

Modeling Interference Using Experiment Roll-out

Ariel Boyarsky¹ Hongseok Namkoong¹ Jean Pouget-Abadie²

¹Decision, Risk, and Operations Division, Columbia Business School, ²Google Research

{aboyarsky26, namkoong}@gsb.columbia.edu, jeanpa@google.com

Abstract

Experiments on online marketplaces and social networks suffer from *interference*, where the outcome of a unit is impacted by the treatment status of other units. We propose a framework for modeling interference using a ubiquitous deployment mechanism for experiments, staggered *roll-out* designs, which slowly increase the fraction of units exposed to the treatment to mitigate any unanticipated adverse side effects. Our main idea is to leverage the temporal variations in treatment assignments introduced by roll-outs to model the interference structure. Since there are often multiple competing models of interference in practice we first develop a model selection method that evaluates models based on their ability to explain outcome variation observed along the roll-out. Through simulations, we show that our heuristic model selection method, LEAVE-ONE-PERIOD-OUT, outperforms other baselines. Next, we present a set of model identification conditions under which the estimation of common estimands is possible and show how these conditions are aided by roll-out designs. We conclude with a set of considerations, robustness checks, and potential limitations for practitioners wishing to use our framework.

1 Introduction

Experimentation is at the core of scientific decision-making in many online platforms. However, many of the experiments run by these platforms suffer from interference, meaning an intervention on a participant might impact the outcome of other participants. For example, interference is problematic in online marketplaces where strategic agents compete for scarce resources (Pouget-Abadie et al., 2018; Ugander and Yin, 2020; Johari et al., 2020; Candogan et al., 2021; Brennan et al., 2022; Bojinov et al., 2022a; Bojinov and Gupta, 2022; Bright et al., 2022), and in online social networks (Eckles et al., 2016; Chang et al., 2022) where peers interact and influence each other. Interference among units leads to a violation of the Stable Unit Treatment Value Assumption (Rubin, 2005; Imbens and Rubin, 2015) and the causal effect may not even be identifiable in a randomized experiment.

A common practical approach to estimating causal effects when interference is present is to assume a specific potential outcome structure based on the network(s) between participants. Together, the postulated potential outcome structure and the underlying network(s) constitute an *interference model*, which can be used to estimate the true causal effect directly (Eckles et al., 2017; Biswas and Airoidi, 2018), or modify the experimental design to mitigate the effects of interference (Basse and Airoidi, 2018). As a motivating example, which we will explore further in Section 2, consider an online auction platform incrementally rolling out an experiment that tests the effectiveness of increasing reserve prices. Treating one auction is likely to affect each participating advertisers, which can in turn affect other auctions in which they participate. Practitioners have postulated several working models of interference for how advertisers and auctions interact with one another. Each modeling choice might lead to a different estimated treatment effect.

Such ad-hoc structural modeling, however, is often unreliable in practice since the interference network is rarely known and the potential outcome structure is almost always misspecified. As a result, selecting and validating models of interference is a crucial—but difficult—step of estimating causal effects when interference is present. Furthermore, even when committing to a fixed interference model, estimating its parameters may be infeasible if the observed data does not contain enough variation in treatment exposure across units.

We propose a framework for modeling interference using a ubiquitous deployment mechanism for experiments, *staggered roll-out designs*. Online platforms rely on staggered roll-out designs (Kohavi et al., 2009; Xu et al., 2015, 2018) as an early detection tool for any unintended consequences caused by a new experiment or product launch, e.g., software bugs or adverse participant responses. Roll-out designs follow a simple principle: instead of intervening on all participants marked for the intervention all at once, the proportion of participants exposed to the intervention is increased incrementally until all participants marked for the intervention have been intervened on. The practice is nearly universal across standard experimentation infrastructures; for example, Xu et al. (2018) notes that out of 5000+ experiments run annually at LinkedIn, “every experiment goes through a *ramp-up/roll-out* process.”

In this paper, we propose to leverage existing roll-out designs for a new purpose: to better select and estimate models of interference in causal estimation when interference is present. Despite its ubiquity, the temporal variation in treatment exposure induced by roll-out designs is a promising yet largely overlooked consequence of how experiments are implemented in practice. Variations in treatment proportions, whether temporal (Bojinov et al., 2022b; Han et al., 2022) or spatial (Athey et al., 2018; Baird et al., 2018), are key to modeling interference and validating modeling choices.

As our first main contribution, we utilize roll-out experimental designs to develop a model selection mechanism, LEAVE-ONE-PERIOD-OUT, to select the best interference model and evaluate this mechanism over a rich set of simulated examples. We show that, given an interference model, roll-out designs allow the identification of interference parameters, which may be unidentified in the absence of any temporal variation in treatment exposure. We theoretically characterize when the causal estimand becomes identifiable with the help of roll-out designs and quantify the level of temporal variation required to identify heterogeneous patterns across units. Finally, we quantify the statistical efficiency gains resulting from a roll-out design.

In Section 2, we introduce our estimand of interest, define roll-out designs, and provide a motivating example. In Section 3, we develop our heuristic modeling framework and provide experimental evidence supporting our model selection algorithm. In Section 4, we study the question of identification, specifically how roll-outs help identify and estimate causal effects in the presence of interference. Finally, in Section 5.1, we address challenges practitioners may face when applying our framework and consider possible robustness checks that could be performed.

Related work

The literature on causal inference in the presence of interference is extensive (e.g., see the work of Rosenbaum (2007); Hudgens and Halloran (2008); Sävje et al. (2017); Leung (2019); Farias et al. (2022); Viviano (2020)). Our work is related to the subset of this literature that focuses on new experimental designs to mitigate the effects of interference (Eckles et al., 2017; Baird et al., 2018; Brennan et al., 2022); we repurpose roll-outs, a common design in practice, to model interference. While several authors have proposed new estimands and estimators to better understand the mechanism of interference and reduce its bias-inducing effects (Yuan et al., 2021; Karrer et al.,

2021; Yu et al., 2022a; Zigler and Papadogeorgou, 2021), we study common models of interference (Aronow and Samii, 2017a; Basse et al., 2016) and focus on the familiar total treatment effect estimand (Chin, 2019). Instead of suggesting new estimators and estimands, we leverage roll-out designs to formulate a new method for validating and choosing from these previously introduced models.

The operational benefits of roll-out designs have been well-documented in the context of online platforms (Kohavi et al., 2009; Xu et al., 2015; Xiong et al., 2020). However, the study of roll-out designs is sparse in the context of interference. Cortez et al. (2022b) recently studied roll-out designs to develop unbiased estimates of treatment effects. The work is similar to ours in that both frameworks do not assume the knowledge of the underlying network but make different assumptions to make inference possible. Cortez et al. (2022b) takes a design-based perspective in which given a low-degree polynomial structure for interference they design a roll-out with an unbiased estimator of the treatment effect. In contrast, our heuristic model selection framework applies to any roll-out design and interference model. When the interference model is well-specified, we provide theoretical guarantees quantifying how roll-out designs boost statistical efficiency.

Validating and choosing from potential outcome models that incorporate interference bears many similarities to the task of detecting interference, which has historically relied on one of two methods. The first set of methods is to compare two designs with different properties under SUTVA and interference, often simultaneously using a hierarchical design structure (Sinclair et al., 2012; Saveski et al., 2017; Pouget-Abadie et al., 2017). A second approach consists in running Fisher-randomized-like tests on observed data to determine the significance of well-chosen estimators that are non-zero if and only if interference is present (Aronow and Samii, 2017b; Athey et al., 2018; Basse et al., 2019b). Both approaches seek to exploit fluctuations in a specific parameter of interference to determine whether (1) interference is present and potentially (2) whether it occurs in the form or through the channel of that parameter.

Roll-outs provide a third paradigm for detecting interference, as they introduce desirable fluctuations in interference-sensitive parameters such as the global treatment fraction. In a concurrent and independent work, Han et al. (2022) study how roll-outs can be used to detect interference. They design randomization tests that can be used to detect cross-unit interference even in the presence of temporal effects. In comparison, the present paper goes beyond detection: at the cost of stronger modeling assumptions, we study the direct modeling of interference effects. Since interference may be common in online platforms, rather than establishing its existence, we propose estimation and model selection methods that allow contextualizing the operational significance of interference effects compared to the direct treatment effect. While our modeling approach can also be used as a heuristic test for interference, we do not establish its theoretical validity and instead focus on identifiability and estimation guarantees.

2 Setting

We use the potential outcomes notation for a finite population of size N . We do not make the Stable Unit Treatment Value Assumption (SUTVA) (Imbens and Rubin, 2015) such that, for any treatment vector $z \in \{0, 1\}^N$, the potential outcome $Y_i(z)$ of each unit i may depend on the treatment status of other units due to interference. For concreteness, we focus on the identification

and estimation of the total treatment effect estimand

$$\text{TTE} := \frac{1}{N} \sum_{i=1}^N (Y_i(\mathbf{1}) - Y_i(\mathbf{0})). \quad (2.1)$$

In practice, the TTE is of particular relevance to online platforms whose goal is to determine whether a product innovation is fruitful when it is completely adopted (Bond et al., 2012; Eckles et al., 2017).

Roll-outs, also known as “ramp-ups”, are a common experimental practice where instead of assigning treatments at once, it is done in incremental steps. While primarily instituted to improve engineering reliability, they induce important temporal variation in a unit’s treatment exposure that can be used to better model the effect of interference in randomized experiments. Formally, a roll-out design with T periods consists of a sequence of treatment assignments $\{Z^t\}_{t=1}^T$ and corresponding observed outcomes $\{Y_{i,t} := Y_i(Z^t)\}_{t \in [T], i \in [N]}$ such that, once a unit is treated, it remains treated for the remainder of the experiment. For typical experiments, T is a small number, often equal to 5 or less. The *completely randomized roll-out design* considers a fixed proportion of units to be newly treated at each period (Cortez et al., 2022a).

Definition 1 (Completely Randomized Roll-outs). *A T -period completely randomized roll-out is an increasing set of random treatment assignments, $\{Z^1, \dots, Z^T\}$, and treatment allocation vector $\vec{p} = \{p_1, \dots, p_T\}$ with $\sum_{t=1}^T p_t \leq 1$ such that in each period, t , $Z_i^t \in \{0, 1\}$ is randomly chosen such that $\sum_{i=1}^N Z_i^t = \lfloor N \sum_{j=1}^t p_j \rfloor$ and if $Z_i^{t-1} = 1$ then $Z_i^t = 1$.*

Another roll-out design is to specify an independent Bernoulli probability to treat each unit, known as a *Bernoulli randomized design*. This design does not meaningfully change the conclusions of our work, and we defer its definition to Section A of the appendix. Roll-out designs are characterized by the proportion of newly treated individuals in each period: “even” (resp. “uneven”) roll-outs treat the same (resp. different) incremental proportions of individuals in each period. Even among even and uneven roll-outs, there are several possible roll-out mechanisms for assigning treatments, corresponding to different joint distributions over Z^t .

Example 1 (Linear-in-Means Models with Heterogeneity): As a motivating example, we consider an advertising auction system where bidders compete for limited items. We are interested in measuring the impact of changing the reserve price—the minimum required bid to participate in the auction—on advertisers’ spend (outcomes). The example models common operational concerns on online platforms (Pouget-Abadie et al., 2018). In this two-sided marketplace with finite resources, interference occurs when changing the reserve price for some items leads bidders to change their bidding strategy, thus affecting the outcome of other auctions they participate in.

While a bipartite graph of bidders and auctions is usually used to represent the full market, in statistical inference, it is common to reduce the bipartite market structure to a single interference graph between the N items (Brennan et al., 2022). In this interference graph, edges represent a notion of competition, e.g., substitutable keywords (goods). As we substantiate further in Section 3, there are multiple ways to construct the item-to-item interference network in practice: we may consider whether differences in advertising budgets should be taken into account or whether two items are considered in competition if their co-bidders achieve a certain activity threshold.

For concreteness, consider two plausible and competing ways of defining “neighboring units” for any given unit i , $\mathcal{G}_1(i)$ and $\mathcal{G}_2(i)$, based on two different interference networks. For each notion

of “neighborhood”, we can posit a simple linear model of interference

$$Y_i^t(Z^t) = \alpha_i^* + \tau^* \cdot Z_i^t + \eta_1^* \cdot \sum_{j \in \mathcal{G}_1(i)} Z_j^t + \epsilon_{i,t} \quad (2.2a)$$

$$Y_i^t(Z^t) = \alpha_i^* + \tau^* \cdot Z_i^t + \eta_2^* \cdot \sum_{j \in \mathcal{G}_2(i)} Z_j^t + \epsilon_{i,t}. \quad (2.2b)$$

Linear models similar to the ones above have been previously studied by [Eckles et al. \(2017\)](#); [Aronow and Samii \(2017a\)](#); [Basse et al. \(2019a\)](#). Given these two competing models of interference (2.2), the better model can be selected by measuring each model’s ability to explain the variation in the outcomes as treatment exposure increases in the roll-out. \diamond

3 Model Selection

To motivate the model selection problem, we again consider the class of models defined in Example 1. This example captures how a standard linear model can be highly flexible, allowing us to incorporate a wide range of interference structures. Its richness highlights the need to distinguish between model instances that are useful in explaining interference and those that are not.

