H_{(l,q_1),t}=k^{'}, H_{(l,q_2),t)}=k^{'})Y_{(l,q_1),t}Y_{(l,q_2),t}}{\mathbb{P}(H_{(l,q_1),t}=k', H_{(l,q_2),t)}=k')}\Bigg)$$
$$+\sum_{l=1}^{n}\sum_{q_1=1}^{r}\sum_{q_2\neq q_1}\frac{\mathbf{1}(H_{(l,q_1),t}=k)Y^2_{(l,q_1),t}}{\mathbb{P}(H_{(l,q_1),t}=k)} + \sum_{l=1}^{n}\sum_{q_1=1}^{r}\sum_{q_2\neq q_1}\frac{\mathbf{1}(H_{(l,q_2),t}=k^{'})Y^2_{(l,q_2),t}}{\mathbb{P}(H_{(l,q_2),t}=k')}\Bigg] \tag{2.9}$$

and

$$\widehat{Cov}(X_{n,t}, X_{n,t+1}) = \frac{1}{nrT}\sum_{l=1}^{n}\sum_{q_1=1}^{r}\sum_{q_2=1}^{r}\Bigg((2^r-1)\frac{\mathbf{1}(H_{(l,q_1),t}=k, H_{(l,q_2),t+1}=k)Y_{(l,q_1),t}Y_{(l,q_2),t+1}}{\mathbb{P}(H_{(l,q_1),t}=k, H_{(l,q_2),t+1}=k)}$$
$$+(2^r-1)\frac{\mathbf{1}(H_{(l,q_1),t}=k^{'}, H_{(l,q_2),t+1}=k^{'})Y_{(l,q_1),t}Y_{(l,q_2),t+1}}{\mathbb{P}(H_{(l,q_1),t}=k', H_{(l,q_2),t+1}=k')}$$
$$+\ \frac{\mathbf{1}(H_{(l,q_1),t}=k^{'})Y^2_{(l,q_1),t}}{\mathbb{P}(H_{(l,q_1),t}=k')} + \frac{\mathbf{1}(H_{(l,q_2),t+1}=k)Y^2_{(l,q_2),t+1}}{\mathbb{P}(H_{(l,q_2),t+1}=k)}$$
$$+\frac{\mathbf{1}(H_{(l,q_1),t}=k)Y^2_{(l,q_1),t}}{\mathbb{P}(H_{(l,q_1),t}=k)} + \frac{\mathbf{1}(H_{(l,q_2),t+1}=k^{'})Y^2_{(l,q_2),t+1}}{\mathbb{P}(H_{(l,q_2),t+1}=k')}\Bigg) \tag{2.10}$$

The expression of the variance is immediate from the setup, (2.34) and (2.35). The estimator is obtained by replacing the non-identifiable terms by an upper bound and estimating the upper bound accordingly.

Remark 2. The difficulty for doing inference under more general setting comes from the fact that it is hard to give explicit expression of the variance. Since there is dependence across time, the variance of $\hat{\hat{\tau}}^{k,k^{'}}$ also involves covariance between Horvitz-Thompson estimators across different times. Hence, in this case, we need at least more assumptions on the assignment mechanism in order to express the variance term explicitly.

2.4 Simulations

In this section, we use simulations to supplement some of our theoretical results, and to provide empirical guidance when theory is lacking. Section 2.4.1 explores some of the finite sample properties of our central limit theorems in different realistic settings. Section 2.4.2 explores empirically some properties of the convex combination estimator proposed in Section 2.2.2: in particular, although we do not prove central limit theorems for this estimator, we show that confidence intervals based on normal approximations behave well in our simulation, and could therefore be reasonable candidates for practical use.

2.4.1 Simulations for central limit theorems

We first explore the finite sample behavior of our central limit theorems. To make our simulations relevant, we consider a version of the popular stratified interference setting (Duflo and Saez [2003]; Basse and Feller [2018]), in which individuals are nested in groups of varying sizes, and interference may occur within but not across groups. Specifically, we consider the exposure mapping $f_i(w_{1:n}) = (w_i, u_i)$, where $u_i = 1$ if unit i has at least one treated neighbor and $u_i = 0$ otherwise, so each unit may receive one of four exposures: (0,0), (0,1), (1,0) and (1,1). Throughout, we consider a two-stage design whereby each group is assigned independently with probability $\frac{1}{2}$ to a high-exposure or low-exposure arm, and then each unit is assigned to treatment independently with probability 0.9 in high-exposure groups, and 0.1 in low-exposure groups.

We focus on the central limit theorems for ATEC. Theorem 10 establishes asymptotic results for ATEC under less constraining assumptions on the interference mechanism than for TEC (Theorem 19 in Appendix 2.7.1). To illustrate this point, we consider the stratified interference setting. We assume that the size of each group is bounded by $n^{1/3}$. In this case, $d_n = n^{1/3}$ and hence $T = \sqrt{n}$ suffices for Theorem 10. Compared to $d_n = o(n^{1/4})$ in the cross-sectional setting, we are able to have larger group size. We consider the exposure mapping $f_i(w_{1:n}) = (w_i, u_i)$ where $u_i = 0$ if less than 25% of the neighbors are treated; $u_i = 1$ if between 25% and 50% of the neighbors are treated; $u_i = 2$ if between 50% and 75% of the neighbors are treated and $u_i = 3$ if more than 75% of the neighbors are treated. We generate the potential outcomes for unit i at time step t according to $\mathcal{N}(3w_i + 2u_i + 5 + \epsilon_t, 1)$, where ϵ_t is uniform$\{-1, 1\}$. Figure 2.1 and 2.2 show that $n = 1000$ suffices for a good approximation. Moreover, the coverage of our 95% confidence interval is 95.4%.

2.4.2 Estimation under the stability assumption

In Section 2.2.2, we showed that with an appropriate choice of weights, the family of convex combination estimators outperforms the Horvitz-Thompson estimator. We illustrate this with a simulation study. We also show that although not supported by theoretical results, naively constructed confidence intervals perform well in our simulated setting.

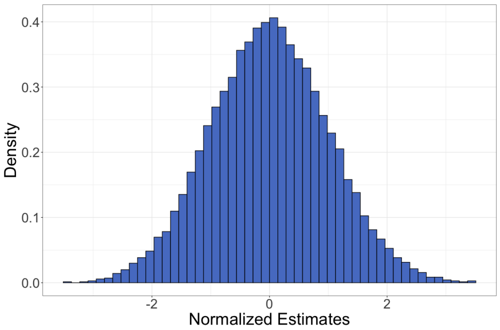


Figure 2.1: Histogram, $n = 1000$

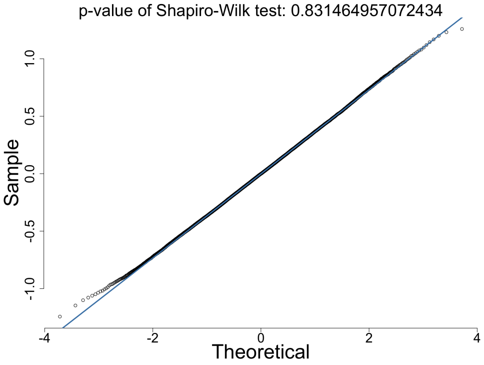


Figure 2.2: Q-Q normal plot, $n = 1000$

Sample Size	n = 50	n = 100	n = 250	n = 500	n = 750	n = 1000
RMSE for $\hat{\tau}_{20}^{TE}$	64.68	28.98	5.80	1.84	0.95	0.70
RMSE for $\hat{\tau}_{20}^{c}$, $k = 2$	14.17	9.18	3.72	1.42	0.68	0.52
RMSE for $\hat{\tau}_{20}^{c}$, $k = 5$	4.39	4.58	3.01	1.17	0.58	0.45

Table 2.1: Root mean squared errors (RMSE) for $\hat{\tau}_{20}^{TE}$, $\hat{\tau}_{20}^{c}$ with $k = 2$ and $\hat{\tau}_{20}^{c}$ with $k = 5$

Estimation under stability assumption for total effects

We consider a social network generated according to an Erdős-Rényi model, in which the units are assigned to treatment or control following a Bernoulli(1/2) design at each time step. We assume a local, pure population form of interference, summarized by the following exposure mappings:

$$f_i(w_{1:n,t}) = \left(w_{i,t}, \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} w_{j,t} \right) \tag{2.11}$$

where $\mathcal{N}_i$ is the neighborhood of the i-th unit; that is, we assume that only direct neighbors affect one's potential outcomes. For each unit i, we generate the potential outcomes at $t = 1$ randomly from $\mathcal{N}(10, 1)$. Then, for each time $t > 1$, we generate the potential outcome $Y_{i,t}(k)$ uniformly from the interval $(Y_{i,t-1}(k) - \epsilon, Y_{i,t-1}(k) + \epsilon)$, so ϵ-stability holds. Throughout our simulations, we assume that $T = 20$ and we are interested in the total effect at time step $t = 20$. We compare the performance of the standard Horvitz-Thompson estimator and the performance of the convex combination estimator for estimating the total effect τ_T^{TE} at time $t = T = 20$, varying both the population size n and the number of time steps k used in the convex combination. We estimate ϵ using Algorithm 1 described in Section 2.2.2; we use Proposition 5 to estimate α when $k = 2$, and solve the optimization problem introduced in Appendix 2.7.3 for $k \geq 3$.

We first fix ϵ to be 3 and vary the sample size. To make each unit have the same expected number of neighbors, we scale the probability p in Erdős-Rényi model accordingly. For each n, we fix the

Confidence Interval	Network 1	Network 2	Network 3
Gaussian CI with variance estimated by $\widehat{\text{Var}}^d$	92.9%	98.4%	95.9%
Gaussian CI with variance estimated by $\widehat{\text{Var}}^u$	97.2%	99.8%	100%
Chebyshev CI with variance estimated by $\widehat{\text{Var}}^d$	91.4%	94.1%	96.4%
Chebyshev CI with variance estimated by $\widehat{\text{Var}}^u$	94.6%	95.6%	97.7%

Table 2.2: Coverage of two approximate confidence intervals for τ_t^{TE} with $k = 2$

graph and generate 100 realizations of assignments. Table 2.1 shows the root mean squared errors for three kinds of estimators for the total effect: the usual Horvitz-Thompson estimator, the convex combination type estimator with $k = 2$, and the convex combination estimator with $k = 5$. We see that the convex combination type estimators effectively reduce the mean squared error. Moreover, when n is relatively small, the reduction in mean squared error is significant.

Coverage of two approximate confidence intervals

Recall that in Section 2.2.2 we gave two approximate confidence intervals of τ_t^{TE} based on our convex combination estimator $\hat{\tau}_t^c$ and variance estimator. We now provide coverage results of these two approximate confidence intervals. We assume a social network generated from the Erdős-Rényi Model with $n = 100$ and $p = 0.05$. We fix the stability parameter ϵ to be 3 and generate the data in the same way as in the previous section. To calculate the coverage, we generate 1000 realizations of the assignments and construct approximate confidence intervals accordingly.

Table 2.2 shows the two approximate confidence intervals provide reasonable coverage across the three different social networks. Although the Gaussian confidence interval ignores the bias of $\hat{\tau}_t^c$, it tends to provide better coverage than the confidence intervals obtained from the Chebyshev approach. Moreover, the Gaussian intervals tend to be shorter, making them practically more useful. Appendix 2.7.5 provides an additional table showing the average lengths of the confidence intervals in Table 2.2.

2.5 Two real data examples

We now apply our methods to two empirical applications. In the first application, we use the convex combination estimator to analyze a panel experiment and show it reduces the variance and leads to more reliable estimates of the temporal exposure contrast. In the second application, we run a semi-synthetic experiment on a social network to demonstrate the necessity of our assumptions for the validity of the analysis and provide further empirical evidence of the advantage of the convex combination estimator.

2.5.1 Rational cooperation

The panel experiment we analyze is from Andreoni and Samuelson [2006]. The authors test a game-theoretic model of "rational cooperation" through a panel experiment. Specifically, in each experiment session, they recruited 22 subjects to play 20 twice-played prisoners' dilemmas, ensuring that no player would meet the same partner twice. The twice-played prisoners' dilemma consists of two periods with different pay-off structures, as shown in Table 2.3. The parameters x_1, x_2 satisfy $x_1, x_2 \geq 0$, $x_1 + x_2 = 10$.

	C	D
C	$(3x_1, 3x_1)$	$(0, 4x_1)$
D	$(4x_1, 0)$	(x_1, x_1)

Period one

	C	D
C	$(3x_2, 3x_2)$	$(0, 4x_2)$
D	$(4x_2, 0)$	(x_2, x_2)

Period two

Table 2.3: Payoff structure in the experiment conducted by Andreoni and Samuelson [2006]. The choice C denotes "cooperate" and the choice D "defect."

Let $\lambda = \frac{x_1}{x_1+x_2}$, then for each round of the experiment, 22 subjects were grouped into 11 pairs, and each pair was randomly assigned with a $\lambda \sim \text{Unif}\{0, 0.1, \cdots, 0.9, 1\}$. The outcomes were the total payoffs. Since there are five sessions in total, we have 110 subjects and 2200 outcomes. We use this panel experiment to illustrate that the convex combination estimator effectively reduces the estimates' variance and thus produces more reliable estimates. To this end, following Bojinov et al. [2021b], we define treatment to be $\lambda > 0.6$ and control to be $\lambda \leq 0.6$. This results in a panel experiment with binary treatments and Bernoulli design with treated probability $\frac{5}{11}$. Under this setup, we generally expect a positive treatment affect as the payoffs are more concentrated in period two.

We next build a social network among all subjects in the experiment. If the players have played each other in the first few rounds, then they should have some influence on each other for the later rounds. Hence, we consider any players that played each other in the first five rounds of the game as being connected. We then use the remaining 15 rounds as our experimental data. So, for our panel experiment, we have $n = 110$ and $T = 15$. As Bojinov et al. [2021b] showed little evidence of carryover effects, we assume there is only population interference. Then, we use the exposure model in (2.11) and the temporal exposure contrast we are interested in is the exposure contrast between $(0, \leq 0.2)$ and $(1, \geq 0.8)$ for each time step. We now report the Horvitz-Thompson estimates of the temporal exposure contrast for the last 10 time steps, the estimates from the 2-step, and the 5-steps convex combination estimator estimates. Table 2.4 shows the results.

In general, we do not expect the temporal exposure contrasts to be different for different time steps since all the 15 rounds of games were done together in one session. And as we can see from the table, the convex combination estimator leads to estimates with much smaller variance. Note that the estimates from 2-step and 5-steps convex combination estimators are similar, illustrating

Time Step	$T=6$	$T=7$	$T=8$	$T=9$	$T=10$	$T=11$	$T=12$	$T=13$	$T=14$	$T=15$	Variance
Horvitz-Thompson	-8.08	11.03	29.39	-3.53	57.48	-3.76	10.29	23.40	-13.70	16.19	452.80
2-step	-6.84	9.48	25.40	-2.99	33.81	-2.98	9.83	22.10	-11.20	14.64	226.13
5-steps	-6.83	9.46	25.38	-2.98	33.74	-2.97	9.82	22.09	-11.17	14.63	225.37

Table 2.4: Estimates for temporal exposure contrasts from the panel experiment in Andreoni and Samuelson [2006]

that the choice of k is not crucial since the estimator itself takes care of it. Moreover, as we pointed out earlier, we would expect positive exposure contrast and the estimates from convex combination estimator are more reliable in the sense that it shrinks the estimates towards zero when the Horvitz-Thompson estimator gives a negative value (this is possible since we only have $n = 110$ subjects which is a small number).

2.5.2 Facebook network semi-synthetic experiment

We now describe a semi-synthetic experiment using the Swarthmore College social network from the Facebook 100 dataset [Traud et al., 2012]. All networks in this dataset are complete online friendship networks for one hundred colleges and universities collected from a single-day snapshot of Facebook in September 2005. The network we use is of size 1657 with 61049 edges. We use this network as the graph that describes population interference among units and generate an assignment vector using a Bernoulli design with a success probability of 1/2. We first show mean squared error reduction of using convex combination estimator to estimate temporal exposure contrast between $(0,0)$ and $(1,1)$ at $T = 20$. Let $\rho_{i,t} = \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} w_{j,t}$, we assume the following exposure mappings:

$$f_i(w_{1:n,t}) = (w_{i,t}, \tilde{\rho}_{i,t}), \quad \text{where } \tilde{\rho}_{i,t} = \begin{cases} 0 & \text{if } \rho_{i,t} \leq 0.3, \\ \rho_{i,t} & \text{if } 0.3 \leq \rho_{i,t} \leq 0.7, \\ 1 & \text{if } \rho_{i,t} > 0.7. \end{cases}$$

Now we make a panel experiment with $T = 20$. We generate the outcomes at each time step according to a linear model that is linear in $w_{i,t}$ and $\tilde{\rho}_{i,t}$ and add a time-varying component ϵ_t that is uniformly distributed on $[-0.5, 0.5]$. Table 2.5 shows the empirical bias, the variance, and the root mean squared errors (RMSE) of the estimates of the temporal exposure contrast at time step $T = 20$ by using Horvitz-Thompson estimator, 2-step convex combination estimator, and 5-step convex combination estimator. As expected, the convex combination estimator reduces the RMSE significantly. Though the biases seem large compared to the Horvitz-Thompson estimator, as we mentioned previously, we can also control the amount of bias we tolerate, which implicitly accounts for the time effect.

The maximal degree for the network is 577, which is far greater than the $\sqrt{n}$. To make Theorem 10

Estimator of TEC	Horvitz-Thompson	2-step CVX	5-step CVX
RMSE	50.48	3.57	3.31
Empirical bias	0.26	2.75	2.80
Empirical variance	2548.495	5.19	3.09

Table 2.5: RMSE for different estimators of temporal exposure contrast at $T = 20$

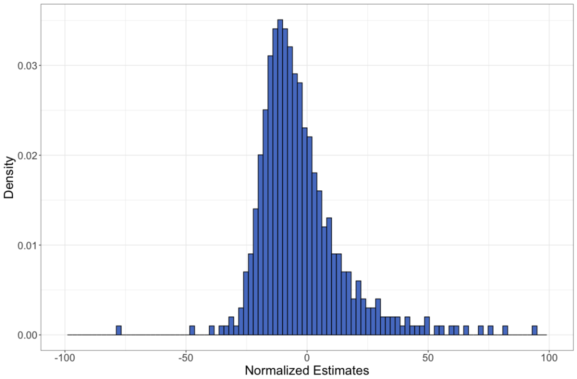


Figure 2.3: Histogram, $T = 20$

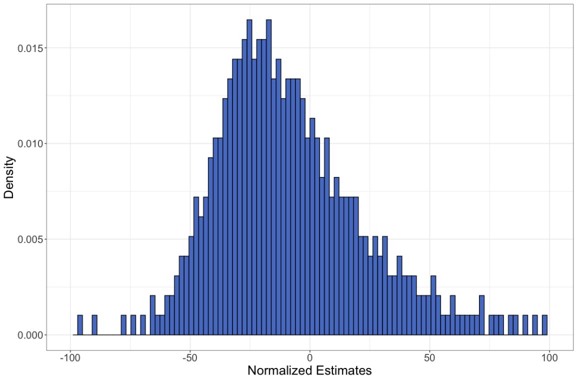


Figure 2.4: Histogram, $T = 100$

hold approximately for this network, we require having an extremely large T. Below we illustrate this empirically through a semi-synthetic experiment. Let

$$f_i(w_{1:n,t}) = (w_{i,t}, \tilde{\rho}_{i,t}), \quad \text{where } \tilde{\rho}_{i,t} = \begin{cases} 0 & \text{if } \rho_{i,t} \leq 0.35, \\ 1 & \text{if } 0.35 < \rho_{i,t} \leq 0.5, \\ 2 & \text{if } 0.5 < \rho_{i,t} \leq 0.65, \\ 3 & \text{if } \rho_{i,t} > 0.65. \end{cases}$$

We are interested in the average temporal exposure contrast between $(1, 3)$ and $(0, 0)$. Since the network is dense, with an average degree of 73.69, we expect the Horvitz-Thompson estimator to be inaccurate since units with exposure values $(1, 3)$ or $(0, 0)$ will unlikely to be those units with many neighbors. Figure 2.3 and 2.4 show the histograms of Horvitz-Thompson estimates for $T = 20$ and $T = 100$ respectively. Here, we calculate ATEC for 10,000 realizations, and since the computation of the variance estimate is time-consuming, we do not rescale the estimates.

Figure 2.3 shows that when $T = 20$ the histogram is far from normally distributed. Figure 2.4 shows that when $T = 100$, although the data are much closer to being normally distributed, they still are not. Also, note that the centers of these two histograms are away from 0; as we stated above, since some of the neighborhoods are extremely large, we cannot observe the exposure value we would need for units with large neighborhoods. This illustrates the necessity of condition (2.6) —

reliable inference requires more experiments if we have a dense network. We also report the coverage using a Gaussian confidence interval here. For both $T = 20$ and $T = 100$, the empirical coverage of naive Gaussian confidence interval is around 80%.

2.6 Conclusion

In this chapter, we have developed estimation and inference results for panel experiments with population interference. In the standard setting with pure population interference, we prove a central limit theorem under weaker conditions than previous results in the existing literature and highlight the trade-off between flexibility in the design and the interference structure. When population interference and carryover effects co-exist, we propose a novel central limit theorem. Finally, we introduce a new type of assumptions —stability assumptions — as an alternative to (or complement of) exposure mappings for controlling interference in temporal settings.

Many interesting avenues of investigation around interference in panel experiments have been left unexplored in this manuscript and will be the object of future work. First, our results only consider the Bernoulli design: this is, of course, limiting, but it does present a useful benchmark. We are particularly interested in exploring how to design panel experiments in the presence of population interference and carryover effects. Basse et al. [2019] study minimax designs with carryover effects, but the symmetries they exploit break under population interference, so new approaches are required. Second, while our simulations show that our convex combination estimators seem to behave well, our formal results under this new stability assumption are still limited. In particular, we plan to study the asymptotic properties of these estimators and provide a firmer theoretical grounding for their inferential properties. Third, explicit discussions on testing are not included. Though our results do provide a way to test certain hypotheses by inverting the confidence intervals and there has been literature [Bojinov et al., 2021b] that discuss how to conduct Fisher randomization test for the sharp null hypothesis in panel experiments, more testing results would be more beneficial to practitioners.

2.7 Appendix

2.7.1 Standard population interference

This appendix focuses on estimating TEC under population interference and assumes that either the experiment was conducted over a single time point or that there are no carryover effects. In both cases, we drop the subscript t for the remainder of the section. Our setup is now equivalent to the one studied in Liu and Hudgens [2014], Aronow and Samii [2017], Chin [2018] and Leung [2022].

Our Horvitz-Thompson type estimator $\hat{\tau}^{k,k'}$ now simplifies to,

$$\hat{\tau}^{k,k'} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbf{1}(H_i = k)}{\pi_i(k)}Y_i(k) - \frac{\mathbf{1}(H_i = k')}{\pi_i(k')}Y_i(k')\right], \tag{2.12}$$

where $\pi_i(k) = \mathbb{P}(H_i = k)$ and $\pi_i(k') = \mathbb{P}(H_i = k')$.

Aronow and Samii [2017] showed that if the potential outcomes and inverse exposure probabilities are bounded, and the number of dependent pairs of H_i's is of order $o(n^2)$, then the estimator $\hat{\tau}^{k,k'}$ is consistent,

$$\left(\hat{\tau}^{k,k'} - \tau^{k,k'}\right) \to_{\mathbb{P}} 0.$$

In addition, the authors provided an asymptotically conservative confidence interval of $\hat{\tau}^{k,k'}$ and implicitly outlined a version of a central limit theorem in the proof. However, the conditions stated in their derivations were sufficient but not necessary. Below, we establish a central limit theorem for $\hat{\tau}^{k,k'}$ under weaker conditions and provide a detailed proof that builds on recent results by Chin [2018]. We then illustrate the trade-offs between the strength of the interference structure assumption and the assignment mechanism's flexibility.

A central limit theorem

Our central limit theorem requires four additional assumptions. The first two assumptions bound the potential outcomes and the inverse probabilities of exposure.

Assumption 16 (Uniformly bounded potential outcomes). Assume that all the potential outcomes are uniformly bounded, i.e., $|Y_i(k)| \leq M$ for some M and for all i and k.

Assumption 17 (Overlap). Assume all the exposure probabilities are bounded away from 0 and 1, i.e., $\exists \eta > 0$ such that $\forall k$ and i, $0 < \eta \leq \pi_i(k) \leq 1 - \eta < 1$.

Assumptions 16 and 17 are standard in the causal inference literature (Aronow and Samii [2017]; Leung [2022]). Assumption 16 holds in most practical applications as realizations of the outcome variables are almost always bounded. Assumption 17 is necessary as vanishing exposure probabilities make the causal question ill-defined as we cannot observe the associated potential outcomes.

The next assumption rules out the existence of a pathological subsequence n_k along which the limiting variance of our estimator is zero.

Assumption 18 (Nondegenerate asymptotic variance). Assume that $\liminf_{n\to\infty} \mathrm{Var}(\sqrt{n}\hat{\tau}_t^{k,k'}) > 0$ for any t.

As a consequence of this assumption, for each t, there exists a constant $c > 0$ such that $\mathrm{Var}(\sqrt{n}\hat{\tau}_t^{k,k'}) \geq c$ for all sufficiently large n. This type of assumption seems unavoidable, even in settings without interference (see, e.g., Corollary 1 in Guo and Basse [2021], and subsequent discussion).

The fourth assumption quantifies the dependence among observations due to interference; to define it, we require a notion of the dependency graph for a collection of random variables. We define the dependency graph G_n for $H_1, \cdots, H_n$ to be the graph with vertices $V = \{1, \cdots, n\}$ and edges E such that $(i, j) \in E$ if and only if H_i and H_j are not independent. The graph G_n models the dependency relationship among n random variables $H_1, \cdots, H_n$. Let d_n be the maximal degree in this graph, which is equal to the maximal number of dependent exposure values for each unit. Notice that the dependency graph depends both on the interference structure and on the assignment mechanism.

We can now state the following central limit theorem for temporal exposure contrast.

Theorem 19. *Under Assumptions 16-18 and the condition that $d_n = o(n^{1/4})$, we have*

$$\frac{\sqrt{n}(\hat{\tau}^{k,k^{'}} - \tau^{k,k^{'}})}{Var(\sqrt{n}\hat{\tau}^{k,k'})^{1/2}} \xrightarrow{d} \mathcal{N}(0,1)$$

as $n \to \infty$.

Theorem 19 strengthens the result of Aronow and Samii [2017] in two ways. First, our Assumption 18 weakens Condition 6 of Aronow and Samii [2017], which requires the convergence of $\text{Var}(\sqrt{n}\hat{\tau}_t^{k,k^{'}})$. Second, we allow for a higher range of dependence ($d_n = o(n^{1/4})$ compared to $d_n = O(1)$ as in Aronow and Samii [2017]) among exposure values. The proof of this theorem relies on recent results in Chin [2018].

Design and interference structure: a trade-off

Intuitively, Theorem 19 asserts that asymptotic normality holds so long as the dependency relations among the H_i's are moderate. However, since $H_i = f_i(W_{1:n})$ is determined by both function f_i and assignment W, the dependence structure among the H_i's — and therefore the value of d_n — depends on both the exposure specification and the assignment mechanism.

This suggests that there exists a trade-off between the strength of the dependence in the W_i's induced by the assignment mechanism and the dependence induced by the interference structure. The less restricted the interference structure is, the more restricted the assignment mechanism must be; in reverse, the more restricted the interference structure, the more flexible one can be with the design. We illustrate these insights with three special cases of Theorem 19, applied to popular settings. We should also note that our condition on d_n is not a sufficient condition for the central limit theorem. For example, if we consider $f_i(W_i) = W_i$ (i.e., there is no interference) and W follows completely randomized design, then the central limit theorem still holds (see Theorem 1 in Ding [2017]). The discussion here mainly illustrates the entanglement between the assignment mechanism and the interference structure from a general perspective.

Example 2. Suppose that the interference structure among n units is adequately described by a social network $\mathcal{A}_n$, and assume that the exposure mapping is of the form $f_i(W_{1:n}) = f_i(W_{\mathcal{N}_i})$; that is, only the neighbors' assignments matter. Let δ_n be the maximal number of neighbors a unit can have in the network $\mathcal{A}_n$ — which is distinct from the dependency graph. Then if $\delta_n = o(n^{1/8})$ and the W_i's are independent (i.e., the design is Bernoulli), then $d_n = o(n^{1/4})$ as required by Theorem 19.

This first example explores one extreme end of the trade-off, in which the assignment mechanism is maximally restricted — the W_i's are independent — which allows for a comparatively large amount of interference.

Example 3. We consider the graph cluster randomization approach (Ugander et al. [2013]) in which case we group units into clusters and randomize at the cluster level. Following the notations in Ugander et al. [2013], we let the vertices be partitioned into n_c clusters $C_1, \cdots, C_{n_c}$. The graph cluster randomization approach assigns either treatment or control to all the units in each cluster. Suppose one's potential outcomes depend only on the assignments of its neighbors. Let δ_n be the maximal number of neighbors one can have and c_n be the maximal size of the cluster. Then $d_n = o(n^{1/4})$ for $\delta_n^2 + \delta_n c_n = o(n^{1/4})$.

Example 4. Another commonly studied scenario is the "household" interference (Basse and Feller [2018]; Duflo and Saez [2003]). In household interference, we assume that each unit belongs to a "household" and their potential outcomes depend only on the assignments of the units within the "household". Suppose we have a two-stage design such that we first assign each household into treatment group or control group independently and then we assign treatments to units in each household depending on the assignment of their associated household. Let r_n be the maximal size of the "household", then $d_n = o(n^{1/4})$ for $r_n = o(n^{1/4})$.

Table 2.6 summarizes the above three examples. In Example 2, to have a general network interference setting with the maximum possible number of neighbors for each unit, we constrain the design to be the Bernoulli design. Further limiting the interference, like in Example 4 where the interference is restricted within households, we can have a more complex two-stage design. In the same spirit, Example 3 shows that for a highly dependent design, we need an even stronger condition on the interference structure, indicated by a stronger rate condition on δ_n. In general, a weaker assumption on the interference structure induces a more complex dependence graph for the exposures, which in turn reduces our flexibility in the choice of design.

Interference	Design	Conditions
Network Interference	Bernoulli Design	$\delta_n = o(n^{1/8})$
Network Interference	Graph Cluster Randomization	$\delta_n^2 + \delta_n c_n = o(n^{1/4})$
Group Interference	Two-stage Design	$r_n = o(n^{1/4})$

Table 2.6: Trade-off between design and interference

Inference

The central limit theorem stated in Theorem 19 serves as our basis for inference.

Proposition 9. *Assuming all the assumptions in Theorem 19, then for any $\delta > 0$,*

$$\mathbb{P}\left(\frac{\widehat{Var}(\hat{\tau}^{k,k'})}{Var(\hat{\tau}^{k,k'})} \geq 1-\delta\right) \to 1.$$

where $\widehat{Var}(\hat{\tau}^{k,k'}) = n^{-1}\widehat{Var}(\sqrt{n}\hat{\tau}^{k,k'})$. Therefore, we can construct asymptotically conservative confidence intervals based on the variance estimator: for any $\delta > 0$,

$$\mathbb{P}\Bigg(\tau^{k,k'} \in \Bigg[\hat{\tau}^{k,k'} - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{1-\delta}}\sqrt{\widehat{Var}(\hat{\tau}^{k,k'})}, \\ \hat{\tau}^{k,k'} + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{1-\delta}}\sqrt{\widehat{Var}(\hat{\tau}^{k,k'})}\Bigg]\Bigg) \geq 1-\alpha$$

for large n.

$\widehat{\mathrm{Var}}(\sqrt{n}\hat{\tau}^{k,k'})$ is the same as the one given in Proposition 2. Once again, this result strengthens that of Aronow and Samii [2017] by both removing the requirement that $n\mathrm{Var}(\hat{\tau}^{k,k'})$ converge, and by relaxing the constraint on the interference mechanism. Note that here $\delta > 0$ is arbitrary and we present detailed simulations in Section 2.4 with $\delta = 0.04$.

2.7.2 Proofs and additional discussions

To begin with, we provide technical tools that we will use in our proofs. We first state a lemma from Ross [2011]:

Lemma 1. *Let $X_1, \cdots, X_n$ be a collection of random variables such that $\mathbb{E}\left[X_i^4\right] < \infty$ and $\mathbb{E}\left[X_i\right] = 0$. 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 $(X_1, \cdots, X_n)$. Then for constants C_1 and C_2 which do not depend on n, d or σ^2,*

$$d_{\mathcal{W}}(S/\sigma) \leq C_1 \frac{d^{3/2}}{\sigma^2}\left(\sum_{i=1}^n \mathbb{E}\left[X_i^4\right]\right)^{1/2} + C_2 \frac{d^2}{\sigma^3}\sum_{i=1}^n \mathbb{E}|X_i|^3, \tag{2.13}$$

where $d_{\mathcal{W}}(S/\sigma)$ is the Wasserstein distance between S/σ and standard Gaussian.

Second, we provide the expression for the variance of $\hat{\tau}^{k,k'}$:

Lemma 2 (Variance of Horvitz-Thompson estimator)**.** *We have that (Aronow and Samii [2017]):*

$$
\begin{aligned}
Var(\sqrt{n}\hat{\tau}^{k,k^{'}}) = \frac{1}{n}\sum_{i=1}^{n}\pi_i(k)(1-\pi_i(k))\left(\frac{Y_i(k)}{\pi_i(k)}\right)^2 \\
+\frac{1}{n}\sum_{i=1}^{n}\pi_i(k^{'})(1-\pi_i(k^{'}))\left(\frac{Y_i(k^{'})}{\pi_i(k^{'})}\right)^2 + \frac{2}{n}\sum_{i=1}^{n}Y_i(k)Y_i(k^{'}) \\
+\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\Bigg\{[\pi_{ij}(k)-\pi_i(k)\pi_j(k)]\frac{Y_i(k)}{\pi_i(k)}\frac{Y_j(k)}{\pi_j(k)} \\
+\left[\pi_{ij}(k^{'})-\pi_i(k^{'})\pi_j(k^{'})\right]\frac{Y_i(k^{'})}{\pi_i(k^{'})}\frac{Y_j(k^{'})}{\pi_j(k^{'})}\Bigg\} \\
-\frac{2}{n}\sum_{i=1}^{n}\sum_{j\neq i}\left\{\left[\pi_{ij}(k,k^{'})-\pi_i(k)\pi_j(k^{'})\right]\frac{Y_i(k)}{\pi_i(k)}\frac{Y_j(k^{'})}{\pi_j(k^{'})}\right\}
\end{aligned}
$$

Here $\pi_{ij}(k) = \mathbb{P}(H_i = k \text{ and } H_j = k)$

Proof of Theorem 19. Note that $\hat{\tau}^{k,k^{'}} = \sum_{i=1}^{n}\tilde{\tau}_i$ where

$$\tilde{\tau}_i = \frac{1}{n}\left[\frac{\mathbf{1}(H_i = k)}{\pi_i(k)}Y_i(k) - \frac{\mathbf{1}(H_i = k^{'})}{\pi_i(k^{'})}Y_i(k^{'})\right]$$

and $\mathbb{E}[\tilde{\tau}_i] = \frac{1}{n}\left[Y_i(k) - Y_i(k^{'})\right]$, hence if we let $X_i = \sqrt{n}(\tilde{\tau}_i - \mathbb{E}[\tilde{\tau}_i])$, then $\sqrt{n}(\hat{\tau}^{k,k^{'}} - \tau^{k,k^{'}}) = \sum_{i=1}^{n} X_i = S$. By Assumption 16 and Assumption 17, we know that $X_i = O_p(n^{-1/2})$, hence there exist some constants C_1 and C_2 such that for sufficiently large n, both

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

and

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

hold. Moreover, by Assumption 18,

$$\sigma^2 = \text{Var}(\sum_i X_i) = n\text{Var}(\hat{\tau}^{k,k^{'}})$$

is bounded away from 0. Note that X_i is a function of H_i, hence X_i and X_j are not independent if and only if H_i and H_j are not independent. Since $d_n = o(n^{1/4})$, we know that the maximal degree of the dependency graph of X_i's is $o(n^{1/4})$. Now we apply Lemma 1. Since σ^2 is bounded away from 0, we get:

$$\text{RHS of (2.13)} = o(n^{-1/8}) + o(1) \to 0$$

We're done. □

Remark 3. In fact, with the tools in Leung [2022], we can prove this theorem with a weaker condition on d_n: $d_n = O(\log n)$.

Proof of Example 2. Note that H_i is a function of W_i and W_j's for j being a neighbor of i. If H_i and H_j are dependent, there must be the case that $(\{i\} \cup \mathcal{N}_i) \cap (\{j\} \cup \mathcal{N}_j)$ is nonempty since we have the Bernoulli design. Hence, for each fixed unit i, there are at most δ_n units such that the above intersection is nonempty. □

Proof of Example 4. We use the same reasoning as in the above proof. The only change is that now we know that each unit is belonged to a group and units in the group are connected. Therefore, for each fixed unit i, all the units outside the group will not have effect on unit i. As a result, we can have $r_n = o(n^{1/4})$. □

Proof of Example 3. Since we do not have Bernoulli design anymore, there might be the case that W_i and W_j are dependent, hence except $(\{i\} \cup \mathcal{N}_i) \cap (\{j\} \cup \mathcal{N}_j)$ is nonempty, there is another case that makes H_i and H_j dependent: a neighbor of i is in the same cluster as a neighbor of j. For this case, we have at most $\delta_n c_n$ such j's for a fixed unit i. Hence, in total, there are at most $\delta_n^2 + \delta_n c_n$ j's such that H_i and H_j are dependent. □

Proof of Proposition 9. We first prove the first part of the proposition. The proof is based on A.7 in Aronow and Samii [2017]. To start with, for any $(i, j) \in \{1, \cdots, \} \times \{1, \cdots, n\}$, we define $e_{ij} = 1$ if H_i and H_j are dependent and 0 otherwise. Let $a_{ij}(H_i, H_j)$ be the sum of the elements in $\widehat{\text{Var}}(\hat{\tau}^{k,k'})$ that incorporate i and j, then

$$\text{Var}\left(\widehat{\text{Var}}(\hat{\tau}^{k,k'})\right) \leq n^{-4}\text{Var}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} e_{ij}a_{ij}(H_i, H_j)\right]$$
$$= n^{-4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \text{Cov}\left[e_{ij}a_{ij}(H_i, H_j), e_{kl}a_{kl}(H_k, H_l)\right]$$

Note that $\text{Cov}\left[e_{ij}a_{ij}(H_i, H_j), e_{kl}a_{kl}(H_k, H_l)\right]$ is nonzero if and only if $e_{ij} = 1, e_{kl} = 1$ and at least one of $e_{ik}, e_{il}, e_{jk}, e_{jl}$ is 1. In total, there are at most $4nd_n^3$ (i, j, k, l)'s satisfying this condition. And by Assumption 16 and 17, each covariance term is bounded, so we know that $\text{Var}\left(\widehat{\text{Var}}(\hat{\tau}^{k,k'})\right) = o(n^{-4} \times n \times n^{3/4}) \to 0$ as $n \to \infty$. Then by Chebyshev's inequality,

$$\left|\widehat{\text{Var}}(\sqrt{n}\hat{\tau}^{k,k'}) - \mathbb{E}\left[\widehat{\text{Var}}(\sqrt{n}\hat{\tau}^{k,k'})\right]\right| = o_p(1).$$

Since $\mathbb{E}\left[\widehat{\mathrm{Var}}(\hat{\tau}^{k,k'})\right] \geq \mathrm{Var}(\hat{\tau}^{k,k'})$,

$$\mathbb{P}\left(\frac{\widehat{\mathrm{Var}}(\hat{\tau}^{k,k'})}{\mathrm{Var}(\hat{\tau}^{k,k'})} \geq 1-\delta\right) \to 1$$

for any $\delta > 0$.

