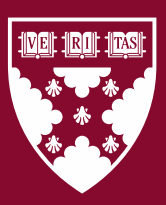
Working Paper 24-058

Design of Panel Experiments with Spatial and Temporal Interference

Tu Ni
Iavor Bojinov
Jinglong Zhao



Harvard
Business
School

Design of Panel Experiments with Spatial and Temporal Interference

Tu Ni
National University of Singapore

Iavor Bojinov
Harvard Business School

Jinglong Zhao
Boston University

Working Paper 24-058

Funding for this research was provided in part by Harvard Business School.

Design of Panel Experiments with Spatial and Temporal Interference

Tu Ni*
Digital, Data, Design Institute, Harvard Business School, Boston, MA 02163, tni@hbs.edu

Iavor Bojinov
Technology and Operations Management Unit, Harvard Business School, Boston, MA 02163, ibojinov@hbs.edu

Jinglong Zhao
Boston University, Questrom School of Business, Boston, MA, 02215, jinglong@bu.edu

Organizations routinely conduct experiments on millions of interconnected people to evaluate new products and services. However, as the practice has proliferated, organizations have encountered a principal challenge: interference, the process in which one person's outcome depends on the treatment assignment of others. The bias from ignoring interference can be substantial, affecting both the magnitude and the sign of the naive difference-in-means estimator. In this paper, we focus on cross-unit interference that is modeled as occurring through edges on a two-dimensional lattice, which is particularly prominent in modern marketplaces, such as ride-sharing, food delivery, and homestay. The leading strategy for overcoming such interference is to combine all experimental units into a single group and perform a switchback (or time series) experiment, in which the treatment assignment is randomized across time periods. However, this approach suffers from low statistical power when the number of time periods is limited. This paper proposes a novel design of panel experiment (the generalization of switchback experiments to multiple units) that allows for both interference and carryover effects (the process whereby past treatments affect current outcomes). Our proposed design has two features: the first is a notion of randomized spatial clustering, which we refer to as random shaking, that partitions units into equal-size clusters; the second is a notion of balanced temporal randomization that extends the classical completely randomized designs to the temporal interference setting. We prove the theoretical performance of our design, develop its inferential techniques, and verify its superior performance by conducting extensive simulations, including simulations using real data from a ride-hailing platform. Practically, our new design can help researchers and practitioners achieve over a 50% increase in statistical power with the same sample size.

Key words: Experimental design, causal inference, network interference, cluster randomized experiments, switchback experiments.

---

* Alphabetical order other than the first author.

1. Introduction

In the past decade, companies have adopted randomized controlled trials (*i.e.,* experiments) as a standard practice for evaluating the impact of changes to products and services before wide-scale release (Thomke 2003, Kohavi and Thomke 2017, Cui et al. 2018, Gupta et al. 2019, Bojinov and Gupta 2022, Larsen et al. 2023). Today, companies conduct thousands of experiments annually on millions of interconnected people (Thomke 2020). The most prominent experimental design is the standard A/B test in which customers are randomly assigned to the change (often called the treatment or "B"), while the rest receive the status quo (often called the control or "A"); the impact is then estimated by comparing the two groups across various business metrics or outcomes (such as revenue, sales, and customer engagement). Unfortunately, A/B tests produce biased results in the presence of interference (Cox 1958) — the processes one unit's treatment impacts another's outcomes — a commonplace phenomenon routinely recognized as one of the main pitfalls of experimentation (Gupta et al. 2019, Bojinov et al. 2020).

Interference is particularly prevalent and challenging in service marketplaces (Taylor 2018) such as DoorDash (Tang et al. 2020), Lyft (Chamandy 2016), and Uber (Farronato et al. 2018), where the interference occurs across units that are in close physical proximity. For a motivating example, suppose that a ride-sharing firm develops a new pricing algorithm. If the firm randomly assigns half of the drivers within a city to the new algorithm, these drivers are likely to alter their behavior, which will affect the drivers assigned to the control through their common pool of passengers. The leading strategy for overcoming such *spatial interference*, has been to aggregate all experimental subjects within a large geographical area into a single group. In the ride-sharing example, all drivers in a specific city could be aggregated into a single unit. Since it is impossible to run an A/B test on a single unit, managers have instead started to use switchback experiments (also known as time series experiments or N-of-1 trials) (Bojinov and Shephard 2019, Bojinov et al. 2022, Glynn et al. 2020, Hu and Wager 2022, Xiong et al. 2022, Chen and Simchi-Levi 2023, Jia et al. 2024).

In a switchback experiment, a single unit is randomly assigned to either treatment or control, its outcome is recorded, and the process is repeated for T time periods; creating a time series of treatment assignments and corresponding outcomes. Inference then focuses on the average responses over time. Although switchback experiments overcome the challenge posed by spatial inference, they create two additional complications. First, time series experiments often suffer from carryover effects (or temporal interference), where past treatments can impact current outcomes (Cox 1958). Prior literature has proposed several ways to incorporate temporal interference. For example, Bojinov and Shephard (2019) focused on a causal estimand that measured the effect of administering an additional treatment, conditional on past assignments; Glynn et al. (2020) proposed using Markov chain modeling to estimate the transition probabilities; and Hu and Wager (2022) suggested the

usage of burn-out periods to stabilize temporal interference, leveraging the rapid mixing condition in Markov chains. Jia et al. (2024) further extended and generalized the ideas in Hu and Wager (2022) with improved performance bound. Second, time series experiments typically have relatively low statistical power, as the variance is inversely proportional to the (few) time points as opposed to the (many) experimental units. Researchers have proposed specific designs that reduce the variance and increase the statistical power; for example, Bojinov et al. (2022) provided a minimax Bernoulli design for switchback experiments and Glynn et al. (2020) proposed optimizing a state-dependent switchback design. Although the above designs help aliviate some of the power conserns, they do not solve the underlying root cause: the aggregation of all subjects into a single unit.

An open problem in this domain is to determine whether it is necessary to aggregate all experimental units in a single group. With less aggregation, it may be possible to achieve substantially higher power by running a panel experiment (also called cross-over experiments), the generalization of switchback experiments to multiple subjects (Cox 1958, Cochran and Cox 1962, Brown Jr 1980, Xiong et al. 2019, Basse et al. 2019, Bajari et al. 2021, Bojinov et al. 2021). Of course, the benefits would depend on the specifics of the interference structure. For instance, Han et al. (2024) showed that for panel experiments with population interference, the additional ability to change units' treatments over time can substantially increase statistical power, but only when the carryover effects are limited. If the carryover effects were long-lasting, the benefits disappeared.

In our motivating ride-sharing example, the spatial interference can be modeled as occurring through a two-dimensional lattice where the vertices represent experiment (grouped) units and the interference is captured by the edges. Specifically, vertices could be geographical locations with an edge connecting two locations if there is a substantial number of rides between them; typically, this would result in connecting regions in close proximity. More generally, it is possible to define these nodes so that the interference only occurs between neighboring nodes (see Figure 2 for visualization of a 6×6 lattice with nearest neighbor interference). This type of interference structure is prevalent beyond ride-sharing. For example, travelers on a homestay platform compete for customers in a specific geographical area – enhancing the picture quality of one house could cannibalize the demand for other nearby properties. Similarly, customers on a meal delivery platform typically order food from local restaurants; testing a new order batching policy in several restaurants may impact the delivery time in the nearby regions, but not beyond.

Main results. The paper proposes a novel design of panel experiments subject to spatial interference modeled as a two-dimensional lattice with limited carryover effects. Our design has two features: *randomized spatial clustering* for determining the appropriate amount of aggregation and *temporal balancing* for ensuring that each aggregated unit is assigned to receive the same number

of treatment and control periods. The two steps in our design are partly inspired by the standard completely randomized design, which ensures that the number of treated and control units are equal and each unit has the same chance of receiving either treatment (Fisher 1936).

In the *randomized spatial clustering* step, we first define spatial regions in which all experimental subjects (*e.g.*, drivers, passengers, houses, restaurant) are aggregated into experimental units so that the interference is restricted to adjoining units. Next, these units are partitioned into equal-sized clusters. We then build on the idea of cluster randomization (Eckles et al. 2017, Holtz et al. 2020, Manski 2013, Ugander et al. 2013, Ugander and Yin 2020, Candogan et al. 2021) to add an outer layer of randomization to the clusters that randomly "shakes" the boundary of all clusters, ensuring that the propensity score is uniform for all interior units. We refer to this outer layer of randomization as *random shaking*, which, to the best of our knowledge, is new to the experimental design literature. As we show in Section 8, random shaking reduces the variance of our estimator by nearly 85%. Finally, we derive the optimal cluster size from a minimax perspective, demonstrating that performing limited aggregations and allowing some cross-unit interference is substantially better than aggregating all units and performing a single switchback experiment.

Next, our design runs a panel experiment in which units in the same cluster receive the same assignment independently of other clusters. Specifically, for each cluster in the *temporal balancing* step, our design ensures that the number of treatment and control periods are equal by performing the analogous completely randomized design, where the time periods are treated as units. Temporal balancing is a key procedure for controlling for carryover effects and preserving a sufficient volume of data. We show provable guarantees of our balanced design from a minimax perspective based on the creative lower bound techniques from Candogan et al. (2021). Finally, we demonstrate both theoretically and empirically that the temporal balanced design, when applied to a single unit, improves on the previous optimal designs for switchback experiments proposed by Bojinov et al. (2022).

For our proposed design, we also provide an inference strategy that leverages the variation introduced by treatment assignments. Specifically, we derive a novel central limit theorem for a fixed number of units as the number of time periods goes to infinity. Our result extends the prior work on completely randomized design (Li and Ding 2017) to panel experiments. The proof builds on earlier ideas from the double-index permutation literature and may be of independent interest (Daniels 1944, Hoeffding 1951, Zhao et al. 1997, Reinert and Röllin 2009).

To demonstrate the robustness of our design, we examine its empirical performance using a comprehensive simulation study, where we consider more general outcome models and interference structures than those required by our theoretical analysis. First, we show that under a stochastic outcome model, our design with the optimal cluster size achieves the smallest variance, even though

our optimal design was not derived under this setting. Second, we demonstrate the robustness of our design by considering more complex networks when spatial interference is present not only within the neighborhood but also among remote units. We consider two simulation setups and construct the remote interference structures using both synthetic networks and a real network abstracted from a ride-hailing platform. We also develop a simple heuristic algorithm for additional aggregation when the interference is more general; our simulations suggest that the algorithm further improves on our design when the interference structure is far from the two-dimensional lattice assumed by our theory. Finally, we show how the optimal cluster sizes depend on the interference structure. When interference is more pervasive, the optimal cluster size should be larger; when interference is less spread, the optimal cluster size should be smaller. Again, this suggests that there are a slew of possible strategies besides clustering all experimental units and performing a single switchback experiment.

Alternative approaches to handling interference. Instead of using switchback or panel experiments, researchers have proposed numerous strategies for overcoming the challenges posed by interference. Typically, these involve performing a subject-level experiment without grouping and adjusting for the interference in the analysis stage by modeling either the outcomes or the exposure probabilities. Although the introduction of modeling assumptions may not precisely characterize the impact of interference, it takes advantage of a high volume of data from thousands or millions of users for inference.

When modeling the outcomes, researchers have proposed methods that utilize the underlying problem contexts (Bright et al. 2022, Farias et al. 2022, Johari et al. 2022b, Li et al. 2021, Munro et al. 2021, Wager and Xu 2019, Li et al. 2023, Dhaouadi et al. 2023, Zhu et al. 2024). For example, Bright et al. (2022) models interference coming from a centralized matching market. Farias et al. (2022), Johari et al. (2022b),Li et al. (2021), Dhaouadi et al. (2023) and Zhu et al. (2024) model interference coming from experimental units sharing a common competitive resource. Munro et al. (2021) and Wager and Xu (2019) propose to use an equilibrium price to capture interference. Li et al. (2023) considers the interference from stochastic congestion when experimental units stay in a queue.

When modeling the exposure probabilities, researchers often search for an underlying structure that limits the scope of interference. This often entails assuming that interference remains confined within groups but not across. Additionally, the indirect effect within a group is typically assumed to be contingent solely on the number of treated individuals (Rosenbaum 2007, Hudgens and Halloran 2008, Sinclair et al. 2012, Li et al. 2019, Sävje et al. 2021). For general interference patterns, Aronow and Samii (2017) presents a comprehensive framework where they introduce exposure mapping, define useful causal estimands, and develop asymptotically valid confidence intervals based on the Horvitz-Thompson estimator.

Roadmap of the paper. The remaining of this paper is structured as follows. In Section 2, we define the treatment assignments and the potential outcomes, discuss the assumptions on interference and introduce the causal effect of interests. In Section 3, we introduce the design of experiments together with the causal estimator. In Section 4, we propose our new design of panel experiments and give an overview of randomized spatial clustering and temporal balancing. In Section 5, we first analyze the temporal balancing by formulating a minimax optimization problem. In Section 6, we then investigate the random shaking idea for randomized spatial clustering. Following a similar minimax optimization framework, we develop the optimal cluster size. In Section 7 we present a new central limit theorem to draw inference. In Section 8, we conduct extensive simulations to evaluate the performance of our new design. In Section 9, we conclude the paper with several practical implications and future directions.

2. Setups, Notations, and Assumptions

2.1. The Assignment Matrix and the Potential Outcomes

Companies regularly run experiments over geographical regions that can be represented as a two-dimensional grid. For example, when data scientists at Lyft (Chamandy 2016) conducted an experiment to test the effectiveness of a surge price subsidy, they partitioned a city into smaller regions to receive different levels of subsidies; see Figure 1 for an illustration. In such experiments, each experimental unit is defined by a local square region as shown in Figure 1.

We model each experimental unit, *i.e.*, each local square region, as a vertex in a lattice, where the edges approximate the interactions (or potential interference) between the local square regions; see Figure 2 for an illustration. For each positive integer $D \in \mathbb{N}$, let there be a $D \times D$ lattice $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ stands for the vertex set with a total of $N = D^2$ units and $\mathcal{E}$ stands for the edge set. For example, the Lyft experiment as shown in Figure 1 could be mapped to a 25×25 grid. For each unit $i \in [N]$, we define its neighborhood $\mathcal{N}(i) = \{j \in [N] | (i,j) \in \mathcal{E}\} \cup \{i\}$ to be the vertex i and the set of vertices connected through edges to the focal vertex i. For an illustration, see the bold edges and units in Figure 2.

Let there be a total of $T \in \mathbb{N}$ periods in the experiment. As suggested by earlier works (Bojinov et al. 2022, Tang et al. 2020), each period is typically selected to be the same length as the carryover effect — the length of time a treatment persists in impacting future outcomes. For on-demand service platforms, one period typically ranges from about 30 minutes to several hours (Chamandy 2016, Tang et al. 2020).

At the design phase of the experiment, the experimenter is presented with T periods and N units. At each time period $t \in [T]$, the experimenter chooses to expose each unit $i \in [N]$ to either treatment (or the new development) or control (the status quo). Let $W_{i,t} \in \{0, 1\}$ be the treatment assignment

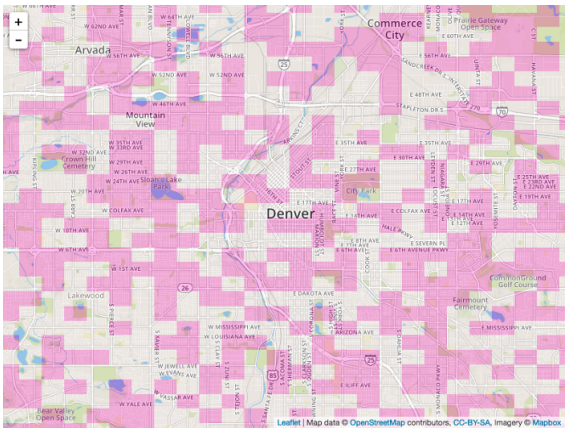


Figure 1 **An illustration of the Geohash system at Lyft (Chamandy 2016).**

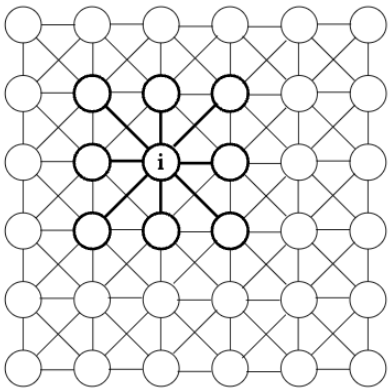


Figure 2 **An illustration of a 6×6 lattice, and the neighborhood of $\mathcal{N}(i)$.**

for unit $i \in [N]$ at time period $t \in [T]$, where $W_{i,t} = 1$ stands for treatment and $W_{i,t} = 0$ stands for control. The assignment matrix is then defined by a collection of assignments across all units and over all time periods, $\boldsymbol{W}_{1:N,1:T} \in \{0,1\}^{N \times T}$. Following convention, we use $\boldsymbol{W}_{1:N,1:T}$ to stand for a random assignment matrix, and $\boldsymbol{w}_{1:N,1:T}$ to stand for one realization.

Once the assignment matrix $\boldsymbol{w}_{1:N,1:T}$ is realized, the experimenter will observe some outcomes of interest. Under the potential outcomes framework (Neyman 1923, Rubin 1974, Robins 1986), we model the observed outcomes to be related to their respective potential outcomes. We denote $Y_{i,t}(\boldsymbol{w}_{1:N,1:T})$ to be the outcome of unit $i \in [N]$ at time period $t \in [T]$ under the assignment matrix $\boldsymbol{w}_{1:N,1:T}$. As a short-hand notation, denote $\mathbb{Y} = \{Y_{i,t}(\boldsymbol{w}_{1:N,1:T})\}_{i,t,\boldsymbol{w}_{1:N,1:T}}$ to be the collection of all potential outcomes. In this paper, we do not assume structural models for the potential outcomes or impose parametric assumptions (Wager and Xu 2019, Glynn et al. 2020, Munro et al. 2021, Johari et al. 2022b, Li et al. 2021, Bright et al. 2022, Farias et al. 2022). Instead, we adopt a design-based perspective (Neyman 1923, Fisher 1936, Kempthorne 1955, Rubin 1980, Imbens and Rubin 2015, Abadie et al. 2020) and treat the potential outcomes as fixed quantities. Or, equivalently, we condition on $\mathbb{Y}$. To make inference possible, we rely on variations introduced by the random assignment matrix.

For any unit and at any time, the experimenter only observes $Y_{i,t}(\boldsymbol{w}_{1:N,1:T})$ under one assignment matrix $\boldsymbol{w}_{1:N,1:T}$, but is unable to observe the potential outcomes under other assignment matrices (Holland 1986). Without making additional assumptions, it is impossible to achieve valid inference (Basse and Airoldi 2018).

2.2. Assumptions on the Spatial and Temporal Interference

We now introduce two practical assumptions that limit the spatial and temporal interference.

First, we assume that the potential outcomes of one unit depend only on the treatment assignments of this unit and its neighboring units. Mathematically, let $\boldsymbol{w}_{\mathcal{N}(i),1:T} \in \{0,1\}^{|\mathcal{N}(i)| \times T}$ be a sub-matrix

that contains the treatment assignments of units $\mathcal{N}(i)$ at all time periods. We introduce the following assumption.

ASSUMPTION 1 **(Neighborhood spatial interference)**. *For any $i \in [N], t \in [T]$ and any two assignment matrices $\boldsymbol{w}_{1:N,1:T}, \boldsymbol{w}'_{1:N,1:T} \in \{0,1\}^{N\times T}$, we have*

$$Y_{i,t}(\boldsymbol{w}_{1:N,1:T}) = Y_{i,t}(\boldsymbol{w}'_{1:N,1:T}) \quad \textit{whenever} \quad \boldsymbol{w}_{\mathcal{N}(i),1:T} = \boldsymbol{w}'_{\mathcal{N}(i),1:T}.$$

The version of this assumption when $T = 1$ is widely adopted in the literature (Ugander et al. 2013, Manski 2013, Aronow and Samii 2017, Ugander and Yin 2020, Han et al. 2024) and is viable in many applications, such as the homestays, ride-sharing, and food deliver, discussed in the introduction. In practice, experimenters are suggested to properly define each unit, relying on their domain knowledge, so that there is only neighborhood interference. However, as we show in Section 8, our design and analysis is somewhat robust to violations of this assumption.

Second, we limit the temporal interfernece, by assuming that the potential outcomes at one time period depend on the treatment assignments at this time period and the preceding time period. Mathematically, let $\boldsymbol{w}_{1:N,t-1:t} \in \{0,1\}^{N\times 2}$ be a sub-matrix that contains the treatment assignments for all units at time periods $t-1$ and t. We introduce the following assumption.

ASSUMPTION 2 **(Limited carryover effects)**. *For any $i \in [N], t \in [T]$ and any two assignment matrices $\boldsymbol{w}_{1:N,1:T}, \boldsymbol{w}'_{1:N,1:T} \in \{0,1\}^{N\times T}$, we have*

$$Y_{i,t}(\boldsymbol{w}_{1:N,1:T}) = Y_{i,t}(\boldsymbol{w}'_{1:N,1:T}) \quad \textit{whenever} \quad \boldsymbol{w}_{1:N,t-1:t} = \boldsymbol{w}'_{1:N,t-1:t}.$$

This assumption is widely adopted in the literature (Laird et al. 1992, Senn and Lambrou 1998, Basse et al. 2019, Bojinov et al. 2022, Han et al. 2024) and viable in many applications as the experiment has control over the length of each period. We can relax this assumption by considering general lengths of carryover effects by allowing the potential outcomes at each time period to depend on the treatment assignments up to $m > 1$ periods ago. The structure of the main results will remain the same; however, for simplicity, in this paper, we focus on the case of $m = 1$ to de-emphasize the importance of the length of carryover effects to our main results. This assumption is in line with prior research that recommends selecting the length of one time period to be the same as the length of the carryover effect (Bojinov et al. 2022, Tang et al. 2020). For example, consider a ride-sharing platform testing a new surge pricing policy with a carryover effect that lasts around 30 to 60 minutes. In this example, setting the length of a period to be one hour would ensure the length of the carryover effect is one period; see Bojinov et al. (2022) for suggestions on how to identify the length of carryover effects.

Combining both assumptions, we have for any $i \in [N], t \in [T]$ and any two assignment matrices $\boldsymbol{w}_{1:N,1:T}, \boldsymbol{w}'_{1:N,1:T} \in \{0,1\}^{N\times T}$,

$$Y_{i,t}(\boldsymbol{w}_{1:N,1:T}) = Y_{i,t}(\boldsymbol{w}'_{1:N,1:T}) \quad \text{whenever} \quad \boldsymbol{w}_{\mathcal{N}(i),t-1:t} = \boldsymbol{w}'_{\mathcal{N}(i),t-1:t}.$$

In this paper, we use the short-hand notation $Y_{i,t}\left(\boldsymbol{w}_{\mathcal{N}(i),t-1:t}\right) = Y_{i,t}\left(\boldsymbol{w}_{1:N,1:T}\right)$ to focus on the dependence of the potential outcomes on the sub-matrices. Using this short-hand notation, the observed outcomes are related to their respective potential outcomes as follows,

$$Y_{i,t} = Y_{i,t}(\boldsymbol{w}_{\mathcal{N}(i),t-1:t}), \quad \text{if} \quad \boldsymbol{W}_{\mathcal{N}(i),t-1:t} = \boldsymbol{w}_{\mathcal{N}(i),t-1:t}.$$

3. The Causal Effect and Causal Estimator

Our primary causal estimand is the *Global Average Treatment Effect*, which measures the difference between the average outcomes when all units are exposed to treatment and when they are all exposed to control, in all time periods (Han et al. 2024). Mathematically, this is defined as

$$\tau(\mathbb{Y}) = \frac{1}{N(T-1)} \sum_{t=2}^{T} \sum_{i=1}^{N} \left[Y_{i,t}\left(\mathbf{1}\right) - Y_{i,t}\left(\mathbf{0}\right) \right], \tag{1}$$

where $\mathbf{1}$ and $\mathbf{0}$ are two assignment sub-matrices with all treatments $\boldsymbol{w}_{\mathcal{N}(i),t-1:t} = \mathbf{1}$ and controls $\boldsymbol{w}_{\mathcal{N}(i),t-1:t} = \mathbf{0}$, respectively. We also introduce a shorthand notation $\tau_{i,t} = Y_{i,t}\left(\mathbf{1}\right) - Y_{i,t}\left(\mathbf{0}\right)$ so that the causal estimand is equivalently $\tau(\mathbb{Y}) = \frac{1}{N(T-1)} \sum_{t=2}^{T} \sum_{i=1}^{N} \tau_{i,t}$.