In this section, we propose a model selection mechanism inspired by leave-one-out cross-validation to choose between models of interference. A unit’s outcome depends on its treatment exposure which varies as we increase the treatment allocation throughout a roll-out. Our main observation is that we can test whether the selected interference model, trained on a subset of treatment periods, is able to extrapolate to different levels of treatment exposure by evaluating its predictive performance on observations from remaining periods. Hence, we use the mean-squared prediction error for a given period t ,

$$\text{MSPE}_t(\hat{\theta}) := \frac{1}{N} \sum_{i=1}^N (X_i^t \hat{\theta} - Y_i^t)^2 = \frac{1}{N} \sum_{i=1}^N (\hat{Y}_i^t(\hat{\theta}) - Y_i^t)^2, \quad (3.1)$$

where $Y_i^t(\hat{\theta}) = X_i^t \hat{\theta}$, $X_i^t \in \mathbb{R}^{1 \times K}$ refers to the row of features for unit i at time t , and $\hat{\theta} \in \mathbb{R}^K$ are estimated parameters. We can relate the MSPE to the estimation error of the TTE. For instance, suppose that we are given data for a new roll-out period, s , of the form (X^s, Y^s) where $X^s \in \mathbb{R}^{N \times K}$ and $Y^s \in \mathbb{R}^N$ (cf. Eq. (4.4)). Using the previous roll-out periods to estimate $\hat{\theta}$, we predict $\hat{Y}^s(\hat{\theta}) = X^s \hat{\theta}$. Define the sample covariance matrix as $\Sigma^{(s)} = \frac{1}{N} \sum_{i=1}^N X_i^s X_i^{s\top}$. Then,

$$\begin{aligned} (\hat{\text{TTE}} - \text{TTE})^2 &= [c^\top (\theta^* - \hat{\theta})]^2 \\ &\leq \|c\|_2^2 \cdot \frac{\|\hat{\theta} - \theta^*\|_{\Sigma^{(s)}}^2}{\lambda_{\min}(\Sigma^{(s)})} = \|c\|_2^2 \frac{\frac{1}{N} \sum_{i=1}^N (X_i^s \hat{\theta} - Y_i^s + \epsilon_{i,s})^2}{\lambda_{\min}(\Sigma^{(s)})} \\ &\leq \frac{2 \|c\|_2^2}{\lambda_{\min}(\Sigma^{(s)})} \cdot \frac{1}{N} \sum_{i=1}^N [(X_i^s \hat{\theta} - Y_i^s)^2 + \epsilon_{i,s}^2] = \frac{2 \|c\|_2^2}{\lambda_{\min}(\Sigma^{(s)})} \cdot \left(\text{MSPE}_s(\hat{\theta}) + \frac{1}{N} \sum_{i=1}^N \epsilon_{i,s}^2 \right) \end{aligned}$$

where $\epsilon_{i,s} = Y_i^s - X_i^s \theta^*$ and the second inequality follows from convexity. This provides a heuristic argument for using the MSPE criteria in our model selection procedure.

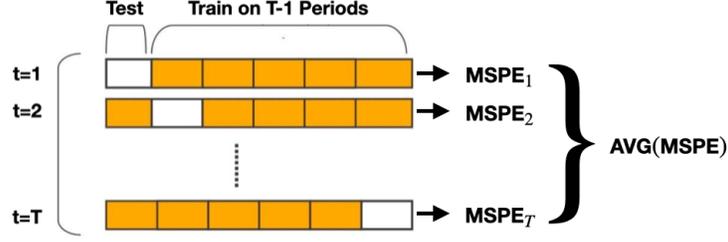


Figure 1: Graphical Representation of Leave-One-Period-Out

Algorithm 1 Leave-One-Period-Out Model Selection

- 1: INPUT: Data $D = \{\{Y_i^t, X_i^{t(m)} = [\mathcal{W}_{i,t}^{(m)}, Z_i^t, \mathbf{f}_i^{(m)}(Z^t)]\}_{i \in [N], t \in [T]}\}_{m \in [M]}$ where $\mathcal{W}_{i,t}^{(m)}$ are other model m specific features
 - 2: **for** $m \in [M]$ **do**
 - 3: **for** $t \in [T]$ **do**
 - 4: Estimate $\hat{\theta}^{(m)}$ using data $D_{-t}^{(m)}$ (excluding period t)
 - 5: Predict $\hat{Y}^t(\hat{\theta}^{(m)}) \leftarrow X^{t(m)} \hat{\theta}^{(m)}$ using data $D_t^{(m)}$
 - 6: Store mean squared prediction error: $\text{MSPE}_t(\hat{\theta}^{(m)}) \leftarrow \frac{1}{N} \sum_{i=1}^N (\hat{Y}_i^t(\hat{\theta}^{(m)}) - Y_i^t)^2$
 - 7: **end for**
 - 8: MODEL_MSPES $_m \leftarrow \text{AVERAGE}(\text{MSPE})$
 - 9: **end for**
 - 10: Return $\arg \min_{m \in [M]} \text{MODEL_MSPES}_m$
-

Proposed method: Leave-One-Period-Out When interference is present, outcomes are non-stationary due to the increasing treatment allocation of a roll-out. For example, an estimator that is fitted on the first period of a roll-out may not extrapolate to the last period. Our proposed procedure, LEAVE-ONE-PERIOD-OUT (LOPO), leverages the fact that every period offers an opportunity to test an interference model’s ability to extrapolate outcomes from different levels of treatment exposure. In parallel, we leave out each period for testing and estimate parameters $\hat{\theta}$ on all other periods. After we have predicted outcomes \hat{Y} for every period, we compute the MSPE over each prediction task and output the model that minimized the average of these MSPEs. This procedure is visualized in Figure 1 and formalized in Algorithm 1. While this is only a heuristic, our empirical results in Section 3.1 demonstrate its effectiveness.

Baseline procedures To evaluate our proposed LEAVE-ONE-PERIOD-OUT (LOPO) procedure, we compare its performance against reasonable baselines outlined in Table 1, including methods that incorporate both temporal and network structure and those that do not. The NO-ROLL-OUT procedure considers what happens when the sample size is increased to match that from a roll-out design, but no temporal variation is generated in the interference structure because treatment status remains constant. The procedure provides a fair comparison against our proposed LOPO method in case the gains our method achieves are simply due to the increased effective sample size from a roll-out. We also include a POOLED \mathcal{K} -FOLD procedure which pools the data across all periods and conducts standard \mathcal{K} -fold cross-validation. POOLED \mathcal{K} -FOLD evaluates how well our proposed methodology works relative to standard cross-validation tools that implicitly assume all

Model Selection Method	Overview	Considers Temporal Variation?	Considers Network Structure?
NO ROLL-OUT	Simulates additional periods all with 50% of units treated. Then applies Algorithm 1.	No	Yes
POOLED \mathcal{K} -FOLD	\mathcal{K} -Fold cross validation over units after pooling all periods together with $\mathcal{K} = 10$.	No	No
TRAIN FIRST	Estimates model on first $T - 1$ periods and evaluates model on period T .	Yes	Yes
TRAIN LAST	Estimates model on last $T - 1$ periods and evaluates model on first period.	Yes	Yes
LOPO (Proposed)	Applies the procedure in Algorithm 1.	Yes	Yes

Table 1: Model Selection Mechanisms

data points are exchangeable. Since the exchangeability assumption is violated due to interference, the LOPO procedure circumvents this problem by exchanging whole periods. Finally, TRAIN FIRST and TRAIN LAST are in the same spirit as LOPO method, except they only consider steps $t = T$ and $t = 0$ respectively of Algorithm 1. They preserve both network and temporal structures of the experiment and are less computationally intensive. There are other model selection methods not mentioned here that are similar to LOPO, e.g., train on the first and last periods and evaluate on all other periods. We find that in most cases LOPO achieves the best performance, and omit them from our comparisons.

3.1 Simulation Setup

Continuing from Example 1, we consider a two-sided marketplace between advertisers and auctions. We anchor our experiments in a previously motivated setting where there are two competing interference graphs that might describe the observed interactions across advertisers (Brennan et al., 2022). For example, one graph might consider an advertiser’s historical spending for a certain window of time, whereas another might consider only certain types of spend or a different window of time. Even if both graphs consider the same bipartite graph to represent interactions between advertisers and auctions, there are different ways to “fold” this graph into a one-sided interference network of advertisers with other advertisers, as suggested by Brennan et al. (2022). Each folded graph leads to different interference neighborhoods, which we can define as $\mathcal{G}_1(\cdot)$ and define $\mathcal{G}_2(\cdot)$. We now generate data according to the *true* model of interference described by $\mathcal{G}_1(i)$

$$Y_i^t(Z^t) = \alpha^* + \tau^* \cdot Z_i^t + \eta_1^* \cdot \sum_{j \in \mathcal{G}_1(i)} Z_j^t + \epsilon_{i,t}. \quad (3.2)$$

Let us define several competing models of interference to predict advertisers’ spend. We want to test how reliably each model selection method can select the true model. We consider the following competing models,

$$Y_i^t(Z^t) = \alpha^* + \tau^* \cdot Z_i^t + \epsilon_{i,t} \quad (3.3)$$

Figures	NO ROLL-OUT	POOLED \mathcal{K} -FOLD	TRAIN FIRST	TRAIN LAST	LOPO (Proposed)
Figure 5a	81.2 (80.4, 82.0)	15.2 (14.5, 15.9)	51.6 (50.6, 52.6)	47.8 (46.8, 48.8)	12.6 (12.0, 13.3)
Figure 5b	81.4 (80.6, 82.2)	19.6 (18.8, 20.4)	45.8 (44.8, 46.8)	46.2 (45.2, 47.2)	15.0 (14.3,15.7)
Figure 6a	21.8 (21.0, 22.6)	19.0 (18.2, 19.8)	39.6 (38.6, 40.6)	44.8 (43.8, 45.8)	19.0 (18.2, 19.8)
Figure 6b	45.8 (44.8, 46.8)	11.0 (10.4, 11.6)	19.8 (19.0, 20.6)	18.6 (17.8, 19.4)	19.6 (18.8, 20.4)

Table 2. Percentage of Models Incorrectly Selected: Each row displays the percentage of incorrect model selections by each procedure for the simulation in the corresponding figure. 95% bootstrapped confidence intervals are displayed in parentheses.

$$Y_i^t(Z^t) = \alpha^* + \tau^* \cdot Z_i^t + \eta_1^* \cdot \sum_{j \in \mathcal{G}_2(i)} Z_j^t + \epsilon_{i,t} \quad (3.4)$$

The model (3.3) assumes no interference and the model (3.4) considers an incorrect interference network defined by $\mathcal{G}_2(\cdot)$ as in Example 1.

In real-world applications, the effect of interference is typically thought to be smaller than the direct effect of treatment (Blake and Coey, 2014). To capture this, we set $\tau^* = 5$, $\eta_1^* = 2$, and simulate data using $\epsilon_{i,t} \stackrel{\text{iid}}{\sim} N(0, 1)$. In each experiment, we observe outcome and treatment data from a 50% completely randomized roll-out (cf. Def. 1), such that 50% of units are treated after the last period. We fit a linear regression model associated with the selected model to estimate the total treatment effect. Throughout, we assume $T = 5$ in each experiment which coincides with many practical applications with few roll-out periods.

3.2 Simulation Results

In the experiments below, we consider several variations for establishing the effectiveness of the LOPO methodology. We first consider both even and uneven 50% roll-outs. We then consider introducing additional interference terms to the true model and allowing for individual heterogeneity and time-varying effects. Lastly, we consider the effect of network sparsity by evaluating the performance of our model selection mechanisms as the underlying network density increases.

We evaluate each method on two metrics: how often it selects the correct model of interference and how well it minimizes the estimation error of the TTE regardless of whether it has chosen the correct model. Table 2 summarizes the percentage of times each model selection procedure selects the *incorrect* model. However, since an incorrectly selected model could still yield similar estimates of the TTE, Figures 2 and 3 present the distribution of the relative absolute percent estimation error of the TTE over 500 runs.