Now we can prove the second part of the proposition. We have that

$$\begin{aligned}
LHS &= \mathbb{P}\left(\left|\frac{\sqrt{n}(\hat{\tau}^{k,k'} - \tau^{k,k'})}{\sqrt{\mathrm{Var}(\sqrt{n}\hat{\tau}^{k,k'})}}\right| \leq \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{1-\delta}}\sqrt{\frac{\widehat{\mathrm{Var}}(\sqrt{n}\hat{\tau}^{k,k'})}{\mathrm{Var}(\sqrt{n}\hat{\tau}^{k,k'})}}\right) \\
&\geq \mathbb{P}\left(\left|\frac{\sqrt{n}(\hat{\tau}^{k,k'} - \tau^{k,k'})}{\sqrt{\mathrm{Var}(\sqrt{n}\hat{\tau}^{k,k'})}}\right| \leq \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{1-\delta}}\sqrt{\frac{\widehat{\mathrm{Var}}(\sqrt{n}\hat{\tau}^{k,k'})}{\mathrm{Var}(\sqrt{n}\hat{\tau}^{k,k'})}} \text{ and } \frac{\widehat{\mathrm{Var}}(\sqrt{n}\hat{\tau}^{k,k'})}{\mathrm{Var}(\sqrt{n}\hat{\tau}^{k,k'})} \geq 1-\delta\right) \\
&\geq \mathbb{P}\left(\left|\frac{\sqrt{n}(\hat{\tau}^{k,k'} - \tau^{k,k'})}{\sqrt{\mathrm{Var}(\sqrt{n}\hat{\tau}^{k,k'})}}\right| \leq z_{1-\frac{\alpha}{2}} \text{ and } \frac{\widehat{\mathrm{Var}}(\sqrt{n}\hat{\tau}^{k,k'})}{\mathrm{Var}(\sqrt{n}\hat{\tau}^{k,k'})} \geq 1-\delta\right) \\
&= \mathbb{P}\left(\left|\frac{\sqrt{n}(\hat{\tau}^{k,k'} - \tau^{k,k'})}{\sqrt{\mathrm{Var}(\sqrt{n}\hat{\tau}^{k,k'})}}\right| \leq z_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(\left|\frac{\sqrt{n}(\hat{\tau}^{k,k'} - \tau^{k,k'})}{\sqrt{\mathrm{Var}(\sqrt{n}\hat{\tau}^{k,k'})}}\right| \leq z_{1-\frac{\alpha}{2}} \text{ and } \frac{\widehat{\mathrm{Var}}(\sqrt{n}\hat{\tau}^{k,k'})}{\mathrm{Var}(\sqrt{n}\hat{\tau}^{k,k'})} < 1-\delta\right)
\end{aligned} \tag{2.14}$$

Now,

$$\begin{aligned}
(2.14) &\geq \mathbb{P}\left(\left|\frac{\sqrt{n}(\hat{\tau}^{k,k'} - \tau^{k,k'})}{\sqrt{\mathrm{Var}(\sqrt{n}\hat{\tau}^{k,k'})}}\right| \leq z_{1-\frac{\alpha}{2}}\right) \\
&\quad - \mathbb{P}\left(\frac{\widehat{\mathrm{Var}}(\sqrt{n}\hat{\tau}^{k,k'})}{\mathrm{Var}(\sqrt{n}\hat{\tau}^{k,k'})} < 1-\delta\right) \\
&= \mathbb{P}\left(\left|\frac{\sqrt{n}(\hat{\tau}^{k,k'} - \tau^{k,k'})}{\sqrt{\mathrm{Var}(\sqrt{n}\hat{\tau}^{k,k'})}}\right| \leq z_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(\frac{\widehat{\mathrm{Var}}(\hat{\tau}^{k,k'})}{\mathrm{Var}(\hat{\tau}^{k,k'})} < 1-\delta\right) \\
&\to 1-\alpha
\end{aligned}$$

as $n \to \infty$ by the first part and Theorem 19. □

Proof of Theorem 8. We use a characteristic function argument. We first note that

$$\begin{aligned}
\frac{\sqrt{nT}(\hat{\bar{\tau}}^{k,k^{'}}-\bar{\tau}^{k,k^{'}})}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\sigma_{n,t}^{2}}} &= \frac{\sqrt{nT}(\frac{1}{T}\sum_{t=1}^{T}\hat{\tau}_t^{k,k^{'}}-\frac{1}{T}\sum_{t=1}^{T}\tau_t^{k,k^{'}})}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\sigma_{n,t}^{2}}} \\
&= \frac{\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\sqrt{n}(\hat{\tau}_t^{k,k^{'}}-\tau_t^{k,k^{'}})}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\sigma_{n,t}^{2}}} \\
&= \frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}X_{n,t}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\sigma_{n,t}^{2}}},
\end{aligned}$$

where $X_{n,t}=\sqrt{n}(\hat{\tau}_t^{k,k^{'}}-\tau_t^{k,k^{'}})$. Now,

$$\mathbb{E}\left[\exp\left\{i\lambda\frac{\sqrt{nT}(\hat{\bar{\tau}}^{k,k^{'}}-\bar{\tau}^{k,k^{'}})}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\sigma_{n,t}^{2}}}\right\}\right] \tag{2.15}$$

$$\begin{aligned}
&= \mathbb{E}\left[\exp\left\{i\lambda\frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}X_{n,t}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\sigma_{n,t}^{2}}}\right\}\right] \\
&= \prod_{t=1}^{T}\mathbb{E}\left[\exp\left\{i\lambda\frac{\frac{1}{\sqrt{T}}X_{n,t}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\sigma_{n,t}^{2}}}\right\}\right] \\
&= \prod_{t=1}^{T}\mathbb{E}\left[\exp\left\{i\frac{\lambda\sigma_{n,t}}{\sqrt{\sum_{t=1}^{T}\sigma_{n,t}^{2}}}\frac{X_{n,t}}{\sigma_{n,t}}\right\}\right]
\end{aligned}$$

$$= \prod_{t=1}^{T}\phi_{\frac{X_{n,t}}{\sigma_{n,t}}}\left(\frac{\lambda\sigma_{n,t}}{\sqrt{\sum_{t=1}^{T}\sigma_{n,t}^{2}}}\right) \tag{2.16}$$

The second equality follows from our assumption that assignment vectors are independent across time and ϕ_X denotes the characteristic function of a random variable X. Pick $\epsilon > 0$. Now, to conclude the proof, we note that

$$\phi_{\frac{X_{n,t}}{\sigma_{n,t}}}(\theta)\to e^{-\frac{\theta^2}{2}}$$

for any $t\in\{1,\cdots,T\}$. Moreover, for each t, the convergence is actually uniform on any bounded interval. Therefore, for any $t\in\{1,\cdots,T\}$,

$$\phi_{\frac{X_{n,t}}{\sigma_{n,t}}}(\theta)\to e^{-\frac{\theta^2}{2}} \text{ uniformly on } (0,1).$$

Note that

$$\frac{\lambda\sigma_{n,t}}{\sqrt{\sum_{t=1}^{T}\sigma_{n,t}^2}} \in (0,1),$$

so for any t, $\exists N_t \in \mathbb{N}$ such that for any $n \geq N_t$,

$$\begin{aligned}\left|\phi_{\frac{X_{n,t}}{\sigma_{n,t}}}\left(\frac{\lambda\sigma_{n,t}}{\sqrt{\sum_{t=1}^{T}\sigma_{n,t}^2}}\right) - \exp\left\{-\frac{1}{2}\frac{\lambda^2\sigma_{n,t}^2}{\sum_{t=1}^{T}\sigma_{n,t}^2}\right\}\right| &= |\epsilon_t| \\ &\leq \frac{1}{2^K}.\end{aligned}$$

Let $N = \max\{N_1, \cdots, N_T\}$, then for all $n \geq N$, and for all $t \in \{1, \cdots, T\}$,

$$\left|\phi_{\frac{X_{n,t}}{\sigma_{n,t}}}\left(\frac{\lambda\sigma_{n,t}}{\sqrt{\sum_{t=1}^{T}\sigma_{n,t}^2}}\right) - \exp\left\{-\frac{1}{2}\frac{\lambda^2\sigma_{n,t}^2}{\sum_{t=1}^{T}\sigma_{n,t}^2}\right\}\right| = |\epsilon_t| \leq \frac{1}{2^K},$$

where K is any big number we want. Now,

$$\begin{aligned}(2.16) &= \prod_{t=1}^{T}\left(\exp\left\{-\frac{1}{2}\frac{\lambda^2\sigma_{n,t}^2}{\sum_{t=1}^{T}\sigma_{n,t}^2}\right\} + \epsilon_t\right) \\ &= \exp\left\{-\frac{1}{2}\lambda^2\right\} + R(\epsilon_t),\end{aligned}$$

where $R(\epsilon_t)$ is a remainder term that is the sum of several monomial terms of ϵ_t's. Note that $\exp\left\{-\frac{1}{2}\frac{\lambda^2\sigma_{n,t}^2}{\sum_{t=1}^{T}\sigma_{n,t}^2}\right\}$ is actually bounded by 1, hence by making K sufficiently large, we can make $R(\epsilon_t)$ arbitrarily small. Pick such K, then we know that for sufficiently large n,

$$\left|\mathbb{E}\left[\exp\left\{i\lambda\frac{\sqrt{nT}(\hat{\bar{\tau}}^{k,k'} - \bar{\tau}^{k,k'})}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\sigma_{n,t}^2}}\right\}\right] - \exp\left\{-\frac{1}{2}\lambda^2\right\}\right| \leq \epsilon.$$

Hence, by standard characteristic function argument, we complete the proof of the theorem. □

To prove Theorem 10, we first state the following version of Lindeberg-Feller central limit theorem.

Lemma 3 (Lindeberg-Feller CLT). *Let $\{k_n\}_{n\geq 1}$ be a sequence of positive integers increasing to infinity. For each n, let $\{X_{n,i}\}_{1\leq i\leq k_n}$ is a collection of independent random variables. Let $\mu_{n,i} := \mathbb{E}(X_{n,i})$ and*

$$s_n^2 := \sum_{i=1}^{k_n} Var(X_{n,i}).$$

Suppose that for any $\epsilon > 0$*,*

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{k_n} \mathbb{E}\left((X_{n,i}-\mu_{n,i})^2; |X_{n,i}-\mu_{n,i}| \geq \epsilon s_n\right) = 0. \tag{2.17}$$

Then the random variable

$$\frac{\sum_{i=1}^{k_n}(X_{n,i}-\mu_{n,i})}{s_n} \xrightarrow{d} \mathcal{N}(0,1)$$

as $n \to \infty$*.*

Proof of Theorem 10. We first prove the theorem with condition (2.5). We note that $\sqrt{nT}(\hat{\bar{\tau}}^{k,k'} - \bar{\tau}^{k,k'}) = \sum_{t=1}^{T} \sqrt{\frac{n}{T}}(\hat{\tau}_t^{k,k'} - \tau_t^{k,k'})$. Let $X_{n,t} = \sqrt{\frac{n}{T}}\hat{\tau}_t^{k,k'}$, then $\mu_{n,t} = \sqrt{\frac{n}{T}}\tau_t^{k,k'}$, so the numerator is exactly $\sum_{t=1}^{T}(X_{n,t}-\mu_{n,t})$. Moreover, note that for any n, $X_{n,1}, \cdots, X_{n,T}$ are independent by the pure population interference assumption. Now,

$$\begin{aligned}
s_n^2 &= \sum_{t=1}^{T} \mathrm{Var}(X_{n,t}) \\
&= \sum_{t=1}^{T} \mathrm{Var}\left(\sqrt{\frac{n}{T}}\hat{\tau}_t^{k,k'}\right) \\
&= \frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\sqrt{n}\hat{\tau}_t^{k,k'}) \\
&= \frac{1}{T}\sum_{t=1}^{T} \sigma_{n,t}^2.
\end{aligned}$$

Hence, to finish the proof, we only need to check (2.17) is satisfied. Notice that for any $\epsilon > 0$,

$$\begin{aligned}
|X_{n,t}-\mu_{n,t}| \geq \epsilon s_n &\Leftrightarrow \left|\sqrt{\frac{n}{T}}\hat{\tau}_t^{k,k'} - \sqrt{\frac{n}{T}}\tau_t^{k,k'}\right| \geq \epsilon\sqrt{\frac{1}{T}\sum_{t=1}^{T}\sigma_{n,t}^2} \\
&\Leftrightarrow \left|\hat{\tau}_t^{k,k'} - \tau_t^{k,k'}\right| \geq \epsilon\sqrt{\frac{1}{n}\sum_{t=1}^{T}\sigma_{n,t}^2}
\end{aligned}$$

By Assumption 18, $\sigma_{n,t}^2 \geq c$ for some $c > 0$ and for all n large. Hence

$$\epsilon\sqrt{\frac{1}{n}\sum_{t=1}^{T}\sigma_{n,t}^2} \geq \epsilon\sqrt{\frac{T}{n}c} \to \infty.$$

Note that by Assumptions 16 and 17, $\left|\hat{\tau}_t^{k,k'} - \tau_t^{k,k'}\right|$ is uniformly bounded. Hence for sufficiently large n, $\left|\hat{\tau}_t^{k,k'} - \tau_t^{k,k'}\right| < \epsilon\sqrt{\frac{1}{n}\sum_{t=1}^T \sigma_{n,t}^2}$ for all t. Therefore, for sufficiently large n,

$$\frac{1}{s_n^2}\sum_{t=1}^T \mathbb{E}\left((X_{n,t}-\mu_{n,t})^2; |X_{n,t}-\mu_{n,t}| \geq \epsilon s_n\right) = 0.$$

As a result, (2.17) is satisfied. We're done. The proof of this theorem with condition (2.6) is exactly the same as in single time step case once we notice that the numerator is just a sum of nT mean 0 dependent random variables. □

To prove Theorem 9, we need the following version of Lyapunov central limit theorem.

Lemma 4 (Lyapunov CLT)**.** *Let $\{X_n\}_{n=1}^\infty$ be a sequence of independent random variables. Let $\mu_i := \mathbb{E}(X_i)$ and*

$$s_n^2 = \sum_{i=1}^n Var(X_i) > 0.$$

If for some $\delta > 0$,

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}|X_i - \mu_i|^{2+\delta} = 0, \tag{2.18}$$

then the random variable

$$\frac{\sum_{i=1}^n (X_i - \mu_i)}{s_n} \xrightarrow{d} \mathcal{N}(0,1)$$

Proof of Theorem 9. This time, we let $X_t = \sqrt{\frac{n}{T}}\hat{\tau}_t^{k,k'}$ then the numerator is $\sum_{t=1}^T (X_t - \mu_t)$. Since we have pure population interference, $\{X_t\}_{t=1}^\infty$ are independent. Now,

$$\begin{aligned} s_T^2 &= \sum_{t=1}^T \mathrm{Var}(X_t) \\ &= \frac{1}{T}\sum_{t=1}^T \sigma_{n,t}^2. \end{aligned}$$

Hence, we only need to check (2.18). We have that

$$\begin{aligned} &\lim_{T\to\infty} \frac{1}{s_T^{2+\delta}} \sum_{t=1}^T \mathbb{E}|X_t - \mu_t|^{2+\delta} \\ &= \lim_{T\to\infty} \frac{1}{s_T^{2+\delta}} \left(\frac{n}{T}\right)^{1+\frac{\delta}{2}} \sum_{t=1}^T \mathbb{E}\left|\hat{\tau}_t^{k,k'} - \tau_t^{k,k'}\right|^{2+\delta} \end{aligned}$$

Now, by Assumptions 16 and 17, $\exists M > 0$ such that $\left|\hat{\tau}_t^{k,k'} - \tau_t^{k,k'}\right| \leq M$ for all t. Hence,

$$\begin{aligned}
&\frac{1}{s_T^{2+\delta}} \left(\frac{n}{T}\right)^{1+\frac{\delta}{2}} \sum_{t=1}^{T} \mathbb{E}\left|\hat{\tau}_t^{k,k'} - \tau_t^{k,k'}\right|^{2+\delta} \\
&\leq \frac{1}{s_T^{2+\delta}} \left(\frac{n}{T}\right)^{1+\frac{\delta}{2}} T M^{2+\delta} \\
&= \frac{1}{s_T^{2+\delta}} \frac{n^{1+\frac{\delta}{2}}}{T} M^{2+\delta}
\end{aligned}$$

If $T \to \infty$, $\frac{1}{s_T^{2+\delta}} \frac{n^{1+\frac{\delta}{2}}}{T} M^{2+\delta} \to 0$. Therefore, (2.18) is satisfied. We're done. □

Proof of Proposition 2. Now we can prove the second part of the proposition. We have that

$$\begin{aligned}
LHS &= \mathbb{P}\left(\left|\frac{\sqrt{nT}(\hat{\bar{\tau}}^{k,k'} - \bar{\tau}^{k,k'})}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\sqrt{n}\hat{\tau}_t^{k,k'})}}\right| \leq \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{1-\delta}} \sqrt{\frac{\frac{1}{T}\sum_{t=1}^{T} \widehat{\mathrm{Var}}(\sqrt{n}\hat{\tau}_t^{k,k'})}{\frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\sqrt{n}\hat{\tau}_t^{k,k'})}}\right) \\
&\geq \mathbb{P}\left(\left|\frac{\sqrt{nT}(\hat{\bar{\tau}}^{k,k'} - \bar{\tau}^{k,k'})}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\sqrt{n}\hat{\tau}_t^{k,k'})}}\right| \leq \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{1-\delta}} \sqrt{\frac{\frac{1}{T}\sum_{t=1}^{T} \widehat{\mathrm{Var}}(\sqrt{n}\hat{\tau}_t^{k,k'})}{\frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\sqrt{n}\hat{\tau}_t^{k,k'})}} \text{ and } \frac{\frac{1}{T}\sum_{t=1}^{T} \widehat{\mathrm{Var}}(\hat{\tau}_t^{k,k'})}{\frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\hat{\tau}_t^{k,k'})} \geq 1-\delta\right) \\
&\geq \mathbb{P}\left(\left|\frac{\sqrt{nT}(\hat{\bar{\tau}}^{k,k'} - \bar{\tau}^{k,k'})}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\sqrt{n}\hat{\tau}_t^{k,k'})}}\right| \leq z_{1-\frac{\alpha}{2}} \text{ and } \frac{\frac{1}{T}\sum_{t=1}^{T} \widehat{\mathrm{Var}}(\hat{\tau}_t^{k,k'})}{\frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\hat{\tau}_t^{k,k'})} \geq 1-\delta\right) \qquad (2.19)
\end{aligned}$$

Furthermore,

$$\begin{aligned}
(2.19) &= \mathbb{P}\left(\left|\frac{\sqrt{nT}(\hat{\bar{\tau}}^{k,k'} - \bar{\tau}^{k,k'})}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\sqrt{n}\hat{\tau}_t^{k,k'})}}\right| \leq z_{1-\frac{\alpha}{2}}\right) \\
&\quad - \mathbb{P}\left(\left|\frac{\sqrt{nT}(\hat{\bar{\tau}}^{k,k'} - \bar{\tau}^{k,k'})}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\sqrt{n}\hat{\tau}_t^{k,k'})}}\right| \leq z_{1-\frac{\alpha}{2}} \text{ and } \frac{\frac{1}{T}\sum_{t=1}^{T} \widehat{\mathrm{Var}}(\hat{\tau}_t^{k,k'})}{\frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\hat{\tau}_t^{k,k'})} < 1-\delta\right) \\
&\geq \mathbb{P}\left(\left|\frac{\sqrt{nT}(\hat{\bar{\tau}}^{k,k'} - \bar{\tau}^{k,k'})}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\sqrt{n}\hat{\tau}_t^{k,k'})}}\right| \leq z_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(\frac{\frac{1}{T}\sum_{t=1}^{T} \widehat{\mathrm{Var}}(\hat{\tau}_t^{k,k'})}{\frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\hat{\tau}_t^{k,k'})} < 1-\delta\right)
\end{aligned}$$

So if we can show $\mathbb{P}\left(\frac{\frac{1}{T}\sum_{t=1}^{T} \widehat{\mathrm{Var}}(\hat{\tau}_t^{k,k'})}{\frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\hat{\tau}_t^{k,k'})} < 1-\delta\right) \to 0$ then we are done. Notice that

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} \widehat{\mathrm{Var}}(\hat{\tau}_t^{k,k'})\right) = \frac{1}{T^2}\sum_{t=1}^{T} \mathrm{Var}\left(\widehat{\mathrm{Var}}(\hat{\tau}_t^{k,k'})\right). \qquad (2.20)$$

If T is fixed (i.e., Theorem 8 holds), then by what we have in Proposition 9, we immediately have that (2.20) → 0 and we are done. Now suppose Theorem 10 holds. Recall that

$$\operatorname{Var}\left(\widehat{\operatorname{Var}}(\hat{\tau}^{k,k'})\right) \leq n^{-4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\operatorname{Cov}\left[e_{ij}a_{ij}(H_i,H_j),\right.$$

$$\left. e_{kl}a_{kl}(H_k,H_l)\right],$$

which implies that $\operatorname{Var}\left(\widehat{\operatorname{Var}}(\hat{\tau}_t^{k,k'})\right)$ is uniformly bounded by a constant M by Assumption 16 and 17. So

$$\frac{1}{T^2}\sum_{t=1}^{T}\operatorname{Var}\left(\widehat{\operatorname{Var}}(\hat{\tau}_t^{k,k'})\right) \leq \frac{1}{T^2}\sum_{t=1}^{T} M = \frac{M}{T} \to 0$$

as $T \to 0$. So in the regime where both n and T go to infinity (i.e., Theorem 10 holds) or T goes to infinity (i.e., Theorem 9 holds), (2.20) → 0 and we are done. □

Proof of Theorem 12. This should be exactly the same as our proof of Theorem 19. □

Proof of Proposition 3.

$$\begin{aligned}
|\mathbb{E}[\hat{\tau}_t^c] - \tau_t^{TE}| &= |\mathbb{E}[\alpha\hat{\tau}_t^{TE} + (1-\alpha)\hat{\tau}_{t-1}^{TE}] - \tau_t^{TE}| \\
&= |\alpha\tau_t^{TE} + (1-\alpha)\tau_{t-1}^{TE} - \tau_t^{TE}| \\
&= |(1-\alpha)(\tau_{t-1}^{TE} - \tau_t^{TE})| \\
&= (1-\alpha)|\tau_{t-1}^{TE} - \tau_t^{TE}|
\end{aligned}$$

The second equality follows from unbiasedness of $\hat{\tau}_t^{TE}$ and $\hat{\tau}_{t-1}^{TE}$. To further bound the bias, we need to bound $|\tau_{t-1}^{TE} - \tau_t^{TE}|$. We do this below.

$$\begin{aligned}
|\tau_{t-1}^{TE} - \tau_t^{TE}| &= \left| \left(\frac{1}{n}\sum_{i=1}^{n} Y_{i,t}(h_i^1) - \frac{1}{n}\sum_{i=1}^{n} Y_{i,t}(h_i^0)\right) - \left(\frac{1}{n}\sum_{i=1}^{n} Y_{i,t-1}(h_i^1) - \frac{1}{n}\sum_{i=1}^{n} Y_{i,t-1}(h_i^0)\right) \right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n}(Y_{i,t}(h_i^1) - Y_{i,t-1}(h_i^1)) - \frac{1}{n}\sum_{i=1}^{n}(Y_{i,t}(h_i^0) - Y_{i,t-1}(h_i^0))\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}|Y_{i,t}(h_i^1) - Y_{i,t-1}(h_i^1)| + \frac{1}{n}\sum_{i=1}^{n}|Y_{i,t}(h_i^0) - Y_{i,t-1}(h_i^0)| \\
&\leq 2\epsilon,
\end{aligned}$$

by our ϵ-weak-stability assumption. Hence,

$$|\mathbb{E}[\hat{\tau}_t^c] - \tau_t^{TE}| \leq 2(1-\alpha)\epsilon.$$

□

Remark 4. Note that following the exact derivation, we can know that

$$|\tau_t^{TE} - \tau_{t'}^{TE}| \leq 2|t - t^{'}|\epsilon \tag{2.21}$$

Proposition 10 (Variance and Covariance of Horvitz-Thompson Type Estimators)**.** *For each $i \in \{1, \cdots, n\}, t \in \{1, \cdots, T\}$, we let $\mathbb{P}(H_{i,t} = h_i^1) = \pi_{i,t}^1$, $\mathbb{P}(H_{i,t} = h_i^0) = \pi_{i,t}^0$, $\mathbb{P}(H_{j,t} = h_j^1) = \pi_{j,t}^1$ and $\mathbb{P}(H_{j,t} = h_j^0) = \pi_{j,t}^0$. Moreover, for each $i \neq j$ and t, we let $\mathbb{P}(H_{i,t} = h_i^1, H_{j,t} = h_j^1) = \pi_{i,j,t}^{1,1}$, $\mathbb{P}(H_{i,t} = h_i^0, H_{j,t} = h_j^1) = \pi_{i,j,t}^{0,1}$, $\mathbb{P}(H_{i,t} = h_i^1, H_{j,t} = h_j^0) = \pi_{i,j,t}^{1,0}$ and $\mathbb{P}(H_{i,t} = h_i^0, H_{j,t} = h_j^0) = \pi_{i,j,t}^{0,0}$, then*

$$\begin{aligned} Var(\hat{\tau}_t^{TE}) = \frac{1}{n^2} \sum_{i=1}^n \left[\frac{Y_{i,t}^2(h_i^1)(1 - \pi_{i,t}^1)}{\pi_{i,t}^1} + \frac{Y_{i,t}^2(h_i^0)(1 - \pi_{i,t}^0)}{\pi_{i,t}^0} + 2Y_{i,t}(h_i^1)Y_{i,t}(h_i^0) \right] \\ + \frac{2}{n^2} \sum_{1 \leq i < j \leq n} \left[\frac{Y_{i,t}(h_i^1)Y_{j,t}(h_j^1)(\pi_{i,j,t}^{1,1} - \pi_{i,t}^1\pi_{j,t}^1)}{\pi_{i,t}^1\pi_{j,t}^1} - \frac{Y_{i,t}(h_i^0)Y_{j,t}(h_j^1)(\pi_{i,j,t}^{0,1} - \pi_{i,t}^0\pi_{j,t}^1)}{\pi_{i,t}^0\pi_{j,t}^1} \right. \\ \left. - \frac{Y_{i,t}(h_i^1)Y_{j,t}(h_j^0)(\pi_{i,j,t}^{1,0} - \pi_{i,t}^1\pi_{j,t}^0)}{\pi_{i,t}^1\pi_{j,t}^0} + \frac{Y_{i,t}(h_i^0)Y_{j,t}(h_j^0)(\pi_{i,j,t}^{0,0} - \pi_{i,t}^0\pi_{j,t}^0)}{\pi_{i,t}^0\pi_{j,t}^0} \right] \end{aligned} \tag{2.22}$$

As for $Cov(\hat{\tau}_t^{TE}, \hat{\tau}_{t'}^{TE})$, if we let $\mathbb{P}(H_{i,t} = h_i^1, H_{i,t'} = h_i^1) = \pi_{i,t,t'}^{1,1}$, $\mathbb{P}(H_{i,t} = h_i^0, H_{i,t'} = h_i^1) = \pi_{i,t,t'}^{0,1}$, $\mathbb{P}(H_{i,t} = h_i^1, H_{i,t'} = h_i^0) = \pi_{i,t,t'}^{1,0}$, $\mathbb{P}(H_{i,t} = h_i^0, H_{i,t'} = h_i^0) = \pi_{i,t,t'}^{0,0}$ and $\mathbb{P}(H_{i,t} = h_i^1, H_{j,t'} = h_j^1) = \pi_{i,t,j,t'}^{1,1}$, $\mathbb{P}(H_{i,t} = h_i^0, H_{j,t'} = h_j^1) = \pi_{i,t,j,t'}^{0,1}$, $\mathbb{P}(H_{i,t} = h_i^1, H_{j,t'} = h_j^0) = \pi_{i,t,j,t'}^{1,0}$, $\mathbb{P}(H_{i,t} = h_i^0, H_{j,t'} = h_j^0) = \pi_{i,t,j,t'}^{0,0}$, then we have the following expression for $Cov(\hat{\tau}_t^{TE}, \hat{\tau}_{t'}^{TE})$:

$$\begin{aligned} \frac{1}{n^2} \sum_{i=1}^n \left[\frac{Y_{i,t}(h_i^1)Y_{i,t'}(h_i^1)(\pi_{i,t,t'}^{1,1} - \pi_{i,t}^1\pi_{i,t'}^1)}{\pi_{i,t}^1\pi_{i,t'}^1} - \frac{Y_{i,t}(h_i^0)Y_{i,t'}(h_i^1)(\pi_{i,t,t'}^{0,1} - \pi_{i,t}^0\pi_{i,t'}^1)}{\pi_{i,t}^0\pi_{i,t'}^1} \right. \\ \left. - \frac{Y_{i,t}(h_i^1)Y_{i,t'}(h_i^0)(\pi_{i,t,t'}^{1,0} - \pi_{i,t}^1\pi_{i,t'}^0)}{\pi_{i,t}^1\pi_{i,t'}^0} + \frac{Y_{i,t}(h_i^0)Y_{i,t'}(h_i^0)(\pi_{i,t,t'}^{0,0} - \pi_{i,t}^0\pi_{i,t'}^0)}{\pi_{i,t}^0\pi_{i,t'}^0} \right] \\ + \frac{2}{n^2} \sum_{1 \leq i < j \leq n} \left[\frac{Y_{i,t}(h_i^1)Y_{j,t'}(h_j^1)(\pi_{i,t,j,t'}^{1,1} - \pi_{i,t}^1\pi_{j,t'}^1)}{\pi_{i,t}^1\pi_{j,t'}^1} - \frac{Y_{i,t}(h_i^0)Y_{j,t'}(h_j^1)(\pi_{i,t,j,t'}^{0,1} - \pi_{i,t}^0\pi_{j,t'}^1)}{\pi_{i,t}^0\pi_{j,t'}^1} \right. \\ \left. - \frac{Y_{i,t}(h_i^1)Y_{j,t'}(h_j^0)(\pi_{i,t,j,t'}^{1,0} - \pi_{i,t}^1\pi_{j,t'}^0)}{\pi_{i,t}^1\pi_{j,t'}^0} + \frac{Y_{i,t}(h_i^0)Y_{j,t'}(h_j^0)(\pi_{i,t,j,t'}^{0,0} - \pi_{i,t}^0\pi_{j,t'}^0)}{\pi_{i,t}^0\pi_{j,t'}^0} \right] \end{aligned} \tag{2.23}$$

Proof of Proposition 10. This can be done by direct calculations. □

Proof of Proposition 4. We'd like to have reduction in MSE by using $\hat{\tau}_t^c$. By the bias-variance decomposition and note that $\hat{\tau}_t^{TE}$ is unbiased, this boils down to

$$Var(\hat{\tau}_t^c) + |\mathbb{E}[\hat{\tau}_t^c] - \tau_t^{TE}|^2 \leq Var(\hat{\tau}_t^{TE})$$

By Proposition 3, it suffices to have

$$Var(\hat{\tau}_t^c) + 4(1-\alpha)^2\epsilon^2 \le Var(\hat{\tau}_t^{TE}),$$

which is further equivalent to

$$\begin{aligned} \alpha^2 Var(\hat{\tau}_t^{TE}) + (1-\alpha)^2 Var(\hat{\tau}_{t-1}^{TE}) \\ +2\alpha(1-\alpha)Cov(\hat{\tau}_t^{TE}, \hat{\tau}_{t-1}^{TE}) + 4(1-\alpha)^2\epsilon^2 \\ \le Var(\hat{\tau}_t^{TE}) \end{aligned} \tag{2.24}$$

Rewrite (2.24), we have

$$\begin{aligned} &\left(4\epsilon^2 + Var(\hat{\tau}_t^{TE}) + Var(\hat{\tau}_{t-1}^{TE}) - 2Cov(\hat{\tau}_t^{TE}, \hat{\tau}_{t-1}^{TE})\right)\alpha^2 \\ &\qquad - \left(8\epsilon^2 + 2Var(\hat{\tau}_{t-1}^{TE}) - 2Cov(\hat{\tau}_t^{TE}, \hat{\tau}_{t-1}^{TE})\right)\alpha \\ &\qquad\qquad + \left(4\epsilon^2 + Var(\hat{\tau}_{t-1}^{TE}) - Var(\hat{\tau}_t^{TE})\right) \le 0 \end{aligned} \tag{2.25}$$

Now we look at the left hand side of (2.25), which is quadratic in α. To ease notations, let $A = Var(\hat{\tau}_t^{TE})$, $B = Var(\hat{\tau}_{t-1}^{TE})$ and $C = Cov(\hat{\tau}_t^{TE}, \hat{\tau}_{t-1}^{TE})$. It's easy to see that the left hand side achieves its minimum at $\alpha = \delta = 1 - \frac{2(A-C)}{8\epsilon^2+2A+2B-4C}$ and is 0 at $\alpha = 1$. So if we have $\delta < 1$, then for some $\alpha \in (0,1)$, we have reduction in MSE. Moreover, if $\delta < \frac{1}{2}$, we then know that for $\alpha = \frac{1}{2}$, we also have smaller MSE by the property of quadratic functions. And simple algebra shows that $\delta < \frac{1}{2}$ is equivalent to $A - B > 4\epsilon^2$. □