The causal effect as defined in (1) captures the effect of permanently implementing a new policy and is the primary estimand of interest for managers. Since the causal effect $\tau(\mathbb{Y})$ is never directly observable, our goal is to estimate the causal effect using observations from the panel experiments efficiently, which requires a careful design of the experiment.

We define the design of experiments (also known as the randomization distribution or simply the design) as a discrete probability distribution $\eta(\cdot) : \{0,1\}^{N\times T} \to [0,1]$ over assignment matrices. For each design, the experimenter first draws one assignment matrix $\boldsymbol{W}_{1:N,1:T} = \boldsymbol{w}_{1:N,1:T}$ from the distribution η, and then implements this assignment matrix to conduct the experiment. During the experiment, the experimenter collects the observed outcomes $\{Y_{i,t}\}_{i\in[N],t\in[T]}$, and uses both the realized assignment matrix and the observed outcomes to estimate the causal estimand.

A commonly used estimator is the Inverse Propensity Weighted (IPW) estimator, which is also referred to as the Horvitz-Thompson estimator (Horvitz and Thompson 1952, Bojinov and Shephard 2019, Han et al. 2024):

$$\widehat{\tau}\left(\mathbb{Y}, \eta, \boldsymbol{w}\right) = \frac{1}{N(T-1)} \sum_{t=2}^{T} \sum_{i=1}^{N} \left\{ Y_{i,t} \ \frac{\mathbb{1}\left\{\boldsymbol{w}_{\mathcal{N}(i),t-1:t} = \mathbf{1}\right\}}{\Pr\left(\boldsymbol{W}_{\mathcal{N}(i),t-1:t} = \mathbf{1}\right)} - Y_{i,t} \ \frac{\mathbb{1}\left\{\boldsymbol{w}_{\mathcal{N}(i),t-1:t} = \mathbf{0}\right\}}{\Pr\left(\boldsymbol{W}_{\mathcal{N}(i),t-1:t} = \mathbf{0}\right)} \right\}. \tag{2}$$

Although this estimator is unbiased over the randomization distribution (that is, $\mathbb{E}_\eta[\widehat{\tau}(\mathbb{Y},\eta,\boldsymbol{w})] = \tau(\mathbb{Y})$) it only uses a subset of the observed data points. This is because the estimator requires an observed outcome $Y_{i,t}$ to be *valid*, meaning that $\boldsymbol{W}_{\mathcal{N}(i),t-1:t} \in \{\mathbf{0},\mathbf{1}\}$. A good design is one that maximizes the number of valid observations, while ensuring there is enough randomness to protect the inference from systematic unobserved factors.

Finally, to evaluate the quality of a design, we adopt the decision-theoretic framework (Berger 2013, Bickel and Doksum 2015) and focus on finding designs that minimize the variance of the IPW estimator, $\mathrm{Var}_\eta(\widehat{\tau}(\mathbb{Y},\eta,\boldsymbol{w})) = \mathbb{E}_\eta[(\widehat{\tau}(\mathbb{Y},\eta,\boldsymbol{w}) - \tau(\mathbb{Y}))^2]$. Since the IPW estimator is unbiased, the variance of the estimator is equivalent to the risk function, or the mean squared error of the estimator.

4. An Overview of Randomized Spatial Clustering and Temporal Balancing

This section provides an intuitive introduction to our proposed design; the theoretical properties are presented in the subsequent sections. Our design belongs to the family of designs H that involve the following two steps:

1. Aggregate adjacent units into clusters of size $d \times d$[1].
2. Conduct an independent time series (or switchback) experiment for each cluster, such that units within the same cluster receive the same treatment assignment at the same time.

Specifically, in the first step, we propose using a *randomized spatial clustering* that ensures uniform exposure probabilities. In the second step, we use a *temporal balanced design* that restricts each group to receiving the same number of treatments as control periods.

Both steps take inspiration from the standard completely randomized design (Fisher 1936), which, in the no-interference setting, ensures that the probabilities of having valid observations are the same for all experimental units and that an equal number of units are assigned to treatment and control. Intuitively, our first step can be through of as the spatial-interference generalization of the completely randomized deisgn and our second step as the temporal-interference generalization.

4.1. Randomized Spatial Clustering

For now, suppose that d, the parameter that determines the cluster sizes, is fixed and known; in Section 6.2, we derive the asymptotically optimal cluster size d within the family of randomized spatial clustering.

[1] Each cluster contains exactly $d \times d$ units as long as it is not on the boundary.

Lattice clustering. Recall that we model the experimental units as vertices on the $D \times D = N$ lattice. Assume there is a Cartesian coordinate system that encodes the vertices as pairs of integers (a_1, a_2), where $a_1, a_2 \in [D]$. For example, in Figure 3, vertex i has a coordinate of $(4, 6)$. The clustering of units can be constructed by partitioning the vertices using straight vertical and horizontal lines. Formally, let a cluster indexed by $\boldsymbol{\psi} = (\underline{\psi}_1, \underline{\psi}_2, \overline{\psi}_1, \overline{\psi}_2)$ be $\mathcal{C}_{\boldsymbol{\psi}} = \left\{ (a_1, a_2) \middle| \underline{\psi}_1 \leq a_1 \leq \overline{\psi}_1, \underline{\psi}_2 \leq a_2 \leq \overline{\psi}_2 \right\}$.

For any d such that (D/d) is an integer, define a spatial clustering $\mathscr{C}(d)$ to be a collection of clusters, such that $\forall\ \mathcal{C}_{\boldsymbol{\psi}} \in \mathscr{C}(d)$, $\overline{\psi}_1 - \underline{\psi}_1 = \overline{\psi}_2 - \underline{\psi}_2 = d$ and that $\bigcup_{\mathcal{C}_{\boldsymbol{\psi}} \in \mathscr{C}_d} \mathcal{C}_{\boldsymbol{\psi}} = [N], \bigcap_{\mathcal{C}_{\boldsymbol{\psi}} \in \mathscr{C}(d)} \mathcal{C}_{\boldsymbol{\psi}} = \emptyset$. In words, $\mathscr{C}(d)$ is a partition of the $D \times D = N$ lattice, such that each cluster $\mathcal{C}_{\boldsymbol{\psi}}$ is a smaller $d \times d$ square. See Figure 3 for an illustration when $d = 3$.

Random shaking. For clustering $\mathscr{C}(d)$, let $\boldsymbol{\omega} = (\omega_1, \omega_2)$ be a pair of shaking parameters whose support is given by

$$\omega_1, \omega_2 \in \Omega := \begin{cases} \left\{ -\dfrac{d-1}{2}, -\dfrac{d-3}{2}, ..., 0, ..., \dfrac{d-1}{2} \right\} & \text{if } d \text{ is odd,} \\ \left\{ -\dfrac{d}{2}, -\dfrac{d-2}{2}, ..., 0, ..., \dfrac{d-2}{2} \right\} & \text{if } d \text{ is even.} \end{cases}$$

Basically, we want to shake the clustering horizontally and vertically for d units, precisely equal to the size of each cluster, resulting in $|\Omega| = d$. For example, when $d = 3$ we have $\Omega = \{-1, 0, 1\}$. Let a cluster $\mathcal{C}_{\boldsymbol{\psi}}$ with the shaking parameters $\boldsymbol{\omega}$ be $\mathcal{C}_{\boldsymbol{\psi}}(\boldsymbol{\omega}) = \left\{ (a_1, a_2) \middle| \underline{\psi}_1 + \omega_1 \leq a_1 \leq \overline{\psi}_1 + \omega_1, \underline{\psi}_2 + \omega_2 \leq a_2 \leq \overline{\psi}_2 + \omega_2 \right\}$. A spatial clustering with shaking parameters $\boldsymbol{\omega} = (\omega_1, \omega_2)$ refers to the collection of clusters denoted by $\mathscr{C}(\boldsymbol{\omega}; d)$. See Figure 4 for an illustration when $d = 3$ and $\boldsymbol{\omega} = (-1, 1)$.

We refer to *random shaking* as a uniform probabilistic distribution over $\{\mathscr{C}(\boldsymbol{\omega}; d), \forall \boldsymbol{\omega} \in \Omega^2\}$. That is, the spatial clustering $\mathscr{C}(\boldsymbol{\omega}; d)$ is generated by sampling $\boldsymbol{\omega}$ from Ω^2 uniformly at random. For example, Figure 3 and Figure 4 each represent one sample of shaking parameters, each with a $1/9$ probability of being sampled.

In Figure 3 with $\boldsymbol{\omega} = (0, 0)$, unit i is connected to three clusters (its own cluster, the one above it, and the one to the left of it) while unit j is only connected to one cluster since it is in the center of its own cluster. The impacts of clustering on interference for these two units are asymmetric because they are located at different (relative) positions within the cluster. In Figure 4 with $\boldsymbol{\omega} = (-1, 1)$, the relative positions of units i and j are swapped. By introducing random shaking, the extra layer of randomization ensures that both units i and j are equally likely to be located at any of the 9 relative positions (in the center, in one of the four edges, or in one of the four corners) in a 3×3 cluster. The additional randomization ensures that each unit has the same propensity scores, simplifying the denominator of our estimator in (2).

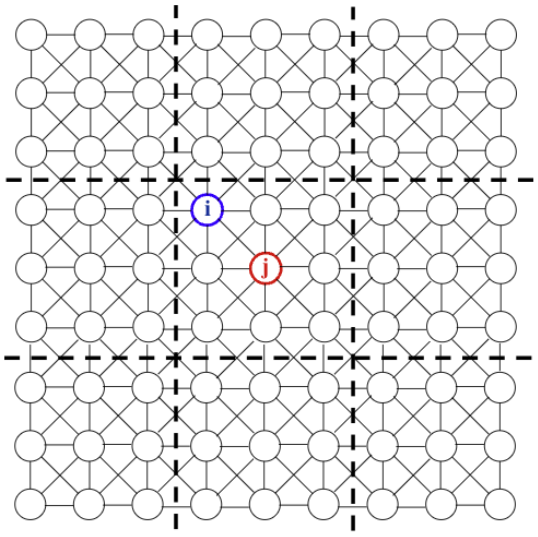


Figure 3 **Random spatial clustering with shaking parameters** $(\omega_1, \omega_2) = (0, 0)$

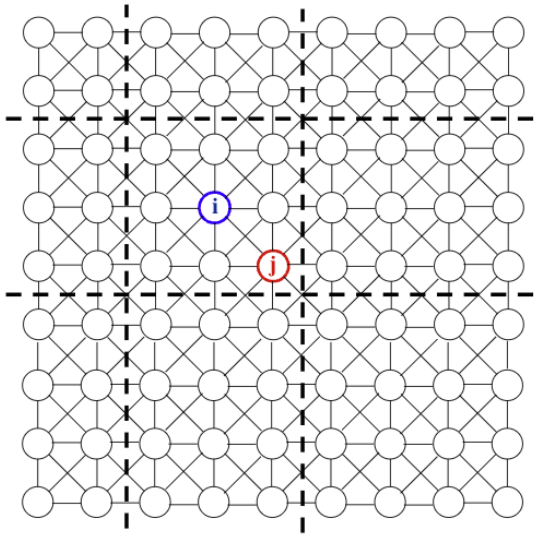


Figure 4 **Random spatial clustering with shaking parameters** $(\omega_1, \omega_2) = (-1, 1)$

Table 1 **An example of assignment path realizations from the balanced design $\eta^\dagger$ with $T = 9$.**

Periods	1	2	3	4	5	6	7	8	9
Assignment Path 1	1	1	0	0	0	1	1	0	1
Assignment Path 2	0	1	1	1	0	0	1	0	0

4.2. Temporal Balancing

After clustering, we run independent switchback experiments for each cluster over T periods. Suppose T is odd, we propose using the following balanced design of switchback experiments:

1. Conduct a complete randomization with an equal number of treatments and controls over the first $T-1$ periods, i.e., there are exactly $(T-1)/2$ treated periods and $(T-1)/2$ control periods.
2. Set the treatment assignment of the last period to be the same as that of the first period, i.e., $W_T = W_1$.

The first step ensures balance, while the second step deals with the boundary conditions of the design and significantly simplifies the subsequent analysis. Table 1 provides two examples of possible assignment path realizations with $T = 9$. We will provide a formal analysis of our balanced design in Section 5.

5. Analysis of Temporal Balancing

We now analyze the design of temporal balancing for switchback experiments conducted in the second step. We will then extend the analysis in the next section by incorporating randomized spatial clustering. Throughout this section, we assume that we have a single unit and drop the subscript i.

5.1. Variance Decomposition

We begin by characterizing the variance $\mathrm{Var}_{\eta^\dagger}(\widehat{\tau}|\mathbb{Y})$ of our temporal balance design. In the IPW estimator (2), whether the observations from two consecutive periods are valid are strongly correlated due to temporal interference. The assignments in all time periods are also weakly correlated due to the balancing structure. Hence, the joint impact of these two types of correlation complicates the analysis.

THEOREM 1. *For any potential outcomes $\mathbb{Y}$, the variance of the balanced design $\eta^\dagger$ can be decomposed as*

$$\mathrm{Var}_{\eta^\dagger}(\widehat{\tau}|\mathbb{Y}) = \frac{(T-2)}{(T-3)(T-4)}\left(\frac{8(T-4)\left(S^t+S^c\right)}{T-3} - 4S^{ct} + R^{ct} - \frac{2\tau^2}{T-2}\right) \tag{3}$$

where

$$S^t = \frac{\sum_{t=2}^{T}\left(Y_t(\mathbf{1}) + Y_{t+1}(\mathbf{1}) - 2\frac{\sum_{t'=2}^{T} Y_{t'}(\mathbf{1})}{T-1}\right)^2}{4(T-1)}, S^c = \frac{\sum_{t=2}^{T}\left(Y_t(\mathbf{0}) + Y_{t+1}(\mathbf{0}) - 2\frac{\sum_{t'=2}^{T} Y_{t'}(\mathbf{0})}{T-1}\right)^2}{4(T-1)}$$

and

$$S^{ct} = \frac{\sum_{t=2}^{T}\left(\tau_t + \tau_{t+1} - 2\tau\right)^2}{4(T-1)}, R^{ct} = \frac{\sum_{t=2}^{T}\tau_t^2}{(T-1)}.$$

We prove Theorem 1 in Appendix C.1.The expression above generalizes the variance decomposition of the classical completely randomized design (Li and Ding 2017) by bridging the outcomes at periods t and $t+1$. In particular, S^t and S^c refer to the variance of the average treatment and control outcomes in every two consecutive periods, respectively, while S^{ct} refers to the variance of average causal effect in every two consecutive periods. In the presence of interference, the number of valid observations is also random, which contributes to another source of variance characterized by R^{ct}. Finally, it leaves a residual term that decays much faster in T.

5.2. Minimax Optimization

To evaluate our proposed design, we adopt the minimax decision rule (Berger 2013, Wu 1981, Li 1983) which entails minimizing the worst-case variance against an adversarial selection of potential outcomes:

$$\min_{\eta\in H}\max_{\mathbb{Y}\in\mathcal{Y}} \mathrm{Var}_\eta(\widehat{\tau}(\mathbb{Y},\eta,\boldsymbol{W})) = \min_{\eta\in H}\max_{\mathbb{Y}\in\mathcal{Y}} \mathbb{E}_\eta\left[\left(\widehat{\tau}\left(\mathbb{Y},\eta,\boldsymbol{W}\right) - \tau(\mathbb{Y})\right)^2\right]. \tag{4}$$

One compelling reason for adopting the minimax decision rule is that we do not impose any parametric or structural model on the potential outcomes. Still, to make the decision-making problem feasible, we impose a bounded support assumption.

ASSUMPTION 3 **(Bounded realized potential outcomes)**. *There exists $B>0$, such that for any $i\in[N], t\in[T], \boldsymbol{w}_{1:N,1:T}\in\{0,1\}^{N\times T}, Y_{i,t}(\boldsymbol{w}_{\mathcal{N}(i),1:t})\in[0,B]$. Equivalently, $\mathcal{Y}=[0,B]^{N\times T}$.*

Assumption 3 is typically satisfied in practice as it assumes that the realized potential outcomes are non-negative and upper bounded by some possibly large constant. For example, revenue, sales, and customer engagement metrics all satisfy the non-negative and upper bounded assumptions. Note that the upper bound B does not impose the boundedness on the underlying random data generation process, but instead on the realization. Moreover, our design does not require knowledge of B.

Since all units from the same cluster share the same assignments in a switchback experiment, whenever it is clear from the context, we drop the unit index and let $\boldsymbol{w}_{1:T}$ denote the assignment vector, which we also refer to as an assignment path of a cluster.

5.3. Performance Analysis

For any design η, let $\overline{\mathrm{Var}}_\eta(\widehat{\tau}) = \max_{\mathbb{Y}\in\mathcal{Y}} \mathbb{E}_\eta\left[\left(\widehat{\tau}\left(\mathbb{Y},\eta,\boldsymbol{W}\right)-\tau(\mathbb{Y})\right)^2\right]$ denote the worst case variance. In particular, let $\overline{\mathrm{Var}}_{\eta^*}(\widehat{\tau})$ be the worst-case variance of an optimal design η^* which solves (4). We study the performance of the balanced design $\eta^\dagger$ by comparing its worst-case variance $\overline{\mathrm{Var}}_{\eta^\dagger}(\widehat{\tau})$ against $\overline{\mathrm{Var}}_{\eta^*}(\widehat{\tau})$.

Directly solving the minimax optimization problem analytically is challenging. Nevertheless, we are still able to compare them in the following theorem.

THEOREM 2. *(1) The worst-case outcomes against the balanced design $\eta^\dagger$ can be characterized by*

$$Y_t(\mathbf{1}) = Y_t(\mathbf{0}) = \begin{cases} B & 2 \le t \le (T+1)/2, \\ 0 & (T+3)/2 \le t \le T, \end{cases} \tag{5}$$

which leads to an explicit expression of the worst-case variance

$$\overline{\mathrm{Var}}_{\eta^\dagger}(\widehat{\tau}) = \frac{4(T-2)}{(T-1)(T-3)}B^2.$$

(2) The worst-case variance of the optimal design is lower bounded by

$$\overline{\mathrm{Var}}_{\eta^*}(\widehat{\tau}) \ge \frac{2}{T-3}B^2.$$

We prove Theorem 2 in Appendix C.2 and C.3. In general, there is no closed form of the worst-case outcomes for an arbitrary design η. Fortunately, the worst-case outcomes against the balanced design $\eta^\dagger$ can be characterized explicitly, which leads to explicit expressions of the worst-case variance. Next, we borrow the proof technique from Candogan et al. (2021) to construct a lower bound for the optimal design. Combining both parts, Theorem 2 directly implies a suboptimality gap of the balanced design.

PROPOSITION 1. *The balanced design $\eta^\dagger$ is a 2-approximation design, i.e.,*

$$\overline{\mathrm{Var}}_{\eta^\dagger}(\widehat{\tau}) \le \frac{2(T-2)}{T-1}\,\overline{\mathrm{Var}}_{\eta^*}(\widehat{\tau}) \le 2\,\overline{\mathrm{Var}}_{\eta^*}(\widehat{\tau}). \tag{6}$$

Finally, our balanced design $\eta^\dagger$ improves on previous designs of switchback experiments. In particular, the optimal Bernoulli design η^{Ber} proposed by Bojinov et al. (2022) could possibly generate imbalanced numbers of treatment and control assignments. Compared with the optimal Bernoulli design η^{Ber}, our balanced design has a much smaller variance; that is, $\lim_{T\to\infty} \overline{\mathrm{Var}}_{\eta^\dagger}(\widehat{\tau})/\overline{\mathrm{Var}}_{\eta^{\mathsf{Ber}}}(\widehat{\tau}) = 1/4$. Detailed discussions and simulations are deferred to Appendix A.1.

6. Analysis of Randomized Spatial Clustering

We now analyze the randomized spatial clustering and investigate how different cluster sizes impact performance. Compared to the last section, the variance derivation needs to incorporate the dependence between periods but also between units. To this end, we will first introduce a notion of the degree of dependence and then derive the optimal cluster size.

6.1. Degree of Dependence

Consider the randomized spatial clustering $\mathscr{C}(\boldsymbol{\omega}; d)$. For each unit $i \in [N]$, denote $\Pi_i(\boldsymbol{\omega}; d)$ to be the cluster it belongs to. For any two units i and j, let $K_{i,j}(\boldsymbol{\omega}; d) \in \mathbb{N}$ be the number of clusters that overlaps with the neighborhoods of both units i and j, so that,

$$K_{i,j}(\boldsymbol{\omega}; d) = \left| \left\{ \mathcal{C} \in \mathscr{C}(\boldsymbol{\omega}; d) \middle| \exists i' \in \mathcal{N}(i), j' \in \mathcal{N}(j), \Pi_{i'}(\boldsymbol{\omega}; d) = \Pi_{j'}(\boldsymbol{\omega}; d) = \mathcal{C} \right\} \right|.$$

Qualitatively, $K_{i,j}(\boldsymbol{\omega}; d)$ implies the degree of dependence between the observations at unit i and j. If $K_{i,j}(\boldsymbol{\omega}; d) = 0$, then the observations at these two units are independent; if $K_{i,j}(\boldsymbol{\omega}; d)$ is large, then they are strongly dependent. Note that $K_{i,j}(\boldsymbol{\omega}; d)$ is a random variable due to the random shaking $\boldsymbol{\omega}$. The randomized spatial clustering effectively determines the distribution of $K_{i,j}(\boldsymbol{\omega}; d)$ that characterizes the dependence between units. Along with the dependence between periods, we derive a variance decomposition in the panel setting with randomized spatial clustering, extending the case of Theorem 1 in switchback experiments; see Appendix A.3 for details.

Note that the degree of dependence $K_{i,j}(\boldsymbol{\omega}; d)$ under $i = j$ reduces to the number of clusters that neighborhood units $\mathcal{N}(i)$ are involved, which is reflected in the magnitude of the propensity score. We visualize the benefits of randomized spatial clustering in Figure 5 by comparing the propensity scores of different units. On the left picture, we evaluate $K_{i,i}(\boldsymbol{\omega}; d)$ for all $i \in [N]$ under the deterministic spatial clustering, where we clearly see the asymmetry between units. On the right picture, $\mathbb{E}_{\boldsymbol{\omega}}[K_{i,i}(\boldsymbol{\omega}; d)]$ is mostly the same between units, except for some boundary ones.

6.2. Optimal Cluster Size

We now study how randomized spatial clustering affects the variance of the IPW estimator using our design and derive the optimal cluster size d. Intuitively, there is a tension between favoring a smaller or larger d. A larger cluster size will ensure that more units receive the same treatment

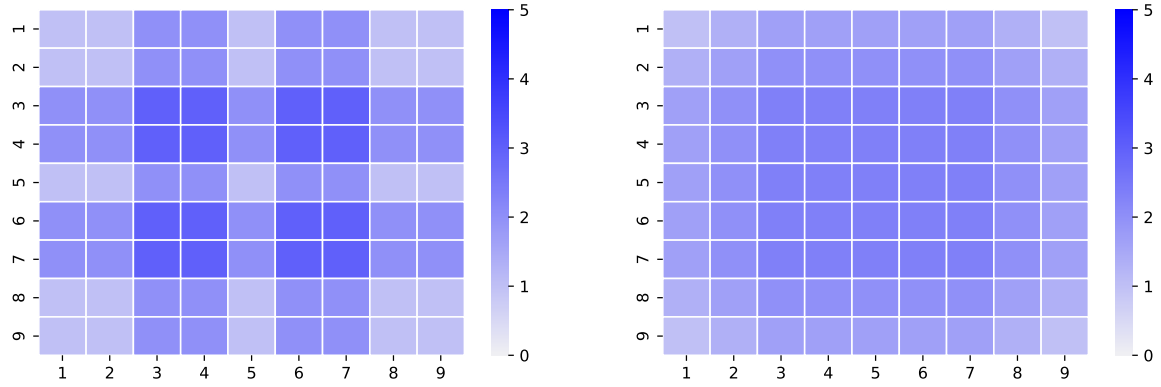


Figure 5 The value of $K_{i,i}(\mathscr{C}(\boldsymbol{\omega};3))$ under cluster size $d=3$ under the naive deterministic clustering (left) and the randomized spatial clustering (right).