Even vs. uneven roll-outs. In this first experiment, we consider model selection in the even vs. uneven roll-out setting. In the even setting, we consider treatment increments of 10%. The uneven setting considers five treatment periods with proportions, $\vec{p} = [0.01, 0.09, 0.10, 0.15, 0.15]$,

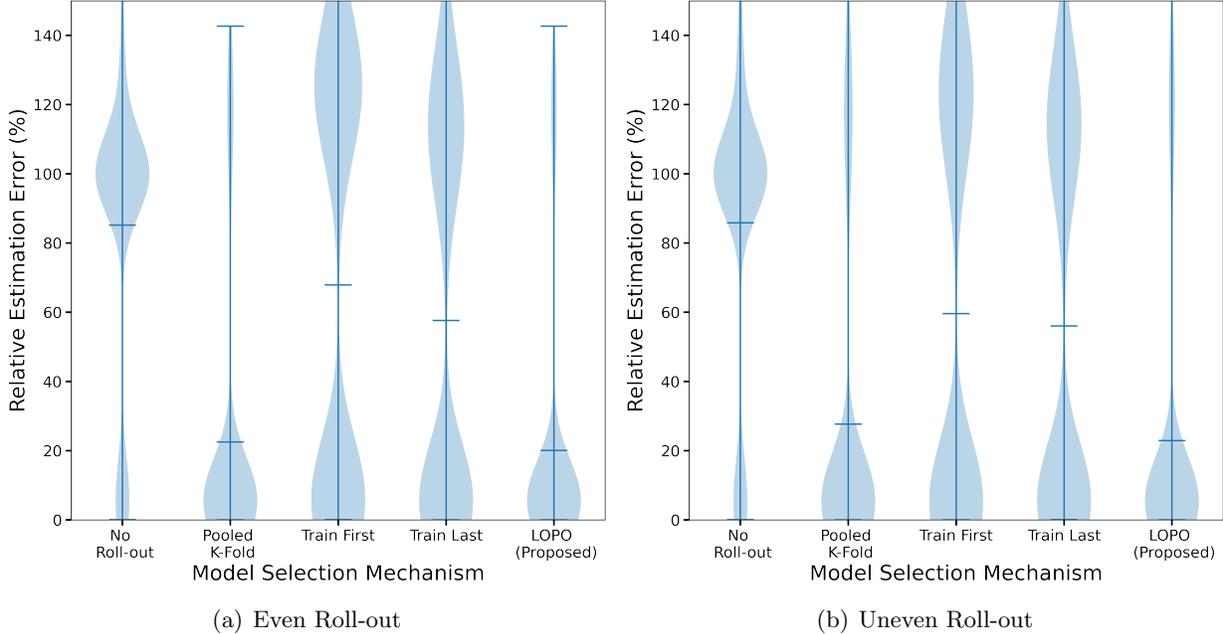


Figure 2. Even vs. Uneven Roll-outs: figures are generated by averaging the results of 500 experiments. Each model is estimated with a sample size of $N = 1000$ and $T = 5$ periods. The plots show the distribution of relative estimation error (%) for the model selection mechanism. The central tick marks represent the median.

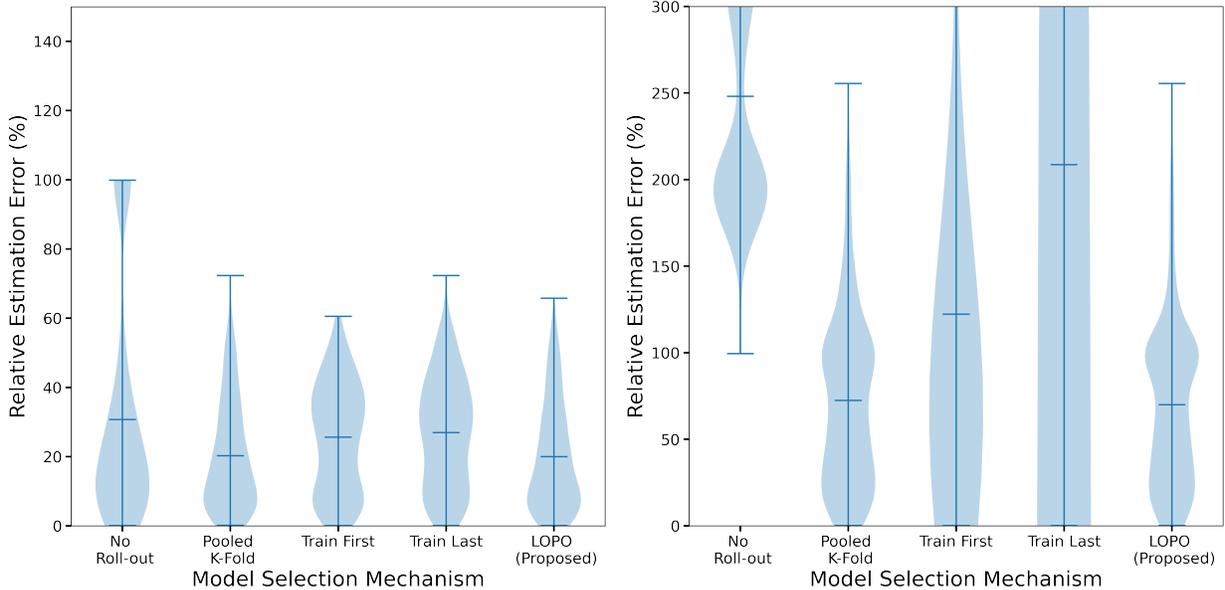
that sum to 0.50. We observe the proposed LOPO procedure performs the best in both the even and uneven settings, verifying our intuition outlined above. The POOLED \mathcal{K} -FOLD procedure also does a good job at minimizing estimation error but performs worse than LOPO on selecting the correct model, as we observe in rows one and two of Table 2. This may emphasize the importance of considering the underlying network structure in our selection procedure. In contrast, TRAIN FIRST, TRAIN LAST, and NO ROLL-OUT have very different performances across simulation settings and usually underperform relative to LOPO. LOPO tends to be more robust to changes in the roll-out schedule, which is useful when the practitioner cannot control the roll-out schedule.

Varying the true interference model We now consider variations to the data-generating model. Figure 3(a) considers a true potential outcome model where interference comes from both first-order and second-order neighbors

$$Y_i^t(Z^t) = \alpha^* + \tau^* \cdot Z_i^t + \eta_1^* \cdot \sum_{j \in \mathcal{G}_1(i)} Z_j^t + \eta_2^* \cdot \sum_{k \in \mathcal{G}_1^{(2)}(i)} Z_k^t + \epsilon_{i,t}, \quad (3.5)$$

where $\mathcal{G}_1^{(2)}(i)$ defines the set of neighbors-of-neighbors of unit i under $\mathcal{G}_1(\cdot)$. Figure 3(b) considers the performance of our model selection methods when adding individual heterogeneity and time-varying effects to the true potential outcomes model:

$$Y_i^t(Z^t) = \alpha_i^* + \gamma_t^* + \tau^* \cdot Z_i^t + \eta_1^* \cdot \sum_{j \in \mathcal{G}_1(i)} Z_j^t + \epsilon_{i,t} \quad (3.6)$$



(a) Neighbors of Neighbors Interference

(b) Individual Heterogeneity and Time Effects

Figure 3. Varying the True Interference Model: Figures are generated by averaging the results of 500 experiments. Each model is estimated with a sample size of $N = 1000$ and $T = 5$ periods. The coefficient on 2nd-order neighbor interference is $\eta_2^* = 2$. The plots show the distribution of relative estimation error (%) for the model selection mechanism. The central tick marks represent the median.

To make the comparison fair, we add these additional terms of the true models in Figure 3 to all the alternate interference models specified in (3.3)-(3.4). In each of these experiments, we use an even roll-out with a 10% per period increment.

In Figure 3(a), all procedures tend to perform better in estimation error when adding the second order neighbors interference term. This is likely because outcomes are now more correlated with the underlying network since interference now has a larger spillover effect. On the other hand, all model selection procedures tend to do worse when considering unit-fixed and time-fixed effects. This is unsurprising since the additional individual and time-varying terms make it difficult to distinguish between spillover effects and individual and temporal heterogeneity. In Figure 3(b), we observe both LOPO and POOLED \mathcal{K} -FOLD perform better in terms of estimation error relative to the other baselines.

Network sparsity Finally, we consider the performance of our procedure when we vary the sparsity of the underlying network. As we have seen in Section 4, sparsity can greatly influence the likelihood that the TTE is identified. Naturally, we expect this parameter to also influence model selection. For example, we might expect graphs that are very sparse to generate little variation preventing us from learning how to extrapolate treatment exposures to outcomes. On the other hand, graphs that are very dense tend to generate colinear data, which complicates parameter estimation.

In Figure 4, we generate a series of Erdos-Renyi graphs with an increasing probability of any two units having an edge, using each graph to define $\mathcal{G}_1(\cdot)$. Figure 4 displays how often each model

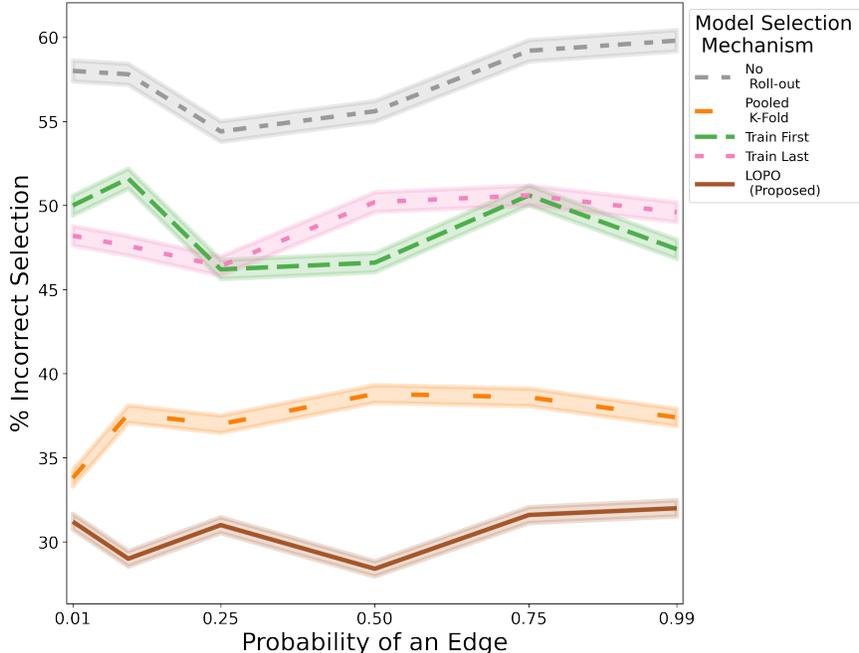


Figure 4. Varying Network Sparsity: Figures are generated by averaging the results of 500 experiments. Each model is estimated with a sample size of $N = 1000$ and $T = 5$ periods. 95% bootstrapped confidence sets are displayed by the shaded region. We exclude probabilities of 0 and 1 due to multicollinearity issues.

selection procedure fails to pick out the true model as we increase this probability. We observe that the procedures are not very sensitive to the underlying sparsity of the graph, with relatively flat selection rates across each graph. Strikingly, the LOPO procedure outperforms every other model selection procedure in selecting the true model at every point.

Polynomial Models In the previous experiments we have generated potential outcomes based on linearly additive models. However, the LOPO procedure can be extended to non-linear outcome models as well. We illustrate this using the polynomial outcome model of Cortez et al. (2022b), which assumes the data-generating process

$$Y_i(\vec{z}) = c_{i,\emptyset} + \sum_{j \in \mathcal{N}(i)} \tilde{c}_{i,j} z_j + \sum_{l=2}^{\beta} \left(\frac{\sum_{j \in \mathcal{N}(i)} \tilde{c}_{i,j} z_j}{\sum_{j \in \mathcal{N}(i)} \tilde{c}_{i,j}} \right)^l, \quad (3.7)$$

where $c_{i,\emptyset} \in U[0, 1]$, $\tilde{c}_{i,i} \sim U[0, 1]$, and for $i \neq j$ $\tilde{c}_{i,j} = v_j |\mathcal{N}(i)| / \sum_{k:(k,j) \in E} |\mathcal{N}(k)|$ and $v_j \sim U[0, r]$. Here, r is a parameter that controls the magnitude of indirect effects; in our simulations, we set $r = 2$. Like Cortez et al. (2022b) we do not include a noise term so that any error is due to the misspecification of the model. We define $\mathcal{N}(i)$ according to a sparse Erdos-Renyi graph where the probability of an edge between any two nodes is 0.1.

An important consideration in Cortez et al. (2022b) is the choice of β controlling the number of higher-order polynomial terms. Cortez et al. (2022b) take a design based perspective where they choose $T = \beta$ periods of roll-out based on a known β . Instead, we consider how a researcher might select β given a T -period roll-out. This complementary perspective is important in many

online platforms where roll-outs are frequently implemented independent of the model specification; researchers are often tasked with evaluating the effects of an intervention ex-post. Figure 5 shows the results of an experiment where we select β by applying *LOPO* to four variations of the interference model (3.2) including a model with no interference terms, a second-order term, and a third-order term. β can then be inferred by inspecting how many higher-order terms are in the selected model.