Proposition 11 (Estimators of variance)**.** *We define two estimators of the variance:*

$$
\begin{aligned}
\widehat{Var}^{u}(\hat{\tau}_t^{TE}) = \frac{1}{n^2}\sum_{i=1}^{n}\Bigg[& \mathbf{1}(H_{i,t}=h_i^1)(1-\pi_{i,t}^1)\left(\frac{Y_{i,t}}{\pi_{i,t}^1}\right)^2 + \mathbf{1}(H_{i,t}=h_i^0)(1-\pi_{i,t}^0)\left(\frac{Y_{i,t}}{\pi_{i,t}^0}\right)^2 \\
& + \frac{Y_{i,t}^2}{\pi_{i,t}^1}\mathbf{1}(H_{i,t}=h_i^1) + \frac{Y_{i,t}^2}{\pi_{i,t}^0}\mathbf{1}(H_{i,t}=h_i^0)\Bigg] \\
+ \frac{2}{n^2}\sum_{1\le i<j\le n}\Bigg[& \mathbf{1}(\pi_{i,j,t}^{1,1}\neq 0)\mathbf{1}(H_{i,t}=h_i^1)\mathbf{1}(H_{j,t}=h_j^1)\frac{(\pi_{i,j,t}^{1,1}-\pi_{i,t}^1\pi_{j,t}^1)Y_{i,t}Y_{j,t}}{\pi_{i,t}^1\pi_{j,t}^1\pi_{i,j,t}^{1,1}} \\
& - \Bigg(\mathbf{1}(\pi_{i,j,t}^{0,1}\neq 0)\mathbf{1}(H_{i,t}=h_i^0)\mathbf{1}(H_{j,t}=h_j^1)\frac{(\pi_{i,j,t}^{0,1}-\pi_{i,t}^0\pi_{j,t}^1)Y_{i,t}Y_{j,t}}{\pi_{i,t}^0\pi_{j,t}^1\pi_{i,j,t}^{0,1}} \\
& \qquad -\mathbf{1}(\pi_{i,j,t}^{0,1}=0)\left(\frac{\mathbf{1}(H_{i,t}=h_i^0)Y_{i,t}^2}{2\pi_{i,t}^0}+\frac{\mathbf{1}(H_{j,t}=h_j^1)Y_{j,t}^2}{2\pi_{j,t}^1}\right)\Bigg) \\
& - \Bigg(\mathbf{1}(\pi_{i,j,t}^{1,0}\neq 0)\mathbf{1}(H_{i,t}=h_i^1)\mathbf{1}(H_{j,t}=h_j^0)\times\frac{(\pi_{i,j,t}^{1,0}-\pi_{i,t}^1\pi_{j,t}^0)Y_{i,t}Y_{j,t}}{\pi_{i,t}^1\pi_{j,t}^0\pi_{i,j,t}^{1,0}} \\
& \qquad -\mathbf{1}(\pi_{i,j,t}^{1,0}=0)\left(\frac{\mathbf{1}(H_{i,t}=h_i^1)Y_{i,t}^2}{2\pi_{i,t}^1}+\frac{\mathbf{1}(H_{j,t}=h_j^0)Y_{j,t}^2}{2\pi_{j,t}^0}\right)\Bigg) \\
& \qquad +\mathbf{1}(\pi_{i,j,t}^{0,0}\neq 0)\frac{\mathbf{1}(H_{i,t}=h_i^0)\mathbf{1}(H_{j,t}=h_j^0)(\pi_{i,j,t}^{0,0}-\pi_{i,t}^0\pi_{j,t}^0)Y_{i,t}Y_{j,t}}{\pi_{i,t}^0\pi_{j,t}^0\pi_{i,j,t}^{0,0}}\Bigg]
\end{aligned}
\tag{2.26}
$$

and

$$
\begin{aligned}
\widehat{Var}^{d}(\hat{\tau}_t^{TE}) = \frac{1}{n^2}\sum_{i=1}^{n}\Bigg[& \mathbf{1}(H_{i,t}=h_i^1)(1-\pi_{i,t}^1)\left(\frac{Y_{i,t}}{\pi_{i,t}^1}\right)^2 + \mathbf{1}(H_{i,t}=h_i^0)(1-\pi_{i,t}^0)\left(\frac{Y_{i,t}}{\pi_{i,t}^0}\right)^2\Bigg] \\
+ \frac{2}{n^2}\sum_{1\le i<j\le n}\Bigg[& \Bigg(\mathbf{1}(\pi_{i,j,t}^{1,1}\neq 0)\frac{\mathbf{1}(H_{i,t}=h_i^1)\mathbf{1}(H_{j,t}=h_j^1)(\pi_{i,j,t}^{1,1}-\pi_{i,t}^1\pi_{j,t}^1)Y_{i,t}Y_{j,t}}{\pi_{i,t}^1\pi_{j,t}^1\pi_{i,j,t}^{1,1}} \\
& \qquad -\mathbf{1}(\pi_{i,j,t}^{1,1}=0)\left(\frac{\mathbf{1}(H_{i,t}=h_i^1)Y_{i,t}^2}{2\pi_{i,t}^1}+\frac{\mathbf{1}(H_{j,t}=h_j^1)Y_{j,t}^2}{2\pi_{j,t}^1}\right)\Bigg) \\
& -\mathbf{1}(\pi_{i,j,t}^{0,1}\neq 0)\frac{\mathbf{1}(H_{i,t}=h_i^0)\mathbf{1}(H_{j,t}=h_j^1)(\pi_{i,j,t}^{0,1}-\pi_{i,t}^0\pi_{j,t}^1)Y_{i,t}Y_{j,t}}{\pi_{i,t}^0\pi_{j,t}^1\pi_{i,j,t}^{0,1}} \\
& -\mathbf{1}(\pi_{i,j,t}^{1,0}\neq 0)\frac{\mathbf{1}(H_{i,t}=h_i^1)\mathbf{1}(H_{j,t}=h_j^0)(\pi_{i,j,t}^{1,0}-\pi_{i,t}^1\pi_{j,t}^0)Y_{i,t}Y_{j,t}}{\pi_{i,t}^1\pi_{j,t}^0\pi_{i,j,t}^{1,0}} \\
& + \Bigg(\mathbf{1}(\pi_{i,j,t}^{0,0}\neq 0)\frac{\mathbf{1}(H_{i,t}=h_i^0)\mathbf{1}(H_{j,t}=h_j^0)(\pi_{i,j,t}^{0,0}-\pi_{i,t}^0\pi_{j,t}^0)Y_{i,t}Y_{j,t}}{\pi_{i,t}^0\pi_{j,t}^0\pi_{i,j,t}^{0,0}} \\
& \qquad -\mathbf{1}(\pi_{i,j,t}^{0,0}=0)\left(\frac{\mathbf{1}(H_{i,t}=h_i^0)Y_{i,t}^2}{2\pi_{i,t}^0}+\frac{\mathbf{1}(H_{j,t}=h_j^0)Y_{j,t}^2}{2\pi_{j,t}^0}\right)\Bigg]\,.
\end{aligned}
\tag{2.27}
$$

Assuming all the potential outcomes are non-negative, we then have that

$$\mathbb{E}\left[\widehat{Var}^{u}(\hat{\tau}_t^{TE})\right] \geq Var(\hat{\tau}_t^{TE})$$

and

$$\mathbb{E}\left[\widehat{Var}^{d}(\hat{\tau}_t^{TE})\right] \leq Var(\hat{\tau}_t^{TE}).$$

Proposition 12 (Estimator of the covariance)**.** *We have the following unbiased estimator of* $Cov(\hat{\tau}_t^{TE}, \hat{\tau}_{t'}^{TE})$*:*

$$\begin{aligned}
\widehat{Cov}(\hat{\tau}_t^{TE}, \hat{\tau}_{t'}^{TE}) = \frac{1}{n^2}\sum_{i=1}^{n} \Bigg[& \frac{\mathbf{1}(H_{i,t} = h_i^1)\mathbf{1}(H_{i,t'} = h_i^1)Y_{i,t}Y_{i,t'}(\pi^{1,1}_{i,t,t'} - \pi^1_{i,t}\pi^1_{i,t'})}{\pi^{1,1}_{i,t,t'}\pi^1_{i,t}\pi^1_{i,t'}} \\
& - \frac{\mathbf{1}(H_{i,t} = h_i^0)\mathbf{1}(H_{i,t'} = h_i^1)Y_{i,t}Y_{i,t'}(\pi^{0,1}_{i,t,t'} - \pi^0_{i,t}\pi^1_{i,t'})}{\pi^{0,1}_{i,t,t'}\pi^0_{i,t}\pi^1_{i,t'}} \\
& - \frac{\mathbf{1}(H_{i,t} = h_i^1)\mathbf{1}(H_{i,t'} = h_i^0)Y_{i,t}Y_{i,t'}(\pi^{1,0}_{i,t,t'} - \pi^1_{i,t}\pi^0_{i,t'})}{\pi^{1,0}_{i,t,t'}\pi^1_{i,t}\pi^0_{i,t'}} \\
& + \frac{\mathbf{1}(H_{i,t} = h_i^0)\mathbf{1}(H_{i,t'} = h_i^0)Y_{i,t}Y_{i,t'}(\pi^{0,0}_{i,t,t'} - \pi^0_{i,t}\pi^0_{i,t'})}{\pi^{0,0}_{i,t,t'}\pi^0_{i,t}\pi^0_{i,t'}} \Bigg] \\
+ \frac{2}{n^2}\sum_{1 \leq i < j \leq n} \Bigg[& \frac{\mathbf{1}(H_{i,t} = h_i^1)\mathbf{1}(H_{j,t'} = h_j^1)Y_{i,t}Y_{j,t'}(\pi^{1,1}_{i,t,j,t'} - \pi^1_{i,t}\pi^1_{j,t'})}{\pi^{1,1}_{i,t,j,t'}\pi^1_{i,t}\pi^1_{j,t'}} \\
& - \frac{\mathbf{1}(H_{i,t} = h_i^0)\mathbf{1}(H_{j,t'} = h_j^1)Y_{i,t}Y_{j,t'}(\pi^{0,1}_{i,t,j,t'} - \pi^0_{i,t}\pi^1_{j,t'})}{\pi^{0,1}_{i,t,j,t'}\pi^0_{i,t}\pi^1_{j,t'}} \\
& - \frac{\mathbf{1}(H_{i,t} = h_i^1)\mathbf{1}(H_{j,t'} = h_j^0)Y_{i,t}Y_{j,t'}(\pi^{1,0}_{i,t,j,t'} - \pi^1_{i,t}\pi^0_{j,t'})}{\pi^{1,0}_{i,t,j,t'}\pi^1_{i,t}\pi^0_{j,t'}} \\
& + \frac{\mathbf{1}(H_{i,t} = h_i^0)\mathbf{1}(H_{j,t'} = h_j^0)Y_{i,t}Y_{j,t'}(\pi^{0,0}_{i,t,j,t'} - \pi^0_{i,t}\pi^0_{j,t'})}{\pi^{0,0}_{i,t,j,t'}\pi^0_{i,t}\pi^0_{j,t'}} \Bigg]
\end{aligned}$$

Proving Theorem 14 relies on results in m-dependence central limit theorem. We need the following result as a lemma.

Lemma 5. *Let* $\{X_{n,i}\}$ *be a triangular array of mean zero random variables. For each* $n = 1, 2, \cdots$ *let* $d = d_n$*, and suppose* $X_{n,1}, \cdots, X_{n,d}$ *is an* m*-dependent sequence of random variables for some* $m \in \mathbb{N}$*. Define*

$$B^2_{n,k,a} = Var\left(\sum_{i=a}^{a+k-1} X_{n,i}\right), B^2_n = B_{n,d,1} = Var\left(\sum_{i=1}^{d} X_{n,i}\right).$$

Assume the following conditions hold. For some $\delta > 0$*,* $-1 \leq \gamma < 1$ *and* $g = g_n > 2m$ *is such that* $\frac{m}{g} \to 0$*:*

$$\mathbb{E}|X_{n,i}|^{2+\delta} \leq \Delta_n \text{ for all } i, \tag{2.28}$$

$$B^2_{n,k,a}/(k^{1+\gamma}) \leq K_n \text{ for all } a \text{ and for all } k \geq m, \tag{2.29}$$

$$B^2_n/(dm^{\gamma}) \geq L_n, \tag{2.30}$$

$$\frac{K_n}{L_n} \cdot \frac{m}{g} \to 0, \tag{2.31}$$

$$\frac{K_n}{L_n} \cdot \left(\frac{m}{g}\right)^{(1-\gamma)/2} \to 0, \tag{2.32}$$

$$\Delta_n L_n^{-(2+\delta)/2} g^{\delta/2+(1-\gamma)(2+\delta)/2} d^{-\delta/2} \left(\frac{m}{g}\right)^{(1-\gamma)(2+\delta)/2} \to 0. \tag{2.33}$$

Then, $B_n^{-1}(X_{n,1} + \cdots X_{n,d}) \xrightarrow{d} \mathcal{N}(0,1)$*.*

Proof. This is essentially Theorem 2.1 in Romano and Wolf [2000]. We replace the original conditions 4, 5 and 6 by the last three conditions. In fact, the last three conditions are needed to establish the theorem and the conditions 4, 5 and 6 in Theorem 2.1 in Romano and Wolf [2000] are sufficient conditions. □

Now, we are ready to prove Theorem 14.

Proof of Theorem 14. We define $\tilde{\tau}_{i,t} = \frac{\mathbf{1}(H_{i,t}=k)}{\mathbb{P}(H_{i,t}=k)} Y_{i,t} - \frac{\mathbf{1}(H_{i,t}=k')}{\mathbb{P}(H_{i,t}=k')} Y_{i,t} = \frac{\mathbf{1}(H_{i,t}=k)}{\mathbb{P}(H_{i,t}=k)} Y_{i,t}(k) - \frac{\mathbf{1}(H_{i,t}=k')}{\mathbb{P}(H_{i,t}=k')} Y_{i,t}(k')$. Then the ATEC can be written as

$$\hat{\bar{\tau}}^{k,k'} = \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{1}{nT} \tilde{\tau}_{i,t}.$$

Similarly, we define $\tau_{i,t} = Y_{i,t}(k) - Y_{i,t}(k')$, which is the true individual exposure contrast. Now,

$$\sqrt{nT}(\hat{\bar{\tau}}^{k,k'} - \bar{\tau}^{k,k'}) = \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{1}{\sqrt{nT}} (\tilde{\tau}_{i,t} - \tau_{i,t}).$$

To proceed, we let $X_{n,i,t} = \frac{1}{\sqrt{nT}}(\tilde{\tau}_{i,t} - \tau_{i,t})$. We view $\{X_{n,i,t}\}$ as a single sequence of random variables by enumerating $X_{n,i,t}$ following the order $X_{n,1,1}, \cdots, X_{n,1,T}, X_{n,2,1}, \cdots, X_{n,2,T}, \cdots, X_{n,n,T}$. Using the language in the lemma, $d = nT$. Since $\{H_{i,t}\}_{i=1}^n$ is a sequence of s-dependent random variables and $X_{n,i,t}$ is a function of $H_{i,t}$, we know that $\{X_{n,i,t}\}$ is a sT-dependent sequence of random variables. In other words, $m = sT$ in the above lemma. Note that $|X_{n,i,t}| \leq \frac{C_1}{\sqrt{nT}}$ by uniform boundedness of potential outcomes. Hence, for any $\delta > 0$, $\Delta_n = C_2(nT)^{-1-\delta/2}$. Now, we calculate $B_{n,k,a}^2$ and B_n^2. We start with $B_{n,k,a}^2$. For all (i_1, t_1) and $k \geq m$, let (i_2, t_2) be the index such that when we order

X's there are exactly k indices from (i_1, t_1) to (i_2, t_2).

$$\begin{aligned} B_{n,k,a}^2 &= \text{Var}\left(\sum_{(i,t)=(i_1,t_1)}^{(i,t)=(i_2,t_2)} X_{n,i,t}\right) \\ &= \frac{1}{nT}\text{Var}\left(\sum_{(i,t)=(i_1,t_1)}^{(i,t)=(i_2,t_2)} \tilde{\tau}_{i,t}\right) \\ &= \frac{1}{nT}\left[\sum_{(i,t)=(i_1,t_1)}^{(i,t)=(i_2,t_2)} \text{Var}(\tilde{\tau}_{i,t}) + 2\sum_{(u,v)\neq(p,q)} \text{Cov}(\tilde{\tau}_{u,v}, \tilde{\tau}_{p,q})\right] \end{aligned}$$

Since $k \geq m = sT$, we know that at most mk covariance terms are non-zero. Given uniform boundedness of potential outcomes and overlap, all the variance and covariance terms are upper bounded by constants $M_1 > 0$ and $M_2 > 0$ respectively. Hence,

$$B_{n,k,a}^2 \leq \frac{1}{nT}(kM_1 + 2mkM_2) \leq M_3\frac{mk}{nT} = M_3\frac{sk}{n}.$$

Therefore,

$$B_{n,k,a}^2/k \leq M_3\frac{sk}{n}/k = M_3\frac{s}{n} = K_n.$$

Now we look at B_n^2. By Assumption 13, $\text{Var}(\sqrt{nT}\hat{\bar{\tau}}^{k,k'}) \geq \epsilon > 0$, hence, for sufficiently large n,

$$B_n^2 = \text{Var}(\sqrt{nT}\hat{\bar{\tau}}^{k,k'}) \geq \epsilon > 0,$$

and

$$B_n^2/d = B_n^2/(nT) \geq \epsilon/(nT) = L_n.$$

We let $\gamma = 0, \delta = 2$. Pick $g = g_n = s^3T^3n^\alpha$. With such g, m/g obviously goes to 0. Now,

$$\frac{K_n}{L_n} \cdot \frac{m}{g} = \epsilon M_3 sT \cdot \frac{1}{s^2T^2n^\alpha} \to 0,$$

$$\frac{K_n}{L_n} \cdot \left(\frac{m}{g}\right)^{(1-\gamma)/2} = \epsilon M_3 sT \cdot \frac{1}{sTn^{0.5\alpha}} \to 0,$$

$$\begin{aligned} &\Delta_n L_n^{-(2+\delta)/2} g^{\delta/2+(1-\gamma)(2+\delta)/2} d^{-\delta/2}\left(\frac{m}{g}\right)^{(1-\gamma)(2+\delta)/2} \\ &= C_2(nT)^{-1-\delta/2}\epsilon^{-(2+\delta)/2}(nT)^{(2+\delta)/2}g^{\delta/2}(nT)^{-\delta/2}(sT)^{1+\delta/2} \\ &= C_4 gs^2T/n \qquad \text{when } \delta = 2 \\ &= C_4 s^5T^4n^\alpha/n. \end{aligned}$$

Since $s^5T^4 = o(n^{1-\alpha})$, $s^5T^4n^\alpha = o(n)$ and hence $\Delta_n L_n^{-(2+\delta)/2} g^{\delta/2+(1-\gamma)(2+\delta)/2} d^{-\delta/2} \left(\frac{m}{g}\right)^{(1-\gamma)(2+\delta)/2} = o(1)$. Having checked all the conditions, by Lemma 5, we are done. □

Proof of Theorem 15. As in the above proof, we check the six conditions in Lemma 5 are satisfied with $\gamma = 0$ and $\delta = 2$. Note that since now $X_{n,i,t}$ and $X_{n,j,t}$ are correlated if and only if i and j are in the same group, we can reorder $X_{n,i,t}$'s as follows:

$$X_{n,1,1}, \cdots, X_{n,r,1}, X_{n,1,2}, \cdots, X_{n,r,2}, \cdots, X_{n,r,T}, X_{n,r+1,1}, \cdots, X_{n,nr,T}.$$

Now, this sequence is actually $(2r)$-dependent, i.e., $m = 2r, s = r$. Then

$$K_n = M_4/(nT), \quad L_n = \epsilon/(nrT).$$

Hence $K_n/L_n = M_5 r$. Pick $g = g_n$ such that $g \to \infty$ and $g = (nT)^{3/4}$. Then with $r = o((nT)^{\frac{1}{4}})$, $r^2/g \to 0$ and $r^3/g \to 0$.

$$\frac{K_n}{L_n} \cdot \frac{m}{g} = M_5 r \cdot \frac{2r}{g} \to 0,$$

$$\frac{K_n}{L_n} \cdot \left(\frac{m}{g}\right)^{(1-\gamma)/2} = M_5 r \cdot \sqrt{\frac{2r}{g}} \to 0$$

and

$$\begin{aligned}
&\Delta_n L_n^{-(2+\delta)/2} g^{\delta/2+(1-\gamma)(2+\delta)/2} d^{-\delta/2} \left(\frac{m}{g}\right)^{(1-\gamma)(2+\delta)/2} \\
&= M_6 g^3 (nrT)^{-1} \left(\frac{r}{g}\right)^2 \\
&= M_6 rg/(nT) \\
&= o(nT)/(nT) \to 0
\end{aligned}$$

Hence all the conditions are satisfied. It is also easy to see that instead of just 2 time steps, any finite p time steps would work. □

Proof of Proposition 8. Let $X_{n,t} = \sqrt{\frac{nr}{T}}(\hat{\tau}_t^{k,k'} - \tau_t^{k,k'})$. The key ingredients are the following two expressions:

$$\begin{aligned}\mathrm{Var}(X_{n,t}) = \frac{1}{nrT}\Bigg[&\sum_{l=1}^{n}\sum_{q=1}^{r}(2^{2r}-1)Y_{(l,q),t}(k)^2 \\ &+\sum_{l=1}^{n}\sum_{q=1}^{r}(2^{2r}-1)Y_{(l,q),t}(k^{'})^2 + 2\sum_{l=1}^{n}\sum_{q=1}^{r}Y_{(l,q),t}(k)Y_{(l,q),t}(k^{'}) \\ &+\sum_{l=1}^{n}\sum_{q_1=1}^{r}\sum_{q_2\neq q_1}\Big((2^{2r}-1)Y_{(l,q_1),t}(k)Y_{(l,q_2),t}(k)+(2^{2r}-1)Y_{(l,q_1),t}(k^{'})Y_{(l,q_2),t}(k^{'})\Big) \\ &+2\sum_{l=1}^{n}\sum_{q_1=1}^{r}\sum_{q_2\neq q_1}Y_{(l,q_1),t}(k)Y_{(l,q_2),t}(k^{'})\Bigg] \qquad (2.34)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}(X_{n,t},X_{n,t+1}) = \frac{1}{nrT}\sum_{l=1}^{n}\sum_{q_1=1}^{r}\sum_{q_2=1}^{r} &\Big((2^{r}-1)Y_{(l,q_1),t}(k)Y_{(l,q_2),t+1}(k) \\ &+(2^{r}-1)Y_{(l,q_1),t}(k^{'})Y_{(l,q_2),t+1}(k^{'}) \\ &+Y_{(l,q_1),t}(k^{'})Y_{(l,q_2),t+1}(k)+Y_{(l,q_1),t}(k)Y_{(l,q_2),t+1}(k^{'})\Big) \qquad (2.35)\end{aligned}$$

We have that

$$B_n^2 = \sum_{t=1}^{T}\mathrm{Var}(X_{n,t}) + 2\sum_{t=1}^{T-1}\mathrm{Cov}(X_{n,t},X_{n,t+1})$$

Plugging in (2.34) and (2.35), we have the expression of B_n^2. The estimator is obtained by replacing the non-identifiable terms by corresponding upper bound. □

2.7.3 $k-$steps convex estimator

The approach we have described in Section 2.2.2 naturally extends to using the $k-1$ previous time steps, yielding the weighted combination estimator:

$$\hat{\tau}_t^c = \alpha_1\hat{\tau}_{t-k+1}^{TE} + \cdots + \alpha_k\hat{\tau}_t^{TE},$$

which exhibits the following absolute bias bound:

Proposition 13 (Bound on the bias of $\hat{\tau}_t^c$)**.**

$$|\mathbb{E}[\hat{\tau}_t^c] - \tau_t^{TE}| \leq 2\left[(k-1)\alpha_1 + (k-2)\alpha_2 + \cdots + \alpha_{k-1}\right]\epsilon$$

As in the previous section, we can estimate $\alpha_1, \cdots, \alpha_k$ by solving the following convex optimization problem:

$$\begin{aligned}\underset{\alpha_1,\cdots,\alpha_k}{\arg\min} \quad & \alpha_1^2 \widehat{\mathrm{Var}}(\hat{\tau}_{t-k+1}^{TE}) + \cdots + \alpha_k^2 \widehat{\mathrm{Var}}(\hat{\tau}_t^{TE}) \\ & + 4\left[(k-1)\alpha_1 + \cdots + \alpha_{k-1}\right]^2 \epsilon^2 \\ \text{subject to} \quad & \alpha_1 + \cdots + \alpha_k = 1,\end{aligned}$$

where $\widehat{\mathrm{Var}}(\hat{\tau}_{t-k+1}^{TE}), \cdots, \widehat{\mathrm{Var}}(\hat{\tau}_t^{TE})$ are estimators of the associated variance terms, and are provided in Appendix 2.7.2. This then suggests the following plug-in estimator:

$$\hat{\tau}_t^c = \hat{\alpha}_1 \hat{\tau}_{t-k+1}^{TE} + \cdots + \hat{\alpha}_k \hat{\tau}_t^{TE}.$$

We can assert stronger control over the bias of $\hat{\tau}^c$ by incorporating an additional constraint to the optimization problem:

$$\begin{aligned}\underset{\alpha_1,\cdots,\alpha_k}{\arg\min} \quad & \alpha_1^2 \widehat{\mathrm{Var}}(\hat{\tau}_{t-k+1}^{TE}) + \cdots + \alpha_k^2 \widehat{\mathrm{Var}}(\hat{\tau}_t^{TE}) \\ & + 4\left[(k-1)\alpha_1 + \cdots + \alpha_{k-1}\right]^2 \epsilon^2 \\ \text{subject to} \quad & \alpha_1 + \cdots + \alpha_k = 1, \\ & 2\left[(k-1)\alpha_1 + (k-2)\alpha_2 + \cdots + \alpha_{k-1}\right] \epsilon \leq \delta.\end{aligned}$$

Numerical solutions for either optimization problem are straightforward to obtain using standard numerical solvers. Variance estimator and confidence interval of τ_t^{TE} can be constructed in exactly the same way as in the case $k = 2$.

Proof of Proposition 13.

$$\begin{aligned}& |\mathbb{E}[\hat{\tau}_t^c] - \tau_t| \\ & = |\alpha_1 \tau_{t-k+1}^{TE} + \cdots + \alpha_k \tau_t^{TE} - \tau_t^{TE}| \\ & = |\alpha_1 \tau_{t-k+1}^{TE} + \cdots + \alpha_{k-1} \tau_{t-1}^{TE} - (1-\alpha_k) \tau_t^{TE}| \\ & = |\alpha_1 \tau_{t-k+1}^{TE} + \cdots + \alpha_{k-1} \tau_{t-1}^{TE} - (\alpha_1 + \cdots + \alpha_{k-1}) \tau_t^{TE}| \\ & = |\alpha_1 (\tau_{t-k+1}^{TE} - \tau_t^{TE}) + \cdots + \alpha_{k-1} (\tau_{t-1}^{TE} - \tau_t^{TE})| \\ & \leq \alpha_1 |\tau_{t-k+1}^{TE} - \tau_t^{TE}| + \cdots + \alpha_{k-1} |\tau_{t-1}^{TE} - \tau_t^{TE}| \\ & \leq 2\alpha_1 (k-1)\epsilon + \cdots + 2\alpha_{k-1}\epsilon \\ & = 2\left[(k-1)\alpha_1 + \cdots + \alpha_{k-1}\right] \epsilon\end{aligned}$$

□

We first give the optimization problem for the general case that assignments may be correlated across time:

$$\begin{aligned}
\underset{\alpha_1,\cdots,\alpha_k}{\arg\min} \quad & \alpha_1^2\widehat{\mathrm{Var}}(\hat{\tau}_{t-k+1}^{TE}) + \cdots + \alpha_k^2\widehat{\mathrm{Var}}(\hat{\tau}_t^{TE}) \\
& + 2\alpha_i\alpha_j \sum_{1\le i<j\le k} \widehat{\mathrm{Cov}}(\hat{\tau}_{t-k+i}^{TE}, \hat{\tau}_{t-k+j}^{TE}) \\
& + 4\left[(k-1)\alpha_1 + \cdots + \alpha_{k-1}\right]^2 \epsilon^2 \\
\text{subject to} \quad & \alpha_1 + \cdots + \alpha_k = 1,
\end{aligned}$$

where $\widehat{\mathrm{Var}}(\hat{\tau}_{t-k+1}^{TE}), \cdots, \widehat{\mathrm{Var}}(\hat{\tau}_t^{TE})$ and $\widehat{\mathrm{Cov}}(\hat{\tau}_{t-k+i}^{TE}, \hat{\tau}_{t-k+j}^{TE})$ can be any estimator in Proposition 11 and 12. Moreover, suppose that the assignments are independent across time, we know that $\mathrm{Cov}(\hat{\tau}_{t-k+i}^{TE}, \hat{\tau}_{t-k+j}^{TE}) = 0$, hence we have an even simpler optimization problem as stated in the main text.

Derivation of the optimization problem. We first calculate the variance.

$$\begin{aligned}
\mathrm{Var}(\hat{\tau}_t^c) &= \mathrm{Var}(\alpha_1\hat{\tau}_{t-k+1}^{TE} + \cdots + \alpha_k\hat{\tau}_t^{TE}) \\
&= \alpha_1^2\mathrm{Var}(\hat{\tau}_{t-k+1}^{TE}) + \cdots + \alpha_k^2\mathrm{Var}(\hat{\tau}_t^{TE}) \\
&\quad + 2\alpha_i\alpha_j \sum_{1\le i<j\le k} \mathrm{Cov}(\hat{\tau}_{t-k+i}^{TE}, \hat{\tau}_{t-k+j}^{TE})
\end{aligned}$$

Suppose we want to have smaller MSE by using $\hat{\tau}_t^c$, we need to have

$$\mathrm{Var}(\hat{\tau}_t^c) + |\mathbb{E}[\hat{\tau}_t^c] - \tau_t^{TE}|^2 \le \mathrm{Var}(\hat{\tau}_t^{TE})$$

By Proposition 13, it suffices to have

$$\begin{aligned}
&\alpha_1^2\mathrm{Var}(\hat{\tau}_{t-k+1}^{TE}) + \cdots + \alpha_k^2\mathrm{Var}(\hat{\tau}_t^{TE}) \\
&\quad + 2\alpha_i\alpha_j \sum_{1\le i<j\le k} \mathrm{Cov}(\hat{\tau}_{t-k+i}^{TE}, \hat{\tau}_{t-k+j}^{TE}) \\
&+ 4\left[(k-1)\alpha_1 + \cdots + \alpha_{k-1}\right]^2 \epsilon^2 \le \mathrm{Var}(\hat{\tau}_t^{TE})
\end{aligned} \tag{2.36}$$

Now, the left hand side of (2.36) is convex in $\alpha_1, \cdots, \alpha_k$. □

Estimate of ϵ	$\hat{\epsilon}$	$1.5\hat{\epsilon}$	$2\hat{\epsilon}$	$2.5\hat{\epsilon}$	$3\hat{\epsilon}$
RMSE for $\hat{\tau}_{20}^{TE}$	33.27	33.27	33.27	33.27	33.27
RMSE for $\hat{\tau}_{20}^{c}$, $k = 2$	8.93	8.81	8.69	8.55	8.42
RMSE for $\hat{\tau}_{20}^{c}$, $k = 5$	5.12	6.20	7.11	7.91	8.64

Table 2.7: Root mean squared errors (RMSE) for $\hat{\tau}_{20}^{TE}$, $\hat{\tau}_{20}^{c}$ with $k = 2$ and $\hat{\tau}_{20}^{c}$ with $k = 5$

2.7.4 Additional simulation results for estimation under stability assumption

Parameters for Erdős-Rényi Model

For the simulation study in Section 2.4.2, we use $p = 0.1$ for $n = 50$ and then scale the probability p accordingly for larger n so that each unit has the same expected number of neighbors.

The effect of estimated stability parameter

Recall that our $\hat{\epsilon}$ is only a lower bound of the true ϵ, hence may underestimate ϵ. To investigate how our estimate of ϵ affects the results, we fix $n = 50$ and generate the social network according to Erdős-Rényi Model with $p = 0.1$. We generate 500 realizations of assignments and plug in $\hat{\epsilon}$, $1.5\hat{\epsilon}$, $2\hat{\epsilon}$, $2.5\hat{\epsilon}$ and $3\hat{\epsilon}$ for three kinds of estimators considered above. Table 2.7 shows the results. We see that the convex combination type estimator with $k = 2$ is not sensitive to the estimate of ϵ while the convex combination type estimator with $k = 5$ is. Even we use $3\hat{\epsilon}$, two convex combination type estimators still show better performance in terms of root mean squared error.

Confidence Interval	Network 1	Network 2	Network 3
Gaussian CI with variance estimated by $\widehat{\text{Var}}^d$	27.38	26.62	27.02
Gaussian CI with variance estimated by $\widehat{\text{Var}}^u$	34.04	32.34	33.33
Chebyshev CI with variance estimated by $\widehat{\text{Var}}^d$	62.47	60.75	61.66
Chebyshev CI with variance estimated by $\widehat{\text{Var}}^u$	77.67	73.79	76.04

Table 2.8: Lengths of two approximate confidence intervals for τ_t^{TE} with $k = 2$

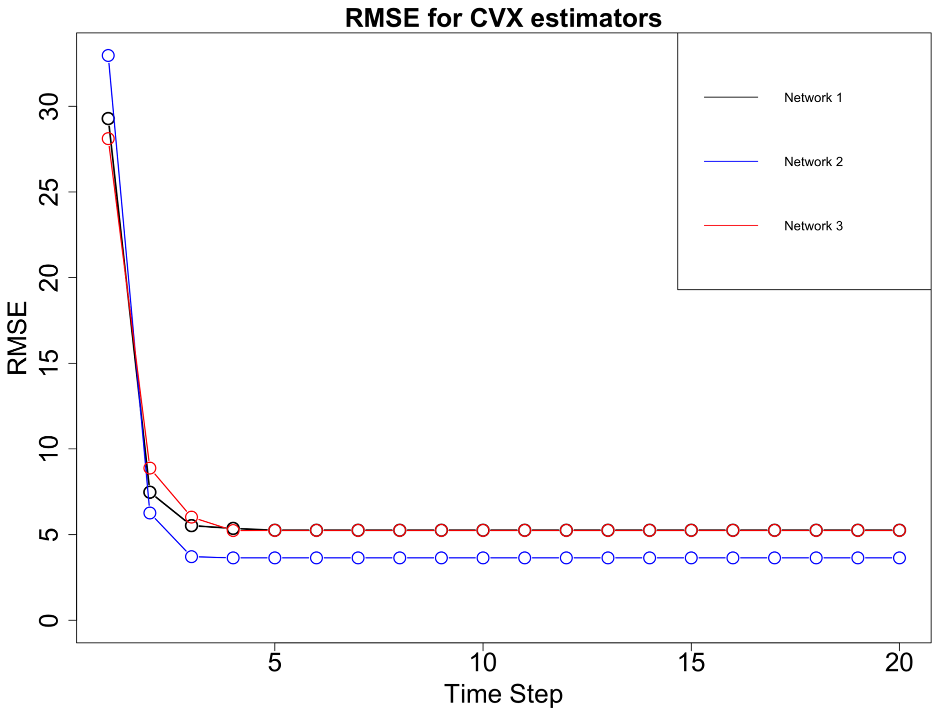


Figure 2.5: Root mean squared errors (RMSE) for $\hat{\tau}_{20}^{TE}$ and $\hat{\tau}^c$

The effect of the number of time steps

Finally, we investigate how k affects the results. We generate three different social networks, and for each one, we plot the root mean squared errors of using 1 time step (i.e., the Horvitz-Thompson type estimator) to 20 time steps (i.e., we use all time steps to estimate the total effect at time step 20). From Figure 2.5 we can see that the RMSE curves stay flat after a certain value of k. Hence, we do not need to worry about using too many time steps as the optimization problem intrinsically pick the right k.

Lengths of approximate confidence intervals

Table 2.8 shows the average lengths of approximate confidence intervals. As expected, Gaussian confidence intervals are shorter.

2.7.5 General Framework

In this chapter, we have a two-dimensional indexing set: one dimension for indexing the multiple units, one dimension for indexing the time. This can be generalized to two arbitrary indices. For example, each place on earth can be indexed by latitude and longitude. We can talk about causal inference in this general case.

We start with an arbitrary indexing set $\mathcal{A}$. Corresponding to each element $a \in \mathcal{A}$, we have an assignment w_a. Hence, there is an assignment array $\underline{w} = (w_a)_{a \in \mathcal{A}}$ associated with $\mathcal{A}$. For each element $a \in \mathcal{A}$, we associate an exposure mapping $f_a : \Omega(\mathcal{A}) \to \Delta$ with it, where $\Omega(\mathcal{A})$ represents all the possible assignment arrays on our indexing set $\mathcal{A}$. Note that although we restrict all the exposure mappings $(f_a)_{a \in \mathcal{A}}$ to have the same range, we do not restrict them to have the same image. We adopt the potential outcome framework and associate each $a \in \mathcal{A}$ a set of potential outcomes $\{Y_a(\underline{w})\}_{\underline{w} \in \Omega(\mathcal{A})}$. Under this general setting, we have the following definition of properly specified exposures:

Definition 20 ($\mathcal{A}$-Properly Specified Exposures)*.* We say that $(f_a)_{a \in \mathcal{A}}$ is $\mathcal{A}$-properly specified if $\forall a \in \mathcal{A}$, $\forall \underline{w}, \underline{w}^{'} \in \Omega(\mathcal{A})$, we have

$$f_a(\underline{w}) = f_a(\underline{w}^{'}) \implies Y_a(\underline{w}) = Y_a(\underline{w}^{'})$$

In the common causal inference literature, such exposure mappings induce interference and thus quantify our belief of the interference mechanism. On the other hand, properly specified exposure mappings reduce the number of possible potential outcomes and hence make inference possible. Two familiar examples are:

Example 5 (Traditional Causal Inference with Interference)*.* This is the setting discussed in Aronow and Samii [2017]. Under this setting, $\mathcal{A} = \mathcal{I} = \{1, \cdots, n\}$, where n is the number of total units in the experiment.

Example 6 (Time Series Experiments)*.* This is the setting discussed in Bojinov and Shephard [2019]. Under this setting, $\mathcal{A} = \mathcal{T} = \{1, \cdots, T\}$, where we have only one unit participating the experiment and we assign treatment or control to this unit at T time points.

The most general causal estimand we are interested in is the following exposure contrast:

Definition 21 (General Exposure Contrast)*.* For $k, k^{'} \in \Delta$ and $\mathcal{A}_0 \subseteq \mathcal{A}$, we define the exposure contrast between k and $k^{'}$ on $\mathcal{A}_0$ as

$$\tau_{k,k'}(\mathcal{A}_0) = \frac{1}{|\mathcal{A}_0|} \sum_{a \in \mathcal{A}_0} (Y_a(k) - Y_a(k^{'}))$$

We have two remarks here. First, this may not be well-defined for all k and $k^{'}$ since the f_a's are not constrained to have the same image. Second, some choices of $\mathcal{A}_0$ do not make sense. We continue our two examples above here. For the traditional causal inference with interference, we average over $\mathcal{A} = \mathcal{I}$ and for time series experiment with one unit, we average over time.

Now, consider the case that $\mathcal{A} = \mathcal{I}_1 \times \mathcal{I}_2$, i.e., we have a two dimensional indexing set. In this case, we have two symmetric parts of the problem: fixing $i \in \mathcal{I}_1$ and do inference on $\mathcal{A}_i = \{i\} \times \mathcal{I}_2$, fixing $j \in \mathcal{I}_2$ and do inference on $\mathcal{A}_j = \mathcal{I}_1 \times \{j\}$. We define two special interference structures on two dimensional indexing sets.

Definition 22 (Purely $\mathcal{I}_1$-level Interference). $\forall t \in \mathcal{I}_2$, $\forall \underline{w}, \underline{w}^{'} \in \Omega(\mathcal{I}_1 \times \mathcal{I}_2)$, we have

$$(w_{i,t})_{i \in \mathcal{I}_1} = (w^{'}_{i,t})_{i \in \mathcal{I}_1} \implies f_{i,t}(\underline{w}) = f_{i,t}(\underline{w}^{'})$$

We can define purely $\mathcal{I}_2$-level interference similarly. We also have two invariant properties of exposure mappings.

Definition 23 ($\mathcal{I}_1$-invariant Exposure Mappings). We say $f_{i,t}$, $(i,t) \in \mathcal{I}_1 \times \mathcal{I}_2$ is $\mathcal{I}_1$-invariant if $\forall t \in \mathcal{I}_2$, $\forall i, i^{'} \in \mathcal{I}_1$, $\forall \underline{w} \in \Omega(\mathcal{I}_1 \times \mathcal{I}_2)$,

$$(w_{i,t})_{t \in \mathcal{I}_2} = (w_{i^{'},t})_{t \in \mathcal{I}_2} \implies f_{i,t}(\underline{w}) = f_{i^{'},t}(\underline{w})$$

Similarly for $\mathcal{I}_2$-invariant exposure mappings.