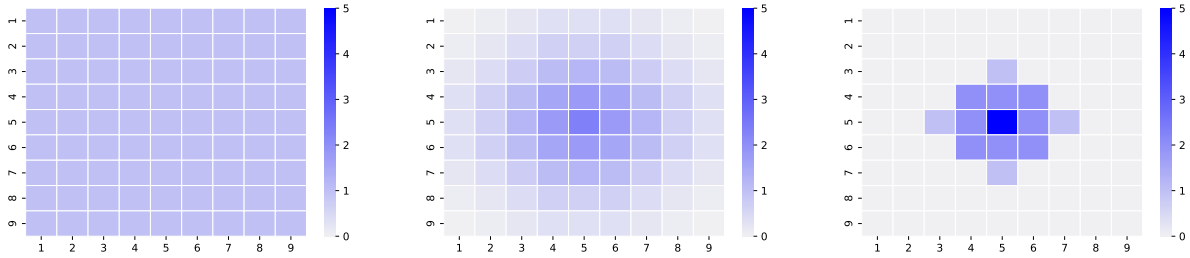


Figure 6 The value of $\mathbb{E}_{\boldsymbol{\omega}}[K_{i,j}(\boldsymbol{\omega};d)]$ when we fix $i=(5,5)$ and move $j\in[D]^2$, under different cluster sizes $d=9$ (left), $d=3$ (middle), and $d=1$ (right).

assignments but increases the pair of units that are dependent. On the other hand, a smaller cluster size produces more units contaminated by spatial interference and decreases the number of valid observations but increases the pair of units that are independent.

Consider a concrete example of $81=9\times 9$ units shown in Figure 6, where we fix $i=(5,5)$ and move $j\in[D]^2$ to evaluate $\mathbb{E}_{\boldsymbol{\omega}}[K_{i,j}(\boldsymbol{\omega};d)]$ — larger value indicates more dependence between units i and j. The left figure shows that when $d=9$, there is a weak global dependence. The right figure demonstrates that when $d=1$, there is a strong local dependence with no global dependence. The middle figure shows that when $d=3$, there is moderate local and global dependence, balancing the number of dependent units and the size of correlations; as our theory below demonstrates, this leads to the optimal design.

With a slight reload of notation, let $\eta_d^\dagger$ denote the design of panel experiments with randomized spatial clustering and temporal balancing using size d. We find the optimal cluster size by referring to the minimax optimization again to study the design $\eta_d^\dagger$ under the worst-case variance. To avoid the boundary cases, we consider a scaling regime when $N\to\infty$ and $T\to\infty$. That is, for any fixed

Table 2 **Worst-case variances and threshold parameters associated with different cluster size d**

d	1	2	3	4	5	6	7	8	D
$\overline{\mathrm{Var}}_{\eta_d^\dagger}(\widehat{\tau}) \cdot NT$	562224	2304	**242.2**	243.4	283.8	340.1	493.6	362.4	$4N$
α_d	1	1	0.620	0.573	0.545	0.523	0.516	0.512	0.5

cluster size d, we evaluate the worst-case variance

$$\overline{\mathrm{Var}}_{\eta_d^\dagger}(\widehat{\tau}) = \max_{\mathbb{Y}\in\mathcal{Y}} \lim_{N\to\infty,T\to\infty} \mathbb{E}_{\eta_d^\dagger}\left[\left(\widehat{\tau}\left(\mathbb{Y}, \eta_d^\dagger, \boldsymbol{W}\right) - \tau(\mathbb{Y})\right)^2\right]. \tag{7}$$

Comparing the worst-case variances with different cluster sizes, we find the optimal one $d^* = \arg\min_d \overline{\mathrm{Var}}_{\eta_d^\dagger}(\widehat{\tau})$. Similar to the analysis in Section 5, it is necessary to characterize the worst-case outcomes, which depend on the cluster size d. Nonetheless, there always exists the same structure that characterizes the worst-case outcomes.

THEOREM 3. *As $N \to \infty$ and $T \to \infty$, the worst-case outcomes against the randomized spatial clustering $\eta_d^\dagger$ under any cluster size d can be characterized by*

$$Y_{i,t}(\mathbf{1}) = Y_{i,t}(\mathbf{0}) = \begin{cases} B & 2 \le t \le \alpha_d T \\ 0 & t > \alpha_d T \end{cases}, \forall i \in [N]. \tag{8}$$

for some constant $\alpha_d \in [0,1]$. This further leads to the asymptotically worst-case variance

$$\overline{\mathrm{Var}}_{\eta_d^\dagger}(\widehat{\tau}) = \frac{\alpha_d B^2}{N^2 T} \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\mathbb{E}[4^{K_{i,j}(\boldsymbol{\omega};d)-K_{j,j}(\boldsymbol{\omega};d)}(2 + 4\cdot 2^{-K_{i,j}(\boldsymbol{\omega};d)} - \alpha_d K_{i,j}(\boldsymbol{\omega};d) 4^{2-K_{i,j}(\boldsymbol{\omega};d)} - 6\mathbb{1}\{K_{i,j}(\boldsymbol{\omega};d)=0\})]}{\mathbb{E}[4^{-K_{j,j}(\boldsymbol{\omega};d)}]}. \tag{9}$$

In Theorem 3, we refer to α_d as a threshold parameter, which does not depend on i, N, or T. The threshold parameter is decreasing in d, see Appendix C.5 for the closed-form expression of α_d and the proof of Theorem 3. The threshold parameter α_d balances the covariance between units and the covariance between periods. When d is small, the positive covariance between units dominates so the worst-case scenario sets more outcomes to be B; when d gets larger, the negative covariance between periods becomes more significant so the worst-case scenario tends to decrease the number of units with outcome B.

Although we derive the worst-case variance (9), we are unable to obtain a simplified closed-form expression because the distribution of $K_{i,j}(\boldsymbol{\omega};d)$ depends on the geographical relations between units i and j. Nevertheless, we can precisely evaluate the worst-case variance with different cluster size d. In Table 2, we tabulate the worst-case variances (normalized by NT) associated with different randomized spatial clustering $\eta_d^\dagger$ and the corresponding threshold parameters α_d. Here we find the optimal cluster size d^*.

PROPOSITION 2. *As $N \to \infty$ and $T \to \infty$, the optimal cluster size is $d^* = 3$, i.e., $d = 3$ gives the smallest objective value in* (7).

First, given that $d = 3$ leads to the worst-case variance, its gap to the second best one $d = 4$ is small. Indeed, we observe from the simulations in Section 8 that these two clustering sizes perform very close in many scenarios with general outcome models, so they are both good choices in practice. Second, we consider the scaling regime for ease of analysis. Although we assume the same structure of neighborhood interference for all units and apply the extra layer of random shaking, a unit in the interior is always involved in more interference than a unit on the boundary due to the geographical nature of units (see Figure 5). This significantly complicates the dependence structure between units and, consequently, the analysis of the worst-case variance. In the scaling regime when $N \to \infty$, the units on the boundary contribute little to the total variance and thus it is sufficient to focus on the interior where units are symmetric. That being said, as the simulations suggest in Section 8, the optimal cluster size $d^* = 3$ has a superior performance as long as N is reasonably large.

7. Inference and Testing

After running an experiment, we observe the assignment vector and the corresponding observed outcomes. Since in these types of experiments N tends to be much smaller than T, we focus on the regime of fixed sample size with a growing number of time periods. Specifically, we derive a central limit theorem (as $T \to \infty$) for the IPW estimator under our design and derive a (conservative) estimator for the variance of the IPW estimator.

We consider the following null hypothesis of no average treatment effect:

$$H_0 : \frac{1}{N(T-1)} \sum_{t=2}^{T} \sum_{i=1}^{N} \left[Y_{i,t}(\mathbf{1}) - Y_{i,t}(\mathbf{0}) \right] = 0. \tag{10}$$

We consider the case with a fixed number of experimental units and rely on variations introduced by random assignments across time periods to make inferences, i.e., we consider a case when N is fixed and $T \to \infty$. To highlight the impact of time periods T on inference, here we present the central limit theorem only for the $N = 1$ case and leave the extension for $N > 1$ cases in Appendix A.3.

THEOREM 4. *Under Assumptions 1 - 3 and if* $\mathrm{Var}_{\eta^\dagger}(\widehat{\tau}|\mathbb{Y}) = \Omega(1/T)$*, the limiting distribution of the IPW estimator is a normal distribution, i.e., as $T \to \infty$,*

$$\frac{\widehat{\tau} - \tau}{\sqrt{\mathrm{Var}_{\eta^\dagger}(\widehat{\tau}|\mathbb{Y})}} \xrightarrow{D} \mathcal{N}(0,1). \tag{11}$$

This gives the nominal coverage that $\lim_{T\to\infty} \mathbb{P}(|\widehat{\tau} - \tau| / \sqrt{\mathrm{Var}_{\eta^\dagger}(\widehat{\tau}|\mathbb{Y})} \geq c(\alpha)) = \alpha$*, where α is the nominal level and $c(\alpha)$ is the critical value of the normal distribution.*

We prove Theorem 4 in Appendix C.6. In Theorem 4, the source of randomness comes from complete randomization, instead of Bernoulli randomization. Central limit theorems of this type have been developed in Li and Ding (2017) to handle multiple treatments and multi-dimension outcomes, yet when there is no interference. With temporal interference, the number of valid observations under treatment and control is no longer a fixed quantity. The random nature of the number of valid observations requires new proof techniques. We adopt the framework of the double-index permutation statistics (Daniels 1944, Hoeffding 1951, Zhao et al. 1997, Reinert and Röllin 2009) to analyze the behavior of the random permutations in the balanced design.

In Theorem 1, the variance of the IPW estimator as shown in (3) involves both the potential outcomes under treatment $Y_t(\mathbf{1})$ and the potential outcomes under control $Y_t(\mathbf{0})$, at all time periods $t \in [T]$. Since we never observe all the potential outcomes, the variance of the IPW estimator can not be directly estimated from the data. As an alternative, we derive an estimable conservative estimator for the variance of the IPW estimator.

PROPOSITION 3. *There exists an upper bound for the variance of the balanced design, i.e.,*

$$\mathrm{Var}_{\eta^\dagger}(\widehat{\tau}|\mathbb{Y}) \leq \frac{(T-2)}{(T-3)(T-4)} \left(\frac{8(T-4)\left(S^t + S^c\right)}{T-3} + \frac{\sum_{t=2}^{T} Y_t^2(\mathbf{1}) + Y_t^2(\mathbf{0})}{(T-1)} \right) \tag{12}$$

where S^t and S^c are defined in Theorem 1. This upper bound $\mathrm{Var}^{\mathrm{U}}_{\eta^\dagger}(\widehat{\tau}|\mathbb{Y})$ can be unbiasedly estimated by

$$\hat{\sigma}^2_{\mathrm{U}} = \frac{(T-2)}{(T-3)^2(T-4)} \left(8(T-4)\left(\widehat{S}^t + \widehat{S}^c\right) + \frac{4(T-2)\sum_{t=2}^{T} Y_t^2 \mathbb{1}\{w_{t-1} = w_t\}}{(T-1)} \right) \tag{13}$$

where

$$\widehat{S}^t = \frac{\sum_{t=2}^{T} \left(Y_{t-1} + Y_t - 2\frac{\sum_{t'=2}^{T} Y_t \cdot \mathbb{1}\{\boldsymbol{w}_{t'-1:t'} = \mathbf{1}\}}{\sum_{t'=2}^{T} \mathbb{1}\{\boldsymbol{w}_{t'-1:t'} = \mathbf{1}\}} \right)^2 \cdot \mathbb{1}\{\boldsymbol{w}_{t-2:t} = \mathbf{1}\}}{4\left(\sum_{t=2}^{T} \mathbb{1}\{\boldsymbol{w}_{t-2:t} = \mathbf{1}\} - 1 \right)}$$

is the sample estimate for S^t, and $\widehat{S}^c$ is defined similarly by replacing $\mathbf{1}$ with $\mathbf{0}$.

We prove Proposition 3 in Appendix C.7. Compared to the true variance $\mathrm{Var}_{\eta^\dagger}(\widehat{\tau}|\mathbb{Y})$, the upper bound is generally not tight. As $T \to \infty$, the coverage probability using this conservative estimator is larger than the nominal value. Nevertheless, we will show in simulations that this conservative estimator admits a good power for testing the null hypothesis.

8. Simulation Study

In this section, we describe an extensive simulation study to analyze the performance and robustness of our design. We first demonstrate that $d=3$ leads to the lowest variance across several realistic scenarios and then show how our method has substantially higher empirical statistical power. To start with, we consider a stochastic data generating process as given by

$$Y_{i,t}(\boldsymbol{w}_{\mathcal{N}(i),t-1:t}) = \alpha_{i,t} + \beta_{i,0} w_{i,t} + \beta_{i,1} \sum_{j \in \mathcal{N}(i)} w_{j,t-1} + \beta_{i,2} \sum_{j \in \mathcal{N}(i) \setminus i} w_{j,t} + \epsilon_{i,t} \tag{14}$$

where $\epsilon_{i,t} \sim \mathcal{N}(0,1)$. In (14), $\beta_{i,0}$ governs the magnitude of the direct effect, $\beta_{i,1}$ governs the magnitude of the temporal interference, and $\beta_{i,2}$ governs the magnitudes of the spatial interference.

8.1. Comparison of Variance Minimization

In our theoretical analysis, we adopted a minimax framework, meaning that our design performs well in worst-case scenarios. Here, we investigate the average performance for different values of d and show that our proposed design performs well in a range of correctly specified and misspecified problems.

Correctly specified interference structure. We set $D=12$ with $N=144$ units in total and $T=201$. We allow the parameters $\alpha_{i,t}, \beta_{i,0}, \beta_{i,1}, \beta_{i,2}$ to be heterogeneous between units. In particular, we set $\alpha_{i,t} = \rho_i(1+\sin(\pi t/4))$ where $\rho_i \sim U(0.5, 1.5)$. We also sample $\beta_{i,0}, \beta_{i,1}, \beta_{i,2}$ from different uniform distributions, allowing us to generate different outcome scenarios to evaluate the average variance; see Table 3, columns 1-3 for the specification. We conducted the experiments using our cluster-based design with different cluster sizes and summarized the average variance over 10000 instances.

Table 3 shows that the cluster size $d=3$ has the smallest variance in all scenarios, which is consistent with our worst-case analysis. The coefficients we set from top to bottom reflect the cases of (almost) no effect (the first row), weak effect (the middle two rows), and substantial effect (the last row). We also observe that the gap between $d=3$ and $d=4$ shrinks as the effect grows more significant. Recall that in our analysis of the worst-case scenario, the potential outcomes imply no effect, so the optimality of $d=3$ diminishes when the true outcome model suggests a strong effect. In fact, as discussed in Bojinov and Gupta (2022), the effect of online experimentation has a light tail, so in most cases, the treatment has a weak effect.

In addition, we also simulate the performance of the fixed clustering without the random shaking. Our results show that the corresponding variances can be around 7 times as large as those in the randomized spatial clustering, which shows that the benefit of additional randomization is very significant. More interestingly, the best cluster size under the fixed clustering is $d=6$, which implies that random shaking admits much smaller cluster sizes and more clusters.

Table 3 **Variances of the cluster-based design under different outcome models. The first three columns specify the distribution that the β coefficients are samples from. The fourth column indicates if the clustering leveraged our random shaking (*Random*) or not (*Fixed*). Switchback is equivalent to grouping all units and setting $d = 12$.**

$\beta_{i,0}$	$\beta_{i,1}$	$\beta_{i,2}$	Clustering	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$	Switchback
$U(0,0.8)$	$U(-0.05,0)$	$U(-0.05,0)$	Random	0.0878	**0.0058**	0.0075	0.0096	0.0140	0.0340
			Fixed	-	0.0471	0.0301	0.0295	0.0234	-
$U(0,0.8)$	$U(-0.05,0)$	$U(-0.02,0)$	Random	0.1088	**0.0065**	0.0081	0.0108	0.0142	0.0344
			Fixed	-	0.0506	0.0321	0.0305	0.0239	-
$U(0,0.8)$	$U(-0.02,0)$	$U(-0.05,0)$	Random	0.1085	**0.007**	0.0085	0.0101	0.0143	0.0347
			Fixed	-	0.0527	0.0334	0.0321	0.0242	-
$U(0,0.8)$	$U(-0.02,0)$	$U(-0.02,0)$	Random	0.1301	**0.0078**	0.0086	0.0125	0.0153	0.0350
			Fixed	-	0.0623	0.0367	0.0344	0.0256	-

Table 4 **Variances of our design with different degrees of remote interference (uniform)**

γ	$\frac{\text{Neighborhood}}{\text{Remote}}$	$d=2$	$d=3$	$d=4$	$d=5$	Switchback
20	24	0.1422	**0.0115**	0.0145	0.0167	0.0344
40	12	0.1831	**0.0208**	0.0223	0.0235	0.0344
60	8	0.2682	**0.0271**	0.0291	0.0342	0.0344
80	6	0.3710	0.0402	0.0401	0.0460	**0.0344**
100	4.8	0.5296	0.0514	0.0490	0.0603	**0.0344**

Performance under more general interference. We now consider a setting where our assumption of neighborhood interference is wrong. Specifically, we investigate how our design performs if, in addition to neighborhood interference, some remote regions may interfere with each other due to underlying phenomena; for example, drivers may regularly drop off customers at an airport far away from their pick-up location. To incorporate this, we make the following adjustments to the original interference network: pick any two remote units at uniformly random and add an edge between them, indicating remote interference. We repeat this procedure until we have added γ remote edges.

In Table 4, we vary γ to reflect different density of remote interference. When conducting the experiments, we update the knowledge about the network so the estimation remains unbiased, but we stick to the same design of experiments and evaluate the performance accordingly. We observe that the cluster size $d = 3$ stays to be the best for $\gamma \leq 60$. Note that when $\gamma = 60$, on average, each local region has a remote edge connected to some non-neighborhood region, which means the neighborhood interference assumption breaks almost everywhere. This shows that our new design is quite robust against deviations to our neighborhood interference structure. When $\gamma \geq 80$, the performance of $d = 3$ becomes worse than that of $d = 4$, but in these cases, all cluster sizes are outperformed by a single switchback design ($d = D$). This implies that the remote interference is so substantial that our randomized spatial clustering does not add value regardless the cluster size.

Table 5 Variances of our design with different degrees of remote interference (non-uniform)

γ	$\frac{\text{Neighborhood}}{\text{Remote}}$	Merge	$d=2$	$d=3$	$d=4$	$d=5$	Switchback
100	4.8	No	2.403(36)	0.0462(16)	0.0278(9)	0.0245(9)	0.0344
		Yes	0.0478(16)	**0.0124**(11)	0.0142(7)	0.0192(7)	-
200	2.4	No	4.231(36)	0.2288(16)	0.0455(9)	0.0478(9)	0.0344
		Yes	0.0439(8)	**0.0212**(9)	0.0256(6)	0.0294(6)	-
300	1.6	No	8.821(36)	0.4129(16)	0.1197(9)	0.0837(9)	0.0344
		Yes	0.0401(2)	**0.0280**(7)	0.0329(5)	0.0345(5)	-
400	1.2	No	19.23(16)	0.8643(16)	0.1615(9)	0.1421(9)	**0.0344**
		Yes	0.0377(1.4)	0.0355(5)	0.0357(4)	0.0356(4)	-
500	1.0	No	115.2(36)	1.351(16)	0.3042(9)	0.2020(9)	**0.0344**
		Yes	0.0348(1.2)	0.0364(3)	0.0360(2)	0.0366(2)	-

In the above simulation, we added edges uniformly at random. In some settings, however, there could be an underlying structure. For instance, there might be a traveling pattern between a business district (contains n_1 units) and a residential district (contains n_2 units), so we should expect $n_1 \times n_2$ edges capturing this remote interference. To this end, we adjust our simulation procedure as follows. We first pick any two remote units randomly as before (e.g., coordinates (i_1, i_2) and (j_1, j_2)), and then construct remote edges between the areas of 2 by 2 units containing (i_1, i_2) and (j_1, j_2). For example, an edge connects $(i_1 - 1, i_2)$ and $(j_1, j_2 - 1)$. Thus, we add 16 ($n_1 = n_2 = 4$) remote edges in this round. We finally iterate over many rounds till reaching γ edges.

In Table 5, we present the results by varying γ from 200 to 1000. We first keep using our proposed design $\eta_d^\dagger$ with different cluster sizes (the first row in each case). Compared to the performance in Table 4, the cluster-based designs are more tolerant to the contamination of remote interference. For example, when $\gamma = 100$, $d = 4$ and $d = 5$ still perform better than a single switchback experiment, while this is not the case in Table 4. The intuition behind this observation is that when remote edges are more concentrated, our clustering approach aggregates local units, and thus, the negative effect of remote interference is naturally offset. That being said, the performance quickly deteriorates as $\gamma \geq 200$ under all cluster sizes, especially for small ones, because the remote interference is so strong that local clustering is insufficient.

To overcome this limitation, we develop a heuristic algorithm that further improves our design by leveraging the interference structure to obtain better clustering. We outline the key idea of the algorithm and leave the details to Appendix B. Given two clusters of local regions that are farther away, their treatment assignments should be sampled independently following our design. Suppose that many regions from these two remote clusters are interfering with each other; we may consider merging these two clusters as a single one so that they share the same treatment assignments. On the one hand, this absorbs the remote interference between two clusters and thus reduces the propensity

score of units; on the other, this introduces excess dependence between units that are not interfering. Our heuristic algorithm carefully examines this tension and makes informed merging decisions.

Table 5, also show the performance using our heuristic algorithm for merging clusters (the second row in each case). We observe a sharp improvement upon our original design, especially for large γ, which justifies the idea behind our simple greedy heuristic. Up to $\gamma = 300$, the performance of $d = 3$ is consistently the best as it strikes a good balance between maintaining sufficient number of clusters to generate more data and absorbing sufficient amount of remote interference to reduce high propensity scores, while all cluster sizes perform similarly as γ gets larger because the number of clusters is small and the difference among clustering is minimal. More importantly, it suggests that choosing our randomized spatial clustering with proper size not only performs well under the neighborhood interference assumption, but also lays a good foundation for more sophisticated designs when the assumption does not hold.

Performance under real networks. To further test our experimental design along with its extension to the heuristic algorithm, we consider a real network structure abstracted from the data of a ride-sharing platform in Singapore. We use a collection of one-month trips to characterize the network structure. We illustrate the geohash map in Figure 7 on a 31×20 lattice. Since we are only interested in the service regions where the platform operates, only 255 local squares involved in the trip data are defined to be valid units.

For each trip record in the data, we have the precise pickup and dropoff locations so that we know exactly the local regions at which the trip starts and ends. For example, we take one CBD region as the pickup region and visualize the density of its trips in Figure 8. We observe that a significant portion of the trips is directed towards neighboring regions, a consideration already integrated through our neighborhood interference assumption. For the remaining trips, their destinations mainly cover the airport (on the right side of the map) as well as some other commercial and residential areas.

To set up the network of spatial interference, besides the edges indicating all neighborhood interference, we construct the remote interference by adding top γ edges that connect two regions with the most frequent trips. Similar to the simulation above, we tabulate the performance of different designs by vary γ in Table 6. It is not surprising to see that the variances without merging explode with a few remote edges, and our heuristic algorithm substantially alleviates the issue for different cluster sizes. The algorithm is also robust against large γ, when the number of clusters shrinks and the performance gradually approaches the switchback one. More importantly, this sheds light on the bias-variance trade-off from the interference structure. On the one hand, remote interference (e.g., trips) could exist between any two regions and thus no clustering could reduce the variance on a

Figure 7 The geohash map of Singapore.

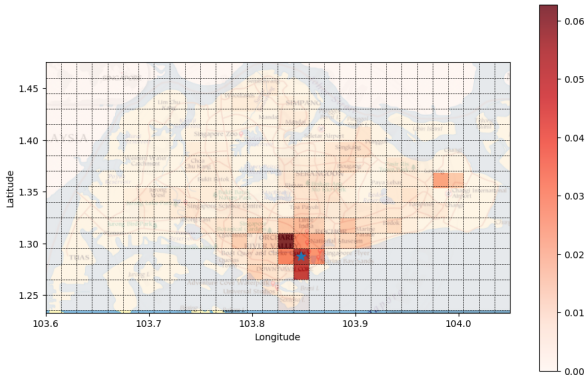


Figure 8 The density of trips that start at one CBD region (the blue star).