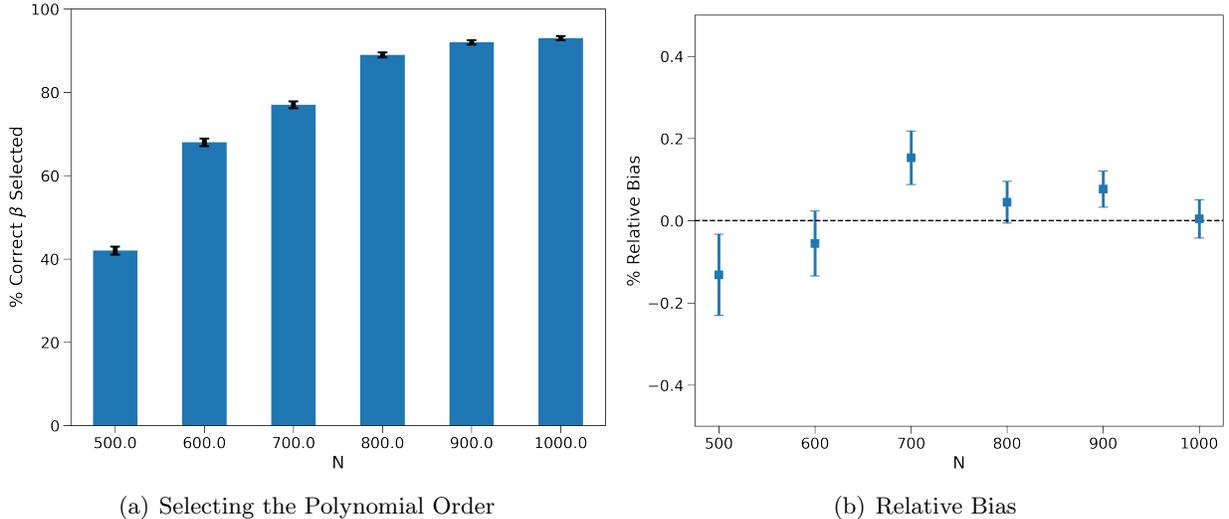


Figure 5. Selecting the Degree of Polynomial Models: figures are generated by averaging the results of 100 experiments. In each experiment, sample size is given by the x-axis, N , with $T = 4$ and an even roll-out. We display bootstrapped 95% confidence intervals of the relative bias. We use the DGP given by Cortez et al. (2022b) in Eq. (3.7) with $r = 2$ and $\beta = 2$.

Notably, LOPO predominantly selects the correct model associated with $\beta = 2$, and the rate of accurate selections rises with growing values of N . After determining the appropriate β value, researchers can employ the Lagrange interpolation estimation technique with $T = \beta$, as elaborated in Section 3 of Cortez et al. (2022b). The results of this procedure are displayed in the second panel of Figure 5 where we find that applying this procedure yields minimal relative bias which is vanishing with the sample size N . Additionally, in Section C, we delve into the impact of model misspecification on the bias of the TTE in the polynomial model.

3.3 Limitations and Extensions

LOPO tends to lead to better model selections and lower absolute error than other methods across each experimental setting we have considered, especially relative to the NO ROLL-OUT procedure. Our empirical findings align with our theoretical study to come, which quantifies the statistical efficiency gains due to roll-outs and confirms our conjecture that temporal variations can be used to better model interference effects. Interestingly, the POOLED \mathcal{K} -FOLD procedure performs very well, even though it does not consider the underlying network structure. One explanation may be that our interference networks induce neighborhoods that are relatively uniform across units. In this case, \mathcal{K} -Fold cross-validation does not risk leaving out highly central units; uniformity effectively implies the particular network we used to generate outcomes satisfies exchangeability.

While the LOPO procedure is reliable overall, there are instances in which it is outperformed by other procedures. As with any model selection procedure, we can always construct adversarial examples in which LOPO will fail. For example, if all the variation in treatment exposure occurs in the first and last periods, we would expect a procedure that only considers estimation and prediction on these periods to do very well while LOPO—which equally weights the middle period—could perform worse. A more challenging example might be the case of threshold interference. In all of our models, we have considered roll-outs only to 50%. If there is a thresholding effect whereby interference only appears at the 51% treated level, then all of our procedures would fail since we would have no variation in the interference effect to learn from.

There are possible extensions to our suggested procedure. For instance, we could consider a weighted average of the MSPEs from each fold. Weights could be proportional to the number of treated observations in the test set, for instance, so that variation in each period is considered. Another approach might be to consider testing on multiple periods at once, e.g., leaving out all combinations of two periods, training on the remaining periods, and testing on these two periods. Such a procedure would maximize the variation we exploit and is computationally very costly for potentially marginal benefit. Still, for any of these extensions, an adversarial example is possible, and we believe the LOPO procedure is a reasonable choice, being both intuitive and reliable in a wide range of environments.

4 Model Identification and Estimation

In the previous section, we discussed how a model of interference may be selected using roll-out experiments. We now take this selected model as given and characterize how roll-outs allow us to identify causal effects. We begin by providing conditions under which the total treatment effect (2.1) can be identified under the presence of interference. We demonstrate theoretically and empirically that roll-outs help satisfy these identification conditions. Furthermore, we prove that when these conditions are met, roll-outs provide gains in statistical efficiency for estimating the TTE.

4.1 Potential Outcomes Model and Estimation Framework

Causal estimation under interference requires structural assumptions. To ground our study, we consider the following model class and associated estimator.

Linear additive models We assume potential outcomes are linearly additive in z_i and a known p -dimensional feature vector $\mathbf{f}_i : \{0, 1\}^N \rightarrow \mathbb{R}^p$

$$Y_i^t(\vec{z}_t) = \alpha_i^* + \gamma_t^* + \psi_{g(i,t)}^* + \tau^* \cdot z_{i,t} + \eta^{*\top} \mathbf{f}_i(\vec{z}_t) + \epsilon_{i,t}. \quad (4.1)$$

The class (4.1) allows flexible modeling of interference effects. It subsumes commonly studied models such as exposure mapping and two-way fixed effect models (Harshaw et al., 2023), as well as models from Example 1.

- $\alpha^* \in \mathbb{R}^N$ is a unit-fixed effects allowing for individual heterogeneity.
- $\gamma^* \in \mathbb{R}^T$ models time-varying trends through period-fixed effects.

- $\psi^* \in \mathbb{R}^G$ models further unit-period heterogeneity where $g : [N] \times [T] \rightarrow [G]$. To make estimation tractable, we assume $G < NT$ is small enough such that we never have more parameters than observations.
- $\tau^* \in \mathbb{R}$ is the direct treatment effect for unit i .
- $\eta^* \in \mathbb{R}^p$ models indirect treatment effects due to interference.
- $\{\epsilon_{i,t}\}_{i \in [N], t \in [T]}$ are idiosyncratic noise, for which we will consider different regimes.

Letting $K = N + T + G + p + 1$, define the vector of model parameters

$$\theta^* := [\alpha^*, \gamma^*, \psi^*, \tau^*, \eta^*] \in \mathbb{R}^K, \quad (4.2)$$

assuming a normalization where $f_i(0) = 0_p$ for all i . Under the data generating process (4.1), the total treatment effect (2.1) can be rewritten as the linear combination

$$\text{TTE} = c^\top \theta^* \quad \text{where} \quad c = \left[0_N, 0_T, 0_G, 1, \overline{f(\mathbb{1})} := \frac{1}{N} \sum_{i=1}^N f_i(\mathbb{1}) \right] \in \mathbb{R}^K. \quad (4.3)$$

Estimation approach To estimate the TTE under the above linearly additive model (4.1), we turn to simple linear regression estimators. We define our matrix of covariates for a single roll-out period as

$$X^t = [I_N, \mathbb{1}_t^\top, \mathbb{1}_{g(i,t)}^\top, Z^t, f(Z^t)] \quad (4.4)$$

where I_N is the $N \times N$ identity matrix representing the individual effects, $\mathbb{1}_t^\top \in \mathbb{R}^{N \times T}$ is a matrix indicating if an observation belongs to period t , $\mathbb{1}_{g(i,t)}^\top \in \mathbb{R}^{N \times G}$ indicates if observation i is in cluster $g = 1, \dots, G$ at period t . Letting $X = [X^1, \dots, X^T]^\top$ and recalling the definition of coefficients c and parameters θ defined in Eq. (4.2), we study the linear regression estimator

$$\hat{\text{TTE}} = c^\top \hat{\theta} \quad \text{where} \quad \hat{\theta} = (X^\top X)^{-1} X^\top Y \quad (4.5)$$

whenever $X^\top X$ is non-singular, in which case $\hat{\text{TTE}}$ is a consistent and unbiased estimator. A major benefit of roll-outs is that they increase the likelihood that $X^\top X$ will be non-singular, which is necessary for the total treatment effect to be identifiable and for $\hat{\text{TTE}}$ to have the aforementioned statistical guarantees.

4.2 Model Based Identification

We now consider conditions that allow us to identify the total treatment effect in the finite population setting and show how they are tied to $X^\top X$ being invertible. We first consider the case of a single interference term to build intuition, and then provide a more general condition for the identification of the TTE. We conclude by showing empirically how roll-outs increase the likelihood that these identification conditions hold in-sample.

In what follows we consider parametric identification of the TTE in the setting of finite populations. We say that a parameter is identified in this sense if there exists a consistent estimator for that parameter where consistency is considered with respect to increasing population sizes. Results of this nature are derived in Section 4.3. In the case of linear models with exogenous covariates, as in our setting, a sufficient condition for identification of the parameter vector is non-singularity of the design matrix, $X^\top X$ (see Section 4.2.1 of Wooldridge (2010)). Section 4.2.1 provides sufficient conditions for the non-singularity condition to hold.

4.2.1 Identification with a Single Interference Term

Proposition 1 below introduces a sufficient condition that ensure $X^\top X$ is invertible when $f_i(\vec{z}) \in \mathbb{R}$. We state the condition in the language of interference networks to illustrate how they relate to roll-outs and interference. The key idea here is that if we can find some unit in the control group connected to a treated unit and observe the spillover effect on this individual, then we have enough information about the interference mechanism to extrapolate to the case of total treatment. Roll-outs increase the probability that this sufficient condition holds by increasing the proportion of treated individuals in each period. The proof of this result is given in Section B.1.

Proposition 1. *Consider a $T > 1$ period roll-out under the following linearly additive model*

$$Y_i^t(\vec{z}) = \alpha^* + \tau^* \cdot z_i + \eta^* \cdot f_i(\vec{z}) + \epsilon_{i,t}.$$

Assume there are $t, t' \in [T]$, $i, j \in [N]$, $(i, t) \neq (j, t')$, such that $Z_i^t = Z_j^{t'} = 0$, $f_i(Z^t) \neq 0$, and $f_i(Z^t) \neq f_j(Z^{t'})$. Then $X^\top X$, as defined in Eq. (4.4), is non-singular.

The condition $f_i(Z^t) \neq f_j(Z^{t'})$ ensures we observe variation in the interference term so that, as the roll-out progresses, the interference effects vary. The condition $f_i(Z^t) \neq 0$ further ensures that we observe a spillover effect on an untreated unit. Together, these conditions are sufficient (but not necessary) for identifying the TTE by ensuring the invertibility of the Gram matrix, $X^\top X$.

Next, we apply Proposition 1 in the context of the interference graph from Example 1.

Corollary 1. *Consider the model from Proposition 1 where the interference term is given by the model (2.2a) in Example 1, i.e., $f(\vec{z}) = \sum_{j \in \mathcal{G}_1(i)} \vec{z}_j$. Assume that neighbors are commutative so that $i \in \mathcal{G}_1(j)$ implies $j \in \mathcal{G}_1(i)$. If at time $t > 1$, there is a treated unit j with an untreated neighbor ($i \in \mathcal{G}_1(j)$ with $Z_i^t = 0$), $X^\top X$ is non-singular.*

4.2.2 Identification with General Interference

In practice, we are only interested in estimating the TTE, not the entire parameter vector θ in the model (4.1). Recalling the linear representation (4.3) for the TTE, we now show we can identify the TTE even when the individual components of θ are not identifiable. Our result shows that the TTE is identified so long as the linear transformation c that maps θ to the TTE (4.3) lies in the space spanned by the covariates (4.4). Intuitively, this shows that we can identify the TTE under general interference patterns whenever the linear transformation (4.3) can be represented by the observed data. This is particularly useful in the small N and T regime where there may not be enough variation to compute a typical least squares estimate.

Theorem 2. *Under the data-generating model (4.1), recall the linear transformation c that maps the vector of parameters to the TTE (4.2). If $c \in \text{span}(X^\top)$, then $\{c^\top \theta : X^\top X \theta = X^\top Y\}$ is a singleton.*

To clarify the importance of Theorem 2, which we prove in Section B.3, we consider the following example where $X^\top X$ is singular but the TTE is still well-defined and identifiable via Theorem 2.