Chapter 3

Model-Based Regression Adjustment with Model-Free Covariates for Network Interference

3.1 Setup

Consider a randomized experiment on n units where there is a simple undirected graph $G = (V, \mathcal{E})$ that describes the social network of interactions among n units. The graph G is associated with a symmetric matrix $A \in \mathbb{R}^n$ so that $A_{ij} = 1$ if $(i, j) \in \mathcal{E}$ and zero otherwise. Let $\mathcal{N}_i^{(k)}$ denote the k-hop neighborhood around each node $i \in V$. We omit the superscript when $k = 1$ and let d_i denote the degree of each node (or equivalently, $d_i = |\mathcal{N}_i|$). We denote by W_i the random assignment and $x_i \in \mathcal{X}$ the pre-treatment covariates for unit i. We assume that the experimental population is the population of interest and hence view pre-treatment covariates as fixed. We only consider binary treatments but note that extensions to non-binary treatments are straightforward. Throughout, we use lower case letters with the appropriate subscript for realizations of the random variables and for non-random quantities.

We work under the Rubin causal model Rubin [1974], Holland [1986], Imbens and Rubin [2015]. For every unit i, we associate it with potential outcomes $Y_i(w) \in \mathbb{R}$ for $w \in \{0, 1\}^n$. We are interested in the following causal estimand that we call the Global Average Treatment Effect (GATE):

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i(\mathbf{1}) - Y_i(\mathbf{0})]. \tag{3.1}$$

Here $\mathbf{1}$ denotes the n-dimensional ones vector and similarly for $\mathbf{0}$. The GATE estimand, also known as the Total Treatment Effect (TTE) in some work Yu et al. [2022], measures the overall effect of the intervention on the experimental units. Under SUTVA, the assignments of other units won't affect one's response and hence there are only two potential outcomes per unit, $Y_i(0)$ and $Y_i(1)$. Under SUTVA, the GATE is then simply the average treatment effect (ATE). When there is interference along a network, there may be up to 2^n different potential outcomes per unit. In the absence of further assumptions, it is impossible to observe $Y_i(\mathbf{1})$ for some unit i and also observe $Y_j(\mathbf{0})$ for any other unit j.

In this work we take a regression perspective and assume two functions f_0 and f_1 such that for each unit i and each assignment vector $w \in \{0,1\}^n$,

$$Y_i(w) = w_i f_1(i, w, x_i, G) + (1 - w_i) f_0(i, w, x_i, G) + \epsilon_i, \tag{3.2}$$

with ϵ_i's being exogenous, i.e. $\mathbb{E}[\epsilon_i|w] = 0$. The functions f_0 and f_1 each take as input the node label i, the assignment vector w, the covariate vector x_i and graph G. This approach uses exposure mappings Aronow and Samii [2017] as functions that map an assignment vector w and x_i to a specific exposure value so that if two assignment vectors w and w' induce the same exposure value for a unit then they have the same value of potential outcome. Since the potential outcomes only depend on the exposure values, we can view them as a function of exposure values and we can rewrite the potential outcomes as in (3.2). Given (3.2), since functions f_1 and f_0 are shared across all units, we can use the treated units to estimate f_1 and control units to estimate f_0. Suppose $\hat{f}_0$ and $\hat{f}_1$ are two estimates of f_0 and f_1 respectively, then a natural estimator of the GATE would be

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}[\hat{f}_1(i, \mathbf{1}, x_i, G) - \hat{f}_0(i, \mathbf{0}, x_i, G)].$$

Unfortunately, estimation of the GATE will be impossible without any further assumptions on the structure of the functions f_0 and f_1[1]. To motivate our structural assumptions on f_0 and f_1, we look at the following example.

Example 7 (Linear-in-means model). Consider the structural model Manski [1993], Moffit [2001], Bramoullé et al. [2009]

$$\mathbf{y} = \alpha\mathbf{1} + \beta\tilde{A}\mathbf{y} + \gamma\mathbf{w} + \delta\tilde{A}\mathbf{w} + \boldsymbol{\epsilon}, \quad \mathbb{E}[\boldsymbol{\epsilon}|\mathbf{w}] = 0, \tag{3.3}$$

[1] Basse and Airoldi Basse and Airoldi [2018] has a discussion from an inference perspective.

where $\mathbf{y}$ is the $n \times 1$ outcome vector, $\tilde{A}$ is the degree-normalized adjacency matrix, i.e., $\tilde{A}_{ij} = A_{ij}/d_i$, $\mathbf{w}$ is the assignment vector, and $(\alpha, \beta, \gamma, \delta)$ are parameters. Bramoullé et al. Bramoullé et al. [2009] show that under some mild conditions on the coefficients and the graph G, we can rewrite the above model as

$$\mathbf{y} = \alpha/(1-\beta)\mathbf{1} + \gamma\mathbf{w} + (\gamma\beta + \delta)\sum_{j=0}^{\infty} \beta^j \tilde{A}^{j+1}\mathbf{w} + \sum_{j=0}^{\infty} \beta^j \tilde{A}^{j+1}\boldsymbol{\epsilon}. \tag{3.4}$$

Note that now the outcome is linear in the assignment vector $\mathbf{w}$ as well as $\{\tilde{A}^{j+1}\mathbf{w}\}_{j=0}^{\infty}$. Let $f_0(i, w, x_i, G) = f_1(i, w, x_i, G) = \alpha/(1-\beta) + \gamma w_i + (\gamma\beta + \delta)\sum_{j=0}^{\infty} \beta^j \tilde{A}^{j+1} w$ and notice that $\mathbb{E}[\sum_{j=0}^{\infty} \beta^j \tilde{A}^{j+1}\boldsymbol{\epsilon}|w] = 0$. Thus, the linear-in-means model (3.3) can be written in the form of (3.2).

While in this example the linear model is infinite-dimensional, the linear structure of (3.4) motivates us to look at linear models for both f_0 and f_1. To make it formal, we make the following definition:

Definition 24 (Linear interference). We say that the model $\mathcal{Y} = \{Y_i(w) : w \in \{0,1\}^n, i \in [n]\}$ exhibits *linear interference* if there exists a function $g : [n] \times \{0,1\}^n \times \mathcal{X} \times \mathcal{G} \to \mathbb{R}^K$ and $\theta_0 \in \mathbb{R}^K$, $\theta_1 \in \mathbb{R}^K$ such that $f_0(i, w, x_i, G) = \theta_0^T g(i, w, x_i, G)$ and $f_1(i, w, x_i, G) = \theta_1^T g(i, w, x_i, G)$. We call each coordinate function g_j of g a *feature* of the interference.

Despite the simplicity of linear interference, from a graph perspective it can be shown that convolutions on graphs can be well-approximated by linear expansion Hammond et al. [2011]. Such a linear interference assumption is not uncommon Deng et al. [2013], Pouget-Abadie et al. [2019a], Chin [2019]. Chin Chin [2019] shows how to do inference once we have access to the oracle g while Pouget-Abadie et al. [2019a] give a testing procedure to detect network interference under linear interference. Moreover, because we are interested in the quality of our estimated functions $\hat{f}_0$ and $\hat{f}_1$ for (only) $w = \mathbf{0}, \mathbf{1}$, we are effectively attempting generalization. Simple models usually generalize well Bousquet et al. [2004], von Luxburg and Schölkopf [2011], and thus linear interference provides credibility of inference without losing flexibility in a world where g can be arbitrarily complex.

Before proceeding, we can simplify (3.2) somewhat. Note that

$$\begin{aligned} Y_i(w) &= w_i f_1(i, w, x_i, G) + (1 - w_i) f_0(i, w, x_i, G) + \epsilon_i \\ &= w_i f_1(i, w^{(i\to 1)}, x_i, G) + (1 - w_i) f_0(i, w^{(i\to 0)}, x_i, G) + \epsilon_i \\ &= w_i \tilde{f}_1(i, w^{(-i)}, x_i, G) + (1 - w_i) \tilde{f}_0(i, w^{(-i)}, x_i, G) + \epsilon_i, \end{aligned} \tag{3.5}$$

where $w^{(i\to t)}$ denotes the $n-$dimensional vector that replaces w_i by t and $\tilde{f}_t$ is a function of $i, w^{(-i)}$, x_i and G only. Therefore, without loss of generality, we assume that the domain of g and hence the domain of f_0 and f_1 is $[n] \times \{0,1\}^{n-1} \times \mathcal{X} \times \mathcal{G}$.

From here on, for presentational simplicity we will omit the pre-treatment covariates x_i in our discussion. Extensions to the case of including pre-treatment covariates will be discussed when not obvious. As a result, g is a function of the node label i, the assignment vector w and the graph G only.

We focus on design that satisfies the following uniformity assumption:

Assumption 25 (Uniformity). We assume that W_i's are independent and $\forall i$, $\mathbb{P}(W_i = 1) = p_i$ for some $0 < p_i < 1$.

We make this assumption to follow the common practice of using Bernoulli randomization in network experiments, e.g., Karrer et al. Karrer et al. [2021]. As an alternative, estimates from designs that accounts for network interference (for example, graph cluster randomization) may suffer from sizable variance Ugander and Yin [2020]. Hereinafter we assume that W_i's are i.i.d. Bernoulli(p) random variables with $0 < p < 1$, i.e., we work with data from experiments under a Bernoulli design.

If we know the function g a priori, Chin Chin [2019] provides a complete solution. However, if we don't know the function g, then there are three significant challenges, all of which we address in this work. First, how should we *construct* g so that the one we construct approximates the true one? Second, suppose we have many candidate functions then how should we *select* among them? Third, even if we have satisfactory answers to the first two questions, how should we do inference? We will address the first two challenges in the next section and the third challenge later.

3.2 Model-free covariates

Now by (3.5), the function g from Definition 24 takes node label $i, w^{(-i)}$ and G as input and outputs a K-dimensional vector, what g essentially does is to produce K covariates based on $w^{(-i)}$ and G for each unit i. In this section, we describe a sequential procedure to generate and select model-free covariates. A high-level description of our method would be that we generate rich candidate features based solely on the graph structure as well as the assignment vector and select among these features based on the observed outcomes. We first give the procedure in Algorithm 2 below and then explain the steps in more detail. We call the procedure ReFeX-LASSO as it builds on the graph mining technique ReFeX Henderson et al. [2011] to generate candidate features while using LASSO Tibshirani [1996] to select features.

ReFeX (Recursive Feature eXtraction) was originally designed to generate features for graph mining tasks and can be viewed as a recursive algorithm that starts with base features of each node in the graph and iteratively (i) adds and (ii) prunes features based on aggregations over features from neighboring nodes. ReFeX can be viewed as a simple early precursor to recent methods for graph representation learning based on graph convolution networks (GCNs) Hamilton et al. [2017], Kipf and Welling [2017]. We adopt the feature generation step in ReFeX algorithm, but replace the feature pruning part of the original algorithm by LASSO, a modification that allows us to more precisely characterize the features that are available at any given step of the algorithm.

ReFeX has two ingredients—base features and aggregation functions. Given w, $\{x_i\}_{i=1}^n$ and

Algorithm 2 ReFeX-LASSO

```
Input: Graph $G = (V, \mathcal{E})$, assignment vector $w \in \{0,1\}^n$, maximum number of iterations $T$.
Output: A set of covariates $S$.
Initialize $S = \{\}$, active feature set $A = \{\}$.
For each node/unit $i$, construct $m$ base features and add $m$ base features to $A$.
for $t = 1$ to $T$ do
    Regress $y$ on $w$ and features from $S$ and $A$ using LASSO with no penalty on features from $S$.
    If no feature in $A$ is selected, return $S$. Otherwise, add selected features from $A$ to $S$.
    Recursively construct features by performing aggregations of features in $A$ over neighbors in 1-hop neighborhood.
    Delete old features in $A$ and add those new features to $A$.
end for
Return $S$.
```

G, base features are those features that can be constructed by only looking at each node's 1-hop neighborhood. They can be arbitrary as long as they satisfy this local look-up constraint. Base features can be purely graph features like degree, centrality, clustering coefficient, etc. They can also be pre-treatment covariates x_i. Often we would also like to have base features that depend on not just one input of the function g but features computed from two inputs of g. For example, features like the number of treated neighbors, which depends on both the assignment vector w as well as the graph G. Or the average feature value over all neighbors, which depends on the pre-treatment covariates and G. With ReFeX, the base features are chosen by the analyst. Aggregation functions are functions that take features from neighboring nodes as inputs and output a single value. Hence, one aggregation function essentially computes a statistic based on the sample of feature values from neighbors. The aggregation functions again can be arbitrary and chosen by the analyst. Some common examples include min, max, sum, mean and variance Henderson et al. [2011].

We are now ready to introduce the ReFeX-LASSO algorithm. The ReFeX-LASSO algorithm starts with two empty feature sets, the target set S and the active feature set A. The first set S stores the selected features and features in S will be used for adjusting the GATE estimate. The active feature set A contains features that were recursively added in the previous step and yet to be selected. At the beginning of the procedure, we construct base features for each unit i. Equipped with a set of base features, each time we regress the outcome vector y on features from both set S and set A using LASSO. The LASSO regularization parameter can be chosen by cross-validation and hence we do not need extra hyper-parameters of the algorithm. Note that we do not put a penalty on features in S since they have already been selected and should be kept. The intuition behind this step is that in general features generated later (pulling information from farther in the graph) should not be more predictive than features selected previously. Next, depending on the number of newly selected features, we either terminate the construction and return the current S or add those selected features to S and proceed with the recursive construction. We then need to generate new features and add them to A. To do so, we now perform aggregations on old features over all neighboring units. Finally, we add those features to A and delete all old features in A.

The maximum number of iterations in Algorithm 2 limits the distance in the graph that we can pull information from. Although each step only performs aggregations over neighbors in the 1-hop neighborhood, by repeatedly performing the aggregations we are able to construct features that are informative for the k-hop neighborhood. To illustrate this point, we give an example.

Example 8 (ReFeX and multi-hop information). Suppose one of the base features we use in ReFeX-LASSO is the fraction of treated neighbors,

$$\rho_i = \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} w_j,$$

and supposed we limit ourselves to mean aggregation, i.e., we look at each unit's neighbors and aggregate their fraction of treated neighbors using a mean function. We call this new feature $\tilde{\rho}_i$. We then have that

$$\begin{aligned}
\tilde{\rho}_i &= \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} \rho_j \\
&= \frac{1}{d_i} \sum_{j \in \mathcal{N}_i} \frac{1}{d_j} \sum_{k \in \mathcal{N}_j} w_k \\
&= \sum_{j=1}^{n} \frac{A_{ij}}{d_i} \sum_{k=1}^{n} \frac{A_{jk}}{d_j} w_k \\
&= \sum_{j=1}^{n} \tilde{A}_{ij} \sum_{k=1}^{n} \tilde{A}_{jk} w_k \\
&= [\tilde{A}^2 w]_i,
\end{aligned}$$

where A and $\tilde{A}$ are the same as defined in the linear-in-means model example from (3.3). Note that the summand is 1 if and only if A_{ij}, A_{jk} and w_k are all 1s. In other words, if we ignore the normalizing terms, the sum essentially represents the number of length-2 paths in G that start at unit i and arrive at a treated unit. With the normalizing terms, it is close to the fraction of such paths among all length-2 paths that start at unit i. Clearly, this feature is informative for unit i's 2-hop neighborhood.

The above example shows the power of recursion. It allows us to have access to information about much larger neighborhoods without actually looking up all units in larger neighborhoods. In fact, the ReFeX component of ReFeX-LASSO is very efficient in terms of computational complexity Henderson et al. [2011], making the procedure ideal for large-scale experiments on online platforms where network interference is ubiquitous. Another advantage of our algorithm is that all the covariates generated are model-agnostic or model-free—we do not generate them according to any particular response model (or graph model). Since the aggregation functions are arbitrary, ReFeX can quickly generate a very large number of features, even for modest iterations budgets T. Despite the fraction of treated neighbors we just saw, we are also able to get the number of treated neighbors for each unit by using sum as the aggregation function. In general, using more complicated aggregation functions yields more complicated features. Thus, the recursive step offers rich features for each unit.

With minor modifications we can see that all pruning steps in our procedure can be grouped together and done *ex ante*, i.e., before running the experiment and observing the outcomes. Then, after the experiment, we use the observed outcomes to select covariates among all the covariates we have generated. This method has certain advantages, so for completeness we give such a modified version of ReFeX-LASSO below in Algorithm 3, calling it post-ReFeX-LASSO.

Algorithm 3 post-ReFeX-LASSO

```
Input: Graph $G = (V, \mathcal{E})$, assignment vector $w \in \{0,1\}^n$, maximum number of iterations $T$.
Output: A set of covariates $S$.
Initialize $S = \{\}$.
For each node/unit $i$, construct $m$ base features and add $m$ base features to $S$.
for $t = 1$ to $T$ do
    Recursively construct features by performing aggregations of features in $S$ that were added
    in the previous iteration over neighbors in 1-hop neighborhood.
    Add those newly constructed features to $S$.
end for
Regress $y$ on $w$ as well as features from $S$ using LASSO.
Keep selected features in $S$ and remove other features from $S$.
Return $S$.
```

An operational advantage of post-ReFeX-LASSO is that two parts of the algorithm, feature generation and selection, can be done separately. However, in practice we find that post-ReFeX-LASSO leads to estimates with larger variance. Our explanation for this increased variance is two-fold. First, since the number of features generated from ReFeX may be large, separating the generation step and the selection step seems to make the selection step unstable. Second, many of the features generated along the way of post-ReFeX-LASSO are correlated and including all of them simultaneously leads to greater uncertainty in terms of features being selected. Hence, it leads to estimates with larger variance and we recommend ReFeX-LASSO over post-ReFeX-LASSO in all use cases when operationally feasible.

3.3 Inference with model-free covariates

In the previous section, we gave a sequential procedure that outputs a set of covariates S that can be used for regression adjustments when estimating GATEs. This section devotes to inference with model-free covariates. We first discuss how to use model-free covariates returned from ReFeX-LASSO or post-ReFeX-LASSO to do regression adjustment. Following that, we show one selection property of ReFeX-LASSO. We then give theoretical properties of regression adjustment estimator of the GATE using model-free covariates as well as a simple way to construct confidence interval for τ.

3.3.1 Estimation

Let $u_i^1, \cdots, u_i^K$ denote the K covariates returned by ReFeX-LASSO or post-ReFeX-LASSO for unit i and let $u_i = \left[u_i^1, \cdots, u_i^K\right]^T \in \mathbb{R}^K$ be the whole feature vector for unit i. We further let $\hat{g}$ be the function that maps (i, w, x_i, G) to u_i for each unit i. Finally, we denote by n_c the number of control units and n_t the number of treated units with $n_c + n_t = n$.

To estimate the GATE, we fit two linear models on control and treated units using u_i's. Ideally, we hope that there exist vectors β_0, β_1 such that $\beta_0^T u_i$ and $\beta_1^T u_i$ are good approximations of f_0 and f_1. To be specific, we first run an ordinary least squares with observations that are from the control group only and obtain $\hat{\beta}_0$. We then run ordinary least squares again, but now with observations that are from treatment group only and obtain $\hat{\beta}_1$. Meanwhile, the features u_i are all features under the treatment assignment w for which the responses were collected. To estimate the GATE, we are interested not in the response under u_i as it was, but u_i as it would be if $w = \mathbf{0}$ or $w = \mathbf{1}$. We thus pass $\mathbf{0}$ and $\mathbf{1}$ to $\hat{g}$ to obtain the feature vectors u_i^{gc} and u_i^{gt} under global control and global treatment, respectively.

Combining the coefficient estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ with the vectors u_i^{gc} and u_i^{gt}, our estimate of the GATE is then simply

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\beta}_1^T u_i^{gt} - \hat{\beta}_0^T u_i^{gc}). \tag{3.6}$$

Though assuming a linear model is restrictive, as we discussed previously, if we are able to generate predictive features then the linear model can be a good approximation to the true model. ReFeX-LASSO or post-ReFeX-LASSO helps us choose good features to adjust for and thus both reduce the variance of the estimate[2] and reduce the bias we typically incur when ignoring interference.

3.3.2 Selection properties

Before we delve into inference details, we first discuss selection properties of ReFeX-LASSO, drawing inspiration from prior work on Sequential LASSO Luo and Chen [2014]. To this end, we introduce some additional notation. For each iteration t, let $\{u_1^t, u_2^t, \cdots, u_{i_t}^t\}$ be the set of features generated in the ReFeX step of ReFeX-LASSO and s_{*t} be the selected features at the t-th iteration (note that s_{*t} may contain features that were selected in previous iterations and thus are not in the set $\{u_1^t, u_2^t, \cdots, u_{i_t}^t\}$). Moreover, we let $\mathcal{R}(s)$ to denote the space spanned by features in s.

[2] In fact, in the case of no interference, Lin Lin [2013] shows that doing linear adjustment can only improve the precision.

Proposition 14. *For $t \geq 1$ and any $j \in \{1, \cdots, i_{t+1}\}$, if $u_j^{t+1} \in \mathcal{R}(s_{*t})$ then $j \notin s_{*(t+1)}$.*

This first proposition implies two things. First, we have a full rank design matrix at each iteration. Second, the subsequent selection will disregard the features that are highly correlated with the existing ones and hence provides intuition for why the post-ReFeX-LASSO leads to estimate with high variance. Without the sequential procedure of (non-post-) ReFeX-LASSO, two highly correlated features may enter the selection stage together.

Proposition 15. *Our selection is nested in the sense that $s_{*1} \subseteq s_{*2} \subseteq \cdots \subseteq s_{*T}$.*

This second proposition is relatively self-explanatory and ensures that the sequential procedure actually provides nested feature sets, i.e., by excluding penalties on selected features, we are able to keep them in our feature set S. Though our selection procedure in ReFeX-LASSO is quite different from Sequential LASSO Luo and Chen [2014], the proofs of the above two propositions are analogous to those in Luo and Chen [2014]. There are two key differences between our selection procedure and Sequential LASSO. First, instead of keeping all the features for every iteration, we throw away non-selected features in previous iterations. Second, the features under consideration at each iteration are newly generated features rather than existing features. Put another way, we find that the analysis in Luo and Chen [2014] is robust to such a change in procedure. Note that Sequential LASSO can be used for post-ReFeX-LASSO (but not ReFeX-LASSO) since for post-ReFeX-LASSO we generate all the candidate features in advance. These two propositions together establish two intuitive properties of our selection step in ReFeX-LASSO that we should expect to hold for our purpose. Their proofs can be found in Appendix 3.7.1.

3.3.3 Consistency

We now prove that post-ReFeX-LASSO leads to a consistent estimator of the GATE under standard assumptions one would require for consistency of LASSO. For each unit i, we denote the set of features generated by the ReFeX step in post-ReFeX-LASSO as $\{u_i^1, \cdots, u_i^M\}$. We drop the subscript i when we refer to the jth feature vector, i.e., $u^j = [u_1^j, \cdots, u_n^j]^T$. Furthermore, we assume that there exists a subset $S_* \subset \{u^1, \cdots, u^M\}$ with $|S_*| = s$ such that both f_0 and f_1 are linear in features in S_* with coefficient vectors β_0 and β_1 respectively. Finally, we denote the design matrix when estimating β_0 by U^0 and the design matrix when estimating β_1 by U^1.

Theorem 26. *Suppose that there exists a constant $C > 0$ such that*

$$\max_{j=1,\cdots,M} \frac{\|u^j\|_2}{\sqrt{n}} \leq C,$$

and the two design matrices U^0 and U^1 satisfy the $(\kappa; 3)$-RE condition over S, then $\hat{\tau}$ is consistent for τ.

A proof of Theorem 26 appears in Appendix 3.7.1 and uses mostly standard tools for the study of LASSO ℓ_2-error bounds Wainwright [2019]. The restricted eigenvalue (RE) condition in Theorem 26 is a standard assumption when proving ℓ_2-error bound on the coefficient vector. It restricts the curvature for a specific subset of vectors in the Euclidean space. It is defined as follows Bickel et al. [2009], van de Geer and Bühlmann [2009], Raskutti et al. [2010]

Definition 27. The matrix $\mathbf{X}$ satisfies the restricted eigenvalue (RE) condition over S with parameters $(\kappa; \alpha)$ if

$$\frac{1}{n}\|\mathbf{X}\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2 \qquad \text{for all } \Delta \in \mathbb{C}_\alpha(S),$$

where $\mathbb{C}_\alpha(S) \coloneqq \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \alpha\|\Delta_S\|_1\}$.

Under the assumptions of Theorem 26, we are now able to prove GATE consistency under LASSO-based feature selection in at least simple settings such as the following, an example setting where our feature generation procedure outputs two simple features.

Proposition 16. *Suppose we run a Bernoulli randomized experiment with treatment probability* $0 < p < 1$ *and we only generate two features, the fraction of treated neighbors* ρ_i *and number of treated neighbors* ν_i*. Furthermore, suppose the graph* G *consists of disjoint cliques of size* $3 \leq m_c \leq M$ *(*m_c *is the size of the c-th cluster) for some positive constant* $M \geq 3$*. If the true* f_0 *and* f_1 *are only linear in* ρ_i*, then* $\hat{\tau}$ *is consistent for* τ*.*

The lower bound on m_c is for identifiability since when all clusters have size 2 then ρ_i and ν_i are essentially the same and we end up with completely duplicated features. Notice also that when all m_c's are equal, we end up with perfect co-linearity so in that case we wouldn't consider distinguishing between these two features. While the above result applies only in a simple setting, it is of its own importance. In practice, it is not uncommon to adjust for fraction of treated neighbors and report the resulting estimate as the estimate of the GATE Saint-Jacques et al. [2019], Karrer et al. [2021]. The above proposition shows that when we only want to distinguish covariates between fraction of treated neighbors and number of treated neighbors, LASSO is a handy tool.

3.3.4 Confidence interval via a block bootstrap

Researchers are usually not just interested in a point estimate of the GATE, they also want to know the uncertainty contained in the estimate, e.g., through confidence intervals. ReFeX-LASSO brings flexibility in doing regression adjustment for GATE estimation, but there is no free lunch and it also brings us difficulty in doing inference, i.e., in constructing confidence interval for τ. First, unlike Chin [2019] where one assumes an oracle model, here the true model is unknown. Second, features constructed in Chin [2019] do not use the observed outcomes. With ReFeX-LASSO, though all the features constructed from ReFeX do not use the outcomes, our selections of covariates *depend* on the realized outcomes. Therefore, ReFeX-LASSO leads to an estimator with no clear variance expression. Moreover, since our final estimate depends on the actual selected covariates, we require some technique analogous to post-selection inference as in Lee et al. [2016]. Lee et al. Lee et al. [2016] consider confidence intervals of coefficients conditional on being selected by LASSO. Yet we are interested in the confidence interval of τ, not the coefficients, where our estimate $\hat{\tau}$ is calculated based on the estimated coefficients as well as selected covariates. Because of the combination of these complexities, we are not able to simply import any known results for inference in this setting.

Let us consider the nature of the inference problem we are facing. In general, the randomness of our estimate is incurred not just by the randomness of the potential outcomes but also by the randomness of the assignment vector. To construct the confidence interval, we need to quantify how these two resources of randomness affect our estimate of the GATE. Note that since we know the distribution of the assignment vector, the distribution of a given feature is in fact known. What we don't have a good characterization of is the randomness of the selection procedure incurred by the randomness of the assignment vector. In other words, we require understanding how the random assignments affect the feature selection procedure.

To tackle this complication, we introduce a way to construct confidence intervals based on a block bootstrap. Ideally if we can do the experiment infinitely many times, we could run 2^n experiments and calculate 2^n estimates of the GATE. A confidence interval for τ could then be derived easily. Our obvious difficulty is then how should we use one single sample to approximate the sample randomness. We turn to the block bootstrap Efron [1979], Efron and Tibshirani [1994], Cameron et al. [2008]. The intuition of this usage is that features of units are correlated according to the particular graph structure of G and hence by sampling clusters (which we expect to be relatively disconnected) we are able to keep the bootstrap sample looking like the original sample. On the other hand, resampling units will fail as it cannot replicate the underlying correlation structure in the data. Though we do not provide theoretical guarantees, we will show that in practice the coverage is good and the resulting confidence intervals are of reasonable width. We also note in passing that recent results in Kojevnikov [2021] demonstrate that there is a version of block bootstrap that does provide theoretical guarantee for certain highlu stylized network processes.

Example 9. Consider the case where our social network G consists of C disjoint cliques $\mathcal{C}_1, \cdots, \mathcal{C}_C$ of size m. Units are fully connected within each clique. This setup can be viewed as a special case of the household experiment studied in Basse and Feller [2018]. In such a case it is natural to consider sampling all C cliques with replacement to get a bootstrap sample. For network dependent processes satisfying certain technical assumptions, this sampling process is the correct thing to do using arguments in Kojevnikov [2021]. Suppose we have a network dependent process $\{Y_n, G_n\}$ that satisfies assumptions in Kojevnikov [2021]. To make block bootstrap consistent, i.e., producing a confidence interval that is consistent in level, Assumption 4.1 in Kojevnikov [2021] needs to hold. We first introduce some notations used in Kojevnikov [2021]. Let $N_n(i; s)$ denote the open neighborhood of radius $s > 0$ around $i \in N_n$, i.e,

$$N_n(i; s) := \{j \in N_n : d_n(i, j) < s\}.$$

We define the following aggregate measures of the network denseness:

$$\delta_n(s) := n^{-1} \sum_{i \in N_n} |N_n(i; s+1)|, \quad D_n(s) := \max_{i \in N_n} |N_n(i; s+1)|.$$

Moreover, let

$$\Delta_n(s; k) := \frac{1}{n} \sum_{i \in N_n} ||N_n(i; s+1)| - \delta_n(s)|^k,$$

which is the k-th absolute central moment of the sizes of the $(s+1)$-neighborhoods. It is easy to verify that in our case, $\delta_n(s_n) = m$, $\Delta_n(s_n, 2) = 0$, and $D_n(s_n) = m$ for $\forall s_n \geq \max_c \operatorname{diam}(\mathcal{C}_c)$, since our graph consists of non-overlapping blocks with equal size m. Now, we let

$$\omega_n(i, j) := \frac{|N_n(i; s_n+1) \cap N_n(j; s_n+1)|}{\delta_n(s_n)}.$$

Then,

$$\omega_n(i, j) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are in the same cluster,} \\ 0 & \text{otherwise.} \end{cases}$$

and $\omega_n(j) = \omega_n(j, j) = 1$ for all $j \in [n]$. With these values, we immediately see that the Assumption 4.1 in Kojevnikov [2021] holds as long as $m = o(n)$. Since the only remaining assumptions needed to make block bootstrap consistent are about the network dependent process itself, we can conclude that block bootstrap would be valid in this toy model for network dependent processes given in Kojevnikov [2021].

We present two versions of block bootstrap here, one for regression adjustment with post-ReFeX-LASSO and one for regression adjustment with ReFeX-LASSO. Before actually giving the two block bootstrap procedures, we first introduce the key ingredient in our block bootstrap procedure, a randomized graph clustering algorithm. Our block bootstrap procedure involves partition the graph into several clusters. The generic algorithm we use is k-hop-max clustering Ugander and Yin [2020], a simple adaptation of the CKR partitioning algorithm Calinescu et al. [2005]. The details are shown in Algorithm 4. The algorithm provides a random clustering of the graph that depends on random initial conditions. The algorithm is light in computation when $k = 1$ as we only need to look at one's direct neighbors. Also, it returns neighborhood-like clusters. As a remark connecting back to above example, if our graph consists of disjoint fully connected clusters then 1-hop max clustering is able to return exactly these clusters as final output. In general, when $k > 1$, we obtain larger clusters that are centered around fewer nodes.

Algorithm 4 k-hop-max graph clustering

```
Input: Graph $G = (V, E)$.
Output: Graph clustering $\mathcal{C}_1, \cdots, \mathcal{C}_c$ where each $\mathcal{C}_j$ contains a collection of nodes.
for $i \in V$ do
    $X_i \leftarrow \mathcal{U}(0, 1)$;
end for
for $i \in V$ do
    $i \leftarrow$ argmax([$X_j$ for $j \in B_k(i)$]) where $B_k(i)$ is the $k$-hop neighborhood of node $i$ (including itself);
end for
Return $\mathcal{C}_1, \cdots, \mathcal{C}_c$.
```

We first present the block bootstrap procedure for post-ReFeX-LASSO, given in Algorithm 5. With post-ReFeX-LASSO, the bootstrap procedure is simpler since the feature generation and selection part are separated. Unlike the usual bootstrap where we sample random individual units with replacement, here we sample random clusters from the graph clustering algorithm with replacement. The intuition is that features u_i of units are correlated according to the particular graph structure of G and hence by sampling clusters, which we expect to be relatively disconnected, we are able to keep the bootstrap sample "looking like" the original sample. As a specific caveat, though in expectation the bootstrap sample has sample size n, if we do not have uniformly sized clusters, then the bootstrap sample may end up with much larger or smaller sample size. Hence we run the graph clustering algorithm l times and for each clustering we run block bootstrap with the number of bootstrap replicates B. We use $k = T^* + 1$ for k-hop-max clustering in Algorithm 5 where T^* is the number of iteration where there were features still got selected (since if no feature got selected in the $(T^* + 1)$-th iteration then interference should happen within $(T^* + 1)$-hop neighborhood).

Next we present the version of block bootstrap with ReFeX-LASSO, given in Algorithm 6. Note that we cannot simply use the same algorithm since it performs feature generation and feature selection concurrently. Compared to Algorithm 5, T^* now represents the stopping time of ReFeX-LASSO. Meanwhile, similar to Algorithm 5, the bootstrap sample is only used in the feature selection step of ReFeX-LASSO. That being said, for each iteration, we still use the same graph G to generate features but then we use the bootstrap sample of these features to do selection. The intuition behind using the original graph is that we view the graph as fixed and the correlation structure of all features are then induced by this graph. Therefore, we do not paste all sampled clusters together to form a new graph to generate features for next iteration. On the other hand, if we do believe that the graph is generated from some random process then we may also reconstruct the graph from sampled units by pasting all sampled clusters together.

In the above two algorithms, we utilize a randomized graph clustering algorithm that can be easily implemented. Of course, this is not the only possible choice for the graph clustering algorithm one can use. We note by passing that there are many graph clustering algorithms available for

Algorithm 5 Block bootstrap for post-ReFeX-LASSO

```
Input: Graph $G = (V, \mathcal{E})$, assignment vector $w \in \{0,1\}^n$, number of bootstrap samples $B$.
Output: Confidence interval for $\tau$.
Collect the assignment $w_i$, features $u_i^1, \cdots, u_i^M$ in $S$ generated before running LASSO, outcome $y_i$ for each unit $i$. Record the maximum iteration number $T^*$ where one of the features generated at that iteration was selected.
for $r = 1$ to $\ell$ do
    Use $k$-hop max clustering algorithm with $k = T^* + 1$ to divide $n$ units into $C$ clusters $\mathcal{C}_1, \cdots, \mathcal{C}_C$.
    for $b = 1$ to $B$ do
        Sample $C$ clusters with replacement from $\mathcal{C}_1, \cdots, \mathcal{C}_C$.
        Construct the $b$-th bootstrap sample $(w^b, u^{1,b}, \cdots, u^{M,b}, y^b)$ with units from sampled clusters.
        Regress $y$ on $w$ as well as $M$ features using LASSO.
        Compute the estimate $\hat{\tau}^b$ using selected features and the bootstrap sample.
    end for
end for
Compute the $\alpha/2$-th quantile $q^*_{\alpha/2}$ and the $(1-\alpha/2)$-th quantile $q^*_{1-\alpha/2}$ of the sample of all bootstrap estimates $\hat{\tau}^1, \cdots, \hat{\tau}^{\ell B}$.
Return $\left[q^*_{\alpha/2}, q^*_{1-\alpha/2}\right]$ as the $(1-\alpha) \times 100\%$ confidence interval for $\tau$.
```

practitioners Nishimura and Ugander [2013], Spielman and Teng [2013], Awadelkarim and Ugander [2020], Shi and Chen [2020] that exhibit various properties.

We conclude this section with a discussion of how to suitably choose the sizes of clusters. We consider three scenarios and show why they may fail with heuristics from Kojevnikov [2021]. Though we are not considering the same problem as in Kojevnikov [2021], given that we have a more complicated setup, we do not expect that weaker assumptions than those in Kojevnikov [2021] would be sufficient for good coverage in our case. Therefore, we view assumptions in Kojevnikov [2021] as what we should expect to have in order to make our block bootstrap consistent.

The first scenario that we consider is when we have $O(n)$ clusters with non-constant sizes. Then the second absolute central moment of block sizes may be non-vanishing as $n \to \infty$ but the average block size is $O(1)$. This implies that unless the clusters are relatively uniform, there would be a violation to Assumption 4.1 in Kojevnikov [2021]. As a second scenario, consider the case when we have $O(1)$ clusters. Now the maximum block size must be of order $O(n)$ and the average block size is at most $O(n)$, hence Assumption 4.1 in Kojevnikov [2021] is certainly violated. In general, we don't want to have too many clusters or too few clusters. Finally, then, consider a scenario where we have $\sqrt{n}-1$ clusters of size $\sqrt{n}$ and $\sqrt{n}$ clusters of size 1. Now the average block size is of order $O(n^{1/2})$ and the second absolute central moment of block sizes is not of lower order, which implies that the ratio does not vanish as $n \to 0$ and again Assumption 4.1 in Kojevnikov [2021] is violated. This last example shows that the cluster sizes are not simply a matter of avoiding too big/small or few/many clusters, but instead here we see we cannot have two groups of clusters with different

Algorithm 6 Block bootstrap for ReFeX-LASSO

```
Input: Graph $G = (V, \mathcal{E})$, assignment vector $w \in \{0,1\}^n$, number of bootstrap samples $B$.
Output: Confidence interval for $\tau$.
Collect the assignment $w_i$ and outcome $y_i$ for each unit $i$. Record the stopping time for ReFeX-LASSO $T^*$.
Use $k$-hop max clustering with $k = T^* + 1$ to divide $n$ units into $C$ clusters $\mathcal{C}_1, \cdots, \mathcal{C}_C$.
for $r = 1$ to $\ell$ do
    for $b = 1$ to $B$ do
        Sample $C$ clusters with replacement from $\mathcal{C}_1, \cdots, \mathcal{C}_C$.
        Construct the $b$-th bootstrap sample with units from sampled clusters.
        Rerun ReFeX-LASSO with the original sample for feature generation and the bootstrap sample for feature selection.
        Use the covariates returned from last step as well as the bootstrap sample to get estimate of $\tau$, $\hat{\tau}^b$.
    end for
end for
Compute the $\alpha/2$-th quantile $q^*_{\alpha/2}$ and the $(1-\alpha/2)$-th quantile $q^*_{1-\alpha/2}$ of the sample of all bootstrap estimates $\hat{\tau}^1, \cdots, \hat{\tau}^{\ell B}$.
Return $\left[q^*_{\alpha/2}, q^*_{1-\alpha/2}\right]$ as the $(1-\alpha) \times 100\%$ confidence interval for $\tau$.
```

size magnitudes. In summary, the advice is to use a reasonable number of clusters that have sizes of roughly the same magnitude. What we present in Algorithm 5 and 6 are good default choices if the network is not very dense.