Table 6 Variances of our design with different degrees of remote interference (real)

γ	$\frac{\text{Neighborhood}}{\text{Remote}}$	Merge	$d=3$	$d=4$	$d=5$	Switchback
0	-	No	**0.0033**(42)	0.0034(27)	0.0044(19)	0.0344
200	5	No	0.3212(42)	0.1857(27)	0.0637(19)	0.0344
		Yes	**0.0096**(23)	0.0109(17)	0.0128(14)	
400	2.5	No	2.412(42)	0.9831(27)	0.1934(19)	0.0344
		Yes	**0.0126**(17)	0.0143(14)	0.0155(12)	
800	1.25	Yes	**0.0195**(15)	0.0196(12)	0.0201(9)	0.0344
1600	0.62	Yes	**0.0238**(12)	0.0247(11)	0.0248(8)	0.0344
3200	0.31	Yes	**0.0271**(10)	0.0283(9)	0.0304(6)	0.0344
6400	0.15	Yes	**0.0320**(7)	0.0322(5)	0.0324(3)	0.0344

Note: the performances without merging under $\gamma = 400$ are significantly worse than a single switchback design and the numbers grow exponentially, so we only present the performance with merging under large γ for clarity.

fully-connected network. On the other hand, ignoring all remote interference leads to the minimal variance, but at the price of a potentially large bias. While a theoretical analysis of this trade-off is beyond the scope of this paper, our new design with the heuristic algorithm essentially provides a spectrum of solutions on the bias-variance frontier.

To conclude, we show in this section the superior performance of our new experimental design along with its extended heuristic approach in different network settings. We demonstrate when the cluster size $d = 3$ remains a good choice and when we should resort to a single switchback experiment with clustering.

8.2. Comparison of Testing Efficiency

We now focus on exploring our ability to draw meaningful inferences after running an experiment with our proposed design.

We first justify the normal approximation following the same outcome model (14) with $T = 401$. More specifically, we generate 100000 samples of the estimator $\widehat{\tau}$ and conduct a Kolmogorov–Smirnov test (Sprent 2012) for the null hypothesis that the samples come from a normal distribution. The test returns an estimated p-value of 0.57, implying the normal approximation is reasonable. Figure 9 shows the histogram and the Q-Q plot that correspond to the distribution induced by $\frac{\widehat{\tau}-\tau}{\sqrt{\mathrm{Var}_{\eta_3^\dagger}(\widehat{\tau}|\mathbb{Y})}}$, for which we numerically compute $\mathrm{Var}_{\eta_3^\dagger}(\widehat{\tau}|\mathbb{Y})$ using samples from the simulation.

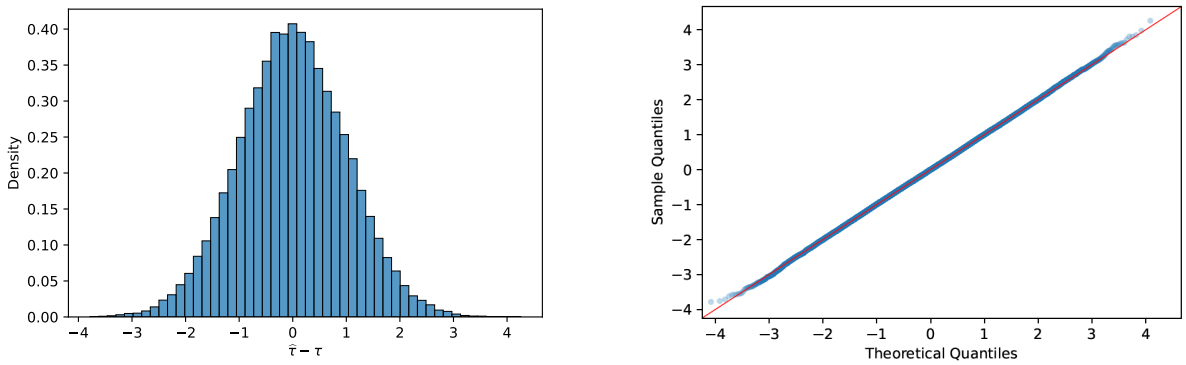
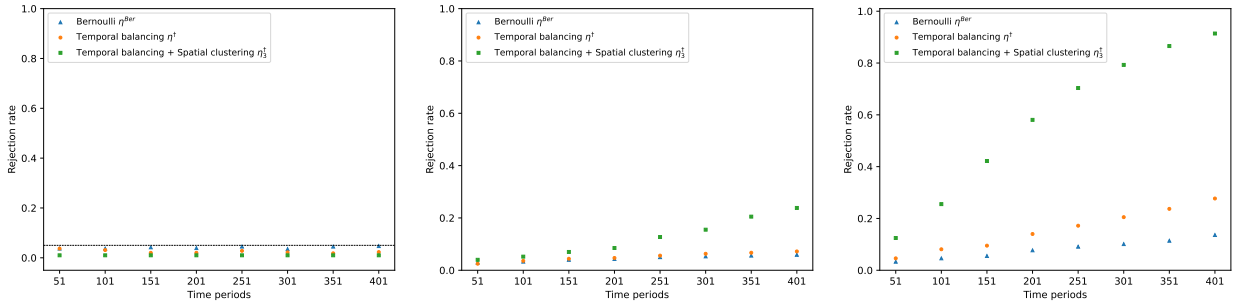


Figure 9 **Normal approximation of $\widehat{\tau}-\tau$ using 100000 samples under the outcomes given in** (14) **with** $\alpha_{i,t} = (1+\sin(\pi t/4))\cdot U(0.5,1.5), \beta_{i,0} = 1, \beta_{i,1} = -0.05, \beta_{i,2} = -0.05, T = 401$.

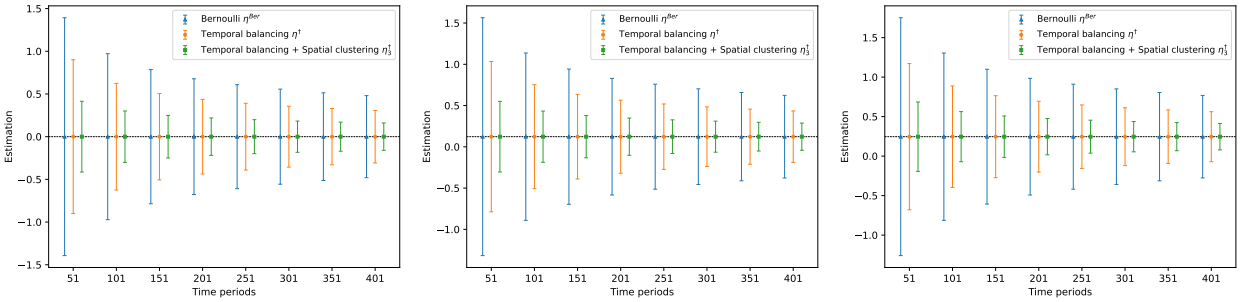
We then examine the effectiveness of inferences using different designs. In particular, we test the null hypothesis (10) using the normal approximation. We not only calculate the IPW estimator (2) based on the observed outcomes, but also estimate the variance by (13) for constructing normal approximation, which further gives an estimated p-value $\hat{p}$. We reject the null hypothesis if $\hat{p} < 0.05$. By repeating this procedure 10000 times, we summarize the frequency of a null hypothesis that is rejected (i.e. rejection rate).

We present the rejection rates as the number of periods T grows for three different designs of experiments: (1) the optimal Bernoulli design η^{Ber} of switchback experiments; (2) the switchback experiments $\eta^\dagger$ with temporal balancing; (3) our design of panel experiments $\eta_3^\dagger$ with temporal balancing and randomized spatial clustering under $d^* = 3$. We compare these designs under three outcome models in Figure 10. First, we set $\beta_{i,0} = \beta_{i,1} = \beta_{i,2} = 0$ such that the true causal effect is $\tau = 0$. All three designs have small rejection rates slightly below the nominal level 0.05. Second, we set $\beta_{i,0} = 0.5, \beta_{i,1} = \beta_{i,2} = -0.025$ which lead to a weak positive causal effect. In this case, the Bernoulli design η^{Ber} almost fails to reject the null. Adding the temporal balancing increases the rejection rate, allowing more chances to detect the causal effect. Adding the randomized spatial clustering further improves the efficiency considerably; Third, we scale the coefficients by 2 with $\beta_{i,0} = 1, \beta_{i,1} = \beta_{i,2} = -0.05$ which lead to a stronger positive causal effect. Similar to before, both



$\beta_{i,0} = \beta_{i,1} = \beta_{i,2} = 0$ $\beta_{i,0} = 0.5, \beta_{i,1} = \beta_{i,2} = -0.025$ $\beta_{i,0} = 1, \beta_{i,1} = \beta_{i,2} = -0.05$

Figure 10 **Rejection rates for testing the null hypothesis** (10)



$\beta_{i,0} = \beta_{i,1} = \beta_{i,2} = 0$ $\beta_{i,0} = 0.5, \beta_{i,1} = \beta_{i,2} = -0.025$ $\beta_{i,0} = 1, \beta_{i,1} = \beta_{i,2} = -0.05$

Figure 11 **Point estimates and confidence intervals for testing the null hypothesis** (10)

temporal balancing and randomized spatial clustering help make the right decision more efficiently than the Bernoulli design η^{Ber}. In particular, our design $\eta_3^{\dagger}$ only needs 15-25% as many time periods as in the Bernoulli design η^{Ber} to achieve the same rejection rate.

Furthermore, we plot the average point estimates and the average confidence intervals in Figure 11. We observe that all the estimates are indeed unbiased and our design consistently achieves much narrower confidence intervals. Specifically, in the third case, for getting the average confidence interval above 0, our new design $\eta_3^{\dagger}$ only needs 201 periods while in contrast, the other two designs fail to do so within 401 periods. This further justifies that our design is more data-efficient.

9. Conclusion, Practical Suggestions, and Future Research Directions

In this paper, we study the design of panel experiments in the presence of spatial and temporal interference. Motivated by ride-sharing applications, we model the spatial interference to happen over a grid structure. We adopt a minimax optimization framework and cast the design of experiments problem as a minimax optimization problem. We solve the problem by proposing a combination of balanced design across time periods, and randomized spatial clustering across units. Our proposal of experimental design is simple in nature, easy to implement in practice, and has good theoretical and numerical performances. We also derive a new central limit theorem to make inferences.

We conclude this paper by providing some practical suggestions for conducting experiments on panels and pointing out some limitations of our methods that could lead to future research directions.

First, we consider the well-known unbiased IPW estimator in this paper. In practice, experimenters could use some biased estimators such as Hajek estimator (Basu 2011), which usually significantly reduces the variances. The key features of our design can still be applied as long as the experimenter trades off between the volume of data and the degree of interference. Moreover, one can also investigate the bias-variance trade-off regarding the assumption of the neighborhood spillover effect. The larger an experimental unit is, the less degree of interference it suffers, which leads to less bias, but at the cost of potentially larger variance. Practitioners should define the experimental unit properly using their domain knowledge to strike a good balance in the spectrum of bias-variance trade-off.

Second, we use the potential outcome framework together with non-parametric modeling of the outcomes. If practitioners have good prior knowledge about the outcome model, such as the system dynamics between units, and over time, they may incorporate such modeling to extend our design. For example, if considering a Markovian system for the time series, one may have judicious use of burn-in periods to considerably reduce the variance (Hu and Wager 2022).

Third, we design the experiment when the number of time periods T is fixed and exogenously given. It is interesting to consider an adaptive version in which the experimenter could sequentially monitor the outcomes of the experiment. This opens the room for adjusting the experiment or even stopping the experiment earlier, especially when some domain knowledge guides the modeling of the outcomes (Glynn et al. 2020, Johari et al. 2022a, Ham et al. 2022). We encourage future research to pursue this direction.

References

Abadie A, Athey S, Imbens GW, Wooldridge JM (2020) Sampling-based versus design-based uncertainty in regression analysis. *Econometrica* 88(1):265–296.

Aronow PM, Samii C (2017) Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics* 11(4):1912–1947.

Bajari P, Burdick B, Imbens GW, Masoero L, McQueen J, Richardson T, Rosen IM (2021) Multiple randomization designs. *arXiv preprint arXiv:2112.13495* .

Basse G, Ding Y, Toulis P (2019) Minimax crossover designs. *arXiv preprint arXiv:1908.03531* .

Basse GW, Airoldi EM (2018) Limitations of design-based causal inference and a/b testing under arbitrary and network interference. *Sociological Methodology* 48(1):136–151.

Basu D (2011) An essay on the logical foundations of survey sampling, part one. *Selected Works of Debabrata Basu*, 167–206 (Springer).

Berger JO (2013) *Statistical decision theory and Bayesian analysis* (Springer Science & Business Media).

Bickel PJ, Doksum KA (2015) *Mathematical statistics: basic ideas and selected topics, volume I*, volume 117 (CRC Press).

Bojinov I, Gupta S (2022) Online experimentation: Benefits, operational and methodological challenges, and scaling guide .

Bojinov I, Rambachan A, Shephard N (2021) Panel experiments and dynamic causal effects: A finite population perspective. *Quantitative Economics* 12(4):1171–1196.

Bojinov I, Saint-Jacques G, Tingley M (2020) Avoid the pitfalls of a/b testing make sure your experiments recognize customers' varying needs. *Harvard Business Review* 98(2):48–53.

Bojinov I, Shephard N (2019) Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association* 114(528):1665–1682.

Bojinov I, Simchi-Levi D, Zhao J (2022) Design and analysis of switchback experiments. *Management Science* .

Bright I, Delarue A, Lobel I (2022) Reducing marketplace interference bias via shadow prices. *arXiv preprint arXiv:2205.02274* .

Brown Jr BW (1980) The crossover experiment for clinical trials. *Biometrics* 69–79.

Candogan O, Chen C, Niazadeh R (2021) Near-optimal experimental design for networks: Independent block randomization. *Chicago Booth Research Paper* (21-17).

Chamandy N (2016) Experimentation in a ridesharing marketplace lyft engineering. *URL: https://eng. lyft. com/experimentation-in-a-ridesharing-marketplace-b39db027a66e* .

Chen H, Simchi-Levi D (2023) Switchback experiments in a reactive environment. *Available at SSRN 4436643* .

Cochran WG, Cox GM (1962) Experimental designs .

Cox DR (1958) Planning of experiments. .

Cui R, Gallino S, Moreno A, Zhang DJ (2018) The operational value of social media information. *Production and Operations Management* 27(10):1749–1769.

Daniels HE (1944) The relation between measures of correlation in the universe of sample permutations. *Biometrika* 33(2):129–135.

Dhaouadi W, Johari R, Weintraub GY (2023) Price experimentation and interference in online platforms. *arXiv preprint arXiv:2310.17165* .

Eckles D, Karrer B, Ugander J (2017) Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference* 5(1).

Farias VF, Li AA, Peng T, Zheng AT (2022) Markovian interference in experiments. *arXiv preprint arXiv:2206.02371* .

Farronato C, MacCormack A, Mehta S (2018) Innovation at uber: The launch of express pool. *Harvard Business School Case* 620(062).

Fisher RA (1936) Design of experiments. *Br Med J* 1(3923):554–554.

Glynn PW, Johari R, Rasouli M (2020) Adaptive experimental design with temporal interference: A maximum likelihood approach. *Advances in Neural Information Processing Systems* 33:15054–15064.

Gupta S, Kohavi R, Tang D, Xu Y, Andersen R, Bakshy E, Cardin N, Chandran S, Chen N, Coey D, et al. (2019) Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter* 21(1):20–35.

Ham DW, Bojinov I, Lindon M, Tingley M (2022) Design-based confidence sequences for anytime-valid causal inference. *arXiv preprint arXiv:2210.08639* .

Han K, Basse G, Bojinov I (2024) Population interference in panel experiments. *Journal of Econometrics* 238(1):105565.

Hoeffding W (1951) A Combinatorial Central Limit Theorem. *The Annals of Mathematical Statistics* 22(4):558 – 566, URL `http://dx.doi.org/10.1214/aoms/1177729545`.

Holland PW (1986) Statistics and causal inference. *Journal of the American statistical Association* 81(396):945–960.

Holtz D, Lobel R, Liskovich I, Aral S (2020) Reducing interference bias in online marketplace pricing experiments. *Available at SSRN 3583836* .

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* 47(260):663–685.

Hu Y, Wager S (2022) Switchback experiments under geometric mixing. *arXiv preprint arXiv:2209.00197* .

Hudgens MG, Halloran ME (2008) Toward causal inference with interference. *Journal of the American Statistical Association* 103(482):832–842.

Imbens GW, Rubin DB (2015) *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press).

Jia S, Kallus N, Yu CL (2024) Clustered switchback experiments: Near-optimal rates under spatiotemporal interference.

Johari R, Koomen P, Pekelis L, Walsh D (2022a) Always valid inference: Continuous monitoring of a/b tests. *Operations Research* 70(3):1806–1821.

Johari R, Li H, Liskovich I, Weintraub GY (2022b) Experimental design in two-sided platforms: An analysis of bias. *Management Science* .

Kempthorne O (1955) The randomization theory of experimental inference. *Journal of the American Statistical Association* 50(271):946–967.

Kohavi R, Thomke S (2017) The surprising power of online experiments. *Harvard business review* 95(5):74–82.

Laird NM, Skinner J, Kenward M (1992) An analysis of two-period crossover designs with carry-over effects. *Statistics in Medicine* 11(14-15):1967–1979.

Larsen N, Stallrich J, Sengupta S, Deng A, Kohavi R, Stevens NT (2023) Statistical challenges in online controlled experiments: A review of a/b testing methodology. *The American Statistician* 1–15.

Li H, Zhao G, Johari R, Weintraub GY (2021) Interference, bias, and variance in two-sided marketplace experimentation: Guidance for platforms. *arXiv preprint arXiv:2104.12222* .

Li KC (1983) Minimaxity for randomized designs: some general results. *The Annals of Statistics* 11(1):225–239.

Li S, Johari R, Wager S, Xu K (2023) Experimenting under stochastic congestion. *arXiv preprint arXiv:2302.12093* .

Li X, Ding P (2017) General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association* 112(520):1759–1769.

Li X, Ding P, Lin Q, Yang D, Liu JS (2019) Randomization inference for peer effects. *Journal of the American Statistical Association* .

Manski CF (2013) Identification of treatment response with social interactions. *The Econometrics Journal* 16(1):S1–S23.

Munro E, Wager S, Xu K (2021) Treatment effects in market equilibrium. *arXiv preprint arXiv:2109.11647* .

Neyman J (1923) On the application of probability theory to agricultural experiments. essay on principles. section 9. masters thesis. portion translated by d. dabrowska and t. speed (1990). *Statistical Science* 465–472.

Reinert G, Röllin A (2009) Multivariate normal approximation with stein's method of exchangeable pairs under a general linearity condition. *The Annals of Probability* 37(6):2150–2173.

Robins J (1986) 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.

Rosenbaum PR (2007) Interference between units in randomized experiments. *Journal of the american statistical association* 102(477):191–200.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66(5):688.

Rubin DB (1980) Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association* 75(371):591–593.

Sävje F, Aronow P, Hudgens M (2021) Average treatment effects in the presence of unknown interference. *Annals of statistics* 49(2):673.

Senn S, Lambrou D (1998) Robust and realistic approaches to carry-over. *Statistics in Medicine* 17(24):2849–2864.

Sinclair B, McConnell M, Green DP (2012) Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science* 56(4):1055–1069.

Sprent P (2012) *Applied nonparametric statistical methods* (Springer Science & Business Media).

Tang Y, Huang C, Kastelman D, Bauman J (2020) Control using predictions as covariates in switchback experiments URL `http://dx.doi.org/10.13140/RG.2.2.34500.04488`.

Taylor TA (2018) On-demand service platforms. *Manufacturing & Service Operations Management* 20(4):704–720.

Thomke SH (2003) *Experimentation matters: unlocking the potential of new technologies for innovation* (Harvard Business Press).

Thomke SH (2020) *Experimentation works: The surprising power of business experiments* (Harvard Business Press).

Ugander J, Karrer B, Backstrom L, Kleinberg J (2013) Graph cluster randomization: Network exposure to multiple universes. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 329–337.

Ugander J, Yin H (2020) Randomized graph cluster randomization. *arXiv preprint arXiv:2009.02297* .

Wager S, Xu K (2019) Experimenting in equilibrium. *arXiv preprint arXiv:1903.02124* .

Wu CF (1981) On the robustness and efficiency of some randomized designs. *The Annals of Statistics* 1168–1177.

Xiong R, Athey S, Bayati M, Imbens GW (2019) Optimal experimental design for staggered rollouts. *Available at SSRN 3483934* .

Xiong R, Chin A, Taylor S, Athey S (2022) Bias-variance tradeoffs for designing simultaneous temporal experiments .

Zhao L, Bai Z, Chao CC, Liang WQ (1997) Error bound in a central limit theorem of double-indexed permutation statistics. *The Annals of Statistics* 25(5):2210–2227.

Zhu Z, Cai Z, Zheng L, Si N (2024) Seller-side experiments under interference induced by feedback loops in two-sided platforms.

A. Appendix for Additional Results

A.1. Comparison between Balanced Design and Bernoulli Design

Theorem 2 in Bojinov et al. (2022) shows that the optimal Bernoulli design η^{Ber} has the worst-case variance $\overline{\text{Var}}_{\eta^{\text{Ber}}}(\widehat{\tau}) = \frac{16T-56}{(T-1)^2}B^2$. This leads to the following corollary which compares our balanced design and the optimal Bernoulli design η^{Ber}.

COROLLARY 1. *The relative performance of the worst-case variances between the balanced design $\eta^{\dagger}$ and the optimal Bernoulli design η^{Ber} is given by*

$$\lim_{T\to\infty} \frac{\overline{\text{Var}}_{\eta^{\dagger}}(\widehat{\tau})}{\overline{\text{Var}}_{\eta^{\text{Ber}}}(\widehat{\tau})} = \lim_{T\to\infty} \frac{(T-2)(T-1)}{(4T-14)(T-3)} = \frac{1}{4}.$$

We also conduct a simulation study to investigate the performance of our balanced design $\eta^{\dagger}$. First of all, we set the outcomes $\mathbb{Y}$ to follow the worst-case structure in (5). For each numerical experiment, we randomly sample an assignment path, compute the IPW estimator (2), and repeat the procedure 10000 times to estimate the performance of the design. In Figure 12, the variance of the balanced design is significantly lower than that of the optimal Bernoulli design. Note that both designs are evaluated under the outcomes in (5), which correspond to the worst-case scenario for the balanced design, but do not correspond to the worst-case scenario for the Bernoulli design. This further justifies the robustness of the balanced design.

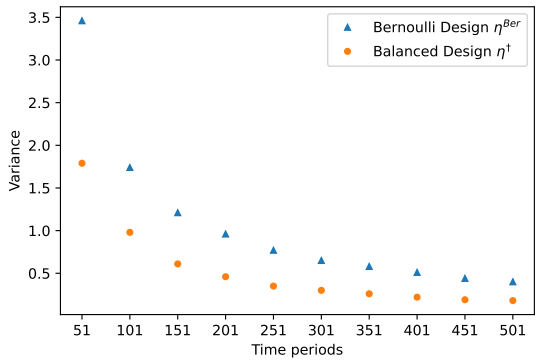


Figure 12 The estimated variance for different experimental periods under the outcomes given in (5) with $B = 3$.

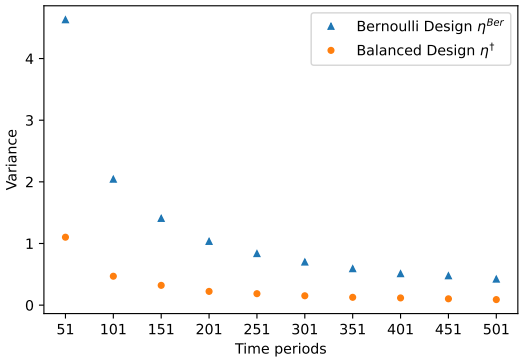


Figure 13 The estimated variance for different experimental periods under the outcomes given in (15) with $\alpha_t = 1 + \sin(\pi t/4), \beta_0 = 1, \beta_1 = 1, \epsilon_t \sim \mathcal{N}(0,1)$.

Next, we consider the following outcome model that is studied in Bojinov et al. (2022):

$$Y_t(\boldsymbol{w}_{t-1:t}) = \alpha_t + \beta_0 w_t + \beta_1 w_{t-1} + \epsilon_t \tag{15}$$

where $\epsilon_t \sim \mathcal{N}(0,1)$. In (15), α_t depicts the base structure of the time series, β_0 governs the direct causal effect of the treatment, and β_1 governs the carryover effect of the treatment. The causal effect

Table 7 **Variances under different outcome models with $T = 101$**

α_t	β_0	β_1	Bernoulli design	Balanced design
$1+\sin(\pi t/4)$	1	1	0.852	0.182(-78.6%)
	1	0	0.533	0.152(-71.5%)
	1	-1	0.314	0.143(-54.4%)
$\log(t)$	1	1	3.706	0.264(-92.8%)
	1	0	2.970	0.233(-92.1%)
	1	-1	2.334	0.223(-90.5%)

of interest is $\tau = \beta_0 + \beta_1$. We first let $\alpha_t = 1 + \sin(\pi t/4), \beta_0 = 1, \beta_1 = 1$. In Figure 13, the balanced design dominates the Bernoulli design. Our new design is an order of magnitude better than the previous Bernoulli style design; for example, the balanced design with $T = 41$ has a lower variance than the Bernoulli design with $T = 181$. This implies that the benefit of the balanced design is also significant beyond the worst-case scenario. Moreover, we test different outcome models by changing the parameters in (15) and lay out the corresponding variances in Table 7.