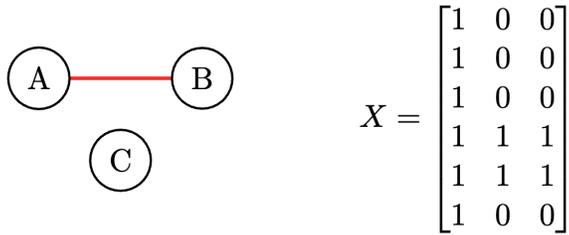


Figure 6. A disconnected network defined by $\mathcal{G}(\cdot)$ and corresponding feature matrix X defined by (4.6).

Example 2 (Identifying the TTE when θ is *not* identified): Consider the linear interference model of Example 1 with no individual heterogeneity, so that $\forall i \in [N]$, $\alpha_i = \alpha$,

$$Y_i^t(Z^t) = \alpha^* + \tau^* \cdot Z_i^t + \eta_1^* \cdot \sum_{j \in \mathcal{G}(i)} Z_j^t + \epsilon_{i,t} \quad (4.6)$$

Let $N = 3$ and define the interference network $\mathcal{G}(\cdot)$ by the graph in Figure 6. Suppose that $T = 2$, such that in the first period $t = 1$ no individuals are treated, and that in period $t = 2$ we treat observations A and B , generating the feature matrix X in Figure 6. We wish to estimate the TTE using the correctly specified model (4.6). Notice that $X^\top X$ is singular since because columns 2 and 3 of X are linearly dependent. The TTE in this model is given by $\tau + \eta$ implying $c = [0, 1, 1]$, and $c \in \text{span}(X^\top)$ because $X^\top v = c$ for $v = [0, 0, 0, 0, 1, -1]^\top \in \mathbb{R}^6$. Theorem 2 shows that the TTE is identifiable; in particular, we can estimate TTE by looking at the difference in outcomes for unit A or B at periods $t = 1$ and $t = 2$. \diamond

Proposition 1 and Theorem 2 consider when it is possible to identify a linear combination of parameters from a linear regression. While it is possible to satisfy the conditions of Proposition 1 and Theorem 2 with variation in interference effects over a single period—sometimes called spatial variation (Aronow and Samii, 2017a)—in many cases, we also need temporal variation to achieve identification, which roll-outs provide. For example, when we have individual heterogeneity parameters, $\{\alpha_i\}_{i \in [N]}$, temporal variation in individual responses is required for identification. Roll-outs should not only help us in identifying individual effects but also enable us to identify interference effects as well as the total treatment effect. Specifically, roll-outs provide added variation that increases the probability that the conditions such as the ones outlined in Proposition 1 and Theorem 2 hold. Figure 7 provides evidence for this idea in a simulated setting where outcomes are sampled according to (2.2a): as the number of roll-out periods increases, the probability of uniquely identifying the total treatment effect increases very quickly, even in extremely sparse models with an Erdos-Renyi parameter of 0.001. As expected, the higher the graph density, the likelier we are to satisfy the conditions in Proposition 1 and Theorem 2. This is because increasing network density also increases the probability that any two units are connected and so generally increases the likelihood that an untreated unit will be connected to a treated unit.

4.3 Estimation of the Total Treatment Effect

In addition to guaranteeing identifiability, temporal variation in the covariates X as measured by the spectrum of $X^\top X$ can also reduce statistical error due to measurement noise. In this section, we

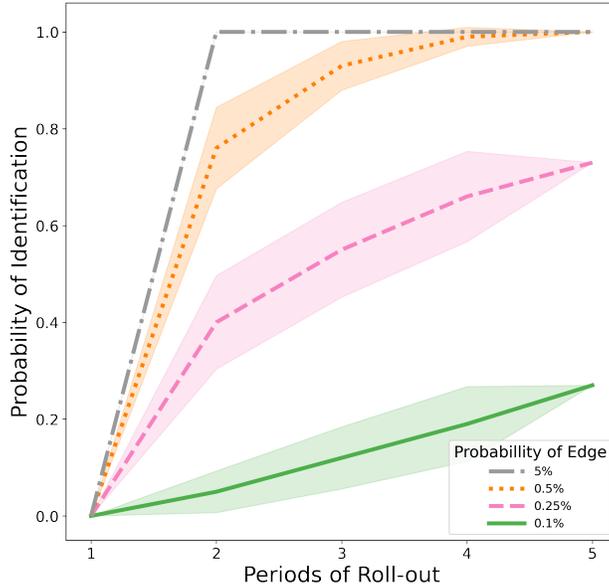


Figure 7. Probability that $c \in \text{span}(X^\top)$ for an underlying Erdos-Renyi graph with varying edge-formation probabilities of an edge. Probabilities are computed over 100 experiments. We display 95% CIs. Over all realizations with a 5% probability of an edge (shown in grey) results are constant, hence there is no visible CI.

quantify how roll-outs provide gains in statistical efficiency under two settings. First, we consider the case of completely correlated noise so that, in each period, outcomes only change as a function of the treatment. This is similar to modeling unobserved individual heterogeneity, which can be fully captured using unit-level fixed effects. Second, and at the other extreme, we consider fully independent noise across individuals and roll-out periods. This setting arises when there are no unobserved temporally persistent events. In both settings we expect roll-outs to improve the final variance bound.

We begin by studying the bias and variance of our estimator under the assumption that our model selection procedure (cf. Section 3) has correctly chosen an interference model. The unbiasedness of the estimator is clear from the randomized design because Z is drawn randomly in each period independent of ϵ . To study the variance of our estimator, we make the following assumption on how unobserved noise enters the model.

Assumption A (Time-invariant Individual Idiosyncrasies). *Suppose $\epsilon_{i,t} = \epsilon_i$ where $\{\epsilon_i\}_{i=1,\dots,N}$, are independent, mean-zero, and satisfies $\mathbb{E}[\epsilon_i^2] \leq \sigma_{\max}^2 < \infty$.*

Applying this assumption, we can control the mean squared error (MSE) of our estimator $T\hat{T}E$. Since our estimator $T\hat{T}E$ is unbiased, we have $\text{MSE}(T\hat{T}E) = \mathbb{E}[(T\hat{T}E - TTE)^2] = \text{Var}(T\hat{T}E)$ and we only have to control the variance term. As noise terms across periods are completely correlated under Assumption A, idiosyncrasies may persist indefinitely across periods and we expect our variance will increase as a function of T . The variance reduction occurs as the population size N grows large, as in the classical linear regression setting. In the below result, this is captured by the geometry of X using $\lambda_{\min}(X^\top X)$, the minimum eigenvalue of $X^\top X$.

Theorem 3. Under the data-generating model (4.1) and Assumption A, we have $\frac{1}{NT} \mathbb{E} \left\| X\hat{\theta} - X\theta^* \right\|_2^2 \leq 4\sigma_{\max}^2 \frac{K}{N}$, and

$$\mathbb{E}[(T\hat{\text{TTE}} - \text{TTE})^2] = \mathbb{E}(c^\top (\hat{\theta} - \theta^*))^2 \leq 4KT \cdot \|c\|_2^2 \cdot \sigma_{\max}^2 \cdot \mathbb{E} \left[\lambda_{\min}(X^\top X)^{-1} \right].$$

It is useful to compare Theorem 3 to the standard linear regression setting with a single period $T = 1$, where a similar analysis yields the same bound but without any dependence on T . Comparing these two results together, it may seem as though roll-outs increase the variance in the case of time-invariant noise. However, we also need to consider how $\lambda_{\min}(X^\top X)$ scales with N and T . By way of illustration, consider the case of a complete graph so that the interference term can be deterministically quantified in relation to Z^t . In that setting, we find that the minimum eigenvalue grows linearly in NT , implying that we recover the classical $\frac{1}{N}$ rate. The following lemma captures this idea that a roll-out allows us to increase our effective data size even with fixed population size and time-invariant errors. In particular, as we have seen in the previous subsection, roll-outs also tend to increase the likelihood that our $X^\top X$ matrix will be full-rank.

Lemma 1. Consider the same model as in Proposition 1 and a completely randomized roll-out (cf. Definition 1) with allocation vector \vec{p} . Let Assumption A hold and let f_i be linear-in-means $f_i(\vec{z}) = \frac{1}{|\mathcal{G}(i)|} \sum_{j \in \mathcal{G}(i)} z_j$ where $\mathcal{G}(i)$ is defined by a complete graph. Then, fixing the sample size at N there exists $M \in \mathbb{R}$ large enough that for $NT > M$

$$\mathbb{E}[(T\hat{\text{TTE}} - \text{TTE})^2] \leq \frac{8\bar{C}_1}{N},$$

where \bar{C}_1 is a constant that depends on K , σ_{\max}^2 from Assumption A, and the allocation vector \vec{p} .

Lemma 1 illustrates how roll-outs even when observations are fully correlated across periods still enable us to obtain the usual $\frac{1}{N}$ decay for the variance through $\lambda_{\min}(X^\top X)$, the minimum eigenvalue of our Gram matrix. The proof can be found in Section B.5.

We now turn to the case where individual noise is independent across periods. As we have noted earlier, this setting arises when unobserved idiosyncrasies do not persist across several periods. Here we make the following analogue to Assumption A.

Assumption B (Time-varying Individual Idiosyncrasies). $\{\epsilon_{i,t}\}_{i \in [N], t \in [T]}$ are independent, mean-zero, and satisfies $\mathbb{E} \left[\epsilon_{i,t}^2 \right] \leq \sigma_{\max}^2 < \infty$.

Because Assumption B requires noise to be independent across time periods, we can achieve tighter control of the variance of our estimator. In particular, roll-outs decrease the MSE at a $\frac{1}{T}$ rate since idiosyncrasies are fully independent across time. Applying the same analysis as in our derivation of Theorem 3, we have the following result which we prove in Section B.4.

Theorem 4. Under the data-generating model (4.1) and Assumption B, we have $\frac{1}{NT} \mathbb{E} \left\| X\hat{\theta} - X\theta^* \right\|_2^2 \leq 4\sigma_{\max}^2 \frac{K}{NT}$, and

$$\mathbb{E}[(T\hat{\text{TTE}} - \text{TTE})^2] = \mathbb{E}(c^\top (\hat{\theta} - \theta^*))^2 \leq 4K \cdot \|c\|_2^2 \cdot \sigma_{\max}^2 \cdot \mathbb{E} \left[\lambda_{\min}(X^\top X)^{-1} \right]$$

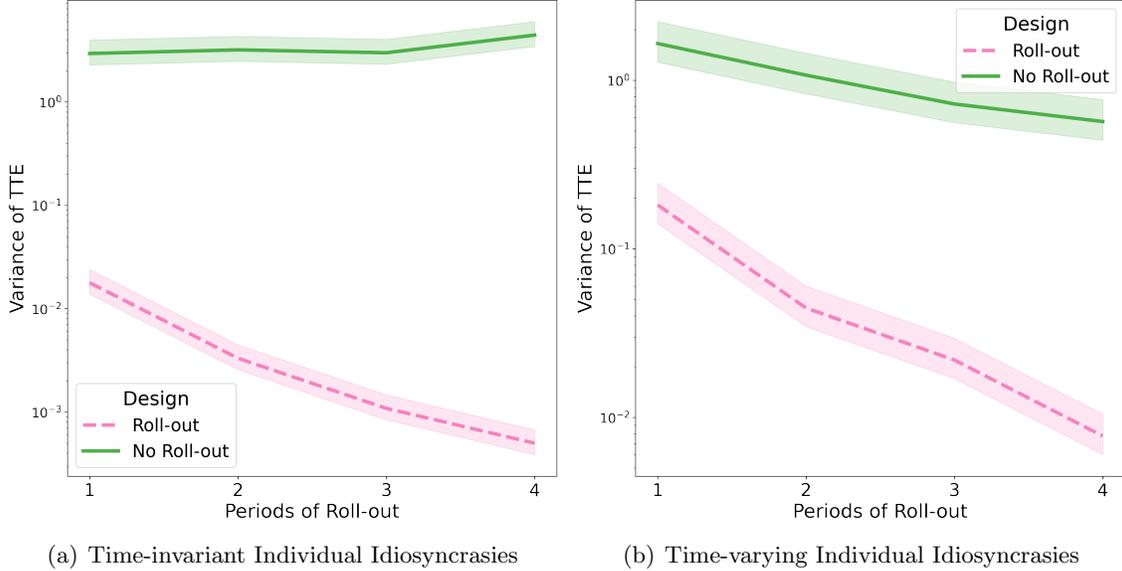


Figure 8. Variance reduction due to roll-out under fixed and i.i.d. errors. The no roll-out baseline is computed by assuming that we observe a 50% treated population at multiple periods in time with both fixed and random potential outcomes. We display 95% χ^2 -confidence intervals.