3.4 Simulation experiments

In this section, we use simulations to provide both empirical guidance on our method when theory is lacking and empirical evidence of the usefulness of our method. We make use of the Facebook 100 dataset Traud et al. [2012] of real-world social networks. The networks in this dataset are complete online friendship networks for one hundred colleges and universities collected from a single-day snapshot of Facebook in September 2005. For our simulations we use the network of Swarthmore college students, being of modest size. We extract the largest connected components of the Swarthmore network, obtaining a social network with 1,657 nodes and 61,049 edges. The diameter of the network is 6 and the average pairwise distance is 2.32. Since this network is quite dense, estimation of the GATE would be very difficult when interference is strong. We use this network to demonstrate that even for such a network, we are still able to get relatively good estimates from (post-) ReFeX-LASSO.

We generate an assignment vector using a Bernoulli design with success probability 0.5 and generate outcome variables according to certain models with varying magnitude of network interference; these models are summarized in Table 3.1 and 3.2. We will discuss in detail about these outcome models in Section 3.4.2. Our simulations can be viewed as semi-synthetic experiments—we use a true social network but we generate outcomes according to specified models.

Section 3.4.1 introduces the baseline estimators that we compare with in our simulations. Section 3.4.2 discusses the outcome models that we use for generating the outcomes with various degree of interference. Section 3.4.3 compares the regression adjustment estimator using model-free covariates with those commonly-used estimators in practice as in Section 3.4.1 and demonstrate that it has good performance in terms of root mean squared error. Section 3.4.4 explores the empirical performance of the confidence interval constructed via block bootstrap and discusses some practical aspects in the procedure.

3.4.1 Estimation of the GATE

Our ultimate goal of constructing model-free covariates is to use them in GATE estimation. We first explore the empirical performance of the regression adjustment estimator using model-free covariates. Specifically, we compare it with two kinds of estimators that are commonly used in practice: (i) the difference-in-mean estimator and (ii) a Hájek estimator under a network exposure model Manski [2013]. Difference-in-mean estimator calculate the difference between average outcome among treated units and average outcome among control units:

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

Obviously this estimator ignores interference and will thus incur large bias when interference is significant.

The basic Hájek estimator for the ATE is defined as

$$\hat{\tau}^{\text{Hájek}} = \frac{\sum_{i=1}^n Y_i W_i/\mathbb{P}(W_i=1)}{\sum_{i=1}^n \mathbb{I}(W_i=1)/\mathbb{P}(W_i=1)} - \frac{\sum_{i=1}^n Y_i(1-W_i)/\mathbb{P}(W_i=0)}{\sum_{i=1}^n \mathbb{I}(W_i=0)/\mathbb{P}(W_i=0)}.$$

Here we will consider a version of Hájek estimator that accounts for interference. Manski Manski [2013] studies identification of potential outcome distributions under interference. One concrete example is when one's outcome only depends on one's own assignment as well as the distribution of assignments for his/her neighbors. Ugander et al. Ugander et al. [2013] further considers a fractional exposure model where it is assumed that if one is treated and a $q > 0.5$ fraction of one's neighbors are treated then one's outcome is equal to the potential outcome associated with the assignment vector $\mathbf{1}$. Similarly, in this exposure model if one is not treated and one's fraction of treated neighbors is at most $1 - q$ then one's outcome is equal to the potential outcome associated with the assignment vector $\mathbf{0}$. Formally, $\forall w, w' \in \{0,1\}^n$, this fractional exposure model assumes:

$$w_i = 1, \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} w_j \geq q \implies Y_i(w) = Y_i(\mathbf{1}),$$

and

$$w_i = 0, \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} w_j \leq 1 - q \implies Y_i(w) = Y_i(\mathbf{0}).$$

We can then use a Hájek estimator that corrects for the probability that these conditions are met under a Bernoulli design. Specifically, we define the events $E_i^{1,q} = \{W_i = 1, \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} w_j \geq q\}$ and $E_i^{0,1-q} = \{W_i = 0, \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} w_j \leq 1 - q\}$. The corresponding Hájek estimator under a fractional exposure model is then

$$\hat{\tau}_{q,1-q}^{\text{Hájek}} = \frac{\sum_{i=1}^n Y_i \mathbb{I}(E_i^{1,q})/\mathbb{P}(E_i^{1,q})}{\sum_{i=1}^n \mathbb{I}(E_i^{1,q})/\mathbb{P}(E_i^{1,q})} - \frac{\sum_{i=1}^n Y_i \mathbb{I}(E_i^{0,1-q})/\mathbb{P}(E_i^{0,1-q})}{\sum_{i=1}^n \mathbb{I}(E_i^{0,1-q})/\mathbb{P}(E_i^{0,1-q})}. \tag{3.7}$$

This estimator accounts for interference by taking the assignments of direct neighbors into consideration. If we still assume local interference in the sense that only one's direct neighbors can impact one's response but want a fully agnostic setting then we could choose $q = 1$ (notice that in this case the Hájek estimator is consistent). In our case, the number of neighbors one has is usually quite large and under independent Bernoulli assignment we wouldn't expect to observe many units with all neighbors being treated or not treated. As a bias-variance compromise, we choose $q = 0.8$.

Finally, we also compare our (post-) ReFeX-LASSO regression adjustment estimator with two linear regression adjustment estimators that adjust for specific features. We will describe these two estimators in detail later when we present the simulation results in Section 3.4.3. For post-ReFeX-LASSO and ReFeX-LASSO, we choose $T = 2$ and the base features to be fraction of treated neighbors, number of treated neighbors, fraction of edges in neighborhood that connects a treated unit and a control unit and also fraction of edges in neighborhood that connects a treated unit and a treated unit. For aggregation functions in (post-) ReFeX-LASSO, we use both the mean and variance.

3.4.2 Outcome models

Here we describing the outcome models we use in our simulation study. We carry forward the notation as in Proposition 16, using ρ_i to denote the fraction of treated direct neighbors for unit i and ν_i to denote number of treated direct neighbors.

We first consider estimation under linear interference. The first model is a linear model in both number of treated neighbors and fraction of treated neighbors. Such model is also considered in

Pouget-Abadie et al. [2019a] and Chin [2019]. Specifically,

$$f_0(w, G) = \alpha_0 + \xi_0 \rho_i + \gamma_0 \nu_i \tag{3.8}$$

and

$$f_1(w, G) = \alpha_1 + \xi_1 \rho_i + \gamma_1 \nu_i. \tag{3.9}$$

The difference $\alpha_1 - \alpha_0$ can be viewed as the primary effect of the treatment and coefficients (ξ_w, γ_w) for $w = 0, 1$ govern how the unit respond to treatment and control, respectively. In particular, if $\xi_w = \gamma_w = 0$ then there is no interference and we are back to usual setup of ATE estimation under SUTVA. Note that for this model, there is no interference beyond the 1-hop neighborhood and hence the estimation problem is considerably easier. We will refer to this response model as simple linear interference.

Building on the discussion of the linear-in-means model in the introduction, we also consider a response model where the interference propagates out to k-hop neighborhoods for $k \geq 2$. This model can be viewed as a truncated linear-in-means model; instead of summing up to infinity, we truncate the model at $j = J$ for some number $J > 1$.

Model type	(α_0, α_1)	(ξ_0, ξ_1)	(γ_0, γ_1)
Model 0	(0, 2)	(0, 0)	(0, 0)
Model 1	(0, 2)	(1, 1.5)	(0.005, 0.0025)
Model 2	(0, 2)	(1, 2)	(0.005, 0.01)

Table 3.1: Parameters of simple linear interference outcome model ((3.8) and (3.9)) used in simulation experiments.

Model type	α	β	γ	J
Model 3	1	5	2	2
Model 4	1	5	3	2
Model 5	1	5	1	3
Model 6	1	5	2	3

Table 3.2: Parameters of truncated linear-in-means outcome model used in simulation experiments.

Overall we consider the following model configurations of linear interference. Table 3.1 and 3.2 summarize the configurations of the models we consider for simulations. Note that model 0 exhibits no interference. For all models, the error terms are independently normally distributed with variance 1. The true GATE in these outcome models (either by an exact calculation or by a Monte Carlo estimate on the Swarthmore network) are 2, 3.69, 4.74, 15, 20, 15 and 35 respectively.

Beyond linear interference, we also examine a slightly more complicated scenario where linear interference is violated. In particular, we consider f_0 and f_1 that are nonlinear in ρ_i and ν_i. The nonlinear functions we use are sigmoid-type so that it is hard to approximate by any linear model[3]. We use the Monte Carlo estimate, 9.55, as the true GATE when reporting the simulation results. Our purpose here is to show that even if we have nonlinear f_0 and f_1 which violates our linear interference assumption, our method still leads to an estimator with reasonable performance. This also echos our previous discussion. In GATE estimation, we are always predicting for a data point that is outside the range of our observed/training data and hence a simple model can be quite reliable.

Estimator	$\hat{\tau}^{\text{DM}}$	$\hat{\tau}^{\text{Hájek}}_{0.8,0.2}$	$\hat{\tau}_{\text{frac}}$	$\hat{\tau}_{\text{num}}$	post-ReFeX-LASSO	ReFeX-LASSO	$\hat{\tau}_{\text{oracle}}$
Model 0	0.05	0.76	0.24	0.07	0.50	0.32	0.05
Model 1	1.53	1.02	0.36	1.22	1.54	0.70	0.25
Model 2	2.06	1.41	0.47	1.49	1.49	0.59	0.24
Model 3	10.02	3.84	0.37	9.86	1.08	0.93	0.37
Model 4	15.02	5.60	0.56	14.72	1.68	1.59	0.56
Model 5	9.92	6.61	4.53	9.82	1.38	1.73	1.98
Model 6	29.67	22.31	18.22	29.46	2.42	2.47	4.45

Table 3.3: RMSE of estimators of the GATE assuming linear interference (simple linear interference and truncated linear-in-means) outcome models.

Estimator	$\hat{\tau}^{\text{DM}}$	$\hat{\tau}^{\text{Hájek}}_{0.8,0.2}$	$\hat{\tau}_{\text{frac}}$	$\hat{\tau}_{\text{num}}$	post-ReFeX-LASSO	ReFeX-LASSO	$\hat{\tau}_{\text{oracle}}$
Model 0	0.004	0.120	0.021	0.009	0.106	0.042	0.004
Model 1	-1.53	-0.68	-0.25	-1.22	-0.72	-0.004	-0.05
Model 2	-2.06	-1.16	-0.39	-1.48	-0.56	-0.29	0.008
Model 3	-10.02	-3.67	-0.01	-9.85	0.28	0.19	-0.01
Model 4	-15.02	-5.46	0.003	-14.72	0.36	0.31	0.03
Model 5	-9.92	-6.55	-4.52	-9.82	-0.01	-0.24	0.68
Model 6	-29.67	-22.28	-18.21	-29.45	-0.21	-0.31	2.77

Table 3.4: Empirical bias of estimators of the GATE assuming linear interference (simple linear interference and truncated linear-in-means) outcome models.

3.4.3 Simulation results

We study both the bias and the root mean squared error (RMSE) of each estimator under these varied models. Table 3.3 and Table 3.4 show the RMSE and bias of several different estimators under linear

[3] We document this model in the Appendix 3.7.2.

interference. In these two tables, we show results of two kinds of regression adjustment estimators. $\hat{\tau}_{\text{frac}}$ is the regression adjustment estimator that adjusts for the fraction of treated neighbors and $\hat{\tau}_{\text{num}}$ adjusts for the number of treated neighbors. They are also considered in Chin [2019]. We also show the oracle adjustment estimator $\hat{\tau}_{\text{oracle}}$ as a reference, which marks the best we can do with full knowledge of the response model. Note that in some cases other estimators can perform better than the oracle since the oracle adjustment estimator only means we use oracle control covariates. The covariates are inevitably random and we are not averaging over all possible assignment vectors. Moreover, for the truncated linear-in-means model, the true covariates are highly correlated, causing the oracle adjustment estimator to have a large variance. Finally, $\hat{\tau}^{\text{DM}}$ and $\hat{\tau}^{\text{Hájek}}_{0.8,0.2}$ refer to the simple difference-in-mean estimator and the Hájek estimator in Equation (3.7) with $q = 0.8$ as we mentioned earlier.

First, if we look at the results for Model 0, i.e., when there is no interference, post-ReFeX-LASSO and ReFeX-LASSO all give better performance compared to the Hájek estimator. Second, for Model 1 and Model 2, the true interference mechanism is simple linear interference. As we can see from the first two rows of Table 3.3 and Table 3.4, if we fail to account for one feature, the bias and/or the RMSE can be large. Also, ReFeX-LASSO is dominating post-ReFeX-LASSO with significantly lower bias and RMSE since for this case ReFeX-LASSO is able to stop considering further features after the first iteration. For Model 3–6, the underlying model is a truncated linear-in-means model and the only difference between them is the stopping number J. For the models with $J = 2$ (Models 3 and 4), the interference is still local, i.e., within one's direct neighbors, but for $J = 3$ (Models 5 and 6), it is crucial to consider information from 2-hop neighbors. Our simulation results verify this intuition. We see that $\hat{\tau}_{\text{frac}}$ is doing well for model 3 and 4 but very poorly for model 5 and 6. Both post-ReFeX-LASSO and ReFeX-LASSO lead to estimators with relatively small bias and small RMSE for these more challenging response models.

Turning to the nonlinear model, Table 3.5 below shows our results there. In this case, $\hat{\tau}_{\text{frac}}$ and $\hat{\tau}_{\text{num}}$ represent the same regression adjustment estimators as in the linear case. Compared to difference-in-means and Hájek, ReFeX-LASSO leads to estimator with much better performance. Also, based on the comparison of $\hat{\tau}_{\text{frac}}$, $\hat{\tau}_{\text{num}}$ and ReFeX-LASSO, we see that, as in the linear interference case, even if we happen to adjust for some feature that is of importance, failing to take all relevant features into account will lead to estimators with either large bias, large variance, or both. In other words, ReFeX-LASSO helps one choose which set of features to adjust for and hence incur much smaller bias or variance.

From these simulations we take away that ReFeX-LASSO is able to identify influential features for regression adjustment and hence produce an estimator with relatively good performance across many model specifications. We also see that ReFeX-LASSO generally, though not always, performs significantly better than post-ReFeX-LASSO. This is due to the fact that we select features sequentially and hence reduce the variance. In contrast, a standard regression adjustment estimator

Estimator	$\hat{\tau}^{\mathrm{DM}}$	$\hat{\tau}^{\text{Hájek}}_{0.8,0.2}$	$\hat{\tau}_{\mathrm{frac}}$	$\hat{\tau}_{\mathrm{num}}$	post-ReFeX-LASSO	ReFeX-LASSO
Bias	-5.54	-2.72	-1.56	-2.72	1.29	1.33
RMSE	5.55	3.73	1.92	2.73	5.68	2.75

Table 3.5: RMSE and empirical bias of estimators of the GATE assuming a nonlinear interference (Appendix 3.7.2) outcome model.

considered in Chin [2019] for some network features ($\hat{\tau}_{\mathrm{frac}}$ and $\hat{\tau}_{\mathrm{num}}$ in our simulations) can be far-off if we fail to choose the right feature. Finally, exposure mapping based estimator like the fractional-exposure-Hájek estimator can also be pretty bad if we have interference that is quite different from the assumptions of the exposure model that such estimators assume.

3.4.4 Confidence interval for the GATE

In Section 3.3.4 we introduced a way to construct a confidence interval for τ via a block bootstrap and gave an explicit algorithm for graph-based block construction. We now evaluate the empirical coverage of the resulting confidence interval from our block bootstrap. Throughout this section, we focus on 90% confidence interval for τ. Instead of using the Swarthmore College network as in the previous section, we use the farmer network in Cai et al. [2015] where we have a larger and sparser network compared to the Swarthmore College network. In fact, the average size of 2-hop neighborhoods in Swarthmore network is 1092.65 and the average size of 3-hop neighborhoods in Swarthmore network is 1622.27. Hence, if we believe that interference is beyond 1-hop neighborhood, bootstrap will not perform well on such a dense graph since it is hard to create bootstrap samples that respect the structure in the original sample[4]. On the other hand, the farmer network in Cai et al. [2015] is less dense with 2-hop neighborhoods having an average size 23.95 and 3-hop neighborhoods having an average size 41.49. We will introduce in more details about the background and the details of this network in Section 3.5. In general, if the network is too dense to produce well-isolated and balanced clusters then the bootstrap would fail. One thing to notice is that the farmer network itself is associated with a natural clustering based on which village the each farmer lives in, namely, each village can be viewed as a cluster in the network. In our simulations here, we thus also show the results of constructing the confidence interval with block bootstrap of ReFeX-LASSO that uses this "oracle clustering" of villages. Finally, since we have a sparser network (making interference easier to manage), we consider two different sets of parameters for linear models that make the effect from number of treated neighbors larger (and thus GATE estimation harder). Table 3.6 shows the values of the parameters, loosely based on Model 2 (thus named 2a and 2b)

We first evaluate the effectiveness of such a bootstrap method. We assume linear interference and

[4] We found that the block bootstrap still gives near to nominal coverage on Swarthmore nwtwork when interference is local, i.e., within direct neighbors.

consider Model 3-6 as well as Model 2a and 2b. We fix $\ell = 3$, $B = 100$ and the coverage is calculated by repeating the whole process 100 times. Table 3.7 and 3.8 show the coverage and the average length of the confidence intervals constructed from our block bootstrap of post-ReFeX-LASSO and ReFeX-LASSO. To show the necessity of using block bootstrap and of considering the randomness of the assignment vector, we also include the result of constructing confidence interval using a naive bootstrap where we just sample each unit with replacement.

Model type	(α_0, α_1)	(ξ_0, ξ_1)	(γ_0, γ_1)
Model 2a	(0, 2)	(1, 3)	(0.01, 0.025)
Model 2b	(0, 2)	(1, 3)	(0.05, 0.15)

Table 3.6: Additional parameters of simple linear interference model ((3.8) and (3.9)) used in simulation experiments.

Model	post-ReFeX-LASSO	ReFeX-LASSO	Naive Bootstrap	Bootstrap with oracle clustering
Model 2a	93%	92%	94%	92%
Model 2b	92%	96%	93%	95%
Model 3	90%	90%	83%	91%
Model 4	88%	87%	80%	91%
Model 5	91%	93%	84%	92%
Model 6	90%	91%	67%	93%

Table 3.7: Coverage of different bootstrap 90% confidence intervals for the GATE with linear interference (simple linear interference and truncated linear-in-means) outcome models.

Model	post-ReFeX-LASSO	ReFeX-LASSO	Naive Bootstrap	Bootstrap with oracle clustering
Model 2a	0.245	0.220	0.235	0.228
Model 2b	0.435	0.384	0.390	0.382
Model 3	0.403	0.380	0.330	0.414
Model 4	0.569	0.534	0.437	0.593
Model 5	0.552	0.549	0.431	0.567
Model 6	1.316	1.316	0.751	1.412

Table 3.8: Average length of 90% confidence intervals for the GATE with linear interference (simple linear interference and truncated linear-in-means) outcome models.

As we can see from the results, our block bootstrap gives us near nominal coverage for ReFeX-LASSO and slightly worse but still close to nominal coverage for post-ReFeX-LASSO. However, the naive bootstrap fails to deliver confidence interval with nominal coverage. In fact, naive bootstrap-based confidence intervals can give us very bad coverage in some cases. We are also able to get good confidence intervals if we use the oracle clustering that is associated with the network. In scenarios where there are clear natural clusters in the network, these clusters can be a good default choice to use for block bootstrap. Moreover, as is shown in Table 3.8, both the block bootstrap confidence interval for ReFeX-LASSO and the block bootstrap confidence interval for post-ReFeX-LASSO are of reasonable length.

We conclude this section with a simulation to show why choosing the k for k-hop max clustering adaptively in our block bootstrap procedure is important and how partitioning the graph into just two clusters fails to give correct coverage. To this end, we consider using 2-hop max and 3-hop max clustering to divide units into clusters as well as randomly divide units into five clusters, i.i.d., without considering the underlying graph structure. We choose to consider 2-hop max and 3-hop max as we found in the simulations that in most of the cases ReFeX-LASSO will stop after selecting features about 2-hop neighborhoods. For Cai network, on average 2-hop max clustering and 3-hop clustering produce 267 and 269 clusters respectively. We choose to compare them with a five-cluster clustering as five is a lot less than the number of clusters we may have using k-hop max clustering. We rerun the block bootstrap procedure with these new clusters for Model 6 using ReFeX-LASSO. Table 3.9 shows the coverage of the confidence intervals. As we can see, contrast to the 91% coverage in Tablr 3.7 provided by the adaptive k-hop max based block bootstrap, all these three clustering methods fail to give us nominal coverage. In particular, completely ignoring the graph structure ("five clusters") leads to confidence intervals with really poor coverage.

Model	2-hop max	3-hop max	Five clusters
Model 6	84%	89%	45%

Table 3.9: Coverage of block bootstrap 90% confidence intervals for the GATE using different graph clustering algorithms with Model 6 as the true outcome model.

3.5 Real data example

In this section, we would like to apply our method to a real experiment where interference is known to exist and simple estimators such as difference-in-means would give poor GATE estimates. We consider data from the intervention in Cai et al. [2015]. They designed a randomized experiment to study the role of social networks on insurance adoption in rural China. Specifically, a random subset of farmers were provided with intensive information sessions about the an insurance product. Cai et al. [2015] found that the diffusion of insurance knowledge drove network effects in product adoption. Hence, this data is ideal for our purpose in the sense that we know for sure that SUTVA is violated and we should not trust the simple difference-in-means estimate for estimating the GATE. Moreover, though we know that network effects do exist, defining an exact exposure model as in Aronow and Samii [2017] is difficult. Hence, analysis done in Chin [2019] is limited since there only four pre-specified features were considered and hence the regression adjustment estimator implicitly assumed a certain exposure model. We revisit this experiment and estimate the GATE using our method.

In the original field experiment in Cai et al. [2015] the intensive information sessions were offered in two separate rounds, leading to four separate treatment arms. For our purpose, following Chin [2019], we simplify the experiment by viewing the two intensive information sessions as the same treatment arm. Hence, we reduce the original field experiment to a binary randomized experiment. As in Cai et al. [2015], the outcome variable is set to be the binary indicator variable for the weather insurance adoption, and we do not include villagers whose treatment or response information was missing as well as villagers whose network information was missing. We also combine all the villages into one social network, denoting this single social network by G. In summary, we have 4,382 nodes and 17,069 edges. This network is also the one that we used in Section 3.4.4.

The first step for our method is generating model-free covariates. We use exactly the same set of base features as in the previous simulation section—fraction of treated neighbors, number of treated neighbors, fraction of edges in neighborhood that connects a treated unit and a control unit and also fraction of edges in neighborhood that connects a treated unit and a treated unit. We then use ReFeX-LASSO to generate a group of covariates, using mean and variance aggregation functions (again, as in the previous simulation section) and estimate the GATE by adjusting for these covariates with a linear model. We compare the standard error estimate from block bootstrap with the one computed in Chin [2019].

Estimator	Estimate	Standard Error
DM	0.078	——
Hájek_1hop ($q = 0.75$)	0.163	——
Hájek_2hop ($q = 0.75$)	0.167	——
$\hat{\tau}_{\text{chin}}$	0.122	0.056
$\hat{\tau}_{\text{num}}$	0.178	0.027
$\hat{\tau}_{\text{refex-lasso}}$	0.178	0.043

Table 3.10: Estimates and standard errors of different estimators for the global average treatment effect on insurance adoption Cai et al. [2015].

Table 3.10 shows the resulting GATE estimates, where $\hat{\tau}_{\text{chin}}$ is the estimator in Chin [2019] that adjusts for four covariates: the fraction of treated neighbors, the number of treated neighbors, the fraction of treated neighbors in 2-hop neighborhoods, the number of treated neighbors in 2-hop neighborhoods. Meanwhile, $\hat{\tau}_{\text{num}}$ only adjusts for the number of treated neighbors and $\hat{\tau}_{\text{refex-lasso}}$ is the ReFeX-LASSO based adjustment estimator. DM refers to the difference-in-means estimator. Hájek_1hop assumes a fractional exposure model for 1-hop neighborhood while Hájek_2hop assumes a fractional exposure model for 2-hop neighborhood, i.e., we use (3.7) but consider 2-hop neighbors instead. The intuition is that sometimes units that are not direct neighbors but neighbors of direct neighbors matter as well and by considering fractional exposure model for 2-hop neighborhood we are able to take these units into account for the exposure model. We notice that $\hat{\tau}_{\text{num}}$ and $\hat{\tau}_{\text{refex-lasso}}$ give us the same estimate and indeed, the only covariate selected from ReFeX-LASSO is the number of treated neighbors. Compared to $\hat{\tau}_{\text{chin}}$, $\hat{\tau}_{\text{refex-lasso}}$ has smaller standard error and a larger estimate of the effect. Finally, though $\hat{\tau}_{\text{num}}$ and $\hat{\tau}_{\text{refex-lasso}}$ give nearly the same estimates (same up to three decimal digits), we see that the former as a smaller standard error. The reasons are twofold. First, bootstrap in general is conservative. Second, ReFeX estimate should have larger variance as we have a random selection procedure involved.

3.6 Discussion

In this chapter, we have developed a method to do estimation and inference for the global average treatment effect (GATE) when network interference is present. We develop a procedure that can be used to estimate the GATE without pre-specifying either exposure mappings or outcome models. We also give a way to construct confidence intervals for the GATE using a block bootstrap. We evaluate our method both through simulations and a real data example.

Many interesting avenues of further investigation have been left unexplored in this manuscript. First, our results only consider designs that satisfy the uniformity assumption (e.g., Bernoulli design): this is, of course, limiting, but it does present a useful benchmark. We are particularly interested in exploring how to extend our work to designs that violate the uniformity assumption such as cluster randomized design. This is challenging since the covariates we adjust for may be correlated with the treatment assignment. Second, while our simulations show that the block bootstrap behaves well in practice, formal results are absent for anything other than a simple toy setting. Third, beyond linear adjustment we may also want to have a completely nonlinear model to estimate the outcomes using the covariates returned from the ReFeX-LASSO feature generation and selection process.

3.7 Appendix

3.7.1 Proofs

The proofs of Proposition 14 and Proposition 15 will be exactly the same as the proofs of Proposition 1 and Proposition 2 in Luo and Chen [2014] once we realize that as long as the features that are included in the penalty do not overlap with the features that have already been selected then we can just use the proofs in Luo and Chen [2014], i.e., though our sequential selection procedure is different from that in Luo and Chen [2014], we share the same properties that make these two propositions hold.

Proof of Proposition 14. We denote by $X(s)$ the design matrix with features in s, i.e., if $|s| = m$

then $X(s)$ is a $n \times m$ matrix. At the $(t+1)-th$ iteration, β will be a $(|s_{*t}| + i_{t+1})$-dimensional vector and we denote by $\beta(s)$ the $|s|$-dimensional vector with only coordinates of β that are in s. Finally, we denote by A_{t+1} the set $\{u_1^{t+1}, u_2^{t+1}, \cdots, u_{i_{t+1}}^{t+1}\}$.

First we note that since $u_j^{t+1} \in \mathcal{R}(s_{*t})$, $\exists v \in \mathbb{R}^{|s_{*t}|}$ such that $u_j^{t+1} = X(s_{*t})v$. We now consider the objective function l_{t+1} at the $(t+1)$-th iteration.

$$\begin{aligned}
l_{t+1} &= \|y - X(s_{*t})(\beta(s_{*t}) + \beta(\{j\})v) - X(A_{t+1}/\{j\})\beta(A_{t+1}/\{j\})\|_2^2 \\
&\qquad + \lambda\left(|\beta(\{j\})| + \|\beta(A_{t+1}/\{j\})\|_1\right) \\
&= \|y - X(s_{*t})\tilde{\beta}(s_{*t}) - X(A_{t+1}/\{j\})\beta(A_{t+1}/\{j\})\|_2^2 \\
&\qquad + \lambda\left(|\beta(\{j\})| + \|\beta(A_{t+1}/\{j\})\|_1\right) \\
&\geq \|y - X(s_{*t})\tilde{\beta}(s_{*t}) - X(A_{t+1}/\{j\})\beta(A_{t+1}/\{j\})\|_2^2 \\
&\qquad + \lambda\|\beta(A_{t+1}/\{j\})\|_1
\end{aligned}$$

Hence, when l_{t+1} is minimized, $\beta(\{j\})$ must be 0 and $j \notin s_{*(t+1)}$. □

Proof of Proposition 15. Again we consider the objective function at the $(t+1)$-th iteration.

$$l_{t+1} = \|y - X(s_{*t})\beta(s_{*t}) - X(A_{t+1})\beta(A_{t+1})\|_2^2 + \lambda\|\beta(A_{t+1})\|_1.$$

Differentiating l_{t+1} with respect to $\beta(s_{*t})$, we have

$$\frac{\partial l_{t+1}}{\partial \beta(s_{*t})} = -2X^T(s_{*t})y + 2X^T(s_{*t})X(s_{*t})\beta(s_{*t}) + 2X^T(s_{*t})X(A_{t+1})\beta(A_{t+1}).$$

Setting the above derivative to zero,

$$\hat{\beta}(s_{*t}) = [X^T(s_{*t})X(s_{*t})]^{-1}X^T(s_{*t})[y - X(A_{t+1})\beta(A_{t+1})]. \tag{3.10}$$

Substituting (3.10) into the objective function, we obtain

$$\begin{aligned}
l_{t+1} &= \|y - X(s_{*t})\beta(s_{*t}) - X(A_{t+1})\beta(A_{t+1})\|_2^2 + \lambda\|\beta(A_{t+1})\|_1 \\
&= \|y - X(s_{*t})[X^T(s_{*t})X(s_{*t})]^{-1}X^T(s_{*t})[y - X(A_{t+1})\beta(A_{t+1})] - X(A_{t+1})\beta(A_{t+1})\|_2^2 \\
&\qquad + \lambda\|\beta(A_{t+1})\|_1 \\
&= \|(I - X(s_{*t})[X^T(s_{*t})X(s_{*t})]^{-1}X^T(s_{*t}))y \\
&\qquad - (I - X(s_{*t})[X^T(s_{*t})X(s_{*t})]^{-1}X^T(s_{*t}))X(A_{t+1})\beta(A_{t+1})\|_2^2 \\
&\qquad + \lambda\|\beta(A_{t+1})\|_1.
\end{aligned}$$

Hence minimizing l_{t+1} does not affect $\hat{\beta}(s_{*t})$ and $\hat{\beta}(s_{*t})$ will be almost surely nonzero. □

Now we show the proof Theorem 26. We will make use of standard results about LASSO ℓ_2-error bounds. Recall the following result Wainwright [2019]:

Lemma 6. *Suppose $y = X\theta^* + w$ ($X \in \mathbb{R}^{n\times d}$) and consider the Lagrangian Lasso with a strictly positive regularization parameter $\lambda_n \geq 2\|\frac{\mathbf{X}^T w}{n}\|_\infty$. Suppose further that θ^* is supported on a subset S of cardinality s, and the design matrix satisfies the $(\kappa; 3)$-RE condition over S, then*

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda_n.$$

We can show that if the design matrix is $C-$column normalized, i.e.,

$$\max_{j=1,\cdots,d} \frac{\|X_j\|_2}{\sqrt{n}} \leq C,$$

then the choice $\lambda_n = 2C\sigma(\sqrt{\frac{2\log d}{n}} + \delta)$ is valid with probability at least $1 - 2e^{-\frac{n\delta^2}{2}}$. We thus proceed with the main proof.

Proof. Notice that $\|\frac{\mathbf{X}^T w}{n}\|_\infty$ corresponds to the absolute maximum of d zero-mean Gaussian random variables by definition of infinity norm and each with variance at most $\frac{C^2\sigma^2}{n}$. Hence, from the Gaussian tail bound, we then have

$$\mathbb{P}\left(\left\|\frac{\mathbf{X}^T w}{n}\right\|_\infty \geq C\sigma\left(\sqrt{\frac{2\log d}{n}} + \delta\right)\right) \leq 2e^{-\frac{n\delta^2}{2}}.$$

□

With this particular choice of λ_n, the lemma implies the upper bound

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{6C\sigma}{\kappa}\sqrt{s}\left(\sqrt{\frac{2\log d}{n}} + \delta\right) \tag{3.11}$$

with the same high probability Wainwright [2019].

Now we are ready to prove consistency. First notice that

$$\begin{aligned}
|\hat{\tau} - \tau| &= \left|\frac{1}{n}\sum_{i=1}^{n}\left[(\hat{\beta}_1 - \beta_1^*)^T u_i^{gt} - (\hat{\beta}_0 - \beta_0^*)^T u_i^{gc}\right]\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|\left[(\hat{\beta}_1 - \beta_1^*)^T u_i^{gt} - (\hat{\beta}_0 - \beta_0^*)^T u_i^{gc}\right]\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}(\|\hat{\beta}_1 - \beta_1^*\|_2\|u_i^{gt}\|_2 + \|\hat{\beta}_0 - \beta_0^*\|_2\|u_i^{gc}\|_2) \\
&\leq C\sqrt{M}(\|\hat{\beta}_1 - \beta_1^*\|_2 + \|\hat{\beta}_0 - \beta_0^*\|_2)
\end{aligned}$$

Let n_0 be the number of control units and n_1 be the number of treated units. Then by strong law of large numbers, $\frac{n_0}{n} \xrightarrow{a.s.} 1-p$ and $\frac{n_1}{n} \xrightarrow{a.s.} p$. Since the design matrices U^0 and U^1 satisfy the RE condition, both $\|\hat{\beta}_1 - \beta_1^*\|_2$ and $\|\hat{\beta}_0 - \beta_0^*\|_2$ converge to 0 in probability by the bound (3.11). Thus $\hat{\tau} \xrightarrow{\mathbb{P}} \tau$.