A.2. Variance Decomposition for Panels

Let us define

$$\widehat{\tau}_i = \frac{1}{T-1}\sum_{t=2}^{T}\left[Y_{i,t}(\mathbf{1})\frac{\mathbb{1}\{\boldsymbol{W}_{\mathcal{N}(i),t-1:t} = \mathbf{1}\}}{\Pr(\boldsymbol{W}_{\mathcal{N}(i),t-1:t} = \mathbf{1})} - Y_{i,t}(\mathbf{0})\frac{\mathbb{1}\{\boldsymbol{W}_{\mathcal{N}(i),t-1:t} = \mathbf{0}\}}{\Pr(\boldsymbol{W}_{\mathcal{N}(i),t-1:t} = \mathbf{0})}\right]. \tag{16}$$

Then the variance could be rewritten as

$$\mathrm{Var}_{\eta_d^\dagger}(\widehat{\tau}|\mathbb{Y}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{Cov}(\widehat{\tau}_i, \widehat{\tau}_j)$$

where $\mathrm{Cov}(\widehat{\tau}_i, \widehat{\tau}_j)$ is the covariance of any two units i and j. Similar to the switchback experiments, we introduce

$$\begin{aligned}
S_{i,j}^{t} &= \frac{1}{T-1}\sum_{t=2}^{T}\left(\frac{Y_{i,t}(\mathbf{1}) + Y_{i,t+1}(\mathbf{1})}{2} - \overline{Y}_i(\mathbf{1})\right)\cdot\left(\frac{Y_{j,t}(\mathbf{1}) + Y_{j,t+1}(\mathbf{1})}{2} - \overline{Y}_j(\mathbf{1})\right),\\
S_{i,j}^{c} &= \frac{1}{T-1}\sum_{t=2}^{T}\left(\frac{Y_{i,t}(\mathbf{0}) + Y_{i,t+1}(\mathbf{0})}{2} - \overline{Y}_i(\mathbf{0})\right)\cdot\left(\frac{Y_{j,t}(\mathbf{0}) + Y_{j,t+1}(\mathbf{0})}{2} - \overline{Y}_j(\mathbf{0})\right),\\
S_{i,j}^{ct} &= \frac{1}{T-1}\sum_{t=2}^{T}\left(\frac{Y_{i,t}(\mathbf{1}) - Y_{i,t}(\mathbf{0}) + Y_{i,t+1}(\mathbf{1}) - Y_{i,t+1}(\mathbf{0})}{2} - (\overline{Y}_i(\mathbf{1}) - \overline{Y}_i(\mathbf{0}))\right)\\
&\qquad\cdot\left(\frac{Y_{j,t}(\mathbf{1}) - Y_{j,t}(\mathbf{0}) + Y_{j,t+1}(\mathbf{1}) - Y_{j,t+1}(\mathbf{0})}{2} - (\overline{Y}_j(\mathbf{1}) - \overline{Y}_j(\mathbf{0}))\right).
\end{aligned}$$

where $\overline{Y}_i(\mathbf{1})$ and $\overline{Y}_i(\mathbf{0})$ refer to the average outcomes for treatment and control.

PROPOSITION 4. *For any potential outcomes* $\mathbb{Y}$*, the covariance of any two units* i *and* j *in the cluster-based design* $\eta_d^\dagger$ *can be decomposed as*

$$\begin{aligned}\mathrm{Cov}(\widehat{\tau}_i,\widehat{\tau}_j) =& \theta_{i,j}^0(K,T)(S_{i,j}^t + S_{i,j}^c)\\ &+\theta_{i,j}^1(K,T)S_{i,j}^{ct}\\ &+\frac{\theta_{i,j}^2(K,T)}{T-1}\sum_{t=2}^{T}(Y_{i,t}(\mathbf{1})Y_{j,t}(\mathbf{1})+Y_{i,t}(\mathbf{0})Y_{j,t}(\mathbf{0}))\\ &+\frac{\theta_{i,j}^3(K,T)}{T-1}\sum_{t=2}^{T}(Y_{i,t}(\mathbf{1})Y_{j,t}(\mathbf{0})+Y_{i,t}(\mathbf{0})Y_{j,t}(\mathbf{1}))\\ &+\frac{\theta_{i,j}^4(K,T)}{T-1}\sum_{t=2}^{T}(Y_{i,t}(\mathbf{1})Y_{j,t+1}(\mathbf{1})+Y_{j,t}(\mathbf{1})Y_{i,t+1}(\mathbf{1})+Y_{i,t}(\mathbf{0})Y_{j,t+1}(\mathbf{0})+Y_{j,t}(\mathbf{0})Y_{i,t+1}(\mathbf{0}))\\ &+\frac{\theta_{i,j}^5(K,T)}{T-1}\sum_{t=2}^{T}(Y_{i,t}(\mathbf{1})Y_{j,t+1}(\mathbf{0})+Y_{j,t}(\mathbf{0})Y_{i,t+1}(\mathbf{1})+Y_{i,t}(\mathbf{0})Y_{j,t+1}(\mathbf{1})+Y_{j,t}(\mathbf{1})Y_{i,t+1}(\mathbf{0})).\end{aligned} \tag{17}$$

We derive this decomposition following the same idea of Theorem 1. Compared to the time series experiments where the coefficients are simply characterized by T, the coefficients $(\boldsymbol{\theta}^0,\boldsymbol{\theta}^1,\boldsymbol{\theta}^2,\boldsymbol{\theta}^3,\boldsymbol{\theta}^4,\boldsymbol{\theta}^5)$ in the panel setting also depend on the randomized spatial clustering(i.e., $K_{i,j}(\boldsymbol{\omega};d)$), which gives a more sophisticated characterization. See the explicit expression of the coefficients and the complete proof in Appendix C.4. We remark that $(\boldsymbol{\theta}^0,\boldsymbol{\theta}^1,\boldsymbol{\theta}^2,\boldsymbol{\theta}^3)$ decrease in T with order $O(1/T)$ while $(\boldsymbol{\theta}^4,\boldsymbol{\theta}^5)$ decrease in T with order $O(1/T^2)$. Despite the absence of closed-form expressions, this leads to a structure similar to the single-unit case in (3). $\boldsymbol{\theta}^0$ and $\boldsymbol{\theta}^1$ are the coefficients we commonly see in complete randomization; $\boldsymbol{\theta}^2$ and $\boldsymbol{\theta}^3$ are the coefficients that refer to the randomness from valid observations on units i and j; $\boldsymbol{\theta}^4$ and $\boldsymbol{\theta}^5$ are the residual coefficients that decay much faster in T.

A.3. Inference for Panels

When it comes to the panel setting with $N>1$, we still rely on variances introduced by random assignments across time periods to make inferences. Note that this is different from the inferential technique that is used for population/network interference (Han et al. 2024, Ugander and Yin 2020). Using the variance $\mathrm{Var}_{\eta_d^\dagger}(\widehat{\tau}|\mathbb{Y})$ we have derived for randomized spatial clustering, we extend the normal approximation in Theorem 4.

COROLLARY 2. *Under Assumptions 1 - 3 and if* $\mathrm{Var}_{\eta_d^\dagger}(\widehat{\tau}|\mathbb{Y})=\Omega(1/T)$*, the limiting distribution of the IPW estimator is a normal distribution, i.e., as* $T\to\infty$*,*

$$\frac{\widehat{\tau}-\tau}{\sqrt{\mathrm{Var}_{\eta_d^\dagger}(\widehat{\tau}|\mathbb{Y})}}\xrightarrow{D}\mathcal{N}(0,1). \tag{18}$$

This gives the nominal coverage that $\lim_{T\to\infty}\mathbb{P}(|\widehat{\tau}-\tau|\geq c(\alpha))=\alpha$*, where* α *is the nominal level and* $c(\alpha)$ *is the critical value of the normal distribution* $\mathcal{N}(0,\mathrm{Var}_{\eta_d^\dagger}(\widehat{\tau}|\mathbb{Y}))$*.*

We prove this in Appendix C.8 and provide an upper bound for the variance with an unbiased estimate.

Algorithm 1 Greedy Algorithm for Merging Clusters

Initialize with our randomized spatial clustering $\{\mathscr{C}(\boldsymbol{\omega};d), \forall \boldsymbol{\omega} \in \Omega^2\}$, and compute the sum of pair-wise propensity scores

$$\Sigma_e = \sum_{i=1}^{N}\sum_{j=1}^{N} e_{i,j} = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\mathbb{E}[0.25^{K_{j,j}(\boldsymbol{\omega};d)-K_{i,j}(\boldsymbol{\omega};d)}(1-\mathbb{1}\{K_{i,j}(\boldsymbol{\omega};d)=0\})]}{\mathbb{E}[0.25^{K_{j,j}(\boldsymbol{\omega};d)}]}.$$

Initialize an indicator $I(\boldsymbol{\omega}) = 1, \forall \boldsymbol{\omega} \in \Omega^2$ to record whether a clustering with certain shaking parameter is allowed to conduct merging.

for iter = 1, 2, 3,... **do**

  for $\boldsymbol{\omega} \in \Omega^2$ **do** ▷ enumerate all shaking parameters

    for $\mathcal{C}_1, \mathcal{C}_2 \in \mathscr{C}(\boldsymbol{\omega};d) \times \mathscr{C}(\boldsymbol{\omega};d)$ **do** ▷ enumerate all pairs of clusters

      Compute the new sum of pair-wise propensity scores $\Sigma'_e(\mathcal{C}_1, \mathcal{C}_2)$ if two clusters $\mathcal{C}_1, \mathcal{C}_2$ were merged, and define $\Delta(\mathcal{C}_1, \mathcal{C}_2) = \Sigma_e - \Sigma'_e(\mathcal{C}_1, \mathcal{C}_2)$.

    end for

    if $\Delta(\mathcal{C}_1, \mathcal{C}_2) \le 0$ for all pairs **then**

      $I(\boldsymbol{\omega}) = 0$

      break

    end if

    Let $\mathcal{C}_1^*, \mathcal{C}_2^* = \arg\min \Delta(\mathcal{C}_1, \mathcal{C}_2)$ and merge clusters $\mathcal{C}_1^*, \mathcal{C}_2^*$ into a single cluster.

    Update $\mathscr{C}(\boldsymbol{\omega};d)$ to be the new clustering after merging, and update $\Sigma_e = \Sigma'_e(\mathcal{C}_1^*, \mathcal{C}_2^*)$

  end for

  if $I(\boldsymbol{\omega}) = 0, \forall \boldsymbol{\omega} \in \Omega^2$ **then**

    break

  end if

end for

Return the new design $\{\mathscr{C}(\boldsymbol{\omega};d), \forall \boldsymbol{\omega} \in \Omega^2\}$ with merged clusters.

B. Appendix for Heuristic Algorithm

We provide the details of the heuristic algorithm. To balance the trade-off in merging two remote clusters and calibrate the merging decision, we introduce the following *pair-wise propensity score*:

$$e_{i,j} = \frac{\mathbb{E}[0.25^{K_{j,j}(\boldsymbol{\omega};d)-K_{i,j}(\boldsymbol{\omega};d)}(1-\mathbb{1}\{K_{i,j}(\boldsymbol{\omega};d)=0\})]}{\mathbb{E}[0.25^{K_{j,j}(\boldsymbol{\omega};d)}]}, \tag{19}$$

which is inspired by the characterization of covariance between two units. See Appendix C.4 for the details. Particularly, in the special case when $i = j$, this can be simplified as $e_{i,i} = \dfrac{1}{\mathbb{E}[0.25^{K_{i,i}(\boldsymbol{\omega};d)}]}$, which reduces to the propensity score of unit i as $T \to \infty$. With the help of this new parameter, we propose a greedy approach for merging clusters in Algorithm 1.

C. Appendix for Complete Proofs

C.1. Proof of Theorem 1

We first compute several joint probabilities that will be used later. The propensity score

$$\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1})=\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{0})=\frac{\binom{T-3}{(T-1)/2}}{\binom{T-1}{(T-1)/2}}=\frac{T-3}{4T-8},$$

where $\binom{n}{k}$ stands for the combinatorial number of choosing k items from a total of n items. Similarly, the probabilities that we can observe three and four consecutive treatment/control are

$$\Pr(\boldsymbol{W}_{t-1:t+1}=\mathbf{1})=\Pr(\boldsymbol{W}_{t-1:t+1}=\mathbf{0})=\frac{\binom{T-4}{(T-1)/2}}{\binom{T-1}{(T-1)/2}}=\frac{(T-3)(T-5)}{(4T-8)(2T-6)},$$

and

$$\Pr(\boldsymbol{W}_{t-1:t+2}=\mathbf{1})=\Pr(\boldsymbol{W}_{t-1:t+2}=\mathbf{0})=\frac{\binom{T-5}{(T-1)/2}}{\binom{T-1}{(T-1)/2}}=\frac{(T-3)(T-5)(T-7)}{(4T-8)(2T-6)(2T-8)}.$$

Furthermore, we have

$$\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1},\boldsymbol{W}_{t+1:t+2}=\mathbf{0})=\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{0},\boldsymbol{W}_{t+1:t+2}=\mathbf{1})=\frac{\binom{T-5}{(T-5)/2}}{\binom{T-1}{(T-1)/2}}=\frac{(T-3)(T-1)(T-3)}{(4T-8)(2T-6)(2T-8)}.$$

Let $\delta_{t,t'}$ denote the distance between two periods. We also introduce the shorthand notations:

$$\lambda_0=\frac{T-3}{4T-8},\lambda_1=\frac{T-5}{2T-6},\lambda_2=\frac{T-7}{2T-8},\lambda_3=\frac{T-1}{2T-6},\lambda_4=\frac{T-3}{2T-8}.$$

Then we define the following:

$$\begin{aligned}q^+(\delta_{t,t'})&=\frac{\operatorname{Cov}\left(\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{1}\},\mathbb{1}\{\boldsymbol{W}_{t'-1:t'}=\mathbf{1}\}\right)}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1})\Pr(\boldsymbol{W}_{t'-1:t'}=\mathbf{1})}=\frac{\operatorname{Cov}\left(\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{0}\},\mathbb{1}\{\boldsymbol{W}_{t'-1:t'}=\mathbf{0}\}\right)}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{0})\Pr(\boldsymbol{W}_{t'-1:t'}=\mathbf{0})}\\&=\begin{cases}1/\lambda_0-1, & \delta_{t,t'}=0\\ \lambda_1/\lambda_0-1, & \delta_{t,t'}=1\\ \lambda_1\lambda_2/\lambda_0-1, & \delta_{t,t'}\geq 2\end{cases}\end{aligned}$$

and

$$\begin{aligned}q^-(\delta_{t,t'})&=-\frac{\operatorname{Cov}\left(\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{1}\},\mathbb{1}\{\boldsymbol{W}_{t'-1:t'}=\mathbf{0}\}\right)}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1})\Pr(\boldsymbol{W}_{t'-1:t'}=\mathbf{0})}=-\frac{\operatorname{Cov}\left(\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{0}\},\mathbb{1}\{\boldsymbol{W}_{t'-1:t'}=\mathbf{1}\}\right)}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{0})\Pr(\boldsymbol{W}_{t'-1:t'}=\mathbf{1})}\\&=\begin{cases}1, & \delta_{t,t'}=0\\ 1, & \delta_{t,t'}=1\\ 1-\lambda_3\lambda_4/\lambda_0, & \delta_{t,t'}\geq 2\end{cases}\end{aligned}$$

Now we are ready to analyze the variance. We first write the estimator as follows,

$$\widehat{\tau}=\frac{1}{T-1}\sum_{t=2}^{T}\left[Y_t(\mathbf{1})\frac{\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{1}\}}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1})}-Y_t(\mathbf{0})\frac{\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{0}\}}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{0})}\right].$$

The variance of the estimator can be decomposed as

$$\begin{aligned}\mathrm{Var}(\widehat{\tau}|\mathbb{Y}) = &\mathrm{Var}\left(\frac{1}{T-1}\sum_{t=2}^{T} Y_t(\mathbf{1})\frac{\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{1}\}}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1})}\right) + \mathrm{Var}\left(\frac{1}{T-1}\sum_{t=2}^{T} Y_t(\mathbf{0})\frac{\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{0}\}}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{0})}\right)\\ &+2\,\mathrm{Cov}\left(\frac{1}{T-1}\sum_{t=2}^{T} Y_t(\mathbf{1})\frac{\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{1}\}}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1})}, -\frac{1}{T-1}\sum_{t=2}^{T} -Y_t(\mathbf{0})\frac{\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{0}\}}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{0})}\right).\end{aligned}$$

We first examine the first part of the variance:

$$\begin{aligned}&\mathrm{Var}\left(\frac{1}{T-1}\sum_{t=2}^{T} Y_t(\mathbf{1})\frac{\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{1}\}}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1})}\right)\\ =&\frac{1}{(T-1)^2}\sum_{t=2}^{T}\sum_{t'=2}^{T}\frac{\mathrm{Cov}\left(\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{1}\},\mathbb{1}\{\boldsymbol{W}_{t'-1:t'}=\mathbf{1}\}\right)}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1})\Pr(\boldsymbol{W}_{t'-1:t'}=\mathbf{1})}Y_t(\mathbf{1})Y_{t'}(\mathbf{1})\\ =&\frac{q^+(0)}{(T-1)^2}\sum_{t=2}^{T}Y_t^2(\mathbf{1}) + \frac{2q^+(1)}{(T-1)^2}\sum_{t=2}^{T}Y_t(\mathbf{1})Y_{t+1}(\mathbf{1}) + \frac{q^+(2)}{(T-1)^2}\sum_{t=2}^{T}\sum_{\delta_{t,t'}\geq 2}Y_t(\mathbf{1})Y_{t'}(\mathbf{1}).\end{aligned}$$

Because of

$$\begin{aligned}\sum_{t=2}^{T}(Y_t(\mathbf{1})-\overline{Y}(\mathbf{1}))^2 &= \sum_{t=2}^{T}Y_t^2(\mathbf{1}) - \frac{1}{T-1}\sum_{t=2}^{T}\sum_{t'=2}^{T}Y_t(\mathbf{1})Y_{t'}(\mathbf{1}),\\ \sum_{t=2}^{T}(Y_t(\mathbf{1})-\overline{Y}(\mathbf{1}))(Y_{t+1}(\mathbf{1})-\overline{Y}(\mathbf{1})) &= \sum_{t=2}^{T}Y_t(\mathbf{1})Y_{t+1}(\mathbf{1}) - \frac{1}{T-1}\sum_{t=2}^{T}\sum_{t'=2}^{T}Y_t(\mathbf{1})Y_{t'}(\mathbf{1}),\end{aligned}$$

we obtain the following reformulation:

$$\begin{aligned}&\mathrm{Var}\left(\frac{1}{T-1}\sum_{t=2}^{T} Y_t(\mathbf{1})\frac{\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{1}\}}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1})}\right)\\ =&\frac{2T^2-13T+17}{(T-1)(T-4)(T-3)^2}\sum_{t=2}^{T}(Y_t(\mathbf{1})-\overline{Y}(\mathbf{1}))^2\\ +&\frac{2T^2-13T+17}{(T-1)(T-4)(T-3)^2}\sum_{t=2}^{T}(Y_t(\mathbf{1})-\overline{Y}(\mathbf{1}))(Y_{t+1}(\mathbf{1})-\overline{Y}(\mathbf{1}))\\ +&\frac{1}{(T-1)(T-4)}\sum_{t=2}^{T}Y_t^2(\mathbf{1}) - \frac{1}{(T-1)(T-4)(T-3)}\sum_{t=2}^{T}Y_t(\mathbf{1})Y_{t+1}(\mathbf{1}).\end{aligned} \tag{20}$$

By symmetry, we can characterize the part for control:

$$\begin{aligned}&\mathrm{Var}\left(\frac{1}{T-1}\sum_{t=2}^{T} Y_t(\mathbf{0})\frac{\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{0}\}}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{0})}\right)\\ =&\frac{2T^2-13T+17}{(T-1)(T-4)(T-3)^2}\sum_{t=2}^{T}(Y_t(\mathbf{0})-\overline{Y}(\mathbf{0}))^2\\ +&\frac{2T^2-13T+17}{(T-1)(T-4)(T-3)^2}\sum_{t=2}^{T}(Y_t(\mathbf{0})-\overline{Y}(\mathbf{0}))(Y_{t+1}(\mathbf{0})-\overline{Y}(\mathbf{0}))\\ +&\frac{1}{(T-1)(T-4)}\sum_{t=2}^{T}Y_t^2(\mathbf{0}) - \frac{1}{(T-1)(T-4)(T-3)}\sum_{t=2}^{T}Y_t(\mathbf{0})Y_{t+1}(\mathbf{0}).\end{aligned} \tag{21}$$

Next, it remains to examine the covariance between the outcomes of treatment and control:

$$\begin{aligned}
&\mathrm{Cov}\left(\frac{1}{T-1}\sum_{t=2}^{T} Y_t(\mathbf{1})\frac{\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{1}\}}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1})}, -\frac{1}{T-1}\sum_{t=2}^{T} Y_t(\mathbf{0})\frac{\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{0}\}}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{0})}\right)\\
&=\frac{q^{-}(\delta_{t,t'})}{(T-1)^2}\sum_{t=2}^{T}\sum_{t'=2}^{T} Y_t(\mathbf{1})Y_{t'}(\mathbf{0}) = \overline{Y}(\mathbf{1})\overline{Y}(\mathbf{0}) - \frac{(T-2)(T-1)}{(T-3)(T-4)(T-1)^2}\sum_{t=2}^{T}\sum_{\delta_{t,t'}\geq 2} Y_t(\mathbf{1})Y_{t'}(\mathbf{0}).
\end{aligned}$$

Because of

$$\begin{aligned}
\sum_{t=2}^{T}(Y_t(\mathbf{1})-\overline{Y}(\mathbf{1}))(Y_t(\mathbf{0})-\overline{Y}(\mathbf{0})) &= \sum_{t=2}^{T} Y_t(\mathbf{1})Y_t(\mathbf{0}) - \frac{1}{T-1}\sum_{t=2}^{T}\sum_{t'=2}^{T} Y_t(\mathbf{1})Y_{t'}(\mathbf{0}),\\
\sum_{t=2}^{T}(Y_t(\mathbf{1})-\overline{Y}(\mathbf{1}))(Y_{t+1}(\mathbf{0})-\overline{Y}(\mathbf{0})) &= \sum_{t=2}^{T} Y_t(\mathbf{1})Y_{t+1}(\mathbf{0}) - \frac{1}{T-1}\sum_{t=2}^{T}\sum_{t'=2}^{T} Y_t(\mathbf{1})Y_{t'}(\mathbf{0})
\end{aligned}$$

we obtain the following reformulation:

$$\begin{aligned}
&\mathrm{Cov}\left(\frac{1}{T-1}\sum_{t=2}^{T} Y_t(\mathbf{1})\frac{\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{1}\}}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1})}, -\frac{1}{T-1}\sum_{t=2}^{T} Y_t(\mathbf{0})\frac{\mathbb{1}\{\boldsymbol{W}_{t-1:t}=\mathbf{0}\}}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{0})}\right)\\
&=\frac{2T-5}{(T-4)(T-3)(T-1)}\sum_{t=2}^{T}(Y_t(\mathbf{1})-\overline{Y}(\mathbf{1}))(Y_t(\mathbf{0})-\overline{Y}(\mathbf{0}))\\
&+\frac{2T-5}{2(T-4)(T-3)(T-1)}\sum_{t=2}^{T}(Y_t(\mathbf{1})-\overline{Y}(\mathbf{1}))(Y_{t+1}(\mathbf{0})-\overline{Y}(\mathbf{0}))\\
&+\frac{2T-5}{2(T-4)(T-3)(T-1)}\sum_{t=2}^{T}(Y_t(\mathbf{0})-\overline{Y}(\mathbf{0}))(Y_{t+1}(\mathbf{1})-\overline{Y}(\mathbf{1}))\\
&-\frac{1}{(T-4)(T-1)}\sum_{t=2}^{T} Y_t(\mathbf{1})Y_t(\mathbf{0})\\
&+\frac{1}{2(T-4)(T-1)(T-3)}\sum_{t=2}^{T} Y_t(\mathbf{1})Y_{t+1}(\mathbf{0}) + Y_t(\mathbf{0})Y_{t+1}(\mathbf{1}).
\end{aligned} \tag{22}$$