Similar to Lemma 1 in the case of Assumption B, the next lemma provides a concrete bound for a complete interference graph. When the noise is fully independent across periods, we gain a reduction in variance as T grows. Hence, in this extreme, roll-outs produce an even faster vaster variance reduction through the geometry of $\lambda_{\min}(X^\top X)$.

Lemma 2. *Consider the same model as in Proposition 1 and a completely randomized roll-out (cf. Definition 1) with allocation vector \vec{p} . Let Assumption B hold and let f_i be linear-in-means $f_i(\vec{z}) = \frac{1}{|\mathcal{G}(i)|} \sum_{j \in \mathcal{G}(i)} z_j$ where $\mathcal{G}(i)$ is defined by a complete graph. Then, fixing the sample size at N there exists $M \in \mathbb{R}$ large enough that for $NT > M$*

$$\mathbb{E}[(\text{T}\hat{\text{T}}\text{E} - \text{TTE})^2] \leq \frac{8\bar{C}_2}{NT},$$

where \bar{C}_2 is a constant that depends on K , σ_{\max}^2 from Assumption B, and the allocation vector \vec{p} .

See Section B.5 for the proof.

Figure 8 shows the implications of Theorems 3 and 4 in both the time-varying and time-invariant noise settings. In both cases, we see a non-trivial variance reduction relative to the no roll-out case, which is emblematic of the variance gains from roll-outs quantified in this section.

5 Discussion

In this work, we leverage a universal experimentation design used throughout online platforms, roll-outs, to model interference effects. We quantify how roll-outs induce temporal variation in treatment exposure that facilitates the identification and estimation of the total treatment effect. We propose a model selection procedure to help practitioners model interference and identify the

total treatment effect. We conclude the paper by discussing robustness checks that can augment our model selection framework, and discuss possible pitfalls practitioners may face when applying our methodology. The heuristics we propose below help practitioners implement our methodology.

5.1 Practical Considerations

Robustness checks A natural question that arises from this analysis is whether there are any robustness checks that can provide evidence that our model selection procedure has chosen the correct model of interference. Fundamentally, testing if a model is correct is not possible. However, there are tests we can perform to build evidence that we are fully accounting for the variation in outcomes caused by interference in our data sample.

The first recommendation we make to practitioners is to include a model without interference terms in the model selection step. Including such a model in our procedure is equivalent to testing for interference. If the model selection procedure chooses the no interference model, when there is a strong prior for interference in the experiment, then this is good evidence that the models that are being tested are inadequately capturing the effects of treatment exposure.

Another possible test uses the interference testing framework of [Han et al. \(2022\)](#). Their work considers what gains roll-outs provide when attempting to detect interference. They provide several permutation tests under a Bernoulli roll-out design that are able to effectively test for the presence of interference. A key component of these tests is the candidate exposure of each unit which is defined in their notation as $h_i(W_{-i,k})$ where, using our notation $W_{i,k} = Z_i^t$. A simple test to consider is to define h_i to be the interference terms in our setting, that is to say, set $h_i(W_{-i,k}) = \mathbf{f}_i(Z^t)$ where $\mathbf{f}(\cdot)$ is given by the selected model, and then conduct the proposed multiple experiment test of [Han et al. \(2022\)](#). If the test finds interference to be statistically significant, then this is good evidence that the selected model of interference is capturing the effect of treatment exposure on outcomes. While this test may still suffer from misspecification issues, comparing its results to the permutation test proposed by [Han et al. \(2022\)](#) again can provide strong evidence in favor of the selected model.

A final approach applies the test proposed by [Pouget-Abadie et al. \(2017\)](#). In this case, after model selection is completed, we compute the new outcome, which subtracts the effects of interference from each outcome at every period. Next, we pool our data across all periods and use the underlying interference network to create clusters of units, allowing us to compute a difference-in-means estimate of the total treatment effect and a Horvitz-Thompson estimate under a cluster-based design. We can now compute Δ as the difference of these estimates and conduct the test proposed in [Pouget-Abadie et al. \(2017\)](#). If we find that the estimates are similar, then this is again evidence that our selected model is accurately capturing interference effects.

Other Considerations Since effect sizes are typically small in online platforms, the lack of statistical power may result in the inability to distinguish between similar potential outcome models. In many of the simulations we considered, we observed different models yield similar estimates of the TTE, somewhat alleviating such concerns. When considering rich interference models, we recommend a LASSO penalty when conducting estimation.

There are often non-stationarities in interference effects that require time to equilibrate, and the length of each period in a roll-out is an important design choice. Our procedure relies on the fact that outcomes are observed after the full interference effects have been experienced. Different experiments and settings will naturally require different time windows, and previous experiments

and domain knowledge should guide these choices. Finally, some practitioners may want to pose auto-regressive models in their experiments. Unfortunately, auto-regressive models pose challenges as they introduce complicated interactions with interference terms, time effects, and individual heterogeneity. While a well-known practical issue in the context of two-way fixed effects (Arellano and Bond, 1991), the consequences of auto-regressive terms are unclear in terms of interference, which we leave as a topic of future work.

5.2 Future Directions

We summarize several future directions of research. First, we have seen how roll-out schedules can influence our ability to conduct model selection effectively. A theoretical study of model selection requires a formal language for model misspecification in the presence of interference, which we leave for future work. A close study of the design-based perspective posed in our work may yield fruit. While requiring more engineering resources, the experimenter may sometimes be able to adaptively choose a roll-out schedule that maximizes the information that can be learned from the experiment before fully launching an intervention. Similarly, she may mitigate the non-stationarity of interference effects by appropriately choosing periods in a roll-out. A third direction may consider how to carefully incorporate auto-regressive terms under the presence of interference, which may prove useful from a modeling perspective. Finally, while we have empirically shown that the LOPO procedure tends to perform reliably, we mention some possible extensions in Section 3.3.

Acknowledgement We thank Kevin Han, Shuangning Li, Jialiang Mao, and Han Wu for their thoughtful feedback.

References

- Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies*, 58(2):277–297.
- Aronow, P. M. and Samii, C. (2017a). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947.
- Aronow, P. M. and Samii, C. (2017b). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912 – 1947.
- Athey, S., Eckles, D., and Imbens, G. W. (2018). Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240.
- Baird, S., Bohren, J. A., McIntosh, C., and Özler, B. (2018). Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*, 100(5):844–860.
- Basse, G., Ding, P., Feller, A., and Toulis, P. (2019a). Randomization tests for peer effects in group formation experiments. arXiv.
- Basse, G. W. and Airoidi, E. M. (2018). Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika*, 105(4):849–858.

- Basse, G. W., Feller, A., and Toulis, P. (2019b). Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494.
- Basse, G. W., Soufiani, H. A., and Lambert, D. (2016). Randomization and the pernicious effects of limited budgets on auction experiments. In *Artificial Intelligence and Statistics*, pages 1412–1420. PMLR.
- Biswas, N. and Airoidi, E. M. (2018). Estimating peer-influence effects under homophily: Randomized treatments and insights. In *Complex Networks IX: Proceedings of the 9th Conference on Complex Networks CompleNet 2018 9*, pages 323–347. Springer.
- Blake, T. and Coey, D. (2014). Why marketplace experimentation is harder than it seems: The role of test-control interference. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, page 567–582, New York, NY, USA. Association for Computing Machinery.
- Bojinov, I. and Gupta, S. (2022). Online experimentation: Benefits, operational and methodological challenges, and scaling guide. *Harvard Data Science Review*, 4(3).
- Bojinov, I., Simchi-Levi, D., and Zhao, J. (2022a). Design and analysis of switchback experiments. *Management Science*.
- Bojinov, I., Simchi-Levi, D., and Zhao, J. (2022b). Design and analysis of switchback experiments. *Management Science*.
- Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., and Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nat.*, 489(7415):295–298.
- Brennan, J. R., Mirrokni, V., and Pouget-Abadie, J. (2022). Cluster randomized designs for one-sided bipartite experiments. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.
- Bright, I., Delarue, A., and Lobel, I. (2022). Reducing marketplace interference bias via shadow prices.
- Candogan, O., Chen, C., and Niazadeh, R. (2021). Correlated cluster-based randomized experiments: Robust variance minimization.
- Chang, S., Vrabac, D., Leskovec, J., and Ugander, J. (2022). Estimating geographic spillover effects of covid-19 policies from large-scale mobility networks.
- Chin, A. (2019). Regression adjustments for estimating the global treatment effect in experiments with interference. *Journal of Causal Inference*, 7(2).
- Cortez, M., Eichhorn, M., and Yu, C. L. (2022a). Exploiting neighborhood interference with low order interactions under unit randomized design.
- Cortez, M., Eichhorn, M., and Yu, C. L. (2022b). Graph agnostic estimators with staggered rollout designs under network interference.

- Eckles, D., Karrer, B., and Ugander, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1).
- Eckles, D., Kizilcec, R. F., and Bakshy, E. (2016). Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, 113(27):7316–7322.
- Farias, V. F., Li, A. A., Peng, T., and Zheng, A. (2022). Markovian interference in experiments.
- Han, K., Li, S., Mao, J., and Wu, H. (2022). Detecting interference in a/b testing with increasing allocation.
- Harshaw, C., Sävje, F., Eisenstat, D., Mirrokni, V., and Pouget-Abadie, J. (2023). Design and analysis of bipartite experiments under a linear exposure-response model. *Electronic Journal of Statistics*, 17(1):464 – 518.
- Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842. PMID: 19081744.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- Johari, R., Li, H., Liskovich, I., and Weintraub, G. (2020). Experimental design in two-sided platforms: An analysis of bias. arXiv.
- Karrer, B., Shi, L., Bhole, M., Goldman, M., Palmer, T., Gelman, C., Konutgan, M., and Sun, F. (2021). Network experimentation at scale. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3106–3116.
- Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181.
- Leung, M. P. (2019). Causal inference under approximate neighborhood interference.
- Pouget-Abadie, J., Mirrokni, V., Parkes, D. C., and Airoidi, E. M. (2018). Optimizing cluster-based randomized experiments under monotonicity. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2090–2099.
- Pouget-Abadie, J., Saveski, M., Saint-Jacques, G., Duan, W., Xu, Y., Ghosh, S., and Airoidi, E. M. (2017). Testing for arbitrary interference on experimentation platforms.
- Rigollet, P. and Hütter, J.-C. (2015). High dimensional statistics.
- Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.
- Saveski, M., Pouget-Abadie, J., Saint-Jacques, G., Duan, W., Ghosh, S., Xu, Y., and Airoidi, E. M. (2017). 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.

- Sinclair, B., McConnell, M., and Green, D. P. (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science*, 56(4):1055–1069.
- Sävje, F., Aronow, P. M., and Hudgens, M. G. (2017). Average treatment effects in the presence of unknown interference.
- Ugander, J. and Yin, H. (2020). Randomized graph cluster randomization.
- Viviano, D. (2020). Policy design in experiments with unknown interference.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.
- Xiong, R., Athey, S., Bayati, M., and Imbens, G. (2020). Optimal experimental design for staggered rollouts. *Management Science*.
- Xu, Y., Chen, N., Fernandez, A., Sinno, O., and Bhasin, A. (2015). 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.
- Xu, Y., Duan, W., and Huang, S. (2018). 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.
- Yu, C. L., Airoidi, E. M., Borgs, C., and Chayes, J. T. (2022a). Estimating the total treatment effect in randomized experiments with unknown network structure. *Proceedings of the National Academy of Sciences*, 119(44):e2208975119.
- Yu, C. L., Airoidi, E. M., Borgs, C., and Chayes, J. T. (2022b). Graph agnostic randomized experimental design. arXiv.
- Yuan, Y., Altenburger, K., and Kooti, F. (2021). Causal network motifs: Identifying heterogeneous spillover effects in a/b tests. In *Proceedings of the Web Conference 2021*, pages 3359–3370.
- Zigler, C. M. and Papadogeorgou, G. (2021). Bipartite causal inference with interference. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 36(1):109.

A Characterizing the Distribution of Roll-out Designs

In this section of the appendix, we analyze two different procedures to implement roll-out designs. The first design, known as the completely randomized roll-out, is defined in Definition 1 and is used throughout our paper and by contemporaneous work, e.g., (Cortez et al., 2022b). In each period t , let $\mathcal{S}_t \subseteq [N]$ be the set of newly treated units. The probability that a unit i is selected for treatment is

$$\mathbb{P}[i \in \mathcal{S}_t] = \frac{Np_t}{N - \lceil N \sum_{j=1}^{t-1} p_j \rceil}. \quad (\text{A.1})$$

Evidently, a completely randomized roll-out is a Markov chain whose state transitions occur according to,

$$Z_i^t = \begin{cases} 1 & Z_i^{t-1} = 1 \text{ or } i \in \mathcal{S}_t \\ 0 & \text{otherwise} \end{cases}.$$

With this in hand, we can compute the marginal distribution of Z_i^t .