Proof of Proposition 16. We show that for the setup in Proposition 16, the design matrices satisfy RE condition with probability going to 1. In our proof, the first column of the design matrix represents the fraction of treated neighbors while the second column represents the number of treated neighbors. We introduce one extra notations: for each unit i, we denote by m_i the size of the cluster unit i belongs to. We show the proof for the design matrix for control units, U^0. Similar proof can be done for U^1. After centering, the design matrix we use for estimating β_0 will be

$$\tilde{U^0} = \begin{bmatrix} \frac{1}{n_0}\sum_{i:W_i=0}(u_i^1 - \bar{u}^1)^2 & \frac{1}{n_0}\sum_{i:W_i=0}(u_i^1 - \bar{u}^1)(u_i^2 - \bar{u}^2) \\ \frac{1}{n_0}\sum_{i:W_i=0}(u_i^1 - \bar{u}^1)(u_i^2 - \bar{u}^2) & \frac{1}{n_0}\sum_{i:W_i=0}(u_i^2 - \bar{u}^2)^2 \end{bmatrix}.$$

Here $\bar{u}^1 = \frac{1}{n_0}\sum_{i:W_i=0} u_i^1$ and $\bar{u}^2 = \frac{1}{n_0}\sum_{i:W_i=0} u_i^2$. Since the true β_0 is non-zero only for the first feature, $\mathbb{C}_3(S) = \{\Delta \in \mathbb{R}^2 : |\Delta_2| \le 3|\Delta_1|\}$. For such Δ,

$$\frac{1}{n_0}\|\tilde{U^0}\Delta\|_2^2 = \Delta_1^2 \frac{1}{n_0}\sum_{i:W_i=0}(u_i^1 - \bar{u}^1)^2 + 2\Delta_1\Delta_2 \frac{1}{n_0}\sum_{i:W_i=0}(u_i^1 - \bar{u}^1)(u_i^2 - \bar{u}^2) + \Delta_2^2 \frac{1}{n_0}\sum_{i:W_i=0}(u_i^2 - \bar{u}^2)^2$$

Note that since $|\Delta_2| \le 3|\Delta_1|$, $\Delta_1\Delta_2 \ge -|\Delta_1||\Delta_2| \ge -\frac{1}{3}\Delta_2^2$. Therefore,

$$\begin{aligned} \frac{1}{n_0}\|\tilde{U^0}\Delta\|_2^2 \ge{} & \frac{1}{n_0}\sum_{i:W_i=0}(u_i^1 - \bar{u}^1)^2\Delta_1^2 \\ & + \left(\frac{1}{n_0}\sum_{i:W_i=0}(u_i^2 - \bar{u}^2)^2 - \frac{1}{3}\frac{1}{n_0}\sum_{i:W_i=0}(u_i^1 - \bar{u}^1)(u_i^2 - \bar{u}^2)\right)\Delta_2^2. \end{aligned} \tag{3.12}$$

To ease notations, we let $\textcircled{1} = \frac{1}{n_0}\sum_{i:W_i=0}(u_i^1 - \bar{u}^1)^2$, $\textcircled{2} = \frac{1}{n_0}\sum_{i:W_i=0}(u_i^2 - \bar{u}^2)^2$ and $\textcircled{3} = \frac{1}{n_0}\sum_{i:W_i=0}(u_i^1 - \bar{u}^1)(u_i^2 - \bar{u}^2)\Delta_2^2$. Now, we analyze each term separately.

$$\begin{aligned} \textcircled{1} &= \frac{1}{n_0}\sum_{i:W_i=0}(u_i^1 - \bar{u}^1)^2 \\ &= \frac{1}{n_0}\sum_{i:W_i=0}(u_i^1)^2 - (\bar{u}^1)^2 \\ &= \frac{n}{n_0}\frac{1}{n}\sum_{i=1}^{n}(1 - W_i)(u_i^1)^2 - (\bar{u}^1)^2 \\ &= \frac{n}{n_0}\frac{1}{n}\sum_{i=1}^{n}(1 - W_i)(u_i^1)^2 - \left(\frac{n}{n_0}\frac{1}{n}\sum_{i=1}^{n}(1 - W_i)u_i^1\right)^2. \end{aligned}$$

Consider the random variables $\{(1-W_i)(u_i^1)^2\}_{i=1}^n$ and $\{(1-W_i)u_i^1\}_{i=1}^n$. Since we have disjoint clusters and the number of units in each cluster is bounded by M, the sum of covariance term is at most $O(n)$ and hence weak law of large numbers applies for both sequences. Therefore,

$$\frac{1}{n}\sum_{i=1}^n (1-W_i)(u_i^1)^2 - \left[p(1-p)^2\frac{1}{n}\sum_{i=1}^n \frac{1}{m_i-1} + p^2(1-p)\right] \xrightarrow{\mathbb{P}} 0.$$

Similarly,

$$\frac{1}{n}\sum_{i=1}^n (1-W_i)u_i^1 \xrightarrow{\mathbb{P}} p(1-p).$$

Note that $n/n_0 \xrightarrow{\mathbb{P}} 1/(1-p)$, we obtain

$$\textcircled{1} - \left[p(1-p)\frac{1}{n}\sum_{i=1}^n \frac{1}{m_i-1}\right] \xrightarrow{\mathbb{P}} 0.$$

Here $\textcircled{2}$ can be done similarly:

$$\textcircled{2} - \left[p(1-p)\frac{1}{n}\sum_{i=1}^n (m_i-1) + p^2\frac{1}{n}\sum_{i=1}^n (m_i-1)^2 - p^2\left(\frac{1}{n}\sum_{i=1}^n (m_i-1)\right)\right] \xrightarrow{\mathbb{P}} 0.$$

For $\textcircled{3}$,

$$\begin{aligned}\textcircled{3} &= \frac{1}{n_0}\sum_{i:W_i=0} (u_i^1-\bar{u}^1)(u_i^2-\bar{u}^2)\\ &= \frac{1}{n_0}\sum_{i:W_i=0} u_i^1u_i^2 - \bar{u}^1\bar{u}^2.\end{aligned}$$

Notice that we have already shown that

$$\bar{u}^1 \xrightarrow{\mathbb{P}} p, \qquad \bar{u}^2 \xrightarrow{\mathbb{P}} p\frac{1}{n}\sum_{i=1}^n (m_i-1).$$

Hence, $\bar{u}^1\bar{u}^2 \xrightarrow{\mathbb{P}} p^2\frac{1}{n}\sum_{i=1}^n (m_i-1)$. Moreover,

$$\frac{1}{n_0}\sum_{i:W_i=0} u_i^1u_i^2 = \frac{n}{n_0}\frac{1}{n}\sum_{i=1}^n (1-W_i)u_i^1u_i^2.$$

Again by the weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^n (1-W_i)u_i^1u_i^2 - \left[p(1-p)^2 + p^2(1-p)\frac{1}{n}\sum_{i=1}^n (m_i-1)\right] \xrightarrow{\mathbb{P}} 0.$$

Hence, $\textcircled{3} - \left[p(1-p) + p^2 \frac{1}{n}\sum_{i=1}^n (m_i - 1)\right] \xrightarrow{\mathbb{P}} 0$. Put all these pieces together, we obtain

$$\begin{aligned}\text{RHS of (3.12)} - \Bigg\{ & \left[p(1-p)\frac{1}{n}\sum_{i=1}^n \frac{1}{m_i - 1}\right] \Delta_1^2 \\ & + \left[p(1-p)\frac{1}{n}\sum_{i=1}^n (m_i - 1) + p^2 \frac{1}{n}\sum_{i=1}^n (m_i - 1)^2 - p^2 \left(\frac{1}{n}\sum_{i=1}^n (m_i - 1)\right)\right. \\ & \left. - \frac{1}{3}\left(p(1-p) + p^2 \frac{1}{n}\sum_{i=1}^n (m_i - 1)\right)\right] \Delta_2^2 \Bigg\} \xrightarrow{\mathbb{P}} 0.\end{aligned}$$

Notice that $m_i \geq 3$ and $m_i \leq M$ for each i, we conclude that for $\kappa = \min\{\frac{p(1-p)}{M-1}, \frac{5}{3}p - \frac{1}{3}p^2\}$,

$$\frac{1}{n_0}\|\tilde{U}^0 \Delta\|_2^2 \geq \kappa \|\Delta\|_2^2 \quad \text{w.p.} \quad \to 1.$$

□

3.7.2 Supplementary Materials

Definition 28 (The nonlinear model in simulations). Suppose the assignment vector is w, then for each unit i, the response is

$$y_i(w) = -5 + 2z_i w_i + 0.03\nu_i + \frac{1}{1 + 0.001\exp(-0.03\nu_i + 9)} + \frac{10}{3 + \exp(-8\rho_i + 3.2)} + \epsilon_i.$$

Here, $z_i, \epsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$, ρ_i is the fraction of treated neighbors for unit i and ν_i is the number of treated neighbors for unit i.

Chapter 4

Detecting Interference in Online Controlled Experiments with Increasing Allocation

4.1 Introduction

4.1.1 Motivations and contributions

A/B testing is a key component in product development, serving as an empirical method to compare two versions of a product or feature. By randomly splitting the user base into two groups, this method allows one group to experience version A (often the current version or "control") and the other to experience version B (the new or "treatment" version). The most straightforward statistical analysis following A/B tests is to compute the difference-in-means estimator, i.e., the difference in the average of outcomes of the treatment group and that of the control group. Under the classical Stable Unit Treatment Value Assumption (SUTVA), which requires that the potential outcomes for any unit do not vary with the treatments assigned to other units, one can easily show that the difference-in-means estimator will be close to the causal effect as long as the sample size is large [Imbens and Rubin, 2015]. This implies that when we compute the difference-in-means estimator for any single randomized experiment in an A/B test with increasing allocation, the value of the estimator should not change by much. However, in some real-world scenarios, we observe drastic change in the difference-in-means estimators throughout the experiments. In Figure 4.1, we show an example from an A/B test implemented by LinkedIn. On the x-axis, we show the percentage of units that are in the treatment group; on the y-axis, we show the value of the difference-in-means estimator. In this example, we see that the difference-in-means estimator decreases as the treatment

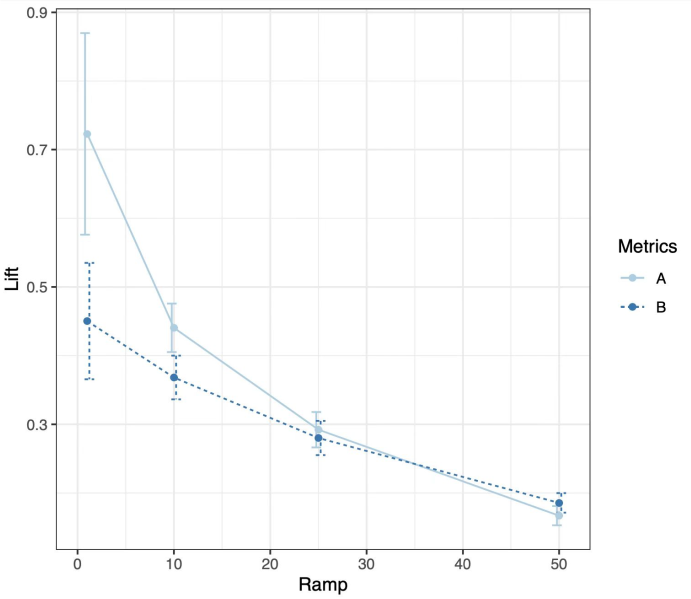


Figure 4.1: An A/B test implemented by LinkedIn with increasing allocation. A and B are different outcome metrics.

is released to more units. We naturally wonder: What causes this phenomenon? Could it be purely due to randomness? Is the SUTVA assumption violated in this case?

One plausible explanation for this phenomenon is the existence of interference, i.e., when treatment assigned to one unit may affect observed outcomes for other units. One form of interference is marketplace competition. Imagine a new treatment that can help units perform better in the market. For any particular unit, the treatment brings benefit, but when more of the other units are treated, the other units become more competitive and thus negatively impact the performance of that particular unit. Therefore, in these cases, we often observe that the difference-in-means estimator decreases with treatment probability. Indeed, the experiments in Figure 4.1 were run in a setting with marketplace competition. One other common form of interference is through social networks. People's behaviors tend to be positively correlated with those of others connected to them in the network. Think about a treatment that encourages users to comment on a social media platform: users tend to comment more when they see comments from friends. In these cases, we usually observe that the difference-in-means estimator increases with treatment probability.

In practice, however, the structure of interference can be more complicated than the two apparent forms discussed in the above paragraph. Often, experimenters manually examine the difference-in-means plot and decide whether to send the job to other experimentation platforms that deal with interference more carefully. We need a way to formally test whether interference exists.

In this chapter, we introduce statistical testing procedures that test for interference in A/B testing with increasing allocation. The methods we propose are scalable and parallelable. They are also agnostic to interference mechanism: even if we have no knowledge of the interference structure, the testing procedure is still valid. Knowledge of the interference structure can, however, be helpful in increasing the power of the testing procedure. We introduce two different testing strategies under different assumptions in Sections 4.3.1 and 4.3.2. In Section 4.3.1, we introduce a general statistical test for interference, a test that requires no additional assumptions. The proposed method is inspired by the testing procedure proposed by [Athey et al., 2018], but it is more powerful than that of [Athey et al., 2018] by making use of multiple experiments. In Section 4.3.2, we introduce a testing procedure that is valid under a time fixed effect assumption. The testing procedure is of very low computational complexity, and it is more powerful than the test proposed in Section 4.3.1. In particular, one special case of this method formalizes a heuristic algorithm discussed above, which decides that interference exists when the difference-in-means estimators are very different.

4.1.2 Related work

The classical literature on causal inference often assumes that there is no cross-unit interference. When interference presents, many classical inference methods break down. Interest in causal inference with interference started in the social and medical sciences [Sobel, 2006b, Hudgens and Halloran, 2008]. Since then, one line of work focuses on estimation and inference of treatment effects under network interference [Tchetgen and VanderWeele, 2012, Toulis and Kao, 2013, Aronow and Samii, 2017, Sussman and Airoldi, 2017, Basse and Feller, 2018, Bhattacharya et al., 2020, Leung, 2020, Sävje et al., 2021, Sävje, 2021, Hu et al., 2022, Li and Wager, 2022]. In order to facilitate estimation, these works either assume that there are special randomization designs or that the interference has some restricted form defined by a given network. Applications to A/B testing are also considered in Ugander et al. [2013], Eckles et al. [2017], and Basse and Airoldi [2018]. One assumption implicitly made in these works is that the experiment is conducted only once. In the multiple experiments regime, Viviano [2020] studies the design of two-wave experiments under interference. Yu et al. [2022] and Cortez et al. [2022] consider estimating the total treatment effects under interference with data from more than two time steps. Bojinov et al. [2021b] and Han et al. [2021] further investigate the problem in panel experiments. Our work differs from the above works for at least two reasons: (1) instead of focusing on estimation, we focus on testing whether interference exists and (2) we do not need to make additional assumptions in order for the testing procedure to be valid.

In the literature of testing for interference, Bowers et al. [2013] consider model-based approaches, Pouget-Abadie et al. [2019c] introduce an experimental design strategy, and Aronow [2012] and Athey et al. [2018] propose conditional randomization tests restricted to a subset of what they call focal units, and a subset of assignments that make the null hypothesis sharp for focal units. Basse et al. [2019] and Puelz et al. [2022] further extend this method by using a conditioning mechanism to allow the selection of focal units to depend on the observed treatment assignment. However, none of these works addresses the problem of multiple experiments, and their methods tend to have lower power when directly applied in our setup. To the best of our knowledge, our work is the first to consider testing interference with a sequence of randomized experiments.

Our work is also related to research on interference in online marketplace experiments (See [Basse et al., 2016, Fradkin, 2019, Holtz et al., 2020, Bajari et al., 2021, Wager and Xu, 2021, Johari et al., 2022, Li et al., 2022] among others). This line of work usually requires careful modeling of the market and the interference mechanism. The testing procedure introduced in this chapter, in contrast, can be applied to arbitrary forms of interference.

4.2 Problem Setup

We work in a setting where we run a sequence of A/B tests with increasing allocations. Formally, suppose that there are K experiments on a population of n units. Let π_k be the marginal treatment probability of the k^{th} experiment. The treatment probabilities satisfy $\pi_1 < \pi_2 < \cdots < \pi_K$. For each experiment $k \in \{1, \ldots, K\}$ and each unit $i \in \{1, \ldots, n\}$, let

$$\begin{aligned} W_{i,k} &:= \text{treatment of unit } i \text{ assigned in the } k^{\text{th}} \text{ experiment,} \\ Y_{i,k} &:= \text{outcome of unit } i \text{ in the } k^{\text{th}} \text{ experiment.} \end{aligned}$$

Here we assume that $W_{i,k} \in \{0, 1\}$ is a binary treatment variable and that a value of 1 corresponds to the treatment group while a value of 0 corresponds to the control group.

The experiments are implemented in the following way. In the first experiment, each unit i is randomly assigned a treatment $W_{i,1}$, where

$$W_{i,1} \sim \text{Bernoulli}(\pi_1) \text{ independently.} \tag{4.1}$$

In the subsequent experiments, more units are assigned to the treatment group. Specifically, conditioning on the previous treatments, each $W_{i,k}$ is sampled from the following distribution independently:

$$\begin{cases} W_{i,k} \sim \text{Bernoulli}\left((\pi_k - \pi_{k-1})/(1 - \pi_{k-1})\right), & \text{if } W_{i,k-1} = 0; \\ W_{i,k} = 1, \quad \text{if } W_{i,k-1} = 1. \end{cases} \tag{4.2}$$

This formulation guarantees that if we look at the k^{th} experiment alone, then the treatments $W_{i,k}$'s are i.i.d. Bernoulli(π_k).

Let $W_{1:n,1:K}$ be the $n \times K$ treatment matrix and $Y_{1:n,1:K}$ be the $n \times K$ outcome matrix of all units and all experiments. Let $X_i \in \mathbb{R}^d$ be the observed covariates of unit i that do not change over the course of the experiments. Correspondingly, let $X_{1:n} \in \mathbb{R}^{n \times d}$ be the matrix of covariates of all units.

Following the Neyman-Rubin causal model, we assume that potential outcomes $Y_{i,k}(w_{1:n,1:K}) \in \mathbb{R}$

exist for all $w_{1:n,1:K} \in \{0,1\}^{n \times K}$ and that the observed outcomes satisfy $Y_{i,k} = Y_{i,k}(W_{1:n,1:K})$.[1] The goal is to test the following hypothesis:

Hypothesis 1 (No cross-unit interference)**.** $Y_{i,k}(w_{1:n,1:K}) = Y_{i,k}(\tilde{w}_{1:n,1:K})$ if $w_{i,1:K} = \tilde{w}_{i,1:K}$.

The hypothesis states that the outcomes of unit i depend only on the treatments of unit i and not on the treatments of others. We call this hypothesis the no cross-unit interference hypothesis.

4.3 Testing for interference

In this section, we introduce methods that test for the existence of cross-unit interference. For brevity's sake, we focus on testing with two experiments. We then discuss further extensions to multiple experiments in Section 4.3.5.

Naturally, the first question that occurs is how interference might arise. To formalize this, we introduce a notion of *candidate exposure* that captures the potential form of interference. Using domain knowledge, experimenters can specify the candidate exposure, which can vary from application to application. When we consider user-level data, we have a natural social network. Here experimenters may suspect that a user's outcome is influenced by treatments of "friends", i.e., users connected through the social network. And thus in this example, some plausible choices of candidate exposures include the fraction of friends who are treated, and the number of friends who are treated. When we consider marketplace competition, advertisers are the subjects of treatment. Here the sales of an advertiser may be impacted by the treatments of competitors, i.e., advertisers that sell similar products. Hence in this application, experimenters can choose candidate exposures to be the number of treated advertisers that sell products of the same category, or an average of treatments given to other advertisers weighted by some product similarity metric.

Formally, for each experiment k and each unit i, we use $H_{i,k} = h_i(W_{-i,k}) \in \mathbb{R}^m$ to denote the candidate exposure. Here $W_{-i,k}$ is the treatments given to all other units except i in the k^{th} experiment. We use the form $h_i(W_{-i,k})$ to emphasize that the candidate exposure depends on other units' treatments. We also write $H_{1:n,k} = (H_{1,k}, H_{2,k}, \ldots, H_{n,k})^\top \in \mathbb{R}^{n \times m}$ to reference the candidate exposures of all units.

We want to emphasize that for all the tests introduced below, we do not require the candidate exposure to be correctly specified in order for the tests to be valid. However, the form of the candidate exposure matters for the power of the tests.

We will then move on to test the hypothesis that no interference exists making use of the candidate exposure $H_{i,k}$. In the following sections, we discuss different strategies to test for interference under different assumptions.

[1] In the literature, a *no anticipation effects* assumption is often made in such potential outcome models. The assumption states that the outcome $Y_{i,k}$ depends only on the treatments assigned during and prior to the k^{th} experiment. With this assumption, the potential outcomes can be written as $Y_{i,k}(w_{1:n,1:k})$ which satisfies $Y_{i,k} = Y_{i,k}(W_{1:n,1:k})$. Here for simplicity, we keep the original notation.

4.3.1 Testing under general assumptions

We start with a setting where we have access to a dataset from *only one* experiment. Suppose that we collect data on units indexed by $i = 1, ..., n$, where each unit is randomly assigned to a binary treatment $W_i \in \{0, 1\}$,

$$W_i \sim \text{Bernoulli}(\pi) \text{ independently} \tag{4.3}$$

for some $0 \leq \pi \leq 1$. For each unit, we observe an outcome of interest $Y_i \in \mathbb{R}$ and some covariates $X_i \in \mathbb{R}^p$. Athey et al. [2018] proposed a method to test for Hypothesis 1 in this setting.[2] We sketch the procedure in Algorithm 7.

Algorithm 7 Testing for interference effect (one experiment).

```
Input: Dataset $\mathcal{D} = (W_{1:n}, X_{1:n}, Y_{1:n}, H_{1:n})$, exposure function $h$, test statistic $T$.

  1. Randomly split the data into two folds. Let $\mathcal{I}_{\text{foc}}$ and $\mathcal{I}_{\text{aux}}$ be the index set for the first fold
     (focal units) and the second fold (auxiliary units). Write the first fold of data as
     $\mathcal{D}_{\text{foc}} = (W_{\text{foc}}, X_{\text{foc}}, Y_{\text{foc}}, H_{\text{foc}})$ and the second as $\mathcal{D}_{\text{aux}} = (W_{\text{aux}}, X_{\text{aux}}, Y_{\text{aux}}, H_{\text{aux}})$.
  2. Compute a test statistic $T^{(0)} = T(W_{\text{foc}}, X_{\text{foc}}, Y_{\text{foc}}, H_{\text{foc}})$ that captures the importance of $H$ in
     predicting $Y$.
  3. For $b = 1, \dots B$:
         Regenerate treatments for the auxiliary units: $\widetilde{W}_i^{(b)} \sim \text{Bernoulli}(\pi)$ for $i \in \mathcal{I}_{\text{aux}}$.
         Recompute the candidate exposure for focal units: $\widetilde{H}_i^{(b)} = h_i(W_{\text{foc}\setminus\{i\}}, \widetilde{W}_{\text{aux}}^{(b)})$ for
         $i \in \mathcal{I}_{\text{foc}}$.
         Recompute the test statistic: $T^{(b)} = T(W_{\text{foc}}, X_{\text{foc}}, Y_{\text{foc}}, \widetilde{H}_{\text{foc}}^{(b)})$.
     End For

Output: The $p$-value
     $$p = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{1}\left\{T^{(0)} \leq T^{(b)}\right\}\right). \tag{4.4}$$
```

Algorithm 7 requires as input a test statistic T that captures the importance of the candidate exposure H in predicting outcome Y. As an illustration, assume for now that $H_i \in \mathbb{R}$. One plausible choice of the test statistic T (when $H_i \in \mathbb{R}$) is the following: we run a linear regression of $Y_{\text{foc}} \sim W_{\text{foc}} + X_{\text{foc}} + H_{\text{foc}}$, extract the coefficient of H_{foc}, and take the test statistic T to be the absolute value of the coefficient. We use this regression coefficient statistic as an example to explain the intuition of the algorithm. Under the null hypothesis, the candidate exposure H has no power to predict the outcome Y before or after regenerating treatments, and thus the distribution of the test

[2] The method proposed by [Athey et al., 2018] is more general. Here we focus on a special case: testing the existence of cross-unit interference in Bernoulli experiments.

statistic T will not change after regenerating treatments. Hence the p-value will be stochastically larger than Unif$[0,1]$. Under the alternative hypothesis, the behavior of the p-value can be very different. Consider a simple example where H_i is the treatment assigned to the closest friend of unit i and $Y_i = \alpha^\top X_i + \beta W_i + \theta H_i + \epsilon_i$ for some i.i.d. zero mean errors ϵ_i. In this example, the original test statistic $T(W_{\text{foc}}, X_{\text{foc}}, Y_{\text{foc}}, H_{\text{foc}}) \approx |\theta|$ when the sample size is large. However, after regenerating treatments, for each focal unit i, if the closest friend of i is among the auxiliary units, then $\widetilde{H}_i$ is marginally a Bern(π) random variable, *independent* of Y_i; and hence the distribution of $T(W_{\text{foc}}, X_{\text{foc}}, Y_{\text{foc}}, \widetilde{H}_{\text{foc}}^{(b)})$ will not concentrate around $|\theta|$. In this case, the p-value is far from Unif$[0,1]$.

In practice, experimenters can use any test statistic T that are suitable for specific applications. For example, if the covariate X is of high dimension, a lasso-type algorithm can be used. One can also run more complicated machine learning algorithms, e.g., random forest and gradient boosting, with Y as a response and X, W, H as predictors, and set the statistic T to be any feature importance statistic of H. Just like the choice of candidate exposure h, the choice of test statistic T will not hurt the validity of the test, but will largely influence the power of the test.

Then a natural question to ask is whether we can make use of information from multiple experiments to further increase the power of the test. Suppose that we collect data from *two* experiments on the same n units indexed by $i = 1, \ldots, n$. In order to increase the power of the previous testing procedure, a natural idea is to reduce the variance in the test statistic computed in Algorithm 7. To do so, instead of focusing on $Y_{i,2}$ itself, we focus on $Y_{i,2} - Y_{i,1}$. This difference is helpful in removing variance of Y_i's that is shared by $Y_{i,1}$ and $Y_{i,2}$ but cannot be explained by the treatment and covariates. If a unit has some hidden individual characteristics, those characteristics could influence both $Y_{i,1}$ and $Y_{i,2}$ in a similar fashion but may not be well captured by the observed covariates. To make this intuition precise, we present Algorithm 8, which makes uses of information from two experiments and tests for the existence of interference effect. We have also included an illustration of the algorithm in Figure 4.2.

Algorithm 8 has a few key differences from Algorithm 7. First, the choices of focal units are different. In Algorithm 7, the choice of focal units cannot depend on the treatment assignments $W_{1:n}$, whereas in Algorithm 8, the focal units are randomly chosen from those whose treatment didn't change. This specific choice guarantees that the treatment of the i^{th} unit will not influence the difference of $Y_{i,2}$ and $Y_{i,1}$ much. Second, as mentioned above, in computing the test statistics, Y^{diff} is used instead of Y itself. As explained above, this helps reduce variance. Third, instead of regenerating treatment, Algorithm 8 permutes the treatment of the auxiliary units. This change is necessary to guarantee the procedure's validity; the choice of focal units depends on the treatment vector, and thus naively regenerating treatments will not give a valid procedure anymore. This will be demonstrated in Section 4.4.

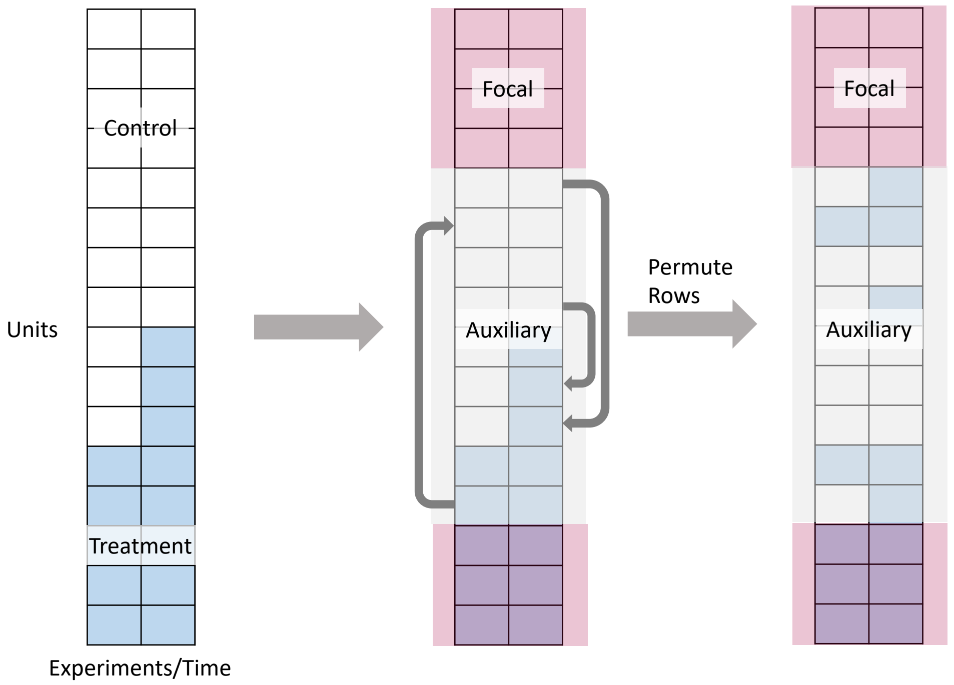


Figure 4.2: An illustration of Algorithm 8. After selecting the set of focal units and auxiliary units, we randomly permute rows of the treatment matrix and compute test statistics and p-values based on the permuted data.

Algorithm 8 Testing for interference effect (two experiments).

```
Input: Datasets $\mathcal{D}_1 = (W_{1:n,1}, X_{1:n}, Y_{1:n,1}, H_{1:n,1})$, $\mathcal{D}_2 = (W_{1:n,2}, X_{1:n}, Y_{1:n,2}, H_{1:n,2})$, exposure function $h$, test statistic $T$.

  1. Let $\mathcal{I}_{\mathrm{nc}} = \{i : W_{i,1} = W_{i,2}\}$ be the set of units whose treatment didn't change over the experiments. Randomly sample a subset of $\mathcal{I}_{\mathrm{nc}}$ of size $n/2$ if $|\mathcal{I}_{\mathrm{nc}}| > n/2$. We call the subset $\mathcal{I}_{\mathrm{foc}}$. Let $\mathcal{I}_{\mathrm{aux}} = [n] \setminus \mathcal{I}_{\mathrm{foc}}$.

  2. Take the difference of $Y_{\mathrm{foc},2}$ and $Y_{\mathrm{foc},1}$: let $Y^{\mathrm{diff}}_{\mathrm{foc}} = Y_{\mathrm{foc},2} - Y_{\mathrm{foc},1}$. Compute a test statistic $T^{(0)} = T(W_{\mathrm{foc},1:2}, X_{\mathrm{foc}}, Y^{\mathrm{diff}}_{\mathrm{foc}}, H_{\mathrm{foc},1:2})$ that captures the importance of $H$ in predicting $Y^{\mathrm{diff}}$.

  3. For $b = 1, \dots B$:

        Randomly permute treatments for the auxiliary units of the data: $\widetilde{W}^{(b)}_{i,1:2} = W_{\sigma^{(b)}(i),1:2}$ for $i \in \mathcal{I}_{\mathrm{aux}}$, for some permutation $\sigma^{(b)}$ of $\mathcal{I}_{\mathrm{aux}}$.

        Recompute the candidate exposure for the focal units: $\widetilde{H}^{(b)}_{i,k} = h_i(W_{\mathrm{foc}\setminus\{i\},k}, \widetilde{W}^{(b)}_{\mathrm{aux},k})$ for $i \in \mathcal{I}_{\mathrm{foc}}$ and $k \in \{1, 2\}$.

        Recompute the test statistic: $T^{(b)} = T(W_{\mathrm{foc},1:2}, X_{\mathrm{foc}}, Y^{\mathrm{diff}}_{\mathrm{foc}}, \widetilde{H}^{(b)}_{\mathrm{foc},1:2})$.

     End For

Output: The $p$-value
    $$p = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{1}\left\{T^{(0)} \leq T^{(b)}\right\}\right). \tag{4.5}$$
```

4.3.2 Testing with a time fixed effect model

In the previous section, we allow the existence of "arbitrary time effect". In particular, Hypothesis 1 allows the outcome $Y_{i,k}$ to depend on the treatments in other experiments, and does not restrict the relationship among outcomes in different experiments. This brings flexibility and generality, but it could reduce the power of the testing procedures. In this section, we make additional assumptions on the structure of time effect and propose a different testing procedure.

Assumption 29 (No temporal interference). $Y_{i,k}(w_{1:n,1:K}) = Y_{i,k}(\tilde{w}_{1:n,1:K})$ if $w_{1:n,k} = \tilde{w}_{1:n,k}$.

Assumption 29 states that the outcomes in experiment k depends only on treatments assigned in experiment k. In other words, the effect of treatment in one experiment will not carry over to the other experiments. Under Assumption 29, we can simplify the notation of potential outcomes: for any $w_{1:n} \in \{0,1\}^n$, we write $Y_{i,k}(w_{1:n})$ as the potential outcome and assume that the observed outcomes satisfy $Y_{i,k} = Y_{i,k}(W_{1:n,k})$. Note the difference from the previous notation. Previously, we wrote the potential outcomes $Y_{i,k}(w_{1:n,1:K})$ for any $w_{1:n,1:K} \in \{0,1\}^{n\times K}$. Here we focus on the potential outcomes $Y_{i,k}(w_{1:n})$ for any $w_{1:n} \in \{0,1\}^n$. Following this new notation, we make an additional assumption.

Assumption 30 (Time fixed effect). For any $w_{1:n} \in \{0,1\}^n$, $i \in \{1,\ldots,n\}$ and $k \in \{1,\ldots,K\}$, $Y_{i,k}(w_{1:n}) = \alpha_i(w_{1:n}) + u_k + \epsilon_{i,k}(w_{1:n})$. The random variables $\epsilon_{i,1}(w_{1:n}),\ldots,\epsilon_{i,K}(w_{1:n})$ are zero mean, and are independently and identically distributed, independently of functions $\alpha_{1:n}$, variables $u_{1:K}$, treatments $W_{1:n,1:K}$, covariates $X_{1:n}$ and other errors $\epsilon_{j,l}$ for $j \neq i$.

Assumption 30 assumes a time fixed effect model. The term u_k captures the time effect: some special events may happen during the k^{th} experiment, and Assumption 30 assumes that the effect of such events is shared by all units in the experiments. The term $\alpha_i(w)$ captures the individual effect, which could depend on the treatment of unit i as well as treatments of other units. Finally, the terms $\epsilon_{i,k}(w_{1:n})$'s are errors that are i.i.d. across experiments.

We also note that the commonly used *no temporal effect* assumption is a special case (stronger version) of Assumption 30. The no temporal effect assumption assumes that $Y_{i,k}(w_{1:n}) = \alpha_i(w_{1:n}) + \epsilon_{i,k}(w_{1:n})$, where the errors $\epsilon_{i,k}(w_{1:n})$'s are zero mean and i.i.d. across experiments. This corresponds to Assumption 30 with all time fixed effects $u_k = 0$. Such an assumption is particularly plausible when all the experiments are implemented within a short period of time, where the distribution of $Y_{i,k}(w_{1:n})$ is not expected to change much.

Assumption 29 and Hypothesis 1 together state that the outcome $Y_{i,k}$ depend only on the treatment of unit i in experiment k. Therefore, under Assumption 29 and Hypothesis 1, we can further simplify the notation of potential outcomes: for any $w \in \{0,1\}$, we write $Y_{i,k}(w)$ as the potential outcome and assume that the observed outcomes satisfy $Y_{i,k} = Y_{i,k}(W_{i,k})$.[3] With this new notation, Assumptions 29, 30 and Hypothesis 1 together become a new hypothesis:

Hypothesis 1'. For any $w \in \{0,1\}$, $i \in \{1,\ldots,n\}$ and $k \in \{1,\ldots,K\}$,

$$Y_{i,k}(w) = \alpha_i(w) + u_k + \epsilon_{i,k}(w), \tag{4.6}$$

such that the vectors $\epsilon_{1:n,1}(w),\ldots,\epsilon_{1:n,K}(w)$ are i.i.d., and independent of functions $\alpha_{1:n}$, vector $u_{1:K}$, treatments $W_{1:n,1:K}$, covariates $X_{1:n}$ and other errors $\epsilon_{j,l}(w)$ for $l \neq k$.

This corresponds to the two-way ANOVA [Yates, 1934, Fujikoshi, 1993] and the two-way fixed effect model [Bertrand et al., 2004, Angrist and Pischke, 2009] in statistics/economics literature.

In the previous section, we conduct some permutation tests that permute the data "vertically", i.e., permute different units. Here with the additional assumptions, we can conduct permutation tests that permute the data "horizontally", i.e., permute different time points or experiments.

To motivate the permutation test, consider two units i and j. Assume that i has been in the treatment group the whole time while j has been in the control group the whole time. Under Hypothesis 1', we have for the first experiment, $Y_{i,1} - Y_{j,1} = (\alpha_i(1) + u_1 + \epsilon_{i,1}(1)) - (\alpha_j(0) + u_1 + \epsilon_{j,1}(0)) =$

[3] Note again the difference with the previous notation. Here we focus on the potential outcomes $Y_{i,k}(w)$ for any $w \in \{0,1\}$, while we consider $w_{1:n,1:K} \in \{0,1\}^{n\times K}$ for the most general case and $w_{1:n} \in \{0,1\}^n$ assuming Assumption 29.

$\alpha_i(1) - \alpha_j(0) + \epsilon_{i,1}(1) - \epsilon_{j,1}(0)$, and for the second experiment, $Y_{i,2} - Y_{j,2} = \big(\alpha_i(1) + u_2 + \epsilon_{i,2}(1)\big) - \big(\alpha_j(0) + u_1 + \epsilon_{j,1}(0)\big) = \alpha_i(1) - \alpha_j(0) + \epsilon_{i,2}(1) - \epsilon_{j,2}(0)$. Thus,

$$\begin{aligned} Y_{i,1} - Y_{j,1} &= \alpha_i(1) - \alpha_j(0) + \epsilon_{i,1}(1) - \epsilon_{j,1}(0) \\ &\overset{d}{=} \alpha_i(1) - \alpha_j(0) + \epsilon_{i,2}(1) - \epsilon_{j,2}(0) = Y_{i,2} - Y_{j,2}. \end{aligned} \tag{4.7}$$

To put it simply, under Hypothesis 1', $Y_{i,1} - Y_{j,1}$ has the same distribution as $Y_{i,2} - Y_{j,2}$. However, the two distributions could be different when there is interference. Consider a simple model:

$$Y_{i,k} = W_{i,k} H_{i,k} + \epsilon_{i,k}, \tag{4.8}$$

where $H_{i,k}$ is the fraction of neighbors of unit i treated in experiment k, and $\epsilon_{i,k}$'s are some i.i.d. zero mean errors. Under this model, $Y_{i,1} - Y_{j,1} = H_{i,1} + \epsilon_{i,1} - \epsilon_{j,1}$ and $Y_{i,2} - Y_{j,2} = H_{i,2} + \epsilon_{i,2} - \epsilon_{j,2}$. When the number of neighbors of unit i is large, by law of large numbers, we have $H_{i,1} \approx \pi_1$ and $H_{i,2} \approx \pi_2$. We can then observe that $Y_{i,1} - Y_{j,1}$ and $Y_{i,2} - Y_{j,2}$ have different distributions; in particular, they have different means.

Given the above observation, we can conduct a permutation test permuting pairs of (i, j) across experiments. We outline the algorithm in Algorithm 9 and provide an illustration in Figure 4.3.

In Algorithm 9, we compare the value of a test statistic to the value of the statistic after permutation. One simple choice of test statistic is the difference-in-differences statistic:

$$T(Y^{\text{diff}}_{\mathcal{I}_1,1:2}, X_{\mathcal{I}_m}, H_{\mathcal{I}_m,1:2}, X_{\mathcal{I}_1}, H_{\mathcal{I}_1,1:2}) = \left|\text{mean}(Y^{\text{diff}}_{\mathcal{I}_1,2}) - \text{mean}(Y^{\text{diff}}_{\mathcal{I}_1,1})\right|, \tag{4.10}$$

where $\mathcal{I}_1$ and $\mathcal{I}_m$ are defined in the first step of Algorithm 9. We use the simple model (4.8) discussed above to explain why this choice of statistic is reasonable. Under model (4.8), the difference-in-differences statistic (without absolute value) will be

$$\text{mean}(Y^{\text{diff}}_{\mathcal{I}_1,2}) - \text{mean}(Y^{\text{diff}}_{\mathcal{I}_1,1}) \approx \text{mean}(H_{\mathcal{I}_1,2}) - \text{mean}(H_{\mathcal{I}_1,1}) \approx \pi_2 - \pi_1. \tag{4.11}$$

However, after permutation, the difference-in-differences statistic (without absolute value) will be mean zero. Therefore, $T^{(0)}$ and $T^{(b)}$ will have different distributions and thus the p-value will be far from the Unif$[0, 1]$ distribution.

One advantage of this difference-in-differences test statistic is its simplicity. To compute this statistic, there is no need of constructing a candidate exposure or any interference graph, and thus the computation cost of the test statistic is very low. This test statistic is also very intuitive to understand. Recall the motivating example in Section 4.1.1: when the difference-in-means estimators are different, the difference-in-differences test statistic is large. With this test statistic, our algorithm formalizes the intuition of the motivating example in Section 4.1.1.