Next, we are going to rewrite (22) in a way that we can combine it with (20) and (21). Since we have the following two equations

$$\begin{aligned}
&\sum_{t=2}^{T}(Y_t(\mathbf{1})-\overline{Y}(\mathbf{1}))(Y_{t+1}(\mathbf{0})-\overline{Y}(\mathbf{0})) + \sum_{t=2}^{T}(Y_t(\mathbf{0})-\overline{Y}(\mathbf{0}))(Y_{t+1}(\mathbf{1})-\overline{Y}(\mathbf{1}))\\
=&\sum_{t=2}^{T}(Y_t(\mathbf{1})-\overline{Y}(\mathbf{1}))(Y_{t+1}(\mathbf{1})-\overline{Y}(\mathbf{1})) + \sum_{t=2}^{T}(Y_t(\mathbf{0})-\overline{Y}(\mathbf{0}))(Y_{t+1}(\mathbf{0})-\overline{Y}(\mathbf{0}))\\
-&\sum_{t=2}^{T}(Y_t(\mathbf{1})-Y_t(\mathbf{0})-\overline{Y}(\mathbf{1})+\overline{Y}(\mathbf{0}))(Y_{t+1}(\mathbf{1})-Y_{t+1}(\mathbf{0})-\overline{Y}(\mathbf{1})+\overline{Y}(\mathbf{0}))
\end{aligned}$$

and

$$2\sum_{t=2}^{T}(Y_t(\mathbf{1})-\overline{Y}(\mathbf{1}))(Y_t(\mathbf{0})-\overline{Y}(\mathbf{0})) = \sum_{t=2}^{T}(Y_t(\mathbf{1})-\overline{Y}(\mathbf{1}))^2 + \sum_{t=2}^{T}(Y_t(\mathbf{0})-\overline{Y}(\mathbf{0}))^2$$

$$-\sum_{t=2}^{T}(Y_t(\mathbf{1})-Y_t(\mathbf{0})-\overline{Y}(\mathbf{1})+\overline{Y}(\mathbf{0}))^2,$$

the covariance (22) can be cast as

$$\begin{aligned}
&\frac{2T-5}{2(T-4)(T-3)(T-1)}\left(\sum_{t=2}^{T}(Y_t(\mathbf{1})-\overline{Y}(\mathbf{1}))^2+\sum_{t=2}^{T}(Y_t(\mathbf{0})-\overline{Y}(\mathbf{0}))^2-\sum_{t=2}^{T}(Y_t(\mathbf{1})-Y_t(\mathbf{0})-\overline{Y}(\mathbf{1})+\overline{Y}(\mathbf{0}))^2\right)\\
&+\frac{2T-5}{2(T-4)(T-3)(T-1)}\left(\sum_{t=2}^{T}(Y_t(\mathbf{1})-\overline{Y}(\mathbf{1}))(Y_{t+1}(\mathbf{1})-\overline{Y}(\mathbf{1}))+\sum_{t=2}^{T}(Y_t(\mathbf{0})-\overline{Y}(\mathbf{0}))(Y_{t+1}(\mathbf{0})-\overline{Y}(\mathbf{0}))\right.\\
&\left.-\sum_{t=2}^{T}(Y_t(\mathbf{1})-Y_t(\mathbf{0})-\overline{Y}(\mathbf{1})+\overline{Y}(\mathbf{0}))(Y_{t+1}(\mathbf{1})-Y_{t+1}(\mathbf{0})-\overline{Y}(\mathbf{1})+\overline{Y}(\mathbf{0}))\right)\\
&-\frac{1}{(T-4)(T-1)}\sum_{t=2}^{T}Y_t(\mathbf{1})Y_t(\mathbf{0})+\frac{1}{2(T-4)(T-1)(T-3)}\sum_{t=2}^{T}Y_t(\mathbf{1})Y_{t+1}(\mathbf{0})+Y_t(\mathbf{0})Y_{t+1}(\mathbf{1}). \quad (23)
\end{aligned}$$

Finally, putting three parts (20), (21), (23) together and using S^c, S^t, S^{ct}, R^{ct}, we derive the variance of the estimator

$$\begin{aligned}
\mathrm{Var}(\widehat{\tau}|\mathbb{Y}) = &\frac{8(T-2)}{(T-3)^2}\left(S^t+S^c\right)-\frac{(4T-10)}{(T-3)(T-4)}S^{ct}+\frac{R^{ct}}{(T-4)}\\
&-\frac{1}{(T-1)(T-4)(T-3)}\sum_{t=2}^{T}\left(Y_t(\mathbf{1})Y_{t+1}(\mathbf{1})+Y_t(\mathbf{0})Y_{t+1}(\mathbf{0})-Y_t(\mathbf{1})Y_{t+1}(\mathbf{0})-Y_t(\mathbf{0})Y_{t+1}(\mathbf{1})\right).
\end{aligned}$$

which can be finally rewritten as

$$\mathrm{Var}_{\eta^\dagger}(\widehat{\tau}|\mathbb{Y})=\frac{(T-2)}{(T-3)(T-4)}\left(\frac{8(T-4)\left(S^t+S^c\right)}{T-3}-4S^{ct}+R^{ct}-\frac{2\tau^2}{T-2}\right).$$

C.2. Proof of Theorem 2 (1)

LEMMA 1. *The worst-case outcomes must satisfy the following structure:*

$$Y_t(\mathbf{1})=Y_t(\mathbf{0})=\begin{cases}B & 2\le t\le s,\\ 0 & t>s.\end{cases} \quad (24)$$

We first expand the variance by definition and have the following:

$$\begin{aligned}
(T-1)^2\cdot\mathrm{Var}_{\eta^\dagger}(\widehat{\tau}|\mathbb{Y})=\sum_{t=2}^{T}\sum_{t'=2}^{T}\Big(&q^+(\delta_{t,t'})Y_t(\mathbf{1})Y_{t'}(\mathbf{1})+q^+(\delta_{t,t'})Y_t(\mathbf{0})Y_{t'}(\mathbf{0})\\
&+q^-(\delta_{t,t'})Y_t(\mathbf{1})Y_{t'}(\mathbf{0})+q^-(\delta_{t,t'})Y_t(\mathbf{0})Y_{t'}(\mathbf{1})\Big).
\end{aligned}$$

This is a quadratic function with variables $Y_t(\mathbf{1}), Y_t(\mathbf{0}), \forall t\in\{2,3,\ldots,T\}$. To show it is also convex, we can rewrite the summation as $\boldsymbol{y}'\boldsymbol{\Sigma}\boldsymbol{y}$, where $\boldsymbol{y}\in\mathbb{R}^{2(T-1)}$ is the vector of all variables and $\boldsymbol{\Sigma}$ is a symmetric matrix of coefficients. Because variance is non-negative, we know that $\boldsymbol{y}'\boldsymbol{\Sigma}\boldsymbol{y}\ge 0$ for any $\boldsymbol{y}$, which implies that $\boldsymbol{\Sigma}$ is PSD and the function is convex in $\boldsymbol{y}$. Since the inner optimization is a

minimization in a bounded feasible region, the worst-case solution can be attained at one of the extreme points. That is,

$$Y_t(\mathbf{1}) \in \{0, B\}, Y_t(\mathbf{0}) \in \{0, B\}, \forall t.$$

Next, given any outcomes at the extreme point, we will argue that transforming into the structure (24) leads to a larger variance. To see this, we need to carefully analyze the coefficients of $\mathbf{\Sigma}$. It is easy to check that both $q^+(\delta_{t,t'})$ and $q^-(\delta_{t,t'})$ are decreasing in $\delta_{t,t'}$. Therefore, the closer two outcomes B, the more they contribute to the variance. Suppose we are given some outcomes at the extreme point and there are T_1 periods whose outcome of treatment is B while T_0 periods whose outcome of control is B. W.L.O.G., assuming $T_1 \geq T_0$, let us consider the following alternative outcomes:

$$Y_t(\mathbf{1}) = \begin{cases} B & 2 \leq t \leq T_1 + 1 \\ 0 & otherwise \end{cases}, \quad Y_t(\mathbf{0}) = \begin{cases} B & \frac{T_1 - T_0}{2} + 2 \leq t \leq \frac{T_1 + T_0}{2} + 1 \\ 0 & otherwise \end{cases} \tag{25}$$

We evaluate the variance of the alternative outcomes using the monotonicity of $q^+(\delta_{t,t'})$ and $q^-(\delta_{t,t'})$. Since the alternative outcomes group B together with the minimal distance, the variance from the outcomes of treatment(control) increases. Moreover, because the alternative outcomes synchronize the outcomes between treatment and control as much as possible, the covariance from the outcomes between treatment and control increases as well. Together, the alternative outcomes achieve a larger variance.

Lastly, it remains to show that further transforming (25) into (24) generates a larger variance. Essentially, the transformation is doing

$$\begin{aligned} Y_t(\mathbf{1}) = B \Longrightarrow Y_t(\mathbf{1}) = 0, \quad & \frac{T_1 + T_0}{2} + 2 \leq t \leq T_1 + 1 \\ Y_t(\mathbf{0}) = 0 \Longrightarrow Y_t(\mathbf{0}) = B, \quad & 2 \leq t \leq \frac{T_1 - T_0}{2} + 1. \end{aligned}$$

This can be illustrated using the examples in Figure 14 with $T = 9$. To see that the variance increases,

Period	2	3	4	5	6	7	8	9
Treatment	B	B	B	B	B	B	B	B
Control	0	0	0	B	B	0	0	0

⇨

Period	2	3	4	5	6	7	8	9
Treatment	B	B	B	B	B	0	0	0
Control	B	B	B	B	B	0	0	0

Figure 14 **Transform outcomes from** (25) **to** (24).

we need to show that the covariance between blue outcomes and red outcomes is getting larger. That is,

$$\sum_{t=2}^{(T_1-T_0)/2+1} \sum_{t'=2}^{(T_1-T_0)/2+1} q^-(\delta_{t,t'}) \geq \sum_{t=(T_1+T_0)/2+2}^{T_1+1} \sum_{t'=2}^{(T_1-T_0)/2+1} q^+(\delta_{t,t'}).$$

It is sufficient to show that

$$\sum_{t'=2}^{(T_1-T_0)/2+1} q^-(\delta_{t,t'}) \geq \sum_{t'=2}^{(T_1-T_0)/2+1} q^+(\delta_{t,t'}), \forall\ 2 \leq t \leq \frac{T_1-T_0}{2}+1.$$

Because of the monotonicity of $q^+(\cdot)$ and $q^-(\cdot)$, it suffices to show that

$$\sum_{\delta_{t,t'}=0}^{(T_1-T_0)/2-1} q^-(\delta_{t,t'}) \geq \sum_{\delta_{t,t'}=1}^{(T_1-T_0)/2} q^+(\delta_{t,t'}).$$

Plugging in the expressions, this is equivalent to

$$\begin{aligned}
&\left(\frac{T_1-T_0}{2}-2\right)\left(q^-(2)-q^+(2)\right)+q^-(0)+q^-(1)-q^+(1)-q^+(2)\geq 0\\
\Longleftrightarrow &-\left(\frac{T_1-T_0}{2}-2\right)\frac{4(T-1)}{(T-3)^2(T-4)}-\frac{(T-7)(T-5)(4T-8)}{(2T-8)(2T-6)(T-3)}-\frac{(T-5)(4T-8)}{(2T-6)(T-3)}+4\geq 0.
\end{aligned}$$

Since T_1-T_0 is bounded above by $\frac{T-1}{2}$, it suffices to show that

$$\begin{aligned}
&-\frac{2(T-5)(T-1)}{(T-3)^2(T-4)}-\frac{(T-7)(T-5)(4T-8)}{(2T-8)(2T-6)(T-3)}-\frac{(T-5)(4T-8)}{(2T-6)(T-3)}+4\geq 0\\
\Longleftrightarrow &-\frac{(T-5)(3T-7)}{(T-3)^2}+4\geq 0\\
\Longleftrightarrow &\frac{(T-1)^2}{(T-3)^2}\geq 0.
\end{aligned}$$

Hence, the worst-case outcomes must obey the structure that:

$$Y_t(\mathbf{1})=Y_t(\mathbf{0})=\begin{cases}B & 2\leq t\leq s,\\ 0 & t>s.\end{cases}$$

From Lemma 1, we know that the worst-case outcomes obey the following structure

$$Y_t(\mathbf{1})=Y_t(\mathbf{0})=\begin{cases}B & 2\leq t\leq s,\\ 0 & t>s.\end{cases}$$

We seek to show that $s=\frac{T+1}{2}$ by contradiction.

Suppose that $s<\frac{T+1}{2}$, then we can set $Y_{s+1}(\mathbf{1})=Y_{s+1}(\mathbf{0})=B$. In this way, we have one more pair of outcomes B which contributes to the variance. To show that the variance increases, it is equivalent to prove that

$$q^+(0)+q^-(0)+2\sum_{\delta_{t,t'}=1}^{s}\left(q^+(\delta_{t,t'})+q^-(\delta_{t,t'})\right)\geq 0$$

Since $q^+(\delta_{t,t'})+q^-(\delta_{t,t'})$ takes negative value when $\delta_{t,t'}\geq 2$, it is sufficient to show that

$$q^+(0)+q^-(0)+2\sum_{\delta_{t,t'}=1}^{(T-3)/2}\left(q^+(\delta_{t,t'})+q^-(\delta_{t,t'})\right)\geq 0$$

Plugging in the expressions of $\boldsymbol{q}^+$ and $\boldsymbol{q}^-$, we have

$$q^+(0)+q^-(0)+2\sum_{\delta_{t,t'}=1}^{(T-3)/2}(q^+(\delta_{t,t'})+q^-(\delta_{t,t'}))=\frac{8(T-2)}{(T-3)^2}>0.$$

In the other way around when $s>\frac{T+1}{2}$, then we can set $Y_{s-1}(\mathbf{1})=Y_{s-1}(\mathbf{0})=0$. Following the similar argument, it is sufficient to show that

$$-q^+(0)-q^-(0)-2\sum_{\delta_{t,t'}=1}^{(T-1)/2}(q^+(\delta_{t,t'})+q^-(\delta_{t,t'}))\geq 0$$

Plugging in the expressions again, we have

$$-q^+(0)-q^-(0)-2\sum_{\delta_{t,t'}=1}^{(T-1)/2}(q^+(\delta_{t,t'})+q^-(\delta_{t,t'}))=\frac{8(T-2)}{(T-3)^2}>0.$$

Hence, the variance reaches the maximum when $s=\frac{T+1}{2}$.

To further derive the worst-case variance, we can simply use the variance decomposition (3).

Note that the last three parts are all zero, the worst-case variance can be calculated by

$$\overline{\mathrm{Var}}_{\eta^\dagger}(\widehat{\tau})=\frac{8(T-2)^2}{(T-3)^2(T-1)}(S^t+S^c)=\frac{4(T-2)}{(T-3)(T-1)}B^2.$$

C.3. Proof of Theorem 2 (2)

LEMMA 2. *Let us consider two consecutive time periods t and $t+1$. For any symmetric design, we have the following inequality:*

$$\begin{aligned}&\frac{1}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1})}-1+\frac{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1},\boldsymbol{W}_{t:t+1}=\mathbf{1})}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1})\Pr(\boldsymbol{W}_{t:t+1}=\mathbf{1})}-1+\\&\frac{1}{\Pr(\boldsymbol{W}_{t:t+1}=\mathbf{1})}-1+\frac{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1},\boldsymbol{W}_{t:t+1}=\mathbf{1})}{\Pr(\boldsymbol{W}_{t-1:t}=\mathbf{1})\Pr(\boldsymbol{W}_{t:t+1}=\mathbf{1})}-1\geq 4\end{aligned}\tag{26}$$

We first reformulate the inequality as

$$\begin{aligned}&\Pr(\boldsymbol{W}_{t-1:t}=1)+2\Pr(\boldsymbol{W}_{t-1:t}=1,\boldsymbol{W}_{t:t+1}=1)+\Pr(\boldsymbol{W}_{t:t+1}=1)\\&\geq 8\Pr(\boldsymbol{W}_{t:t+1}=1)\Pr(\boldsymbol{W}_{t:t+1}=1).\end{aligned}\tag{27}$$

Now let us focus on three time periods: $t-1$, t and $t+1$. There are 8 possible assignment paths. We layout 4 of them and the remaining ones are just symmetric:

$$\boldsymbol{W}_{t-1:t+1}\in\{(1,1,1),(1,1,0),(0,1,1),(1,0,1)\}.$$

with their probability mass denoted as $a_3,a_{2,1},a_{2,2},a_{2,3}$ respectively. Then we can characterize the probabilities in the inequality using these $\boldsymbol{a}$:

$$a_3+a_{2,1}+2a_3+a_3+a_{2,2}\geq 8(a_3+a_{2,1})(a_3+a_{2,2}).$$

Since $a_3 + a_{2,1} + a_{2,2} + a_{2,3} = 0.5$, it is equivalent to show

$$2(a_3 + a_{2,1} + a_{2,2} + a_{2,3})(4a_3 + a_{2,1} + a_{2,2}) \geq 8(a_3 + a_{2,1})(a_3 + a_{2,2}).$$

Notice that $a_{2,3}$ only appears on the left-hand-side, so it is sufficient to show

$$2(a_3 + a_{2,1} + a_{2,2} + a_{2,3})(4a_3 + a_{2,1} + a_{2,2}) \geq 8(a_3 + a_{2,1})(a_3 + a_{2,2}).$$

which can be further simplified as

$$(a_{2,1} - a_{2,2})^2 + a_3(a_{2,1} + a_{2,2}) \geq 0.$$

This is true for any $\boldsymbol{a}$.

In the proof of Lemma 1, we rewrite the variance by introducing $\boldsymbol{y} \in \mathbb{R}^{2(T-1)}$ to denote the vector of outcomes. Our original minimax problem is equivalent to

$$(T-1)^2 \cdot \mathrm{Var}(\widehat{\tau}) = \min_{\Sigma} \max_{\boldsymbol{y} \in [0,B]} \boldsymbol{y}^T \Sigma \boldsymbol{y},$$

where Σ is some covariance matrix that can be mapped from a feasible design and the adversary finds an outcome vector to maximize the variance. To get a lower bound of the optimal worst-case variance, we consider a randomized feasible solution $\tilde{\boldsymbol{y}}$ regardless of the covariance matrix in the outer optimization. We first combine every two time periods as a group, so we have overall $n = \frac{T-1}{2}$ groups. We then randomly pick half of the groups and set their corresponding outcomes to B and others to 0. Let $h(i)$ denote the group of some outcome $\tilde{y}_i$. In this way, for any two outcomes $\tilde{y}_i$ and $\tilde{y}_j$ from the same group(i.e. $h(i) = h(j)$), $\mathbb{E}[\tilde{y}_i \tilde{y}_j] = \frac{1}{2}B^2$; for any two outcomes from different groups, $\mathbb{E}[\tilde{y}_i \tilde{y}_j] = \frac{n-2}{4(n-1)}B^2$. Now we can bound the inner optimization as follows:

$$\begin{aligned}
\max_{\boldsymbol{y} \in [0,B]} \boldsymbol{y}^T \Sigma \boldsymbol{y} \geq \mathbb{E}\left[\tilde{\boldsymbol{y}}^{\mathrm{T}} \Sigma \tilde{\boldsymbol{y}}\right] &= \sum_{h(i)=h(j)} \Sigma_{i,j} \frac{1}{2} B^2 + \sum_{h(i) \neq h(j)} \Sigma_{i,j} \frac{n-2}{4(n-1)} B^2 \\
&= \sum_{h(i)=h(j)} \Sigma_{i,j} \frac{n}{4(n-1)} B^2 + \sum_{\forall i,j} \Sigma_{i,j} \frac{n-2}{4(n-1)} B^2
\end{aligned}$$

Note that the last term is non-negative, so it implies that

$$\max_{\boldsymbol{y} \in [0,B]} \boldsymbol{y}^T \Sigma \boldsymbol{y} \geq \sum_{h(i)=h(j)} \Sigma_{i,j} \frac{n}{4(n-1)} B^2.$$

Then it remains to investigate $\Sigma_{i,j}$ when two outcomes $\tilde{y}_i$ and $\tilde{y}_j$ are from the same group(i.e. two consecutive periods). Let us focus on what will happen in one group. First of all, it is easy to observe that the following is true for any design:

$$1 - \frac{\Pr(\boldsymbol{W}_{t-1:t} = \boldsymbol{0}, \boldsymbol{W}_{t:t+1} = \boldsymbol{1})}{\Pr(\boldsymbol{W}_{t-1:t} = \boldsymbol{0}) \Pr(\boldsymbol{W}_{t:t+1} = \boldsymbol{1})} = 1 - \frac{\Pr(\boldsymbol{W}_{t-1:t} = \boldsymbol{0}, \boldsymbol{W}_{t-1:t} = \boldsymbol{1})}{\Pr(\boldsymbol{W}_{t-1:t} = \boldsymbol{0}) \Pr(\boldsymbol{W}_{t-1:t} = \boldsymbol{1})} = 1$$

We have 4 such pairs in one group, so they contribute $\frac{8n}{4(n-1)}B^2$ to the variance. Next, if we set the assignments in the above equation to be jointly $\mathbf{0}$ or $\mathbf{1}$, we are not able to know the exact values. Fortunately, based on Lemma 2, we can still bound the variance to which they contribute by $\frac{8n}{4(n-1)}B^2$. Lastly, as we have n groups, we get the lower bound

$$\max_{\boldsymbol{y}\in[0,B]} \boldsymbol{y}^T\Sigma\boldsymbol{y} \geq n\frac{(8+8)n}{4(n-1)}B^2 = \frac{4n^2}{n-1}B^2 = \frac{2}{T-3}B^2.$$

C.4. Proof of Proposition 4

For ease of exposition, we use $K_{i,j}$ as a shorthand notation for $K_{i,j}(\boldsymbol{\omega}; d)$. This is a random variable whose distribution is governed by cluster size d and realization is determined by shaking parameters $\boldsymbol{\omega}$. We first extend our notations to the panel setting. Specifically, for each pair of units $i \in [N]$ and $j \in [N]$, we introduce $q^+_{i,j}(\delta_{t,t'})$ and $q^-_{i,j}(\delta_{t,t'})$ as functions of $\delta_{t,t'}$:

$$\begin{aligned} q^+_{i,j}(\delta_{t,t'}) &= \frac{\mathrm{Cov}(\mathbb{1}\{\boldsymbol{W}_{\mathcal{N}(i),t-1:t} = \mathbf{1}\}, \mathbb{1}\{\boldsymbol{W}_{\mathcal{N}(j),t'-1:t'} = \mathbf{1}\})}{\Pr(\boldsymbol{W}_{\mathcal{N}(i),t-1:t} = \mathbf{1})\Pr(\boldsymbol{W}_{\mathcal{N}(j),t'-1:t'} = \mathbf{1})} \\ &= \begin{cases} \dfrac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]} - 1, & \delta_{t,t'} = 0 \\ \dfrac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}\lambda_1^{K_{i,j}}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]} - 1, & \delta_{t,t'} = 1 \\ \dfrac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}(\lambda_1\lambda_2)^{K_{i,j}}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]} - 1, & \delta_{t,t'} \geq 2 \end{cases} \end{aligned}$$

and

$$\begin{aligned} q^-_{i,j}(\delta_{t,t'}) &= -\frac{\mathrm{Cov}(\mathbb{1}\{\boldsymbol{W}_{\mathcal{N}(i),t-1:t} = \mathbf{1}\}, \mathbb{1}\{\boldsymbol{W}_{\mathcal{N}(j),t'-1:t'} = \mathbf{0}\})}{\Pr(\boldsymbol{W}_{\mathcal{N}(i),t-1:t} = \mathbf{1})\Pr(\boldsymbol{W}_{\mathcal{N}(j),t'-1:t'} = \mathbf{0})} \\ &= \begin{cases} 1 - \dfrac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}\mathbb{1}\{K_{i,j}=0\}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]}, & \delta_{t,t'} = 0 \\ 1 - \dfrac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}\mathbb{1}\{K_{i,j}=0\}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]}, & \delta_{t,t'} = 1 \\ 1 - \dfrac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}(\lambda_3\lambda_4)^{K_{i,j}}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]}, & \delta_{t,t'} \geq 2 \end{cases} \end{aligned}$$