Lemma 3. *The distribution of Z^t under Definition 1 is given by*

$$yuetal2022rollout \mathbb{P}[Z_i^t = 1] = \sum_{k=1}^t p_k \quad (\text{A.2})$$

Proof Let $\mathcal{S}_t \subseteq [N]$ be the set of newly treated units in period t .

$$\begin{aligned} 1 - \mathbb{P}[Z_i^t = 0] &= \mathbb{P}[Z_i^t = 1] \\ &= \mathbb{P}[Z_i^{t-1} = 1] + \mathbb{P}[i \in \mathcal{S}_t] \mathbb{P}[Z_i^{t-1} = 0] \\ &= \mathbb{P}[i \in \mathcal{S}_1] + \mathbb{P}[i \in \mathcal{S}_2] \mathbb{P}[Z_i^1 = 0] + \dots + \mathbb{P}[i \in \mathcal{S}_t] \mathbb{P}[Z_i^{t-1} = 0] \\ &= p_1 + \sum_{k=2}^t \frac{Np_k}{N - \lceil N \sum_{j=1}^{k-1} p_j \rceil} \left(1 - \mathbb{P}[Z_i^{k-1} = 1]\right). \end{aligned}$$

Conclude

$$\begin{aligned} \mathbb{P}[Z_i^1 = 1] &= p_1 \\ \mathbb{P}[Z_i^2 = 1] &= p_1 + \frac{Np_2}{N(1-p_1)}(1-p_1) = p_1 + p_2 \\ &\vdots \\ \mathbb{P}[Z_i^t = 1] &= \sum_{k=1}^t p_k. \end{aligned}$$

□

We now turn to the second roll-out implementation that is based on Bernoulli trials. This design appears in Han et al. (2022) and much of the analysis in this paper can also be carried through in this implementation.

Definition 2 (Bernoulli Roll-outs). A *T-period Bernoulli roll-out* is an increasing set of treatment assignments, $\{Z^1, \dots, Z^T\}$ and treatment proportions $p = \{p_1, \dots, p_T\}$ where $\sum_{t=1}^T p_t = \bar{P} \leq 1$ such that the distribution of Z follows

$$Z_i^1 \sim \text{Bernoulli}(p_1) \quad \text{and} \quad Z_i^t \sim \begin{cases} 1 & \text{if } Z_i^{t-1} = 1 \\ \text{Bernoulli}(p_t) & \text{otherwise} \end{cases}$$

Lemma 4. Suppose that Z is a roll-out given by Definition 2. Then, Z^t is a Markov chain with

1. $\mathbb{P}[Z_i^t = 0] = \prod_{j=1}^t (1 - p_j)$
2. $\mathbb{P}[Z_i^t = 1] = p_1 + p_2(1 - p_1) + \dots + p_t \prod_{j=1}^{t-1} (1 - p_j)$.

Proof To see Property 1, apply the law of total probability and $\mathbb{P}[Z_i^t = 0 | Z_i^{t-1} = 1] = 0$

$$\begin{aligned} \mathbb{P}[Z_i^1 = 0] &= 1 - p_1 \\ \mathbb{P}[Z_i^2 = 0] &= \mathbb{P}[Z_i^2 = 0 | Z_i^1 = 0] \mathbb{P}[Z_i^1 = 0] = (1 - p_2)(1 - p_1) \\ &\vdots \\ \mathbb{P}[Z_i^t = 0] &= \prod_{j=1}^t (1 - p_j). \end{aligned}$$

To see Property 2, we again apply the law of total probability

$$\begin{aligned} \mathbb{P}[Z_i^1 = 1] &= p_1 \\ \mathbb{P}[Z_i^2 = 1] &= \mathbb{P}[Z_i^2 = 1 | Z_i^1 = 0] \mathbb{P}[Z_i^1 = 0] + \mathbb{P}[Z_i^2 = 1 | Z_i^1 = 1] \mathbb{P}[Z_i^1 = 1] \\ &= p_2(1 - p_1) + 1 \cdot p_1 = p_1 + p_2(1 - p_1) \\ &\vdots \\ \mathbb{P}[Z_i^t = 1] &= p_1 + p_2(1 - p_1) + \dots + p_t \prod_{j=1}^{t-1} (1 - p_j). \end{aligned}$$

□

B Proof of Identification Results in Section 4

B.1 Proof of Proposition 1

We begin by writing out X

$$X = \begin{bmatrix} 1 & Z_1^1 & f_1(Z^1) \\ \vdots & \vdots & \vdots \\ 1 & Z_N^1 & f_N(Z^1) \\ \vdots & \vdots & \vdots \\ 1 & Z_1^T & f_1(Z^T) \\ \vdots & \vdots & \vdots \\ 1 & Z_N^T & f_N(Z^T) \end{bmatrix} \in \mathbb{R}^{TN \times 3}.$$

$X^\top X$ is non-singular when the columns of X are linearly independent. Suppose by way of contradiction that the columns X are dependent: there is a nonzero vector $\lambda \in \mathbb{R}^3$ such that

$$\lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 = 0.$$

Recall from our hypothesis that there is a $i \in [N]$ and $t \in [T]$ with $Z_i^t = 0$ and $f_i(Z^t) \neq 0$. From the linear dependence of the columns, we have

$$\lambda_1 + \lambda_3 f_i(Z^t) = 0 \implies \lambda_1 = -\lambda_3 f_i(Z^t). \quad (\text{B.1})$$

Now, take $j \in [N]$, $t' \in [T]$ such that $Z_j^{t'} = 0$ and $f_i(Z^t) \neq f_j(Z^{t'})$ as assumed in our hypothesis. Linear dependence again implies

$$\lambda_1 + \lambda_3 f_j(Z^{t'}) = 0 \Leftrightarrow -\lambda_3 f_i(Z^t) + \lambda_3 f_j(Z^{t'}) = 0 \Leftrightarrow \lambda_3 (f_j(Z^{t'}) - f_i(Z^t)) = 0 \Rightarrow \lambda_3 = 0,$$

where we use $f_j(Z^{t'}) \neq f_i(Z^t)$ in the final line. Conclude $\lambda_1 = 0$ from Eq. (B.1).

Now, take any $k \in [N]$ and $q \in [T]$ such that $Z_k^q = 1$, which is guaranteed to exist by the definition of the roll-out. Notice that if the columns of X are linearly dependent, $\lambda_1 = \lambda_3 = 0$ implies

$$\lambda_1 + \lambda_2 + \lambda_3 f_j(Z^{t'}) = 0 \Rightarrow \lambda_2 = 0.$$

This is a contradiction since $\lambda \neq \mathbf{0}$.

B.2 Proof of Corollary 1

Since $Z_j^t = 1$ and $i \in \mathcal{G}_1(j) \Leftrightarrow j \in \mathcal{G}_1(i)$, unit i 's interference term in model (2.2a) is positive: $f_i(\vec{z}) = \sum_{j \in \mathcal{G}_1(i)} \vec{z}_j > 0$. Noting that unit i is untreated $Z_i^t = 0$, the hypothesis of Proposition 1 is satisfied for (i, t) and $(i, 1)$.

B.3 Proof of Theorem 2

To prove this result, we consider two optimization problems that solve for the upper and lower bounds of the total treatment effect.

$$\begin{aligned} \max_{\theta \in \mathbb{R}^k} c^\top \theta & & \min_{\theta \in \mathbb{R}^k} c^\top \theta \\ \text{s.t. } X^\top X \theta = X^\top Y & & \text{s.t. } X^\top X \theta = X^\top Y \end{aligned} \quad (\text{B.2})$$

The dual problems are given by

$$\begin{aligned} \min_{p \in \mathbb{R}^k} p^\top X^\top Y & & \max_{p \in \mathbb{R}^k} p^\top X^\top Y \\ \text{s.t. } p^\top X^\top X = c^\top & & \text{s.t. } p^\top X^\top X = c^\top \end{aligned} \quad (\text{B.3})$$

The dual problems are feasible if and only if c lies within $\text{span}(X^\top X) = \text{span}(X^\top)$ where $\text{span}(\cdot)$ refers to the column span. We show that this condition implies that the lower and upper bounds (B.2) are equal, yielding a unique estimate of the total treatment effect. If $X^\top X$ is nonsingular, this is evident. Otherwise, let

$$\Theta_0 = \{\theta \in \mathbb{R}^k : X^\top X \theta = X^\top Y\}.$$

Suppose $X^\top X$ is singular so that there may exist $\theta, \theta' \in \Theta_0$ with $\theta \neq \theta'$. Let $\tilde{\theta}, \tilde{p}$ and $\underline{\theta}, \underline{p}$ be the optimal primal-dual pair for the upper and lower bound problems respectively. From primal feasibility $\tilde{\theta}, \underline{\theta} \in \Theta_0$ and dual feasibility $\tilde{p}, \underline{p} \in \{p : X^\top X p = c\}$, strong duality gives

$$\begin{aligned} c^\top \tilde{\theta} &= \tilde{p}^\top X^\top Y = \tilde{p}^\top X^\top X \underline{\theta} && \underline{\theta} \in \Theta_0 \\ &= (X^\top X \tilde{p})^\top \underline{\theta} && X^\top X \text{ is symmetric} \\ &= c^\top \underline{\theta} && X^\top X \tilde{p} = c \text{ by dual feasibility} \end{aligned}$$

B.4 Proof of Theorems 3 and 4

By the definition of the minimum eigenvalue, we have the tautological bound

$$\left\| \hat{\theta} - \theta^* \right\|_2^2 \leq \lambda_{\min} \left(X^\top X \right)^{-1} \left\| X \hat{\theta} - X \theta^* \right\|_2^2.$$

Conditioning on X so that $\lambda_{\min}(X^\top X)$ is deterministic, Cauchy-Schwarz gives

$$\mathbb{E}_X |c^\top (\hat{\theta} - \theta^*)|^2 \leq \|c\|_2^2 \mathbb{E}_X \left\| \hat{\theta} - \theta^* \right\|_2^2 \leq \|c\|_2^2 \lambda_{\min} \left(X^\top X \right)^{-1} \cdot \mathbb{E}_X \left\| X \hat{\theta} - X \theta^* \right\|_2^2. \quad (\text{B.4})$$

To bound the final term in the preceding display, we use an adaptation of [Rigollet and Hütter \(2015, Theorem 2.2\)](#); recall that $X \in \mathbb{R}^{NT \times K}$ and $Y = X \theta^* + \epsilon$.

Lemma 5. *Under the linear model of (4.1), suppose either Assumption A or B hold. Then, the least squares estimator $\hat{\theta} = (X^\top X)^{-1} X^\top Y$ satisfies*

$$\mathbb{E}_X \left\| X \hat{\theta} - X \theta^* \right\|_2^2 \leq \begin{cases} 4\sigma_{\max}^2 K T & \text{under Assumption A} \\ 4\sigma_{\max}^2 K & \text{under Assumption B} \end{cases}. \quad (\text{B.5})$$

Applying the lemma, we have the desired result.

In the remainder of the section, we prove the bound (B.5). Since $\hat{\theta}$ is the least squares estimator, we have

$$\left\| Y - X \hat{\theta} \right\|_2^2 \leq \left\| Y - X \theta^* \right\|_2^2 = \|\epsilon\|_2^2.$$

As we assume a well-specified linear model $Y = X \theta^* + \epsilon$, this implies

$$\|\epsilon\|_2^2 \geq \left\| Y - X \hat{\theta} \right\|_2^2 = \left\| X \theta^* + \epsilon - X \hat{\theta} \right\|_2^2 = \left\| X \hat{\theta} - X \theta^* \right\|_2^2 - 2\epsilon^\top X (\hat{\theta} - \theta^*) + \|\epsilon\|_2^2.$$

Rearranging terms, we get

$$\left\| X \hat{\theta} - X \theta^* \right\|_2 \leq 2 \frac{\epsilon^\top X (\hat{\theta} - \theta^*)}{\left\| X (\hat{\theta} - \theta^*) \right\|_2}. \quad (\text{B.6})$$

Let $\Phi \in \mathbb{R}^{NT \times K}$ be an orthonormal basis for the column span of X . Then, there exists $\nu \in \mathbb{R}^K$ such that $X (\hat{\theta} - \theta^*) = \Phi \nu$. Letting $B_K = \{u \in \mathbb{R}^K : \|u\|_2 \leq 1\}$ be the unit ball in \mathbb{R}^K , conclude

$$\begin{aligned} \frac{\epsilon^\top X (\hat{\theta} - \theta^*)}{\left\| X (\hat{\theta} - \theta^*) \right\|_2} &= \frac{\epsilon^\top \Phi \nu}{\|\Phi \nu\|_2} = \frac{\epsilon^\top \Phi \nu}{\|\nu\|_2} && \text{by orthonormality of } \Phi \\ &\leq \sup_{\mu \in B_K} \epsilon^\top \Phi \mu && \text{since } \nu / \|\nu\|_2 \in B_K. \end{aligned}$$

From Equation (B.6), we arrive at

$$\|X\hat{\theta} - X\theta^*\|_2^2 \leq 4 \cdot \left(\sup_{\mu \in B_k} \epsilon^\top \Phi \mu \right)^2. \quad (\text{B.7})$$

Since the last expression is maximized at $\mu = \Phi^\top \epsilon / \|\Phi^\top \epsilon\|_2$

$$\mathbb{E}_X \left[\left(\sup_{\mu \in B_k} \epsilon^\top \Phi \mu \right)^2 \right] = \mathbb{E}_X \left[\sum_{i=1}^K \left([\Phi^\top \epsilon]_i \right)^2 \right] \leq K \cdot \max_{i=1, \dots, K} \text{Var}_X([\Phi^\top \epsilon]_i).$$

We now consider the case of perfectly correlated and fully independent errors separately.