The difference-in-differences statistic is not the only one we can choose. Indeed, just as for

Algorithm 9 Testing for interference effect (two experiments, time fixed effect model).

```
Input: Datasets $\mathcal{D}_1 = (W_{1:n,1}, X_{1:n}, Y_{1:n,1}, H_{1:n,1})$, $\mathcal{D}_2 = (W_{1:n,2}, X_{1:n}, Y_{1:n,2}, H_{1:n,2})$, matching algorithm $m$, test statistic $T$.

1. Let $\mathcal{I}_0 = \{i : W_{i,1} = W_{i,2} = 0\}$ and $\mathcal{I}_1 = \{i : W_{i,1} = W_{i,2} = 1\}$.
2. For each $i$ in $\mathcal{I}_1$, match an index $j \in \mathcal{I}_0$ to $i$ (with no repeat): let $m(i)$ be the matched index of $i$. Let $\mathcal{I}_m = \{m(i) : i \in \mathcal{I}_1\}$ be the set of matched indices. Here we assume that $|\mathcal{I}_1| < |\mathcal{I}_0|$. If $|\mathcal{I}_1| \geq |\mathcal{I}_0|$, we start with $\mathcal{I}_1$ instead.
3. For each $k \in \{1,2\}$, compute $Y^{\text{diff}}_{\mathcal{I}_1,k} = \left(Y_{i,k} - Y_{m(i),k}\right)_{i \in \mathcal{I}_1}$, which is the vector of differences between the outcomes of the treated units and those of the matched units.
   Compute a test statistic $T^{(0)} = T(Y^{\text{diff}}_{\mathcal{I}_1,1:2}, X_{\mathcal{I}_m}, H_{\mathcal{I}_m,1:2}, X_{\mathcal{I}_1}, H_{\mathcal{I}_1,1:2})$.
4. For $b = 1, \ldots B$:
       For each $i \in \mathcal{I}_1$:
           Randomly permute outcomes across experiments: $\widetilde{Y}^{(b)}_{i,k} = Y_{i,\sigma_{i,b}(k)}$ and
           $\widetilde{Y}^{(b)}_{m(i),k} = Y_{m(i),\sigma_{i,b}(k)}$ for some permutation $\sigma_{i,b}$ of $\{1,2\}$.
       End For
       Recompute $\widetilde{Y}^{\text{diff},(b)}_{\mathcal{I}_1,k} = (\widetilde{Y}^{(b)}_{i,k} - \widetilde{Y}^{(b)}_{m(i),k})_{i \in \mathcal{I}_1}$.
       Recompute the test statistic: $T^{(b)} = T(\widetilde{Y}^{\text{diff}(b)}_{\mathcal{I}_1,1:2}, X_{\mathcal{I}_m}, H_{\mathcal{I}_m,1:2}, X_{\mathcal{I}_1}, H_{\mathcal{I}_1,1:2})$.
   End For

Output: The $p$-value
    $p = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{1}\left\{T^{(0)} \leq T^{(b)}\right\}\right).$    (4.9)
```

Algorithms 7 and 8, we have full flexibility in choosing the test statistic. For example, we can add covariate adjustment into the test statistics: instead of taking the difference of mean($Y^{\text{diff}}_{\mathcal{I}_1,2}$) and mean($Y^{\text{diff}}_{\mathcal{I}_1,1}$), we can take the difference of the fitted intercepts after regressing Y^{diff}_1 (and Y^{diff}_2) on $X_{\mathcal{I}_m}$ and $X_{\mathcal{I}_1}$. We can also bring the candidate exposure H into the picture. For example, we can similarly define $H^{\text{diff}}_{\mathcal{I}_1,k} = \left(H_{i,k} - H_{m(i),k}\right)_{i \in \mathcal{I}_1}$ for $k \in \{1,2\}$ and consider the test statistic (when $H_{i,k} \in \mathbb{R}$):

$$\left|\text{Corr}\left[Y^{\text{diff}}_{\mathcal{I}_1,2} - Y^{\text{diff}}_{\mathcal{I}_1,1}, H^{\text{diff}}_{\mathcal{I}_1,2} - H^{\text{diff}}_{\mathcal{I}_1,1}\right]\right|. \tag{4.12}$$

Finally, we want to comment on the matching algorithm m used in Algorithm 9. We would first like to stress that as long as the matching algorithm only looks at the covariates X, the test will be valid regardless of the quality of matching. In the most extreme case, we can simply conduct a random matching, and the test will remain valid. More ideally, we would hope each i is matched to an $m(i)$ such that X_i is close to $X_{m(i)}$. This matching step helps reduce variance due to the covariates

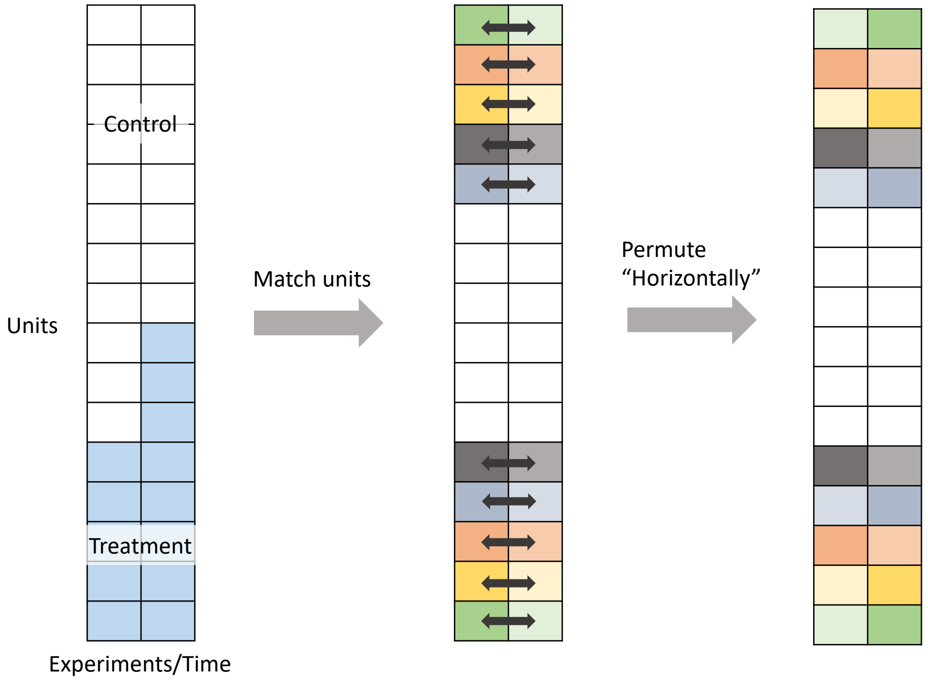


Figure 4.3: An illustration of Algorithm 9. Algorithm 9 permutes the outcomes across experiments, whereas Algorithm 8 permutes the treatments across units.

and thus increase the power of the test. In the causal inference literature, matching algorithms have been widely studied [Rubin, 1973, Stuart, 2010], and we recommend that experimenters choose from existing algorithms based on their needs and the computational resources available.

4.3.3 Usage of graphs of experimental units

In implementing the proposed algorithms, we often find it helpful to construct a graph of the n experimental units. Formally, let $G = (V, E)$, with vertex set $V = \{1, 2, \dots, n\}$ and edge set $E = \{E_{ij}\}_{i,j=1}^{n}$. We now discuss a few different ways of using graphs to test and learn interference structure.

Interference graph. A graph can be constructed to model interference and to help compute candidate exposure. We call such a graph an *interference graph*. When experimental units are users, it is plausible to assume that a user’s behavior is mostly influenced by friends in a social network. In this case, we can simply take the interference graph to be the social network, i.e., we set $E_{ij} = 1$ if user i and j are friends on the social network. With this graph, many candidate exposures can be computed easily: number of treated friends $H_{i,k}^{\text{numFrds}} = \sum_{j:E_{ij}=1} W_{j,k}$, fraction of friends that are treated $H_{i,k}^{\text{fracFrds}} = \sum_{j:E_{ij}=1} W_{j,k} / \left|\{j : E_{ij} = 1\}\right|$, number of treated two-hop friends $H_{i,k}^{\text{num2Frds}} = \sum_{l:\exists j \text{ s.t.} E_{ij}E_{jl}=1} W_{j,k}$. The interference graph can be constructed differently in other settings. When experimental units are advertisers, there is no natural social network. However, we can construct a "competition network" based on the similarity of the covariates. For a similarity measure s and a threshold ϵ, we can define $E_{ij} = \mathbb{1}\left\{s(X_i, X_i) \geq \epsilon\right\}$. Such a graph reflects that an advertiser is mostly influenced by its competitors, especially those that are similar to it. Candidate exposures can then be computed based on this interference graph: number of treated competitors $H_{i,k}^{\text{numCpt}} = \sum_{j:E_{ij}=1} W_{j,k}$, weighted average of competitors' treatments: $H_{i,k}^{\text{wAvgCpt}} = \sum_{j:E_{ij}=1} s(X_i, X_i) W_{j,k}$.

The interference graph also helps experimenters to understand the nature of interference. Imagine we have two different interference graphs G_1 and G_2 and we apply the testing procedure separately using G_1 and G_2. If we observe a much smaller p-value for the procedure using G_1 than that we obtain using G_2, then we have some evidence suggesting that the interference in the form of G_1 is much stronger than in that of G_2. In particular, the units that are connected to unit i in G_1 might be the most influential in impacting the outcome of unit i. This kind of analysis, though not fully rigorous, can help experimenters to build better intuitions for modelling in subsequent analysis. For example, once the interference effect is statistically significant, experimenters may consider re-running experiments with a cluster randomized controlled trial. Understanding the structure of interference can be helpful in constructing better clusters.

Graph in matching. A graph can also be helpful in the matching step in Algorithm 9. In the causal inference literature, matched pairs are often constructed using a minimum cost flow algorithm on a bipartite graph with treated units on one side and control units on the other side [Rosenbaum, 1989, Hansen and Klopfer, 2006]. Here, the cost of flow from unit i to j can be defined as some dissimilarity metric between X_i and X_j. For example, the Mahalanobis distance is a common choice of such a dissimilarity metric [Rubin, 1980]. The bipartite graph may not always be a complete bipartite graph: sometimes a caliper can be applied to the graph resulting in the removal of edges. A caliper based on covariates limits with which a unit can be paired [Mahmood, 2018].[4] For example, researchers may only want advertisers to be matched/paired with advertisers who sell products of the same category; in such cases, there is an edge between i and j only if they sell products of the same category.

Interestingly, calipered graphs may correspond to the interference graph introduced in the above section, and thus we only need to construct the graph once and use it in both the step of computing candidate exposure and the step of matching. This is especially relevant in a market competition application: a company is expected to be mostly influenced by companies selling similar products,

[4] In the observational study literature, calipers are often applied on the propensity score [Cochran and Rubin, 1973, Rosenbaum and Rubin, 1985]. Here we are in an experimental setting instead, where the propensity score is known and it is the same for all units.

and thus we put edges in the interference graph; in the mean time, we would like to match companies selling similar products, and thus we put edges in the bipartite graph used in matching.

4.3.4 Aggregating p-values

One issue with the algorithms proposed is that randomly splitting the data (Algorithms 7, 8 and 10) or the random matching step (Algorithms 9 and 11) can inject randomness into the p-value. In order to derandomize the procedure, we can run the algorithms many times and aggregate the p-values. Since the p-values can be arbitrarily dependent on each other, we cannot use Fisher's method to aggregate the p-values, which requires independence [Fisher, 1925]. Some possible ways include, e.g., setting $p = 2\sum p_i/n$ (See [Vovk and Wang, 2020] for more details).

In the previous section, we discuss the usage of an interference graph in constructing candidate exposure. In practice, experimenters may construct several interference graphs with different sparsity or structure. We can make use of information from different graphs and construct an "aggregated p-value". We can run the algorithms separately for each graph, and compute an "aggregated test statistic". For example, we can choose $T^{\text{aggre}} = \sum_m T(G_m)$, where G_m is the m^{th} interference graph considered. Then we can compute an aggregated p-value in the following way:

$$p^{\text{aggre}} = \frac{1}{B+1}\left(1 + \sum_{m=1}^{B} \mathbb{1}\left\{T^{\text{aggre}} \leq T^{\text{aggre}(b)}\right\}\right). \tag{4.13}$$

4.3.5 Extension to three or more experiments

More generally, experiments may be conducted more than two times. Formally, suppose that we run K experiments where treatments are randomly assigned according to (4.1) and (4.2). To test for interference, we can adopt a similar strategy as in Section 4.3.1. We outline the general algorithm in Algorithm 10. We note that Algorithm 8 is a special case of Algorithm 10. In practice, we recommend computing the test statistic using the difference of outcomes between experiments (as emphasized in Algorithm 8), since this helps remove common variance shared by outcomes in the experiments. One example of such statistic is the following.

$$\sum_{(k,l):k\neq l} \left|\text{Corr}\left[Y_{\text{foc},k} - Y_{\text{foc},l}, H_{\text{foc},l} - H_{\text{foc},l}\right]\right|. \tag{4.14}$$

If we assume a time fixed effect model as in Section 4.3.2, we can then extend Algorithm 9 to settings with more experiments. We outline the algorithm in Algorithm 11. Again, we note that Algorithm 9 is a special case of Algorithm 11. Algorithm 11 allows permutation over more experiments then Algorithm 9 does. In particular, if unit i is treated in experiments $K_1, K_1 + 1, \ldots, K$, then the algorithm permutes outcome for unit i and its matched unit over experiments $K_1, K_1 + 1, \ldots, K$. Permuting over more experiments helps the test to leverage information from

Algorithm 10 Testing for interference effect (multiple experiments).

```
Input: Datasets $\mathcal{D}_k = (W_{1:n,k}, X_{1:n}, Y_{1:n,k}, H_{1:n,k})$ for $k = 1, \ldots, K$, exposure function $h$, test statistic $T$.

1. Let $\mathcal{I}_{\text{nc}} = \{i : W_{i,1} = \cdots = W_{i,K}\}$ be the set of units whose treatment didn't change over the experiments. Randomly sample a subset of $\mathcal{I}_{\text{nc}}$ of size $n/2$. We call the subset $\mathcal{I}_{\text{foc}}$. Let $\mathcal{I}_{\text{aux}} = [n] \setminus \mathcal{I}_{\text{foc}}$.
2. Compute a test statistic $T^{(0)} = T(W_{\text{foc},1:K}, X_{\text{foc}}, Y_{\text{foc},1:K}, H_{\text{foc},1:K})$ that captures the importance of $H$ in predicting $Y$.
3. For $b = 1, \ldots B$:
        Randomly permute treatments for the auxiliary units of the data: $\widetilde{W}^{(b)}_{i,1:K} = W_{\sigma^{(b)}(i),1:K}$ for $i \in \mathcal{I}_{\text{aux}}$, for some permutation $\sigma^{(b)}$ of $\mathcal{I}_{\text{aux}}$.
        Recompute the candidate exposure for the focal units: $\widetilde{H}^{(b)}_{i,k} = h_i(W_{\text{foc}\setminus\{i\},k}, \widetilde{W}^{(b)}_{\text{aux},k})$, for $i \in \mathcal{I}_{\text{foc}}$ and $k \in \{1, 2, \ldots, K\}$.
        Recompute the test statistic: $T^{(b)} = T(W_{\text{foc},1:K}, X_{\text{foc}}, Y_{\text{foc},1:K}, \widetilde{H}^{(b)}_{\text{foc},1:K})$.
   End For

Output: The $p$-value
    $$p = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{1}\left\{T^{(0)} \leq T^{(b)}\right\}\right). \tag{4.15}$$
```

more experiments and thus increases power of the test. We have included an illustration of this algorithm in Figure 4.4.

Algorithm 11 Testing for interference effect (multiple experiments, time fixed effect model).

Input: Datasets $\mathcal{D}_k = (W_{1:n,k}, X_{1:n}, Y_{1:n,k}, H_{1:n,k})$ for $k = 1, \ldots, K$, matching algorithm m, test statistic T.

```
1. Let $\mathcal{I}_0 = \{i : W_{i,1} = \cdots = W_{i,k} = 0\}$ be the set of units that are in the control group in all
   experiments. Let $\mathcal{I}_1 = \{i : W_{i,K-1} = W_{i,K} = 1\}$ be the set of units that are in the treatment
   group in the last two experiments (this is to ensure that we can permute the treatment
   assignments across time).
2. For each $i$ in $\mathcal{I}_1$, match an index $j \in \mathcal{I}_0$ to $i$ (with no repeat): let $m(i)$ be the matched index
   of $i$. Let $\mathcal{I}_m = \{m(i) : i \in \mathcal{I}_1\}$ be the set of matched indices. Here we assume that $|\mathcal{I}_0| \geq n/2$.
3. For each $k \in \{1, \ldots, K\}$, compute $Y^{\text{diff}}_{\mathcal{I}_1,k} = \left(Y_{i,k} - Y_{m(i),k}\right)_{i \in \mathcal{I}_1}$, which is the vector of
   differences between the outcomes of the units in $\mathcal{I}_0$ and those of the matched units.
   Compute a test statistic $T^{(0)} = T(Y^{\text{diff}}_{\mathcal{I}_1,1:K}, X_{\mathcal{I}_m}, H_{\mathcal{I}_m,1:K}, X_{\mathcal{I}_1}, H_{\mathcal{I}_1,1:K})$.
4. For $b = 1, \ldots B$:
       For each $i \in \mathcal{I}_1$:
           Let $S_i = \{k : W_{i,k} = 1\}$ be the set of experiments in which unit $i$ is treated.
           Randomly permute outcomes across $S_i$: $\widetilde{Y}^{(b)}_{i,k} = Y_{i,\sigma_{i,b}(k)}$ and $\widetilde{Y}^{(b)}_{m(i),k} = Y_{m(i),\sigma_{i,b}(k)}$
           for all $k \in S_i$, where $\sigma_{i,b}$ is a random permutation of $S_i$.
       End For
       Recompute $\widetilde{Y}^{\text{diff},(b)}_{\mathcal{I}_1,k} = (\widetilde{Y}^{(b)}_{i,k} - \widetilde{Y}^{(b)}_{m(i),k})_{i \in \mathcal{I}_1}$.
       Recompute the test statistic: $T^{(b)} = T(\widetilde{Y}^{\text{diff}(b)}_{\mathcal{I}_1,1:K}, X_{\mathcal{I}_m}, H_{\mathcal{I}_m,1:K}, X_{\mathcal{I}_1}, H_{\mathcal{I}_1,1:K})$.
   End For
```

Output: The p-value

$$p = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{1}\left\{T^{(0)} \leq T^{(b)}\right\}\right). \tag{4.16}$$

4.4 Validity of the testing procedures

In this section, we establish validity of the above proposed algorithms. We make use of the following theorem in [Hemerik and Goeman, 2018a,b, Theorem 2].

Theorem 31 (Random permutations)**.** *Let $A_1, A_2, \ldots, A_n \in \mathcal{A}$ be n random variables. Let $\mathcal{S}_n$ denote the set of all permutations on $[n]$. Assume that*

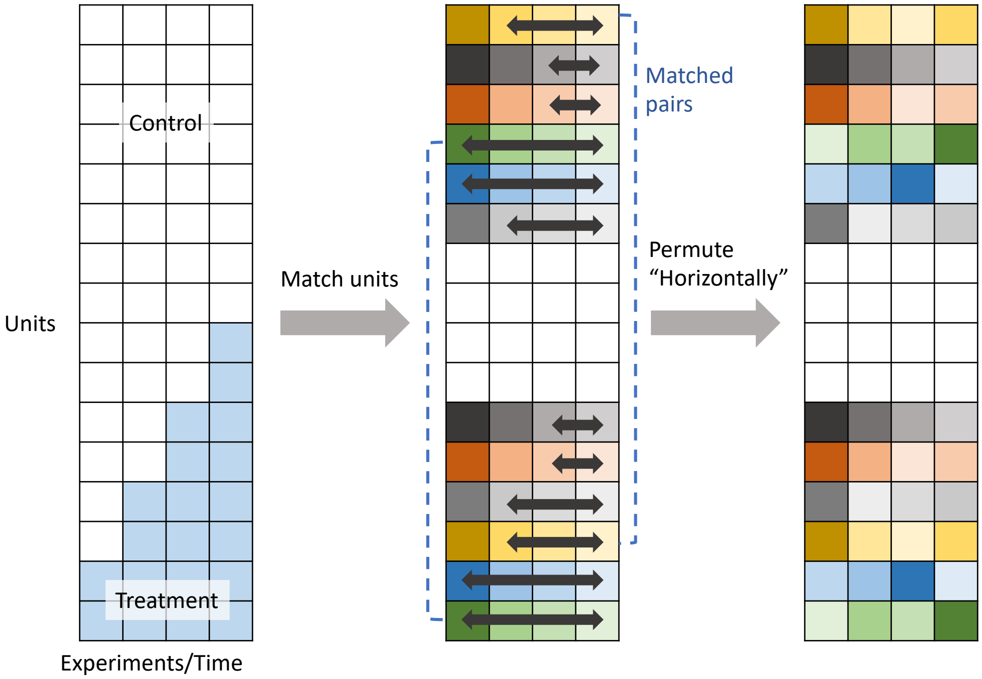


Figure 4.4: An illustration of Algorithm 11. Pairs of units are matched and the outcomes of paired units are permuted together across experiments.

1. *$G \subset \mathcal{S}_n$ is a subgroup;*
2. *For any $\sigma \in G$, $A = (A_1, \ldots, A_n) \overset{d}{=} (A_{\sigma(1)}, \ldots, A_{\sigma(n)}) = A_\sigma$.*

*If $\sigma_1, \ldots, \sigma_B$ are drawn independently uniformly from G, then for any test statistic T, the p-*value

$$p = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{1}\left\{T(A) \leq T(A_\sigma)\right\}\right) \tag{4.17}$$

satisfies

$$\mathbb{P}\left[p \leq \alpha\right] \leq \alpha. \tag{4.18}$$

for any $\alpha \in (0,1)$.

We start with establishing the validity of Algorithms 7, 8 and 10 under general assumptions.

Theorem 32 (General assumptions)**.** *Assume that the treatments are assigned according to rules defined in* (4.1) *and* (4.2)*. Under Hypothesis 1, the p-values produced by Algorithms 7, 8 and 10 are valid in the following sense: for any $\alpha \in (0,1)$,*

$$\mathbb{P}\left[p \leq \alpha\right] \leq \alpha. \tag{4.19}$$

Proof. Algorithm 7 has been shown to provide valid p-values in [Athey et al., 2018]. Since Algorithm 8 is a special case of Algorithm 10, it suffices to prove that the p-values produced by Algorithm 10 are valid. We will be making use of Theorem 31 to show the result.

We start by noting that since $H_{\text{foc},1:K}$ is a function of $W_{\text{foc},1:K}$ and $W_{\text{aux},1:K}$, the test statistic $T(W_{\text{foc},1:K}, X_{\text{foc}}, Y_{\text{foc},1:K}, H_{\text{foc},1:K})$ can be rewritten as

$$\begin{aligned} &T(W_{\text{foc},1:K}, X_{\text{foc}}, Y_{\text{foc},1:K}, H_{\text{foc},1:K}) \\ &\qquad = \check{T}(W_{\text{foc},1:K}, X_{\text{foc}}, Y_{\text{foc},1:K}, W_{\text{aux},1:K}) \end{aligned} \tag{4.20}$$

for some function $\check{T}$. Thus we can also rewrite

$$\begin{aligned} &T(W_{\text{foc},1:K}, X_{\text{foc}}, Y_{\text{foc},1:K}, \widetilde{H}^{(b)}_{\text{foc},1:K}) \\ &\qquad = \check{T}(W_{\text{foc},1:K}, X_{\text{foc}}, Y_{\text{foc},1:K}, \widetilde{W}^{(b)}_{\text{aux},1:K}). \end{aligned} \tag{4.21}$$

By construction, $\widetilde{W}^{(b)}_{\text{aux},1:K}$ is a random permutation of the rows of $W_{\text{aux},1:K}$. Thus we can take the permutation group G to be the set of all permutation on $\mathcal{I}_{\text{aux}}$. By Theorem 31, it suffices to establish that

$$\begin{aligned} &W_{\sigma(\text{aux}),1:K} \mid W_{\text{foc},1:K}, X_{\text{foc}}, Y_{\text{foc},1:K}, \mathcal{I}_{\text{foc}} \\ &\qquad \overset{d}{=} W_{\text{aux},1:K} \mid W_{\text{foc},1:K}, X_{\text{foc}}, Y_{\text{foc},1:K}, \mathcal{I}_{\text{foc}}, \end{aligned} \tag{4.22}$$

for any permutation $\sigma(\text{aux})$ on $\mathcal{I}_{\text{aux}}$. The above is equivalent to

$$\begin{aligned}&\left(W_{\sigma(\text{aux}),1:K}, W_{\text{foc},1:K}, X_{\text{foc}}, Y_{\text{foc},1:K}, \mathbb{1}\left\{\mathcal{I}_{\text{foc}} = \mathcal{I}_{\text{fix}}\right\}\right)\\ &\quad\overset{d}{=} \left(W_{\text{aux},1:K}, W_{\text{foc},1:K}, X_{\text{foc}}, Y_{\text{foc},1:K}, \mathbb{1}\left\{\mathcal{I}_{\text{foc}} = \mathcal{I}_{\text{fix}}\right\}\right),\end{aligned} \tag{4.23}$$

for any fixed subset $\mathcal{I}_{\text{fix}} \subset [n]$ of size $n/2$. Let $\mathcal{I}_{\text{fix}^c} = [n] \setminus \mathcal{I}_{\text{foc}}$. Then, under the null hypothesis 1,

$$\begin{aligned}&p\left(W_{\text{aux},1:K}, W_{\text{foc},1:K}, X_{\text{foc}}, Y_{\text{foc},1:K}, \mathbb{1}\left\{\mathcal{I}_{\text{foc}} = \mathcal{I}_{\text{fix}}\right\}\right)\\ &\quad= p\left(W_{\text{fix}^c,1:K}, W_{\text{fix},1:K}, X_{\text{fix}}, Y_{\text{fix},1:K}, \mathbb{1}\left\{\mathcal{I}_{\text{foc}} = \mathcal{I}_{\text{fix}}\right\}\right)\\ &\quad= p(W_{\text{fix}^c,1:K}) p(W_{\text{fix},1:K}, X_{\text{fix}}, Y_{\text{fix},1:K})\\ &\qquad\qquad\qquad \mathbb{P}\left[\mathcal{I}_{\text{foc}} = \mathcal{I}_{\text{fix}} \mid W_{\text{fix}^c,1:K}, W_{\text{fix},1:K}\right],\end{aligned} \tag{4.24}$$

where the last line follows from the no cross-unit interference hypothesis and the fact that treatments are sampled independently across units. Note also that permuting $\mathcal{I}_{\text{fix}^c}$ will not change the selection probability of the focal units, i.e., $\mathbb{P}\left[\mathcal{I}_{\text{foc}} = \mathcal{I}_{\text{fix}} \mid W_{\text{fix}^c}, W_{\text{fix}}\right] = \mathbb{P}\left[\mathcal{I}_{\text{foc}} = \mathcal{I}_{\text{fix}} \mid W_{\sigma(\text{fix}^c)}, W_{\text{fix}}\right]$, and thus

$$\begin{aligned}&p(W_{\text{fix}^c,1:K}) p(W_{\text{fix},1:K}, X_{\text{fix}}, Y_{\text{fix},1:K})\\ &\qquad\qquad\qquad \mathbb{P}\left[\mathcal{I}_{\text{foc}} = \mathcal{I}_{\text{fix}} \mid W_{\text{fix}^c,1:K}, W_{\text{fix},1:K}\right]\\ &\quad= p(W_{\sigma(\text{fix}^c),1:K}) p(W_{\text{fix},1:K}, X_{\text{fix}}, Y_{\text{fix},1:K})\\ &\qquad\qquad\qquad \mathbb{P}\left[\mathcal{I}_{\text{foc}} = \mathcal{I}_{\text{fix}} \mid W_{\sigma(\text{fix}^c),1:K}, W_{\text{fix},1:K}\right]\\ &\quad= p\left(W_{\sigma(\text{fix}^c),1:K}, W_{\text{fix},1:K}, X_{\text{fix}}, Y_{\text{fix},1:K}, \mathbb{1}\left\{\mathcal{I}_{\text{foc}} = \mathcal{I}_{\text{fix}}\right\}\right)\\ &\quad= p\left(W_{\sigma(\text{aux}),1:K}, W_{\text{foc},1:K}, X_{\text{foc}}, Y_{\text{foc},1:K}, \mathbb{1}\left\{\mathcal{I}_{\text{foc}} = \mathcal{I}_{\text{fix}}\right\}\right),\end{aligned} \tag{4.25}$$

and thus proving (4.23). □

Theorem 33 (Time fixed effect model)**.** *Assume that the treatments are assigned according to rules defined in* (4.1) *and* (4.2)*. Under Assumptions 29- 30 and Hypothesis 1, the p-values produced by Algorithms 9 and 11 are valid in the following sense: for any* $\alpha \in (0, 1)$*,*

$$\mathbb{P}\left[p \le \alpha\right] \le \alpha. \tag{4.26}$$

Proof. Algorithm 9 is a special case of Algorithm 11, and thus we will only work with Algorithm 11 here. We will again make use of Theorem 31 to show the result.

By construction, the elements in $\widetilde{Y}^{\text{diff},(b)}_{\mathcal{I}_1,1:K}$ are a random permutation of the elements in $Y^{\text{diff}}_{\mathcal{I}_1,1:K}$. The allowed permutations in Algorithm 11 clearly form a group. Specifically, the allowed permutations are defined by $\sigma = (\sigma_i)_{i\in\mathcal{I}_1}$, where each σ_i is a permutation of $S_i = \{k : W_{i,k} = 1\}$, and $\sigma(Y^{\text{diff}}_{i,k}) = Y^{\text{diff}}_{i,\sigma_i(k)}$. Following this notation, by Theorem 31, it suffices to show that for any allowed permutation σ,

$$\sigma(Y^{\text{diff}}_{\mathcal{I}_1,1:K}) \mid W_{1:n,1:K}, X_{1:n}, \mathcal{I}_m \overset{d}{=} Y^{\text{diff}}_{\mathcal{I}_1,1:K} \mid W_{1:n,1:K}, X_{1:n}, \mathcal{I}_m. \tag{4.27}$$

Under Assumptions 29 - 30 and Hypothesis 1, following (4.6), we can write $Y_{i,k}(w) = \alpha_i(w) + u_k + \epsilon_{i,k}(w)$. Therefore, for any $i \in \mathcal{I}_1$ and $k \in S_i$, $Y_{i,k} = Y_{i,k}(1) = \alpha_i(1) + u_k + \epsilon_{i,k}(1)$. At the same time, for the matched unit of i, we have $W_{m(i),k} = 0$, and thus $Y_{m(i),k} = Y_{m(i),k}(0) = \alpha_{m(i)}(0) + u_k + \epsilon_{m(i),k}(0)$. The difference of the two satisfies

$$\begin{aligned} Y^{\text{diff}}_{i,k} &= Y_{i,k} - Y_{m(i),k} \\ &= \alpha_i(1) + u_k + \epsilon_{i,k}(1) - \big(\alpha_{m(i)}(0) + u_k + \epsilon_{m(i),k}(0)\big) \\ &= \alpha_i(1) + \epsilon_{i,k}(1) - \alpha_{m(i)}(0) + \epsilon_{m(i),k}(0). \end{aligned} \tag{4.28}$$

Under Assumption 30,

$$\begin{aligned} &\big(\alpha_i(1) + \epsilon_{i,k}(1) - \alpha_{m(i)}(0) + \epsilon_{m(i),k}(0)\big) \mid W_{1:n,1:K}, X_{1:n}, \mathcal{I}_m, \alpha_{1:n} \\ &\overset{d}{=} \Big(\alpha_i(1) + \epsilon_{i,\sigma_i(k)}(1) - \alpha_{m(i)}(0) + \epsilon_{m(i),\sigma_i(k)}(0)\Big) \\ &\qquad\qquad\qquad\qquad \mid W_{1:n,1:K}, X_{1:n}, \mathcal{I}_m, \alpha_{1:n} \end{aligned} \tag{4.29}$$

for any permutation σ_i of S_i, because the errors $\epsilon_{i,k}$ and $\epsilon_{i,\sigma_i(k)}$ are i.i.d conditioning on $W_{1:n,1:K}, X_{1:n}$ and $\alpha_{1:n}$ (and same for $\epsilon_{m(i),k}$ and $\epsilon_{m(i),\sigma_i(k)}$). In addition, since all the errors $\epsilon_{i,k}$'s are independent conditioning on $W_{1:n,1:K}, X_{1:n}$ and $\alpha_{1:n}$, we have that

$$\begin{aligned} &\big(\alpha_i(1) + \epsilon_{i,k}(1) - \alpha_{m(i)}(0) + \epsilon_{m(i),k}(0)\big)_{i\in\mathcal{I}_1} \mid W_{1:n,1:K}, X_{1:n}, \mathcal{I}_m, \alpha_{1:n} \\ &\overset{d}{=} \Big(\alpha_i(1) + \epsilon_{i,\sigma_i(k)}(1) - \alpha_{m(i)}(0) + \epsilon_{m(i),\sigma_i(k)}(0)\Big)_{i\in\mathcal{I}_1} \\ &\qquad\qquad\qquad\qquad \mid W_{1:n,1:K}, X_{1:n}, \mathcal{I}_m, \alpha_{1:n}. \end{aligned} \tag{4.30}$$

Rewriting the above, we get

$$\begin{aligned} &Y^{\text{diff}}_{\mathcal{I}_1,1:K}, \alpha_{1:n} \mid W_{1:n,1:K}, X_{1:n}, \mathcal{I}_m \\ &\qquad \overset{d}{=} \sigma(Y^{\text{diff}}_{\mathcal{I}_1,1:K}) \mid W_{1:n,1:K}, X_{1:n}, \mathcal{I}_m, \alpha_{1:n}, \end{aligned} \tag{4.31}$$

which further implies (4.27) and hence gives the desired result. □

4.5 Simulations

In this section, we focus on a form of network interference. Specifically, we use a real-life social network to describe social interactions among units. We generate outcomes with some magnitude of network interference and evaluate our methods based on these generated outcomes. Our simulations can be viewed as semi-synthetic experiments—we use a real-life network but we generate outcomes according to some model.

We consider the Swarthmore network in the Facebook 100 dataset [Traud et al., 2012]. All networks in this dataset are complete online friendship networks for one hundred colleges and universities collected from a single-day snapshot of Facebook in September 2005. Here we focus on the Swarthmore college network in our simulation. To make the social network connected, we extract the largest connected component of the Swarthmore network. To summarize, the network we use is of size 1657 with 61049 edges. The diameter of the network is 6 and the average pairwise distance is 2.32.

Throughout this section, we assume that we have access to the data of three randomized experiments. We take treatment probabilities $\pi_1 = 10\%$, $\pi_2 = 25\%$ and $\pi_3 = 50\%$. In the following simulation studies, we consider level of significance $\alpha = 0.05$. Every dot on each plot is an average over 500 replications. We take $B = 200$.

4.5.1 Under general assumptions

We compare the power of the tests given in Algorithms 7, 8 and 10. We run Algorithm 10 using all three experiments, run Algorithm 8 using the second and the third experiments, and run Algorithm 7 using the third experiment, i.e., we always use experiments with the largest treatment probabilities. We discuss the choice of test statistics in Appendix 4.8. In Figure 4.5a, we assume a linear model of the outcome Y; in Figure 4.5b, we assume a nonlinear model. The details of the generating model can also be found in Appendix 4.8.

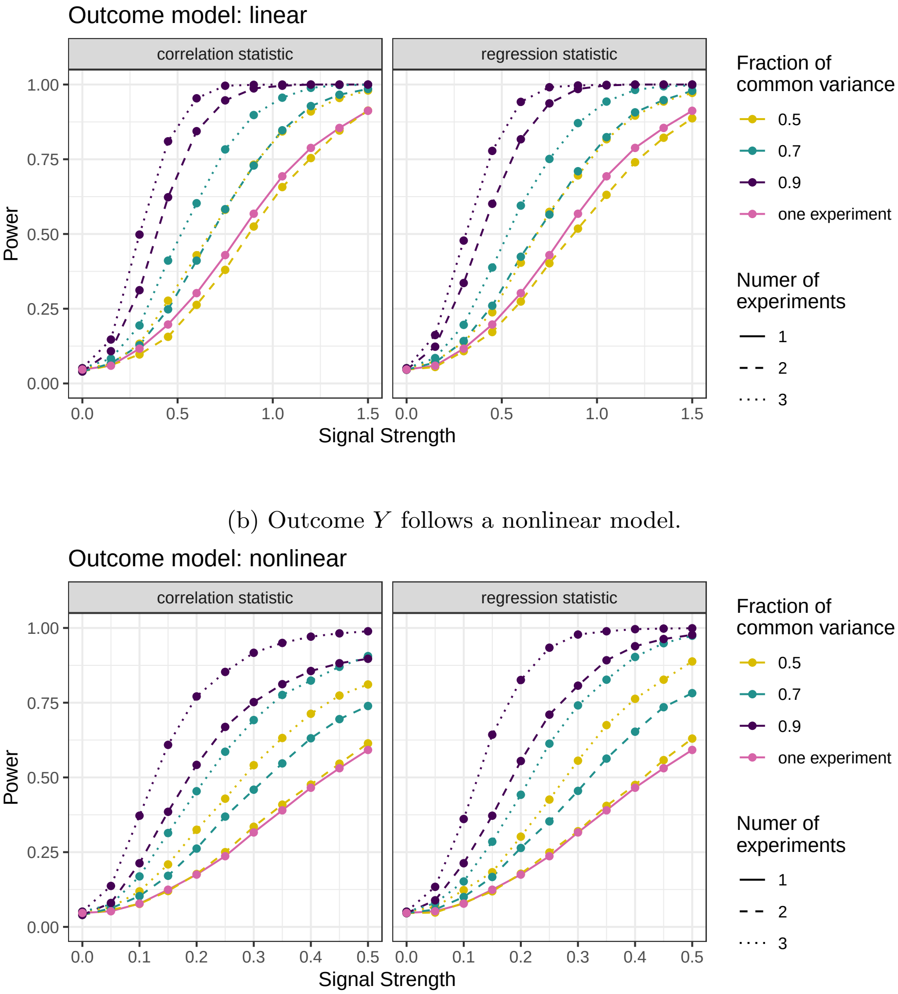
In Figures 4.5a and 4.5b, we plot the power of the testing algorithms 7, 8 and 10 at different levels of interference effects (signal strengths). In the figures, the fraction of common variance controls the correlation of the individual outcomes across experiments.

We observe from Figures 4.5a and 4.5b that utilizing more experiments helps our algorithms become more powerful, especially when the fraction of common variance is high. As discussed in Section 4.1.2, our work is the first to consider testing interference with multiple randomized experiments. Therefore, we can treat the algorithm utilizing one experiment as the *baseline method* that represents the state-of-the-art. Our algorithms appear to have a clear advantage over the baseline in terms of the power.