Then, similar to the single-unit case, we have

$$\begin{aligned} (T-1)^2 \cdot \mathrm{Cov}(\widehat{\tau}_i, \widehat{\tau}_j) = \sum_{t=2}^{T}\sum_{t'=2}^{T}\Big(& q^+_{i,j}(\delta_{t,t'})Y_{i,t}(\mathbf{1})Y_{j,t'}(\mathbf{1}) + q^+_{i,j}(\delta_{t,t'})Y_{i,t}(\mathbf{0})Y_{j,t'}(\mathbf{0}) \\ & + q^-_{i,j}(\delta_{t,t'})Y_{i,t}(\mathbf{1})Y_{j,t'}(\mathbf{0}) + q^-_{i,j}(\delta_{t,t'})Y_{i,t}(\mathbf{0})Y_{j,t'}(\mathbf{1})\Big). \end{aligned}$$

This is equivalent to

$$(T-1)^2 \cdot \mathrm{Cov}(\widehat{\tau}_i, \widehat{\tau}_j) = \sum_{t=2}^{T}\left(q^+_{i,j}(0)(Y_{i,t}(\mathbf{1})Y_{j,t}(\mathbf{1}) + Y_{i,t}(\mathbf{0})Y_{j,t}(\mathbf{0})) + q^-_{i,j}(0)(Y_{i,t}(\mathbf{1})Y_{j,t}(\mathbf{0}) + Y_{i,t}(\mathbf{0})Y_{j,t}(\mathbf{1}))\right)$$

$$
\begin{aligned}
&+\sum_{t=2}^{T} q_{i,j}^{+}(1)\left(Y_{i,t}(\mathbf{1})Y_{j,t+1}(\mathbf{1})+Y_{j,t}(\mathbf{1})Y_{i,t+1}(\mathbf{1})+Y_{i,t}(\mathbf{0})Y_{j,t+1}(\mathbf{0})+Y_{j,t}(\mathbf{0})Y_{i,t+1}(\mathbf{0})\right)\\
&+\sum_{t=2}^{T} q_{i,j}^{-}(1)\left(Y_{i,t}(\mathbf{1})Y_{j,t+1}(\mathbf{0})+Y_{i,t}(\mathbf{0})Y_{j,t+1}(\mathbf{1})+Y_{j,t}(\mathbf{1})Y_{i,t+1}(\mathbf{0})+Y_{j,t}(\mathbf{0})Y_{i,t+1}(\mathbf{1})\right)\\
&+\sum_{t=2}^{T}\sum_{t':\delta_{t,t'}\geq 2}^{T} q_{i,j}^{+}(2)\left(Y_{i,t}(\mathbf{1})Y_{j,t'}(\mathbf{1})+Y_{i,t}(\mathbf{0})Y_{j,t'}(\mathbf{0})\right)\\
&+\sum_{t=2}^{T}\sum_{t':\delta_{t,t'}\geq 2}^{T} q_{i,j}^{-}(2)\left(Y_{i,t}(\mathbf{1})Y_{j,t'}(\mathbf{0})+Y_{i,t}(\mathbf{0})Y_{j,t'}(\mathbf{1})\right)
\end{aligned}
$$

which can be reformulated as

$$
\begin{aligned}
&(T-1)^2\cdot\mathrm{Cov}(\widehat{\tau}_i,\widehat{\tau}_j)\\
&=\sum_{t=2}^{T}(q_{i,j}^{+}(0)-q_{i,j}^{+}(2))(Y_{i,t}(\mathbf{1})Y_{j,t}(\mathbf{1})+Y_{i,t}(\mathbf{0})Y_{j,t}(\mathbf{0}))\\
&+\sum_{t=2}^{T}(q_{i,j}^{-}(0)-q_{i,j}^{-}(2))(Y_{i,t}(\mathbf{1})Y_{j,t}(\mathbf{0})+Y_{i,t}(\mathbf{0})Y_{j,t}(\mathbf{1}))\\
&+\sum_{t=2}^{T}(q_{i,j}^{+}(1)-q_{i,j}^{+}(2))(Y_{i,t}(\mathbf{1})Y_{j,t+1}(\mathbf{1})+Y_{j,t}(\mathbf{1})Y_{i,t+1}(\mathbf{1})+Y_{i,t}(\mathbf{0})Y_{j,t+1}(\mathbf{0})+Y_{j,t}(\mathbf{0})Y_{i,t+1}(\mathbf{0}))\\
&+\sum_{t=2}^{T}(q_{i,j}^{-}(1)-q_{i,j}^{-}(2))(Y_{i,t}(\mathbf{1})Y_{j,t+1}(\mathbf{0})+Y_{i,t}(\mathbf{0})Y_{j,t+1}(\mathbf{1})+Y_{j,t}(\mathbf{1})Y_{i,t+1}(\mathbf{0})+Y_{j,t}(\mathbf{0})Y_{i,t+1}(\mathbf{1}))\\
&+q_{i,j}^{+}(2)(T-1)^2(\overline{Y}_i(\mathbf{1})\overline{Y}_j(\mathbf{1})+\overline{Y}_i(\mathbf{0})\overline{Y}_j(\mathbf{0}))+q_{i,j}^{-}(2)(T-1)^2(\overline{Y}_i(\mathbf{1})\overline{Y}_j(\mathbf{0})+\overline{Y}_i(\mathbf{0})\overline{Y}_j(\mathbf{1})).
\end{aligned}
$$

We further rewrite this by absorbing the terms in the last line with $S_{i,j}^{c}, S_{i,j}^{t}, S_{i,j}^{ct}$:

$$
\begin{aligned}
&(T-1)^2\cdot\mathrm{Cov}(\widehat{\tau}_i,\widehat{\tau}_j)\\
&=(T-1)^2(-q_{i,j}^{+}(2)-q_{i,j}^{-}(2))(S_{i,j}^{t}+S_{i,j}^{c})\\
&+(T-1)^2 q_{i,j}^{-}(2)S_{i,j}^{ct}\\
&+(q_{i,j}^{+}(0)+\frac{T-3}{2}q_{i,j}^{+}(2))\sum_{t=2}^{T}(Y_{i,t}(\mathbf{1})Y_{j,t}(\mathbf{1})+Y_{i,t}(\mathbf{0})Y_{j,t}(\mathbf{0}))\\
&+(q_{i,j}^{-}(0)+\frac{T-3}{2}q_{i,j}^{-}(2))\sum_{t=2}^{T}(Y_{i,t}(\mathbf{1})Y_{j,t}(\mathbf{0})+Y_{i,t}(\mathbf{0})Y_{j,t}(\mathbf{1}))\\
&+(q_{i,j}^{+}(1)+\frac{T-5}{4}q_{i,j}^{+}(2))\sum_{t=2}^{T}(Y_{i,t}(\mathbf{1})Y_{j,t+1}(\mathbf{1})+Y_{j,t}(\mathbf{1})Y_{i,t+1}(\mathbf{1})+Y_{i,t}(\mathbf{0})Y_{j,t+1}(\mathbf{0})+Y_{j,t}(\mathbf{0})Y_{i,t+1}(\mathbf{0}))\\
&+(q_{i,j}^{-}(1)+\frac{T-5}{4}q_{i,j}^{-}(2))\sum_{t=2}^{T}(Y_{i,t}(\mathbf{1})Y_{j,t+1}(\mathbf{0})+Y_{j,t}(\mathbf{0})Y_{i,t+1}(\mathbf{1})+Y_{i,t}(\mathbf{0})Y_{j,t+1}(\mathbf{1})+Y_{j,t}(\mathbf{1})Y_{i,t+1}(\mathbf{0})).
\end{aligned}
$$

Plugging in the expressions of $\boldsymbol{q}^+$ and $\boldsymbol{q}^-$, we finally have the variance decomposition shown in Theorem 4 where the coefficients are given by

$$\theta^0_{i,j}(K,T) = \frac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}((\lambda_3\lambda_4)^{K_{i,j}} - (\lambda_1\lambda_2)^{K_{i,j}})]}{\mathbb{E}[\lambda_0^{K_{j,j}}]}$$

$$\theta^1_{i,j}(K,T) = \left(1 - \frac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}(\lambda_3\lambda_4)^{K_{i,j}}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]}\right)$$

$$\theta^2_{i,j}(K,T) = \frac{1}{(T-1)}\left(\frac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]} + \frac{T-3}{2}\frac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}(\lambda_1\lambda_2)^{K_{i,j}}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]} - \frac{T-1}{2}\right)$$

$$\theta^3_{i,j}(K,T) = \frac{1}{(T-1)}\left(\frac{T-1}{2} - \frac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}\mathbb{1}\{K_{i,j}=0\}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]} - \frac{T-3}{2}\frac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}(\lambda_3\lambda_4)^{K_{i,j}}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]}\right)$$

$$\theta^4_{i,j}(K,T) = \frac{1}{(T-1)}\left(\frac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}\lambda_1^{K_{i,j}}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]} + \frac{T-5}{4}\frac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}(\lambda_1\lambda_2)^{K_{i,j}}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]} - \frac{T-1}{4}\right)$$

$$\theta^5_{i,j}(K,T) = \frac{1}{(T-1)}\left(\frac{T-1}{4} - \frac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}\mathbb{1}\{K_{i,j}=0\}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]} - \frac{T-5}{4}\frac{\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}(\lambda_3\lambda_4)^{K_{i,j}}]}{\mathbb{E}[\lambda_0^{K_{j,j}}]}\right).$$

C.5. Proof of Theorem 3

We prove the results by first generalizing Lemma 1 to the panel setting.

LEMMA 3. *The worst-case outcomes for our design $\eta_d^\dagger$ must satisfy the following structure: $\forall i \in [N]$, there is some $\alpha_{d,i} > 0$ such that*

$$Y_{i,t}(\mathbf{1}) = Y_{i,t}(\mathbf{0}) = \begin{cases} B & 2 \le t \le \alpha_{d,i}T, \\ 0 & t > \alpha_{d,i}T. \end{cases} \tag{28}$$

Note that the variance is still a convex function of potential outcomes so the worst-case solution remains to be one of the extreme points that $Y_{i,t}(\mathbf{1}), Y_{i,t}(\mathbf{0}) \in \{0, B\}, \forall i, t$. Similar to the proof of Theorem 2, our goal is to show that transforming any extreme point to the one with the structure (28) gives a larger variance.

First of all, we notice that both $q^+_{i,j}(\delta_{t,t'})$ and $q^-_{i,j}(\delta_{t,t'})$ are decreasing in $\delta_{t,t'}$ for all $i, j \in [N]$. Therefore, following the same argument in Appendix C.2, grouping and synchronizing treatment and control outcomes with B for both units i and j enhances their covariance $\mathrm{Cov}(\widehat{\tau}_i, \widehat{\tau}_j)$. Since the variance could be written as $\mathrm{Var}(\widehat{\tau}) = \frac{1}{N^2}\sum_i^N\sum_j^N \mathrm{Cov}(\widehat{\tau}_i, \widehat{\tau}_j)$, the total variance is maximized as well. Precisely, we obtain the following extreme point for each unit i:

$$Y_{i,t}(\mathbf{1}) = \begin{cases} B & 0.5(1-\alpha^t_{d,i})T \le t \le 0.5(1+\alpha^t_{d,i})T \\ 0 & \textit{otherwise} \end{cases}, \quad Y_{i,t}(\mathbf{0}) = \begin{cases} B & 0.5(1-\alpha^c_{d,i})T \le t \le 0.5(1+\alpha^c_{d,i})T \\ 0 & \textit{otherwise} \end{cases} \tag{29}$$

where $\alpha^t_{d,i}$ and $\alpha^c_{d,i}$ indicate the proportion of periods whose outcomes are B for treatment and control, respectively. Compared to the single-unit case, now we have multiple units with potentially different $\alpha^t_{d,i}$ and $\alpha^c_{d,i}$.

Next, we are going to modify the outcomes to meet the desired structure. We still borrow the idea in Appendix C.2 that suggests for unit $i \in [N]$ with $\alpha_{d,i}^t \geq \alpha_{d,i}^c$, we do the following transformation:

$$Y_{i,t}(\mathbf{1}) = B \Longrightarrow Y_{i,t}(\mathbf{1}) = 0, \quad 0.5(1+\alpha_{d,i}^t)T \leq t \leq 0.5(1+\alpha_{d,i}^c)T$$
$$Y_{i,t}(\mathbf{0}) = 0 \Longrightarrow Y_{i,t}(\mathbf{0}) = B, \quad 0.5(1-\alpha_{d,i}^t)T \leq t \leq 0.5(1-\alpha_{d,i}^c)T.$$

Period	2	3	4	5	6	7	8	9
Treatment	B	B	B	B	B	B	B	B
Control	0	0	0	B	B	0	0	0

⇨

Period	2	3	4	5	6	7	8	9
Treatment	B	B	B	B	B	0	0	0
Control	B	B	B	B	B	0	0	0

Figure 15 **Transforming outcomes of unit** i

Period	2	3	4	5	6	7	8	9
Treatment	0	B	B	B	B	B	B	0
Control	0	0	0	B	B	0	0	0

⇨

Period	2	3	4	5	6	7	8	9
Treatment	0	B	B	B	B	0	0	0
Control	0	B	B	B	B	0	0	0

Figure 16 **Transforming outcomes of unit** j

There are two cases for analyzing the change of covariance $\text{Cov}(\widehat{\tau}_i, \widehat{\tau}_j)$ for each pair of units i, j.

Case 1: $\alpha_{d,i}^t \geq \alpha_{d,i}^c, \alpha_{d,j}^t \geq \alpha_{d,j}^c$**.** We illustrate the transformation in this case using Figure 15 and 16 for instance. In this way, the only parts that may change $\text{Cov}(\widehat{\tau}_i, \widehat{\tau}_j)$ are (i) the covariance between the blue outcomes and the green outcomes; (ii) the covariance between the red outcomes and the yellow outcomes. These two parts are actually equal due to the symmetry, so we just need to show that (i) increases, which is true if

$$\sum_{t=0.5(1-\alpha_{d,j}^t)T}^{0.5(1-\alpha_{d,j}^c)T} \sum_{t'=0.5(1-\alpha_{d,i}^t)T}^{0.5(1-\alpha_{d,i}^c)T} q_{i,j}^-(\delta_{t,t'}) \geq \sum_{t=0.5(1-\alpha_{d,j}^t)T}^{0.5(1-\alpha_{d,j}^c)T} \sum_{t'=0.5(1+\alpha_{d,i}^c)T}^{0.5(1+\alpha_{d,i}^t)T} q_{i,j}^+(\delta_{t,t'}). \tag{30}$$

It suffices to show that

$$\sum_{t'=0.5(1-\alpha_{d,i}^t)T}^{0.5(1-\alpha_{d,i}^c)T} q_{i,j}^-(\delta_{t,t'}) \geq \sum_{t'=0.5(1+\alpha_{d,i}^c)T}^{0.5(1+\alpha_{d,i}^t)T} q_{i,j}^+(\delta_{t,t'}), \forall 0.5(1-\alpha_{d,j}^t)T \leq t \leq 0.5(1-\alpha_{d,j}^c)T.$$

Note that we are not able to use the sufficient condition in Appendix C.2 because $q_{i,j}^+(\delta_{t,t'})$ could be larger than $q_{i,j}^-(\delta_{t,t'})$ if $K_{i,j}$ is large. Nonetheless, the asymptotic view of T simplifies the problem. It is sufficient to show that as $T \to \infty$,

$$q_{i,j}^-(0) + q_{i,j}^-(1) + ((\alpha_{d,i}^t - \alpha_{d,i}^c)T - 2)q_{i,j}^-(2) \geq (\alpha_{d,i}^t - \alpha_{d,i}^c)Tq_{i,j}^+(2). \tag{31}$$

Expanding the expressions of $\boldsymbol{q}^-_{i,j}$ and $\boldsymbol{q}^+_{i,j}$, we have

$$\begin{aligned}
&q^-_{i,j}(2) - q^+_{i,j}(2) = O(1/T^2)\\
&q^-_{i,j}(2) = O(1/T), q^+_{i,j}(2) = O(1/T)\\
&q^-_{i,j}(0) \geq 0, q^-_{i,j}(1) \geq 0.
\end{aligned}$$

Therefore, the condition (31) holds as $T \to \infty$.

Case 2: $\alpha^t_{d,i} \geq \alpha^c_{d,i}, \alpha^t_{d,j} \leq \alpha^c_{d,j}$**.** We follow the same argument above by modifying the coefficients in (30), which gives

$$\sum_{t=0.5(1-\alpha^c_{d,j})T}^{0.5(1-\alpha^t_{d,j})T} \sum_{t'=0.5(1-\alpha^t_{d,i})T}^{0.5(1-\alpha^c_{d,i})T} q^+_{i,j}(\delta_{t,t'}) \geq \sum_{t=0.5(1-\alpha^c_{d,j})T}^{0.5(1-\alpha^t_{d,j})T} \sum_{t'=0.5(1+\alpha^c_{d,i})T}^{0.5(1+\alpha^t_{d,i})T} q^-_{i,j}(\delta_{t,t'}). \tag{32}$$

It is sufficient to show that as $T \to \infty$,

$$q^+_{i,j}(0) + q^+_{i,j}(1) + ((\alpha^t_{d,i} - \alpha^c_{d,i})T - 2)q^+_{i,j}(2) \geq (\alpha^t_{d,i} - \alpha^c_{d,i})Tq^-_{i,j}(2).$$

This is true following a similar argument as above.

Therefore, after the transformation, treatment and control have the same proportion of outcomes with B: $\alpha_{d,i} = 0.5(\alpha^t_{d,i} + \alpha^c_{d,i})$. This shows that the worst-case outcomes must satisfy the structure in (28).

Now we are ready to conclude the theorem by refining the structure derived in Lemma 3. In particular, we aim to show that there is a common threshold parameter α_d such that

$$\alpha_{d,1} = \alpha_{d,2} = \cdots = \alpha_{d,N} = \alpha_d.$$

Suppose this is not true and units have different threshold parameters. For example, we have $\alpha_{d,j} = \min_l \alpha_{d,l} < \alpha_{d,i} = \max_l \alpha_{d,l}$. In this case, we do the following transformation for some small constant $\epsilon > 0$:

$$\begin{aligned}
&Y_{i,t}(\mathbf{1}) = Y_{i,t}(\mathbf{0}) = B \Longrightarrow Y_{i,t}(\mathbf{1}) = Y_{i,t}(\mathbf{0}) = 0, (\alpha_{d,i} - \epsilon)T < t \leq \alpha_{d,i}T,\\
&Y_{j,t}(\mathbf{1}) = Y_{j,t}(\mathbf{0}) = 0 \Longrightarrow Y_{j,t}(\mathbf{1}) = Y_{j,t}(\mathbf{0}) = B, \alpha_{d,j}T < t \leq (\alpha_{d,j} + \epsilon)T.
\end{aligned}$$

By doing so, for each unit $l \notin \{i, j\}$, we transfer part of the covariance $\mathrm{Cov}(\widehat{\tau}_i, \widehat{\tau}_l)$ to the covariance $\mathrm{Cov}(\widehat{\tau}_j, \widehat{\tau}_l)$. We analyze them accordingly.

Change of $\mathrm{Cov}(\widehat{\tau}_j, \widehat{\tau}_l)$**:** Adding ϵT periods of B to the unit j gives the change of covariance

$$\frac{1}{(T-1)^2} \sum_{t=\alpha_{d,j}T}^{(\alpha_{d,j}+\epsilon)T} 2(q^+_{j,l}(0) + q^-_{j,l}(0)) + 4(q^+_{j,l}(1) + q^-_{j,l}(1)) + 2(\alpha_{d,l}T - 3)(q^+_{j,l}(2) + q^-_{j,l}(2)).$$

As $T\to\infty$, since $q^+_{j,l}(2)+q^-_{j,l}(2)=\dfrac{\mathbb{E}[0.25^{K_{l,l}-K_{j,l}}(-8\cdot K_{j,l}0.25^{K_{j,l}})]}{\mathbb{E}[0.25^{K_{l,l}}]\cdot T}+o(1/T)$, this becomes

$$\begin{aligned}
&\frac{2\epsilon T}{(T-1)^2}\left(\frac{\mathbb{E}[0.25^{K_{l,l}-K_{j,l}}]}{\mathbb{E}[0.25^{K_{l,l}}]}-\frac{\mathbb{E}[0.25^{K_{l,l}-K_{j,l}}\mathbb{1}\{K_{j,l}=0\}]}{\mathbb{E}[0.25^{K_{l,l}}]}\right)\\
+&\frac{4\epsilon T}{(T-1)^2}\left(\frac{\mathbb{E}[0.25^{K_{l,l}-K_{j,l}}0.5^{K_{j,l}}]}{\mathbb{E}[0.25^{K_{l,l}}]}-\frac{\mathbb{E}[0.25^{K_{l,l}-K_{j,l}}\mathbb{1}\{K_{j,l}=0\}]}{\mathbb{E}[0.25^{K_{l,l}}]}\right)\\
+&\frac{\epsilon T}{(T-1)^2}\frac{\mathbb{E}[0.25^{K_{l,l}-K_{j,l}}(-8\cdot K_{j,l}0.25^{K_{j,l}})]}{\mathbb{E}[0.25^{K_{l,l}}]}+o(1/T)\\
=&\frac{\epsilon}{T}\frac{\mathbb{E}[0.25^{K_{l,l}-K_{j,l}}(2+4\cdot 0.5^{K_{j,l}}-6\mathbb{1}\{K_{j,l}=0\})-16\cdot\alpha_{d,l}K_{j,l}0.25^{K_{j,l}}]}{\mathbb{E}[0.25^{K_{l,l}}]}+o(1/T).
\end{aligned}$$

Notice that for any K, we have

$$0.25^{K_{l,l}-K_{j,l}}(2+4\cdot 0.5^{K_{j,l}}-6\mathbb{1}\{K_{j,l}=0\}-16\cdot\alpha_{d,l}K_{j,l}0.25^{K_{j,l}})\geq 0,$$

which implies that the change of covariance is non-negative.

Change of $\mathrm{Cov}(\widehat{\tau}_i,\widehat{\tau}_l)$**:** Removing ϵT periods of B from the unit i gives the change of covariance

$$-\frac{1}{(T-1)^2}\sum_{t=(\alpha_{d,i}-\epsilon)T}^{\alpha_{d,i}T}2\alpha_{d,l}T(q^+_{i,l}(2)+q^-_{i,l}(2))\geq 0$$

Period	2	3	4	5	6	7	8	9	…
Unit i	B	B	B	B	B	B	B	B	…
Unit j	B	B	0	0	0	0	0	0	…

⇨

Period	2	3	4	5	6	7	8	9	…
Unit i	B	B	B	B	B	B	B	0	…
Unit j	B	B	B	0	0	0	0	0	…

Figure 17 **Transforming outcomes of units i and j.**

The discussion above handles the impact of our transformation on the covariance of units other than i and j. Then it remains to consider the change of $\mathrm{Cov}(\widehat{\tau}_i,\widehat{\tau}_i)$, $\mathrm{Cov}(\widehat{\tau}_j,\widehat{\tau}_j)$ and $\mathrm{Cov}(\widehat{\tau}_i,\widehat{\tau}_j)$, which is illustrated in Figure 17 for instance. When making the transformation, the covariance between blue outcomes and black outcomes remains unchanged due to the symmetry of units, and the only part we need to study is the covariance between blue outcomes and red outcomes, which is precisely

$$\begin{aligned}
&\frac{\epsilon T}{(T-1)^2}2(q^+_{i,j}(0)+q^-_{i,j}(0))+4(q^+_{i,j}(1)+q^-_{i,j}(1))+2((\alpha_{d,i}-\alpha_{d,j}-\epsilon)T-3)(q^+_{i,j}(2)+q^-_{i,j}(2))\\
&-\frac{\epsilon T}{(T-1)^2}2(\alpha_{d,i}-\alpha_{d,j}-\epsilon)T(q^+_{i,i}(2)+q^-_{i,i}(2)).
\end{aligned}$$

Again, following the same argument as before, we can show it is non-negative. Therefore, the transformation does not decrease the total variance.

To prove that all threshold parameters must be equal, we do the following. Suppose $\alpha_{d,j'}$ is the second smallest threshold and $\alpha_{d,i'}$ is the second largest threshold. Let $\epsilon=\min\{\alpha_{d,i}-\alpha_{d,i'},\alpha_{d,j'}-\alpha_{d,j}\}$.