Case for Assumption A (Theorem 3) Recall that under Assumption A, $\epsilon_{i,t}$ are perfectly correlated across t . For convenience, we define the vector $\vec{\sigma}_i = [\sigma_{i,1}, \dots, \sigma_{i,T}]^\top \in \mathbb{R}^T$ and let $\Sigma_i = \vec{\sigma}_i (\vec{\sigma}_i)^\top$. We begin by noting that

$$\text{Cov}_X(\Phi^\top \epsilon) = \Phi^\top \text{Cov}(\epsilon) \Phi \preceq \lambda_{\max}(\text{Cov}(\epsilon)) I_{K \times K},$$

where Assumption A imposes a block-diagonal structure on $\text{Cov}(\epsilon)$ such that each block is given by Σ_i . Letting $\eta = [\eta_1, \dots, \eta_N]^\top \in \mathbb{R}^{NT}$ so that $\eta_i \in \mathbb{R}^T$, we bound the variational representation for the maximum eigenvalue of $\text{Cov}(\epsilon)$

$$\begin{aligned} \lambda_{\max}(\text{Cov}(\epsilon)) &= \max_{\|\eta\|_2=1} \eta^\top \text{Cov}(\epsilon) \eta = \max_{\|\eta\|_2=1} \sum_{i=1}^N \eta_i^\top \Sigma_i \eta_i = \max_{\|\eta\|_2=1} \sum_{i=1}^N \left(\eta_i^\top \vec{\sigma}_i \right)^2 \\ &\leq \max_{\|\eta\|_2=1} \sum_{i=1}^N \|\eta_i\|_2^2 \|\vec{\sigma}_i\|_2^2 \leq T \sigma_{\max}^2 \cdot \max_{\|\eta\|_2=1} \sum_{i=1}^N \|\eta_i\|_2^2 = T \cdot \sigma_{\max}^2. \end{aligned}$$

Noting that $\text{Cov}_X(\Phi^\top \epsilon) \preceq T \cdot \sigma_{\max}^2 \cdot I_{K \times K}$, conclude

$$\mathbb{E}_X \|X\hat{\theta} - X\theta^*\|_2^2 \leq 4\sigma_{\max}^2 KT.$$

Case for Assumption B (Theorem 4) Under Assumption B, $\epsilon_{i,t}$ is independent across both i and t . In this case, $\text{Cov}(\epsilon)$ is a diagonal matrix with entries $\text{Var}(\epsilon_{i,t})$ and evidently, $\lambda_{\max}(\text{Cov}(\epsilon)) \leq \sigma_{\max}^2$. As before, $\text{Cov}_X(\Phi^\top \epsilon) \preceq \sigma_{\max}^2 \cdot I_{K \times K}$, which implies

$$\mathbb{E}_X \|X\hat{\theta} - X\theta^*\|_2^2 \leq 4\sigma_{\max}^2 K.$$

B.5 Proof of Lemma 1 and 2

Under the model in Proposition 1, $X^\top X$ can be computed as

$$X^\top X = \begin{pmatrix} 1 & \cdots & 1 \\ Z^1 & \cdots & Z^T \\ f(Z^1) & \cdots & f(Z^T) \end{pmatrix} \cdot \begin{pmatrix} 1 & Z^1 & f(Z^1) \\ \vdots & \vdots & \\ 1 & Z^T & f(Z^T) \end{pmatrix} = \begin{pmatrix} NT & \sum_{i,t} Z_i^t & \sum_{i,t} f_i(Z^t) \\ \sum_{i,t} Z_i^t & \sum_{i,t} Z_i^t Z_i^t & \sum_{i,t} Z_i^t f_i(Z^t) \\ \sum_{i,t} f_i(Z^t) & \sum_{i,t} Z_i^t f_i(Z^t) & \sum_{i,t} (f_i(Z^t))^2 \end{pmatrix}$$

where $Z^t = [Z_1^t, \dots, Z_N^t]^\top$ and $f(Z^t) = [f_1(Z^t), \dots, f_N(Z^t)]^\top$. Under a completely randomized design with allocation vector \vec{p} and linear-in-means interference, we can fully characterize the design matrix X . Specifically, we know that the first column is the vector $\mathbf{1} \in \mathbb{R}^{NT}$ (i.e. the intercept term), and the second column is the stacked treatment vectors $[Z^1, \dots, Z^T]^\top \in \mathbb{R}^{NT}$. The last column is the interference term which can be computed exactly because we know since $f_i(Z^t) = \frac{1}{|\mathcal{G}(i)|} \sum_{j \in \mathcal{G}(i)} Z_j^t$ and \mathcal{G} is defined by a fully connected network. Therefore, $|\mathcal{G}(i)| = N - 1$ for every $i \in [N]$. Then if unit i is treated ($Z_i^t = 1$) at period t we have

$$f_i(Z^t) = \frac{(\sum_{k=1}^t p_k)N - 1}{N - 1}.$$

Otherwise, if unit i is untreated ($Z_i^t = 0$) at period t we have

$$f_i(Z^t) = \frac{(\sum_{k=1}^t p_k)N}{N - 1}.$$

From here we can easily compute the elements of the matrix $X^\top X$ which are given by

$$\begin{aligned} \sum_{i,t} Z_i^t &= \sum_{i,t} f_i(Z^t) = N(Tp_1 + (T-1)p_2 + \dots + 1 \cdot p_T) = N \sum_{t=0}^{T-1} t \cdot p_{T-t}, \\ \sum_{i,t} Z_i^t f_i(Z^t) &= \frac{p_1 N - 1}{N - 1} p_1 N + \dots + \frac{(N \sum_{t=1}^{T-1} p_t - 1)}{N - 1} \left(N \sum_{t=1}^{T-1} p_t \right) + \frac{(N \sum_{t=1}^T p_t - 1)}{N - 1} \left(N \sum_{t=1}^T p_t \right) \\ &= \sum_{J=1}^T \left(\frac{(\sum_{t=1}^J p_t) N - 1}{(N - 1)} \right)^2 \left(N \sum_{t=1}^J p_t \right), \text{ and} \\ \sum_{i,t} f_i(Z^t)^2 &= \sum_{J=1}^T \left(\frac{(\sum_{t=1}^J p_t) N - 1}{(N - 1)} \right)^2 \left(N \sum_{t=1}^J p_t \right) + \left(\frac{(\sum_{t=1}^J p_t) N}{(N - 1)} \right)^2 \left(N \left(1 - \sum_{t=1}^J p_t \right) \right). \end{aligned}$$

Plugging these into our matrix $X^\top X$, we can now solve the characteristic polynomial, $\det(\lambda I - X^\top X) = 0$ as $N \rightarrow \infty$ and $T \rightarrow \infty$ and examine how the eigenvalues scale with NT . Due to the analytic complexity of the problem, we compute this in Mathematica using the `ASYMPTOTICSOLVE` method which yields that for NT large enough

$$\lambda_i \asymp \frac{NT}{C_i(\vec{p})} \text{ for all } i$$

where $C_i(\vec{p})$ is a function that only depends on the increment vector \vec{p} .

Case for Assumption A (Lemma 1) Define $\bar{C}_1 = K \cdot \sigma_{\max}^2 \cdot C_{i^*}(\vec{p})$ where $i^* = \arg \min_i \lambda_i$ and K is fixed under the model. Under this model $c = [0, 1, 1] \implies \|c\|_2^2 = 2$. Applying Theorem 3 and plugging in these values yields the desired result: for some $M \in \mathbb{R}$ and $NT > M$

$$\mathbb{E} \left[|c^\top (\theta^* - \hat{\theta})|^2 \right] \leq KT \cdot \sigma_{\max}^2 \cdot \left(\frac{NT}{C_i(\vec{p})} \right)^{-1} = \frac{8\bar{C}_1}{N}.$$

Case for Assumption B (Lemma 2) Define $\bar{C}_2 = K \cdot \sigma_{\max}^2 \cdot C_{i^*}(\vec{p})$ where $i^* = \arg \min_i \lambda_i$ and σ_{\max}^2 is from Assumption B. Again, $c = [0, 1, 1] \implies \|c\|_2^2 = 2$. Plugging in these values and noting that T when using Theorem 4. Conclude that for some $M \in \mathbb{R}$ large enough, whenever $NT > M$

$$\mathbb{E} \left[|c^\top(\theta^* - \hat{\theta})|^2 \right] \leq K \cdot \sigma_{\max}^2 \cdot \left(\frac{NT}{C_i(\vec{p})} \right)^{-1} = \frac{8\bar{C}_2}{NT}.$$

C Estimation of the TTE under Model Misspecification

In Figure 9, we assess the performance of our estimator based on the potential outcomes model specified in Cortez et al. (2022b) (Sec. 5 Eq. 4) and reproduced in (3.7). We apply LOPO to variations of the model in 3.2 including a model with a second-order term. In the following experiment, we generate data using the model (3.7) with $\beta = 2$. We also consider a misspecified Lagrange interpolation estimator of Cortez et al. (2022b) with $\beta \neq 2$. To facilitate comparison, we follow Cortez et al. (2022b) and do not include a noise term so that any error is due to model misspecification alone.

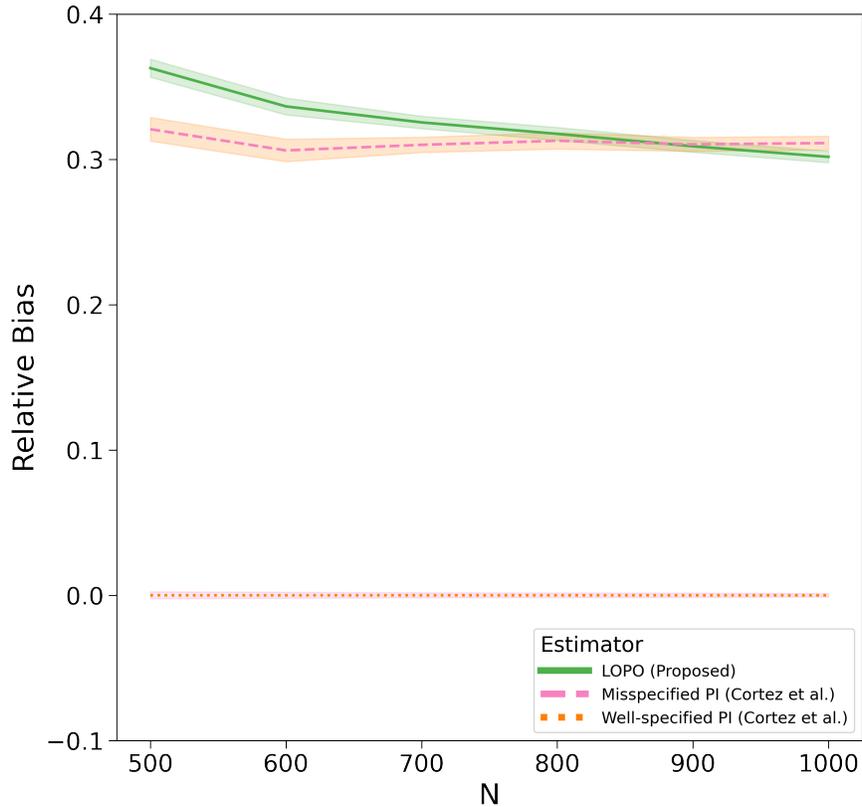


Figure 9. Relative Bias of Estimators of the TTE: figures are generated by averaging the results of 100 experiments. In each experiment sample size is given by the x-axis, N , with $T = 4$ and an uneven roll-out. We display bootstrapped 95% confidence intervals of the relative bias. We use the DGP given by Cortez et al. (2022b) in Eq. (3.7) with $r = 2$ and $\beta = 2$. Note that the correctly specified PI estimator (with $\beta = 2$) has constant zero bias, hence there is no visible CI.

Firstly, we validate that in the well-specified case, we reproduce the results of Yu et al. (2022a),

which demonstrates that their estimator has zero bias. As the sample size increases, our misspecified estimator using LOPO preforms similarly to the Lagrange interpolation estimator. The phenomenon shows how linear models are able to approximate polynomial models. This also underscores the need to consider model misspecification in practice as even slight changes in the interference model estimated can result in possibly large relative biases.