We also find that the regression statistic performs better than the correlation statistic, because the regression step helps reduce variance caused by the observed covariates.

4.5.2 Time fixed effect model

We compare the power of the tests given in Algorithms 10 and 11. We run both algorithms using all three experiments. We use a regression test statistic in both algorithms. We discuss the choice of



(a) Outcome Y follows a linear model.

(b) Outcome Y follows a nonlinear model.

Figure 4.5: Power of Algorithms 7, 8 and 10.

test statistics and matching algorithms in Appendix 4.8. In Figure 4.6a, we assume a linear model of the outcome Y, whereas in Figure 4.6b, we assume a nonlinear model. The details of the generating model can also be found in Appendix 4.8.

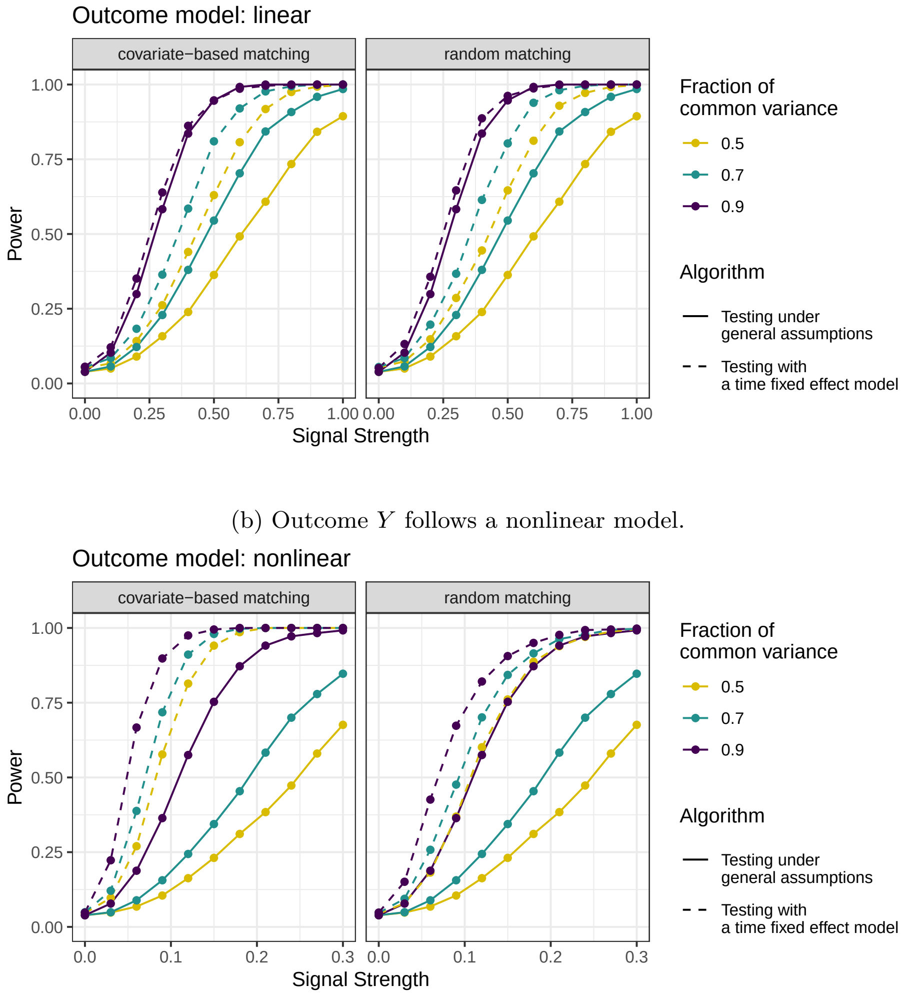
In Figures 4.6a and 4.6b, we plot the power of the testing algorithms 10 and 11 at different levels of interference effects (signal strengths). Algorithm 11 (testing with a time fixed effect model) appears more powerful than Algorithm 10 (testing under general assumptions). To understand this phenomenon, we recall that Algorithm 10 permutes data across experiments, whereas Algorithm 11 permutes data across units. Due to the nature of A/B tests, there is more variability in treatment allocation across experiments than across units. For example, assume that all units have around n_{ngb} neighbors in the social network. Looking at the fraction of neighbors in the treatment group, we find that the variation of this quantity across units is of scale $1/\sqrt{n_{\text{ngb}}}$, whereas the variation of this quantity across experiments is of constant scale. By permuting over data points that are more different, Algorithm 11 gains extra power.

Recall that there is a matching step in Algorithm 11. We find from Figure 4.6a and 4.6b that covariate-based matching outperforms random matching, especially under a nonlinear outcome model. In a linear model, the regression step has already removed almost all of the variance caused by observed covariates. In a nonlinear model, nevertheless, the regression step cannot fully remove all variance and the matching step can help further reduce variance.

4.6 Applications

In this section, we illustrate how the proposed procedure has been successfully implemented at LinkedIn as an add-on to their experimentation toolkit. Like other firms in the technology sector such as Google and Meta, LinkedIn makes business decisions in a data-driven manner and has a culture to "test everything". To support the needs to run concurrent A/B tests at scale, LinkedIn built an in-house experimentation platform, called T-REX (Targeting, Ramping, and Experimentation), which provides end-to-end experimentation supports [Xu et al., 2015, Ivaniuk, 2020]. Regardless of the application, T-REX implements simple Bernoulli randomization and relies on t-test for readout without taking into account potential interactions among experimental units.

This becomes a major limitation for experimentation in a marketplace environment, including the ads marketplace, where units on either side of the marketplace (advertisers and ad viewers) can interfere with each other [Basse et al., 2016, Pouget-Abadie et al., 2019b, Liu et al., 2021, Johari et al., 2022]. For example, ad campaigns that share the targeting audiences interfere with each other by competing in auctions for ad slots; different ad viewers with similar attributes are connected through the finite budget of certain ad campaigns. To remove bias in experiments caused by interference, LinkedIn has implemented the Budget-split platform on top of T-REX for experimentation in their ads marketplace [Liu et al., 2021].



(a) Outcome Y follows a linear model.

(b) Outcome Y follows a nonlinear model.

Figure 4.6: Power of Algorithms 10 and 11.

However, since Budget-split uses two halves of the marketplace to simulate the counterfactuals under different treatment variants, it does not support the classic factorial design. Under the current implementation, the platform only runs one experiment at a time, which is much smaller than the total number of experiments they need to run. This limitation in Budget-split capacity severely delays innovation: teams need to wait for weeks for a Budget-split slot in order to get an accurate measurement of their feature ramp before product launch. Nevertheless, not all ramps suffer from unit interaction, even in the ads marketplace setting. Running Budget-split experiments with negligible interference incurs a huge opportunity cost. Ideally, the Budget-split platform wants to prioritize tests that are impacted the most by the interference effects.

At LinkedIn, all feature launches start with small percentage ramps for risk mitigation and gradually increase the treatment percentage (i.e., 1%, 5%, 10%, 25%) before reaching the iteration for treatment effect measurement (50%) [Xu et al., 2018, Mao and Bojinov, 2021]. Specifically, Budget-split amounts to a 50% ramp on the viewers' side. This increasing allocation scheme provides us information to detect potential interference. With the algorithms proposed in this chapter, we implemented a screening step for each feature after the 25% iteration. The experiments are then ranked by the p-value in the interference test to determine their priority on the Budget-split platform.

It is important to note that the screening module was designed as an add-on to the system without touching LinkedIn's existing experimentation solution such as T-REX. By default, the interference detector only requires experimentation data in two previous iterations and runs Algorithm 9. Users have the option to provide additional network information that characterizes the potential interference mechanism among units and run other algorithms in this chapter. Because of this standalone nature, a similar interference detector can be readily added to any existing experimentation platforms to trigger alerts when interference might cause a problem.

As an illustration, we consider an online controlled experiment implemented by LinkedIn. The treatment in this experiment corresponds to a new feature that improves the quality of LinkedIn members' attribute for ads targeting. We run a series of experiments with increasing allocation with the members as the randomization units. Interference effect is expected in these experiments: when the allocation percentage is small, only a small set of members have the updated attributes, making them easier to be targeted by ad campaigns. Thus, when comparing metrics such as total ad impressions, these members tend to have larger average results than members in the control group. When the treatment allocation increases, more members get the improved attributes. Since the total ad budget does not increase much, the average difference between treatment and control units becomes smaller. Figure 4.1 shows the average differences between treatment and control units in the experiment series. Figure 4.7 shows the output from the interference detector after running Algorithm 9 based on the 10% and 25% iterations with respect to two different metrics. The p-values of the permutation test confirm the strong interference effects in these experiments.

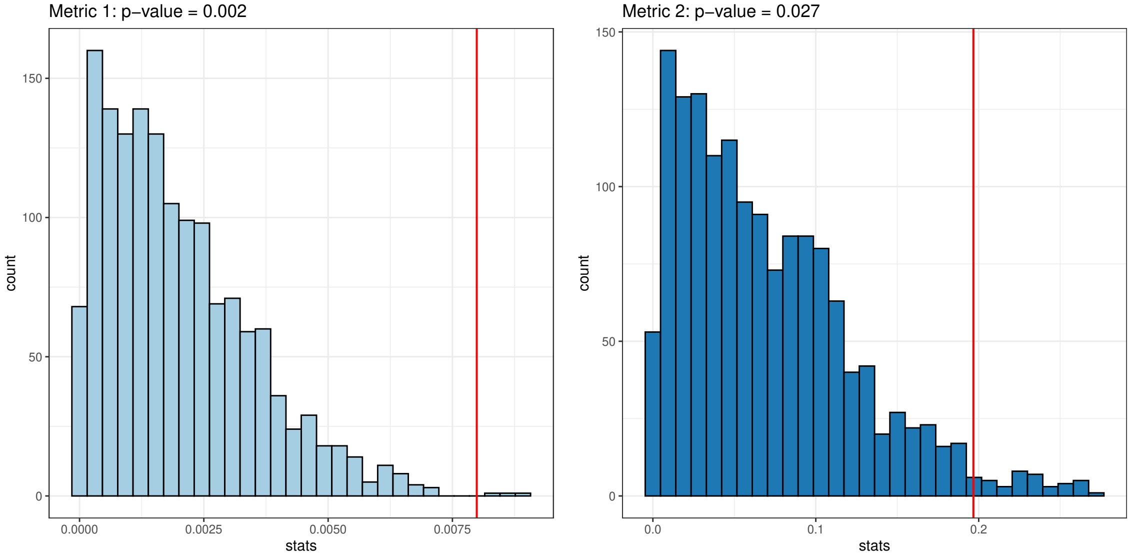


Figure 4.7: Example experiment: Test statistics and p-values from the permutation test. Results on two metrics are shown.

4.7 Discussion

Missingness In this chapter, we make the assumption that the dataset is complete. A natural future direction of work is to extend the current methods to scenarios with missing data. It is not hard to show that if the data is missing completely at random (MCAR), then the proposed testing procedures are still valid. When MCAR is unrealistic, it will be interesting to study whether our methods can still be applied under certain conditions. In practice, experimenters need to carefully examine the possible causes and consequences of missingness and make decisions correspondingly.

Selective inference We propose to use our testing procedure as a screening step for A/B testing: if the test suggests that no interference exists, then the experimenter can proceed with classical causal inference analysis. Strictly speaking, the data is used twice here—in the screening step and in the follow-up analysis. It would be of interest to understand the impact of the screening step on the follow-up analysis, and to develop valid statistical inference methods conditioning on the result of the screening step.

Sequential Testing Another question left open by this chapter is whether the proposed methods can be extended to the sequential testing setting. Our current procedure fixes the number of experiments a priori and constructs a single p-value from the permutation test. In real life, the treatment probability increases gradually and it would be of practical interest to end the experiment early as soon as we detect any interference. In that scenario, we need to take into account the randomness in stopping time and construct always valid p-values [Johari et al., 2017].

4.8 Appendix: simulation details

4.8.1 Under general assumptions

In Section 4.5.1, we compare the power of the tests given in Algorithms 7, 8 and 10.

Test statistics

Here, we discuss the test statistics used by the algorithms. Let $H_{i,k}$ be the fraction of treated neighbors of unit i in experiment k. Let N_i be the number of neighbors of unit i in the social network.

One experiment. For Algorithm 7, we use the following test statistic: run a linear regression of

$$Y_{\text{foc}} \sim W_{\text{foc}} + X_{\text{foc}} + N_{\text{foc}} + H_{\text{foc}}, \tag{4.32}$$

extract the regression coefficient of H and take the absolute value of the coefficient.

Two experiments. For Algorithm 8, we consider two different test statistics, a correlation statistic and a regression statistic. For the correlation statistic, we take

$$T(W_{\text{foc},1:2}, X_{\text{foc}}, Y_{\text{foc}}^{\text{diff}}, H_{\text{foc},1:2}) = \left|\text{Corr}\left[Y_{\text{foc}}^{\text{diff}}, H_{\text{foc},2} - H_{\text{foc},1}\right]\right|. \tag{4.33}$$

For the regression statistic, we run a regression of

$$Y_{\text{foc}}^{\text{diff}} \sim X_{\text{foc}} + N_{\text{foc}} + H_{\text{foc},1} + (H_{\text{foc},2} - H_{\text{foc},1}), \tag{4.34}$$

extract the regression coefficient of $(H_{\text{foc},2} - H_{\text{foc},1})$ and take the absolute value of the coefficient.

Three experiments. Let $T_{k,l}$ be the test statistic (regression or correlation) defined above when only two experiments are utilized (the k-th and l-th experiments are utilized). We then simply use $T_{1,2} + T_{2,3} + T_{1,3}$ as the test statistic for Algorithm 10 with $K = 3$.

Outcome models

We consider two different outcome models. For the linear model, let $H_{i,k}$ be the fraction of treated neighbors of unit i in experiment k. We assume

$$Y_{i,k} = (\text{signal strength})H_{i,k} + 2W_{i,k} + X_{i,1} + X_{i,2} + \varepsilon_{i,k}, \tag{4.35}$$

where $k \in \{1,2,3\}$ and $X_{i,1} \sim \mathcal{N}(0.5, 1)$, $X_{i,2} \sim \text{Poisson}(3)$ independently. The errors $\varepsilon_{i,k}$'s are such that $(\varepsilon_{i,1}, \ldots, \varepsilon_{i,K})$ is distributed as multivariate Gaussian with $\mathbb{E}[\varepsilon_{i,k}] = 0$, $\text{Var}[\varepsilon_{i,k}] = 1$ and $\text{Cov}[\varepsilon_{i,k}, \varepsilon_{i,l}] = (\text{fraction of common variance})$ for $k \neq l$.

For the non-linear model, let $M_{i,k}$ be the number of treated neighbors of unit i in experiment k. We assume

$$\begin{aligned} Y_{i,k} = (\text{signal strength}) \left(\frac{M_{i,k}}{20} + 5 \exp \left(\frac{1}{50} \min \left(M_{i,k}, 20 \right) \right) \right) + \\ 2W_{i,k} + X_{i,1} + X_{i,2} + \varepsilon_{i,k}, \end{aligned} \tag{4.36}$$

where $k \in \{1,2,3\}$ and $X_{i,1} \sim \mathcal{N}(0.5, 1)$, $X_{i,2} \sim \text{Poisson}(3)$ independently. The errors $\varepsilon_{i,k}$'s are such that $(\varepsilon_{i,1}, \ldots, \varepsilon_{i,K})$ is distributed as multivariate Gaussian with $\mathbb{E}[\varepsilon_{i,k}] = 0$, $\text{Var}[\varepsilon_{i,k}] = 1$ and $\text{Cov}[\varepsilon_{i,k}, \varepsilon_{i,l}] = (\text{fraction of common variance})$ for $k \neq l$.

4.8.2 Time fixed effect model

In Section 4.5.2, we compare the power of the tests given in Algorithms 10 and 11.

Test statistics

Here, we discuss the test statistics used by the algorithms. Let $H_{i,k}$ be the fraction of treated neighbors of unit i in experiment k. Let N_i be the number of neighbors of unit i in the social network.

Algorithm 10. We use the regression statistic defined in Section 4.5.1.

Algorithm 11. For Algorithm 11, we use an "anova" statistic. Let $\mathcal{I}_1' = \{i \in \mathcal{I}_1 : W_{i,1} = 1\}$ and let $\mathcal{I}_m' = \{m(i) : i \in \mathcal{I}_1'\}$. We start with concatenate $Y^{\text{diff}}_{\text{concat}} = \left(Y^{\text{diff}}_{\mathcal{I}_1',1}, Y^{\text{diff}}_{\mathcal{I}_1,2}, Y^{\text{diff}}_{\mathcal{I}_1,3} \right)$. Similarly, let $N_{\text{concat}} = (N_{\text{concat},1}, N_{\text{concat},m})$, where $N_{\text{concat},1} = \left(N_{\mathcal{I}_1',1}, N_{\mathcal{I}_1,2}, N_{\mathcal{I}_1,3} \right)$ and $N_{\text{concat},m} = \left(N_{\mathcal{I}_1',1}, N_{\mathcal{I}_1,2}, N_{\mathcal{I}_1,3} \right)$. We do the same concatenation for X and H. The reason we take the subset $\mathcal{I}_1'$ of $\mathcal{I}_1$ in the first experiment is that we want $Y^{\text{diff}}_{\text{concat}}$ to be a pure contrast of treatment group and control group. Without the subsetting step, Y^{diff} contains both treatment-control differences and control-control differences. Let Ind_2 be the indicator of the second experiment and Ind_3 be the indicator of the third experiment. We then run two regressions:

$$\begin{aligned} &\text{Model 1: } Y^{\text{diff}}_{\text{concat}} \sim X_{\text{concat}} + H_{\text{concat}} + N_{\text{concat}} + \text{Ind}_2 + \text{Ind}_3, \\ &\text{Model 2: } Y^{\text{diff}}_{\text{concat}} \sim X_{\text{concat}} + N_{\text{concat}}. \end{aligned} \tag{4.37}$$

Finally, we let the test statistic be the F-statistic from the anova testing of contrasting Model 1 with Model 2.

Matching algorithms

Random matching. We sample $m(i)$ uniformly at random without replacement.

Covariate-based matching. We use optimal matching based on the Mahalanobis distance of observed covariates and N_i Sekhon [2011].

Outcome models

We consider two different outcome models. For the linear model, let $H_{i,k}$ be the fraction of treated neighbors of unit i in experiment k. We assume

$$Y_{i,k} = (\text{signal strength})(2W_i + 1)H_{i,k} + 2W_{i,k} + X_{i,1} + X_{i,2} + \varepsilon_{i,k}, \tag{4.38}$$

where $k \in \{1, 2, 3\}$ and $X_{i,1} \sim \mathcal{N}(0.5, 1)$, $X_{i,2} \sim$ Poisson(3) independently. The errors $\varepsilon_{i,k}$'s are such that $(\varepsilon_{i,1}, \ldots, \varepsilon_{i,K})$ is distributed as multivariate Gaussian with $\mathbb{E}[\varepsilon_{i,k}] = 0$, $\text{Var}[\varepsilon_{i,k}] = 1$ and $\text{Cov}[\varepsilon_{i,k}, \varepsilon_{i,l}] = (\text{fraction of common variance})$ for $k \neq l$.

For the non-linear model, let $M_{i,k}$ be the number of treated neighbors of unit i in experiment k. We assume

$$\begin{aligned} Y_{i,k} = (\text{signal strength})(2W_i + 1)\left(\frac{M_{i,k}}{20} + 5\exp\left(\frac{1}{50}\min(M_{i,k}, 20)\right)\right) \\ + 2W_{i,k} + X_{i,1}X_{i,2} + \mathbb{1}\{X_{i,1} > 0.5, X_{i,2} > 3.5\} + \varepsilon_{i,k}, \end{aligned} \tag{4.39}$$

where $k \in \{1, 2, 3\}$ and $X_{i,1} \sim \mathcal{N}(0.5, 1)$, $X_{i,2} \sim$ Poisson(3) independently. The errors $\varepsilon_{i,k}$'s are such that $(\varepsilon_{i,1}, \ldots, \varepsilon_{i,K})$ is distributed as multivariate Gaussian with $\mathbb{E}[\varepsilon_{i,k}] = 0$, $\text{Var}[\varepsilon_{i,k}] = 1$ and $\text{Cov}[\varepsilon_{i,k}, \varepsilon_{i,l}] = (\text{fraction of common variance})$ for $k \neq l$.

Bibliography

Alberto Abadie, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296, 2020. doi: 10.3982/ECTA12675.

James Andreoni and Larry Samuelson. Building rational cooperation. *Journal of Economic Theory*, 127(1):117–154, 2006.

Joshua D Angrist and Guido M Kuersteiner. Causal effects of monetary shocks: Semiparametric conditional independence tests with a multinomial propensity score. *Review of Economics and Statistics*, 93(3):725–747, 2011.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

Peter M. Aronow. A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, 41(1):3–16, 2012.

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

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

Amel Awadelkarim and Johan Ugander. Prioritized restreaming algorithms for balanced graph partitioning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1877–1887, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403239.

Patrick Bajari, Brian Burdick, Guido W Imbens, Lorenzo Masoero, James McQueen, Thomas Richardson, and Ido M Rosen. Multiple randomization designs. *arXiv preprint arXiv:2112.13495*, 2021.

Eytan Bakshy, Dean Eckles, and Michael S Bernstein. Designing and deploying online field experiments. In *Proceedings of the 23rd international conference on World wide web*, pages 283–292, 2014.

G W Basse, A Feller, and P Toulis. Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494, 02 2019.

Guillaume Basse and Iavor Bojinov. A general theory of identification. *arXiv preprint arXiv:2002.06041*, 2020.

Guillaume Basse and Avi Feller. Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, 113(521):41–55, 2018. doi: 10.1080/01621459.2017.1323641.

Guillaume Basse, Yi Ding, and Panos Toulis. Minimax designs for causal effects in temporal experiments with treatment habituation. *arXiv e-prints*, art. arXiv:1908.03531, 2019.

Guillaume W. Basse and Edoardo M. Airoldi. Limitations of design-based causal inference and a/b testing under arbitrary and network interference. *Sociological Methodology*, 48(1):136–151, 2018. doi: 10.1177/0081175018782569.

Guillaume W Basse, Hossein Azari Soufiani, and Diane Lambert. Randomization and the pernicious effects of limited budgets on auction experiments. In *Artificial Intelligence and Statistics*, pages 1412–1420. PMLR, 2016.

GW Basse, A Feller, and P Toulis. Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494, 2019.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275, 2004.

Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. Causal inference under interference and network uncertainty. In *Uncertainty in Artificial Intelligence*, pages 1028–1038. PMLR, 2020.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705 – 1732, 2009. doi: 10.1214/08-AOS620.

Matthew Blackwell and Adam N Glynn. How to make causal inferences with time-series cross-sectional data under selection on observables. *American Political Science Review*, 112(4):1067–1082, 2018.

Iavor Bojinov and Neil Shephard. Time series experiments and causal estimands: Exact randomization tests and trading. *Journal of the American Statistical Association*, 114(528):1665–1682, 2019. doi: 10.1080/01621459.2018.1527225.

Iavor Bojinov, Prithwiraj Choudhury, and Jacqueline N Lane. Virtual watercoolers: A field experiment on virtual synchronous interactions and performance of organizational newcomers. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (21-125), 2021a.

Iavor Bojinov, Ashesh Rambachan, and Neil Shephard. Panel experiments and dynamic causal effects: A finite population perspective. *Quantitative Economics*, 12(4):1171–1196, 2021b.

Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *Management Science*, 2022.

Audrey Boruvka, Daniel Almirall, Katie Witkiewitz, and Susan A Murphy. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113 (523):1112–1121, 2018.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. *Introduction to Statistical Learning Theory*, pages 169–207. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-28650-9. doi: 10.1007/978-3-540-28650-9_8.

Jake Bowers, Mark M. Fredrickson, and Costas Panagopoulos. Reasoning about interference between units: A general framework. *Political Analysis*, 21(1):97–124, 2013. doi: 10.1093/pan/mps038.

Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Peer effects in networks: A survey. *Annual Review of Economics*, 12:603–629, 2020.

Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55, 2009. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2008.12.021.

Jing Cai, Alain De Janvry, and Elisabeth Sadoulet. Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108, April 2015. doi: 10.1257/app.20130442.

Gruia Calinescu, Howard Karloff, and Yuval Rabani. Approximation algorithms for the 0-extension problem. *SIAM Journal on Computing*, 34(2):358–372, 2005.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 08 2008. ISSN 0034-6535. doi: 10.1162/rest.90.3.414.

Alex Chin. Central limit theorems via Stein's method for randomized experiments under interference. *arXiv e-prints*, art. arXiv:1804.03105, 2018.

Alex Chin. Regression adjustments for estimating the global treatment effect in experiments with interference. *Journal of Causal Inference*, 7(2):20180026, 2019. doi: doi:10.1515/jci-2018-0026.

William G Cochran and Donald B Rubin. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446, 1973.

Mayleen Cortez, Matthew Eichhorn, and Christina Lee Yu. Graph agnostic estimators with staggered rollout designs under network interference. *arXiv preprint arXiv:2205.14552*, 2022.

D. R. Cox. *Planning of Experiments*. New York, Wiley, 1958.

Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 123–132, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318693. doi: 10.1145/2433396.2433413.

Peng Ding. A Paradox from Randomization-Based Causal Inference. *Statistical Science*, 32(3):331 – 345, 2017. doi: 10.1214/16-STS571.

DoorDash. Switchback tests and randomized experimentation under network effects at doordash, 2018. URL `https://medium.com/@DoorDash/switchback-tests-and-randomized-experimentation-under-network-effects-at-doordash-f1d938ab7c2a`

Esther Duflo and Emmanuel Saez. The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment. *The Quarterly Journal of Economics*, 118(3):815–842, 08 2003. ISSN 0033-5533. doi: 10.1162/00335530360698432.

Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1):20150021, 2017. doi: doi: 10.1515/jci-2015-0021.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979. doi: 10.1214/aos/1176344552.

B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994. ISBN 9780412042317.

Ronald Aylmer Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, 1925.

Andrey Fradkin. A simulation approach to designing digital matching platforms. *Boston University Questrom School of Business Research Paper Forthcoming*, 2019.

Yasunori Fujikoshi. Two-way anova models with unbalanced data. *Discrete Mathematics*, 116(1-3): 315–334, 1993.

Kevin Guo and Guillaume Basse. The generalized Oaxaca-Blinder estimator. *Journal of the American Statistical Association*, pages 1–13, 2021. doi: 10.1080/01621459.2021.1941053.

M. Elizabeth Halloran and Claudio J. Struchiner. Causal inference in infectious diseases. *Epidemiology*, 6(2):142–151, 1995. ISSN 10443983.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

David K. Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2010.04.005.

Kevin Wu Han, Iavor Bojinov, and Guillaume Basse. Population interference in panel experiments, 2021. URL `https://arxiv.org/abs/2103.00553`.

Ben B Hansen and Stephanie Olsen Klopfer. Optimal full matching and related designs via network flows. *Journal of computational and Graphical Statistics*, 15(3):609–627, 2006.

Christopher Harshaw, Fredrik Sävje, David Eisenstat, Vahab Mirrokni, and Jean Pouget-Abadie. Design and analysis of bipartite experiments under a linear exposure-response model. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, page 606, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391504. doi: 10.1145/3490486.3538269.

Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *Test*, 27(4):811–825, 2018a.

Jesse Hemerik and Jelle J Goeman. False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):137–155, 2018b.

Keith Henderson, Brian Gallagher, Lei Li, Leman Akoglu, Tina Eliassi-Rad, Hanghang Tong, and Christos Faloutsos. It's who you know: Graph mining using recursive structural features. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 663–671, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450308137. doi: 10.1145/2020408.2020512.

Henning Hohnhold, Deirdre O'Brien, and Diane Tang. Focusing on the long-term: It's good for users and business. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1849–1858, 2015.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ISSN 01621459.

David Holtz, Ruben Lobel, Inessa Liskovich, and Sinan Aral. Reducing interference bias in online marketplace pricing experiments. *arXiv preprint arXiv:2004.12489*, 2020.

Guanglei Hong and Stephen W Raudenbush. Evaluating kindergarten retention policy. *Journal of the American Statistical Association*, 101(475):901–910, 2006. doi: 10.1198/016214506000000447.

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. doi: 10.1080/01621459.1952.10483446.

Yuchen Hu, Shuangning Li, and Stefan Wager. Average direct and indirect causal effects under interference. *Biometrika*, 02 2022. ISSN 1464-3510. doi: 10.1093/biomet/asac008. URL `https://doi.org/10.1093/biomet/asac008`. asac008.

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

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Alexander Ivaniuk. Our evolution towards t-rex: The prehistory of experimentation infrastructure at linkedin. *LinkedIn Engineering Blog*, 2020.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1517–1525, New York, NY, USA, 2017. Association for Computing Machinery.

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

Ruth A Judson and Ann L Owen. Estimating dynamic panel data models: a guide for macroeconomists. *Economics Letters*, 65(1):9–15, 1999. ISSN 0165-1765. doi: https://doi.org/10.1016/S0165-1765(99)00130-5.

Brian Karrer, Liang Shi, Monica Bhole, Matt Goldman, Tyrone Palmer, Charlie Gelman, Mikael Konutgan, and Feng Sun. Network experimentation at scale. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 3106–3116, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467091.

Oscar Kempthorne. The randomization theory of experimental inference. *Journal of the American Statistical Association*, 50(271):946–967, 1955. ISSN 01621459.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 1168–1176, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450321747. doi: 10.1145/2487575.2488217. URL `https://doi.org/10.1145/2487575.2488217`.

Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, 2020. doi: 10.1017/9781108653985.

Denis Kojevnikov. The bootstrap for network dependent processes. *arXiv preprint arXiv:2101.12312*, 2021. doi: 10.48550/ARXIV.2101.12312.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907 – 927, 2016. doi: 10.1214/15-AOS1371.

Michael P. Leung. Treatment and spillover effects under network interference. *The Review of Economics and Statistics*, 102(2):368–380, 05 2020. ISSN 0034-6535. doi: 10.1162/rest_a_00818.

Michael P. Leung. Causal inference under approximate neighborhood interference. *Econometrica*, 90(1):267–293, 2022. doi: https://doi.org/10.3982/ECTA17841.

Hannah Li, Geng Zhao, Ramesh Johari, and Gabriel Y Weintraub. Interference, bias, and variance in two-sided marketplace experimentation: Guidance for platforms. In *Proceedings of the ACM Web Conference 2022*, pages 182–192, 2022.

Shuangning Li and Stefan Wager. Random graph asymptotics for treatment effect estimation under network interference. *The Annals of Statistics*, 50(4):2334 – 2358, 2022. doi: 10.1214/22-AOS2191. URL `https://doi.org/10.1214/22-AOS2191`.

Xinran Li and Peng Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769, 2017. doi: 10.1080/01621459.2017.1295865.

Xinran Li, Peng Ding, Qian Lin, Dawei Yang, and Jun S. Liu. Randomization inference for peer effects. *Journal of the American Statistical Association*, 114(528):1651–1664, 2019. doi: 10.1080/01621459.2018.1512863.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, 7(1):295–318, 2013. doi: 10.1214/12-AOAS583.

Lan Liu and Michael 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. doi: 10.1080/01621459.2013.844698.

Min Liu, Jialiang Mao, and Kang Kang. Trustworthy and powerful online marketplace experimentation with budget-split design. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3319–3329, 2021.

Shan Luo and Zehua Chen. Sequential lasso cum EBIC for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association*, 109(507):1229–1240, 2014. doi: 10.1080/01621459.2013.877275.

Sharif Mahmood. The performance of largest caliper matching: A monte carlo simulation approach. *arXiv preprint arXiv:1806.02149*, 2018.

Charles F. Manski. Identification of Endogenous Social Effects: The Reflection Problem. *The Review of Economic Studies*, 60(3):531–542, 07 1993. ISSN 0034-6527. doi: 10.2307/2298123.

Charles F. Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013. doi: 10.1111/j.1368-423X.2012.00368.x.

Jialiang Mao and Iavor Bojinov. Quantifying the value of iterative experimentation. *arXiv preprint arXiv:2111.02334*, 2021.

Robert A. Moffit. Policy Interventions, Low-Level Equilibria, and Social Interactions. In *Social Dynamics*. The MIT Press, 04 2001. ISBN 9780262272056. doi: 10.7551/mitpress/6294.003.0005.

Joel Nishimura and Johan Ugander. Restreaming graph partitioning: Simple versatile algorithms for advanced balancing. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 1106–1114, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450321747. doi: 10.1145/2487575.2487696.

J Pouget-Abadie, G Saint-Jacques, M Saveski, W Duan, S Ghosh, Y Xu, and E M Airoldi. Testing for arbitrary interference on experimentation platforms. *Biometrika*, 106(4):929–940, 09 2019a. ISSN 0006-3444. doi: 10.1093/biomet/asz047.

Jean Pouget-Abadie, Kevin Aydin, Warren Schudy, Kay Brodersen, and Vahab Mirrokni. Variance reduction in bipartite experiments through correlation clustering. *Advances in Neural Information Processing Systems*, 32, 2019b.

Jean Pouget-Abadie, Guillaume Saint-Jacques, Martin Saveski, Weitao Duan, S Ghosh, Y Xu, and Edoardo M Airoldi. Testing for arbitrary interference on experimentation platforms. *Biometrika*, 106(4):929–940, 2019c.

David Puelz, Guillaume Basse, Avi Feller, and Panos Toulis. A graph-theoretic approach to randomization tests of causal effects under general interference. *Journal of the Royal Statistical Society Series B*, 84(1):174–204, February 2022.

Ashesh Rambachan and Neil Shephard. Econometric analysis of potential outcomes time series: instruments, shocks, linearity and the causal response function. *arXiv preprint arXiv:1903.01637*, 2019.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(78):2241–2259, 2010.

James 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.

James M. Robins, Sander Greenland, and Fu-Chang Hu. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association*, 94(447):687–700, 1999. doi: 10.1080/01621459.1999.10474168.

Joseph P. Romano and Michael Wolf. A more general central limit theorem for m-dependent random variables with unbounded m. *Statistics & Probability Letters*, 47(2):115 – 124, 2000. ISSN 0167-7152. doi: https://doi.org/10.1016/S0167-7152(99)00146-7.

Paul R Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032, 1989.

Paul R Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200, 2007. doi: 10.1198/016214506000001112.

Paul R Rosenbaum and Donald B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

Nathan Ross. Fundamentals of stein's method. *Probability Surveys*, 8:210–293, 2011. doi: 10.1214/11-PS182.

Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Donald B Rubin. Bias reduction using mahalanobis-metric matching. *Biometrics*, pages 293–298, 1980.

Guillaume Saint-Jacques, Maneesh Varshney, Jeremy Simpson, and Ya Xu. Using ego-clusters to measure network effects at linkedin, 2019.

Martin Saveski, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M Airoldi. Detecting network effects: Randomizing over randomized experiments. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1027–1035, 2017.

Fredrik Sävje, Peter M. Aronow, and Michael G. Hudgens. Average treatment effects in the presence of unknown interference. *The Annals of Statistics*, 49(2):673 – 701, 2021. doi: 10.1214/20-AOS1973. URL `https://doi.org/10.1214/20-AOS1973`.

Jasjeet S. Sekhon. Multivariate and propensity score matching software with automated balance optimization: The matching package for r. *Journal of Statistical Software*, 42(7):1–52, 2011. doi: 10.18637/jss.v042.i07. URL `https://www.jstatsoft.org/index.php/jss/article/view/v042i07`.

Lizhen Shi and Bo Chen. Comparison and benchmark of graph clustering algorithms. *arXiv preprint arXiv:2005.04806*, 2020.

Betsy Sinclair, Margaret McConnell, and Donald P. Green. Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science*, 56(4):1055–1069, 2012. doi: 10.1111/j.1540-5907.2012.00592.x.

Michael E Sobel. What do randomized studies of housing mobility demonstrate? *Journal of the American Statistical Association*, 101(476):1398–1407, 2006a. doi: 10.1198/016214506000000636.

Michael E Sobel. What do randomized studies of housing mobility demonstrate? *Journal of the American Statistical Association*, 101(476):1398–1407, 2006b.

Daniel A. Spielman and Shang-Hua Teng. A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM Journal on Computing*, 42(1):1–26, 2013. doi: 10.1137/080744888.

Elizabeth 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.

Daniel L Sussman and Edoardo M Airoldi. Elements of estimation theory for causal effects in the presence of network interference. *arXiv preprint arXiv:1702.03578*, 2017.

Fredrik Sävje. Causal inference with misspecified exposure mappings, 2021. URL `https://arxiv.org/abs/2103.06471`.

Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 17–26, 2010.

Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246.

Panos Toulis and Edward Kao. Estimation of causal peer influence effects. In *International conference on machine learning*, pages 1489–1497. PMLR, 2013.

Amanda L. Traud, Peter J. Mucha, and Mason A. Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2011.12.021.

Johan Ugander and Hao Yin. Randomized graph cluster randomization. *arXiv preprint arXiv:2009.02297*, 2020. doi: 10.48550/ARXIV.2009.02297.

Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 329–337, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450321747.

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3(none):1360 – 1392, 2009. doi: 10.1214/09-EJS506.

Gonzalo Vazquez-Bare. Identification and estimation of spillover effects in randomized experiments. *Journal of Econometrics*, 2022. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2021.10.014.

Davide Viviano. Experimental design under network interference. *arXiv preprint arXiv:2003.08421*, 2020.

U. von Luxburg and B. Schölkopf. *Statistical Learning Theory: Models, Concepts, and Results*, volume 10, pages 651–706. Elsevier North Holland, Amsterdam, Netherlands, 2011. doi: 10.1016/B978-0-444-52936-7.50016-1.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.

Stefan Wager and Kuang Xu. Experimenting in equilibrium. *Management Science*, 67(11):6694–6715, 2021.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Jefrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data.* The MIT Press, 2010. ISBN 9780262232586.

Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. From infrastructure to culture: A/b testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2227–2236, 2015.

Ya Xu, Weitao Duan, and Shaochen Huang. Sqr: Balancing speed, quality and risk in online experiments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 895–904, 2018.

Frank Yates. The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Association*, 29(185):51–66, 1934.

Christina Lee Yu, Edoardo M Airoldi, Christian Borgs, and Jennifer T Chayes. Estimating the total treatment effect in randomized experiments with unknown network structure. *Proceedings of the National Academy of Sciences*, 119(44):e2208975119, 2022.