After doing the transformation above, we will update the threshold parameters so that two units will have equal smallest(or largest) threshold parameters. Then we could "merge" these two units for later transformation. For instance, if $\alpha_{d,j'} = \alpha_{d,j}$, then for the next step, we will add $\frac{\epsilon T}{2}$ periods of B to both units i and j. The analysis above can still be applied and we keep doing the transformation until all units have the same threshold parameter α_d satisfying the structure in (8).

Finally, we will derive the size of α_d in the worst-case outcomes. Suppose that we are given the outcomes that satisfy the worst-case structure with threshold $\alpha_d T$. Let us investigate the change of variance if we add one more period with B for all units, which is precisely

$$\begin{aligned}
&\frac{1}{N^2(T-1)^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(2(q_{i,j}^{+}(0)+q_{i,j}^{-}(0))+4\sum_{\delta_{t,t'}=1}^{\alpha_d T}(q_{i,j}^{+}(\delta_{t,t'})+q_{i,j}^{-}(\delta_{t,t'}))\right)\\
=&\frac{1}{N^2(T-1)^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(2(q_{i,j}^{+}(0)+q_{i,j}^{-}(0))+4(q_{i,j}^{+}(1)+q_{i,j}^{-}(1))+4(\alpha_d T-3)(q_{i,j}^{+}(2)+q_{i,j}^{-}(2))\right).
\end{aligned}$$

As $T\to\infty$, this could be simplified as

$$\frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\mathbb{E}[0.25^{K_{j,j}-K_{i,j}}(2+4\cdot 0.5^{K_{i,j}}-32\cdot\alpha_d K_{i,j}0.25^{K_{i,j}}-6\mathbb{1}\{K_{i,j}=0\})]}{\mathbb{E}[0.25^{K_{j,j}}]}+o(1/T^2).$$

One can check this by revisiting the switchback experiments (i.e. $d=D$) when $K_{i,j}\equiv 1$. When $\alpha_d=0.5$, the change of variance is zero so it hits the worst-case variance as we showed in Theorem 2. In general, to derive the value of α_d for attaining the asymptotically worst-case variance, we simply solve the following linear equation:

$$\alpha_d\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\mathbb{E}[0.25^{K_{j,j}-K_{i,j}}16K_{i,j}0.25^{K_{i,j}}]}{\mathbb{E}[0.25^{K_{j,j}}]}=\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\mathbb{E}[0.25^{K_{j,j}-K_{i,j}}(1+2\cdot 0.5^{K_{i,j}}-3\mathbb{1}\{K_{i,j}=0\})]}{\mathbb{E}[0.25^{K_{j,j}}]}.$$

Since we consider the scaling regime when $N\to\infty$, units are homogeneous with equal $\mathbb{E}[0.25^{K_{j,j}}]$. This leads to

$$\alpha_d=\min\left\{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[0.25^{K_{j,j}-K_{i,j}}(1+2\cdot 0.5^{K_{i,j}}-3\mathbb{1}\{K_{i,j}=0\})]}{\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[0.25^{K_{j,j}}16K_{i,j}]},1\right\}.$$

Finally, we compute the asymptotically worst-case variance

$$\overline{\mathrm{Var}}_{\eta_d^{\dagger}}(\widehat{\tau})=\frac{\alpha_d B^2}{N^2T}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\mathbb{E}[4^{K_{i,j}-K_{j,j}}(2+4\cdot 2^{-K_{i,j}}-16\cdot\alpha_d K_{i,j}4^{-K_{i,j}}-6\mathbb{1}\{K_{i,j}=0\})]}{\mathbb{E}[4^{-K_{j,j}}]}.$$

C.6. Proof of Theorem 4

Let us first define

$$\xi_T(t,t',s,s') = \begin{cases} \frac{4T-8}{(T-1)(T-3)} Y_{t'}(\mathbf{1}) & t' = t+1, 2 \le s \ne s' \le \frac{T+1}{2}, \\ -\frac{4T-8}{(T-1)(T-3)} Y_{t'}(\mathbf{0}) & t' = t+1, \frac{T+1}{2} \le s \ne s' \le T, \\ 0 & \textit{otherwise.} \end{cases} \tag{33}$$

Let π be a random permutation that shuffles the original indices:

$$\{2,3,\ldots,T-1,T\} \to \{\pi(2),\pi(3),\ldots,\pi(T-1),\pi(T)\}.$$

Given these, we can rewrite the estimator as

$$\widehat{\tau} = \sum_{t \ne t'}^{T} \xi_T(t,t',\pi(t),\pi(t')). \tag{34}$$

where $\sum_{t\ne t'}^{T}$ indicates $\sum_{t=2}^{T}\sum_{t'=2:t\ne t'}^{T}$. To derive the normal approximation of this, we adopt Stein's method of exchange pairs for double-index permutation statistics proposed in Reinert and Röllin (2009). Specifically, they construct an exchangeable pair as follows. Let t and t' be distributed uniformly over $1,\ldots,T-1$ conditioned that $t \ne t'$. Define the permutation $\pi' = (\pi(t)\pi(t')) \circ \pi$ so that π' is the permutation where $\pi'(s) = \pi(s)$ for all $k \ne t,t'$, and where $\pi'(t) = \pi(t')$ and $\pi'(t') = \pi(t)$. Let $V_1 = \widehat{\tau}$, and we define the other two random variables for proof purposes:

$$V_2 = \frac{1}{T-1}\sum_{t=2}^{T}\sum_{s,s'}^{T} \xi_T(t,s,\pi(t),s'), V_3 = \frac{1}{T-1}\sum_{t=2}^{T}\sum_{s,s'}^{T} \xi_T(s,t,s',\pi(t)).$$

Then we have $\boldsymbol{V}' = (V_1',V_2',V_3') = \boldsymbol{V}(\pi')$ to be the estimators with the exchange pair. For the random exchange pair (t,t'), we have the following equations:

$$\begin{aligned}
V_1' - V_1 &= \xi_T(t,t+1,\pi(t'),\pi(t+1)) + \xi_T(t',t'+1,\pi(t),\pi(t'+1)) \\
&\quad + \xi_T(t-1,t,\pi(t-1),\pi(t')) + \xi_T(t'-1,t',\pi(t'-1),\pi(t)) \\
&\quad - \xi_T(t,t+1,\pi(t),\pi(t+1)) - \xi_T(t',t'+1,\pi(t'),\pi(t'+1)) \\
&\quad - \xi_T(t-1,t,\pi(t-1),\pi(t)) - \xi_T(t'-1,t',\pi(t'-1),\pi(t'), \\
V_2' - V_2 &= \frac{1}{T-1}\sum_{s=2}^{T} \xi_T(t,t+1,\pi(t'),s) + \frac{1}{T-1}\sum_{s=2}^{T} \xi_T(t',t'+1,\pi(t),s) \\
&\quad - \frac{1}{T-1}\sum_{s=2}^{T} \xi_T(t,t+1,\pi(t),s) - \frac{1}{T-1}\sum_{s=2}^{T} \xi_T(t',t'+1,\pi(t'),s), \\
V_3' - V_3 &= \frac{1}{T-1}\sum_{s=2}^{T} \xi_T(t-1,t,s,\pi(t')) + \frac{1}{T-1}\sum_{s=2}^{T} \xi_T(t'-1,t',s,\pi(t)) \\
&\quad - \frac{1}{T-1}\sum_{s=2}^{T} \xi_T(t-1,t,s,\pi(t)) - \frac{1}{T-1}\sum_{s=2}^{T} \xi_T(t'-1,t',s,\pi(t')).
\end{aligned}$$

They further satisfy that

$$\mathbb{E}^{\boldsymbol{V}}\left(\boldsymbol{V}'-\boldsymbol{V}\right)=-\boldsymbol{\Lambda}\boldsymbol{V}+\boldsymbol{R} \tag{35}$$

where

$$\boldsymbol{\Lambda}=\frac{2}{T-2}\begin{pmatrix}\frac{2T-3}{T-1} & -1 & -1\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix},\quad \boldsymbol{R}=\left(-\frac{2}{(T-1)(T-2)}\sum_{t,t'}^{T}\xi_T(t,t',\pi(t'),\pi(t)),0,0\right).$$

To be self-contained, we re-state the following theorem to show the asymptotic normality.

Theorem 2 in Reinert and Röllin (2009). Assume that $(\boldsymbol{V},\boldsymbol{V}')$ is an exchangeable pair of random vectors such that

$$\mathbb{E}[\boldsymbol{V}]=\boldsymbol{0},\quad \mathbb{E}[\boldsymbol{V}\boldsymbol{V}^t]=\boldsymbol{\Sigma},$$

with $\boldsymbol{\Sigma}\in\mathbb{R}^{3\times 3}$ symmetric and positive definite. If (35) holds and $\boldsymbol{Z}$ has a 3-dimensional standard normal distribution, we have for every three times differentiable function h,

$$\left|\mathbb{E}h(\boldsymbol{V})-\mathbb{E}h\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}\right)\right|\le\frac{|h|_2}{4}A+\frac{|h|_3}{12}B+\left(|h|_1+\frac{3}{2}\|\boldsymbol{\Sigma}\|^{1/2}|h|_2\right)$$

where

$$\begin{aligned}\gamma^{(i)}&=\sum_{m=1}^{3}\left|\left(\boldsymbol{\Lambda}^{-1}\right)_{m,i}\right|\\ A&=\sum_{i,j=1}^{3}\gamma^{(i)}\sqrt{\operatorname{Var}\mathbb{E}^{\boldsymbol{V}}\left(V_i'-V_i\right)\left(V_j'-V_j\right)},\\ B&=\sum_{i,j,k=1}^{3}\gamma^{(i)}\mathbb{E}\left|\left(V_i'-V_i\right)\left(V_j'-V_j\right)\left(V_k'-V_k\right)\right|,\\ C&=\sum_{i=1}^{3}\gamma^{(i)}\sqrt{\operatorname{Var}R_i}.\end{aligned}$$

To apply the theorem, we first note that $\mathbb{E}[\boldsymbol{V}]=\boldsymbol{0}$ may not hold. Nevertheless, we can simply de-mean $\boldsymbol{V}$ by $\mathbb{E}[\boldsymbol{V}]$, and thus the condition is satisfied. Next, it is easy to see $\boldsymbol{\gamma}=O(T)$ and we need to characterize A,B,C using (33):

- A: Let us use the analysis of $\operatorname{Var}\mathbb{E}^{\boldsymbol{V}}(V_1'-V_1)^2$ as instance. First of all, we have

$$\begin{aligned}\mathbb{E}^{\boldsymbol{V}}(V_1'-V_1)^2&=\frac{1}{(T-1)(T-2)}\sum_{t\neq t'}^{T}(V_1(\pi')-V_1))^2\\ &=\frac{1}{(T-1)(T-2)}\sum_{t\neq t'}^{T}(\xi_T(t,t+1,\pi(t'),\pi(t+1))+\xi_T(t',t'+1,\pi(t),\pi(t'+1))\\ &+\xi_T(t-1,t,\pi(t-1),\pi(t'))+\xi_T(t'-1,t',\pi(t'-1),\pi(t))\\ &-\xi_T(t,t+1,\pi(t),\pi(t+1))-\xi_T(t',t'+1,\pi(t'),\pi(t'+1))\\ &-\xi_T(t-1,t,\pi(t-1),\pi(t))-\xi_T(t'-1,t',\pi(t'-1),\pi(t')))^2.\end{aligned}$$

Let π' be the permutation with the exchange pair t, t' and π'' be the permutation with the exchange pair s, s'. To analyze the variance of $\mathbb{E}^{\boldsymbol{V}}(V_1' - V_1)^2$, it suffices to see that

$$\mathrm{Cov}((V_1(\pi') - V_1))^2, (V_1(\pi'') - V_1))^2) = O\left(\frac{1}{T^4}\right).$$

This further leads to

$$\sqrt{\mathrm{Var}\,\mathbb{E}^{\boldsymbol{V}}(V_1' - V_1)^2} = O\left(\frac{1}{T^2}\right)$$

Following the same procedure, we can obtain that

$$\sqrt{\mathrm{Var}\,\mathbb{E}^{\boldsymbol{V}}\left(V_i' - V_i\right)\left(V_j' - V_j\right)} = O\left(\frac{1}{T^2}\right).$$

- B: Let us use the analysis of $\mathbb{E}[|(V_1' - V_1)^3|]$ as instance. We take the conditioning on the exchange pair (t, t'), which gives

$$\begin{aligned}\mathbb{E}[|(V_1' - V_1)^3|] = \frac{1}{(T-1)(T-2)} \sum_{t \neq t'}^{T} \mathbb{E}[|&(\xi_T(t, t+1, \pi(t'), \pi(t+1)) + \xi_T(t', t'+1, \pi(t), \pi(t'+1)) \\ &+ \xi_T(t-1, t, \pi(t-1), \pi(t')) + \xi_T(t'-1, t', \pi(t'-1), \pi(t)) \\ &- \xi_T(t, t+1, \pi(t), \pi(t+1)) - \xi_T(t', t'+1, \pi(t'), \pi(t'+1)) \\ &- \xi_T(t-1, t, \pi(t-1), \pi(t)) - \xi_T(t'-1, t', \pi(t'-1), \pi(t')))^3|] \leq \left(\frac{8B}{T}\right)^3 = O\left(\frac{1}{T^3}\right)\end{aligned}$$

  Following the same procedure, we can obtain that $\mathbb{E}\left|\left(V_i' - V_i\right)\left(V_j' - V_j\right)\left(V_k' - V_k\right)\right| = O\left(\frac{1}{T^3}\right)$.
- C: Since $R_2 = R_3 = 0$, we simply need to consider R_1.

$$\begin{aligned}\sqrt{\mathrm{Var}\,R_1} &= \frac{2}{(T-1)(T-2)} \sqrt{\mathrm{Var}\left(\sum_{t \neq t'}^{T} \xi_T(t, t', \pi(t'), \pi(t))\right)} \\ &= \frac{2}{(T-1)(T-2)} \sqrt{\mathrm{Var}\,(V_1)} = O\left(\frac{1}{T^{2.5}}\right)\end{aligned}$$

Putting A, B, C together, we have

$$\left|\mathbb{E}h(\boldsymbol{V}) - \mathbb{E}h\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}\right)\right| = O\left(\frac{1}{T}\right).$$

Note that $\Sigma_{1,1}^{1/2} = \sqrt{\mathrm{Var}_{\eta^\dagger}(\widehat{\tau})}$ has an order of $\frac{1}{\sqrt{T}}$. If we normalize V_1 by the standard deviation, this leads to the typical rate of convergence $O\left(\frac{1}{\sqrt{T}}\right)$ for asymptotic normality.

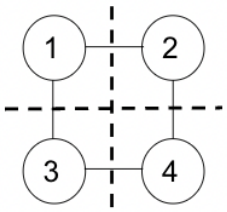


Figure 18 **The 2×2 example**

C.7. Proof of Proposition 3

We first write the original variance decomposition (3):

$$\mathrm{Var}_{\eta^\dagger}(\widehat{\tau}|\mathbb{Y}) = \frac{(T-2)}{(T-3)(T-4)}\left(\frac{8(T-4)\,(S^t+S^c)}{T-3} - 4S^{ct} + R^{ct} - \frac{2\tau^2}{T-2}\right)$$

Removing the non-positive parts gives

$$\mathrm{Var}_{\eta^\dagger}(\widehat{\tau}|\mathbb{Y}) \leq \frac{(T-2)}{(T-3)(T-4)}\left(\frac{8(T-4)\,(S^t+S^c)}{T-3} + R^{ct}\right)$$

Next, because of the non-negative outcomes, we have

$$R^{ct} = \frac{\sum_{t=2}^{T}\tau_t^2}{(T-1)} \leq \frac{\sum_{t=2}^{T} Y_t^2(\mathbf{1}) + Y_t^2(\mathbf{0})}{(T-1)}.$$

This finally leads to the upper bound

$$\mathrm{Var}_{\eta^\dagger}(\widehat{\tau}|\mathbb{Y}) \leq \frac{(T-2)}{(T-3)(T-4)}\left(\frac{8(T-4)\,(S^t+S^c)}{T-3} + \frac{\sum_{t=2}^{T} Y_t^2(\mathbf{1}) + Y_t^2(\mathbf{0})}{(T-1)}\right)$$

Following the similar argument in Imbens and Rubin (2015), we obtain an unbiased estimate for the upper bound

$$\hat{\sigma}^2_{\mathrm{U}} = \frac{(T-2)}{(T-3)^2(T-4)}\left(8(T-4)\left(\widehat{S}^t + \widehat{S}^c\right) + \frac{4(T-2)\sum_{t=2}^{T} Y_t^2 \mathbb{1}\{w_{t-1} = w_t\}}{(T-1)}\right)$$

where $\widehat{S}^t, \widehat{S}^c$ are the sample estimates and Y_t is the observed outcome.

C.8. Proof of Corollary 2

Central Limit Theorem To do the inference over the panel, we follow the proof idea of the normal approximation C.6 derived for the single-unit case. Let us consider the toy example (Figure 18) to illustrate how we extend the double-index permutation statistics in (33). We construct the random permutations for each unit:

$$\{2, 3, \ldots, T-1, T\} \to \{\pi_i(2), \pi_i(3), \ldots, \pi_i(T-1), \pi_i(T)\}.$$

and we set $\boldsymbol{\pi}(t) = (\pi_1(t), \pi_2(t), \pi_3(t), \pi_4(t))$. We rewrite the global average treatment effect as

$$\widehat{\tau} = \frac{1}{4}\sum_{i=1}^{4}\sum_{t\neq t'}^{T} \xi_T^i(t, t', \boldsymbol{\pi}(t), \boldsymbol{\pi}(t'))$$

where

$$\xi_T^i(t,t',\boldsymbol{s},\boldsymbol{s}') = \begin{cases} \frac{(4T-8)^3}{(T-1)(T-3)^3} Y_{i,t'}(\mathbf{1}) & t'=t+1, 2\le s_j\neq s_j'\le \frac{T+1}{2}, \forall j\in\mathcal{N}(i), \\ -\frac{(4T-8)^3}{(T-1)(T-3)^3} Y_{i,t'}(\mathbf{0}) & t'=t+1, \frac{T+1}{2}\le s_j\neq s_j'\le T, \forall j\in\mathcal{N}(i), \\ 0 & otherwise. \end{cases} \tag{36}$$

We define $\boldsymbol{V}$ as follows:

$$\begin{aligned} V_1 &= \frac{1}{4}\sum_{i=1}^{4}\sum_{t\neq t'}^{T}\xi_T^i(t,t',\boldsymbol{\pi}(t),\boldsymbol{\pi}(t')) \\ V_2 &= \frac{1}{4(T-1)}\sum_{i=1}^{4}\sum_{s=2}^{T}\sum_{t\neq t'}^{T}\xi_T^i(s,t,\boldsymbol{\pi}(s),\boldsymbol{\pi}(t')) \\ V_3 &= \frac{1}{4(T-1)}\sum_{i=1}^{4}\sum_{s=2}^{T}\sum_{t\neq t'}^{T}\xi_T^i(t,s,\boldsymbol{\pi}(t'),\boldsymbol{\pi}(s)) \end{aligned}$$

Now for the exchange pair (t,t'), we do the swapping for all permutations such that $\boldsymbol{\pi}'(t)=\boldsymbol{\pi}(t')$ and we define $\boldsymbol{V}'$ accordingly. After some calculations, we could derive the same structure $\mathbb{E}^{\boldsymbol{V}}(\boldsymbol{V}'-\boldsymbol{V}) = \boldsymbol{\Lambda}\boldsymbol{V}+\boldsymbol{R}$ as in the single-unit case. Since $\xi_T^i(t,t',\boldsymbol{s},\boldsymbol{s}')$ preserves the same order $\frac{1}{T}$ as in the single-unit case, following the same argument in A.3 gives:

$$\left|\mathbb{E}h(\boldsymbol{V}) - \mathbb{E}h\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}\right)\right| = O\left(\frac{1}{T}\right). \tag{37}$$

We can easily generalize the argument to more units with any clustering. The only difference is that we need to calibrate $\xi_T^i(t,t',\boldsymbol{s},\boldsymbol{s}')$ based on its propensity score $\mathbb{P}(\boldsymbol{W}_{\mathcal{N}(i),t-1:t}=\mathbf{1})$. Finally, we have

$$\frac{\widehat{\tau}-\tau}{\sqrt{\mathrm{Var}_{\eta_d^\dagger}(\widehat{\tau}|\mathbb{Y})}} \xrightarrow{D} \mathcal{N}(0,1), \quad \text{as } T\to\infty.$$

Variance Estimation Similar to the single-unit case, we further derive a conservative variance estimate for the panel setting. To this end, we write down the variance as:

$$\begin{aligned} \mathrm{Var}_{\eta_d^\dagger}(\widehat{\tau}|\mathbb{Y}) = {} & \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\theta_{i,j}^0(K,T)(S_{i,j}^t+S_{i,j}^c) \\ & + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\theta_{i,j}^1(K,T)S_{i,j}^{ct} \\ & + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\theta_{i,j}^2(K,T)}{T-1}\sum_{t=2}^{T}(Y_{i,t}(\mathbf{1})Y_{j,t}(\mathbf{1})+Y_{i,t}(\mathbf{0})Y_{j,t}(\mathbf{0})) \\ & + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\theta_{i,j}^3(K,T)}{T-1}\sum_{t=2}^{T}(Y_{i,t}(\mathbf{1})Y_{j,t}(\mathbf{0})+Y_{i,t}(\mathbf{0})Y_{j,t}(\mathbf{1})). \\ & + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\theta_{i,j}^4(K,T)}{T-1}\sum_{t=2}^{T}(Y_{i,t}(\mathbf{1})Y_{j,t+1}(\mathbf{1})+Y_{j,t}(\mathbf{1})Y_{i,t+1}(\mathbf{1})+Y_{i,t}(\mathbf{0})Y_{j,t+1}(\mathbf{0})+Y_{j,t}(\mathbf{0})Y_{i,t+1}(\mathbf{0})) \\ & + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\theta_{i,j}^5(K,T)}{T-1}\sum_{t=2}^{T}(Y_{i,t}(\mathbf{1})Y_{j,t+1}(\mathbf{0})+Y_{j,t}(\mathbf{0})Y_{i,t+1}(\mathbf{1})+Y_{i,t}(\mathbf{0})Y_{j,t+1}(\mathbf{1})+Y_{j,t}(\mathbf{1})Y_{i,t+1}(\mathbf{0})). \end{aligned}$$

In order to provide an upper bound for the variance, we need to analyze the sign of all coefficients. We first prove that $\theta^1_{i,j}(K,T) \leq 0$. To see this, it suffices to show that

$$\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}(\lambda_3\lambda_4)^{K_{i,j}}] \geq \mathbb{E}[\lambda_0^{K(j,j)}].$$

We derive this inequality by conditioning on $K_{i,j}$. When $K_{i,j}=0$, we have

$$\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}(\lambda_3\lambda_4)^{K_{i,j}}|K_{i,j}=0] = \mathbb{E}[\lambda_0^{K_{j,j}}|K_{i,j}=0].$$

When $K_{i,j} \geq 1$, as $\lambda_3\lambda_4 \geq \lambda_0$, we have

$$\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}(\lambda_3\lambda_4)^{K_{i,j}}|K_{i,j}\geq 1] \geq \mathbb{E}[\lambda_0^{K_{j,j}}|K_{i,j}\geq 1].$$

Putting two scenarios together, we show that $\theta^1_{i,j}(K,T) \leq 0$.

We then prove that $\theta^2_{i,j}(K,T) \geq 0, \theta^3_{i,j}(K,T) \leq 0, \theta^4_{i,j}(K,T) \leq 0, \theta^5_{i,j}(K,T) \geq 0$. Let us take $\theta^3_{i,j}(K,T)$ as an example, and the argument for the remaining ones are almost the same. To see $\theta^3_{i,j}(K,T) \leq 0$, it suffices to show that

$$\frac{T-1}{2}\mathbb{E}[\lambda_0^{K_{j,j}}] - \mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}\mathbb{1}\{K_{i,j}=0\}] - \frac{T-3}{2}\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}(\lambda_3\lambda_4)^{K_{i,j}}] \leq 0.$$

We take the conditioning on $K_{i,j}$. When $K_{i,j}=0$, LHS becomes

$$\frac{T-1}{2}\mathbb{E}[\lambda_0^{K_{j,j}}|K_{i,j}=0] - \mathbb{E}[\lambda_0^{K_{j,j}}|K_{i,j}=0] - \frac{T-3}{2} = 0.$$

When $K_{i,j} \geq 1$, LHS becomes

$$\begin{aligned}
&\frac{T-1}{2}\mathbb{E}[\lambda_0^{K_{j,j}}|K_{i,j}\geq 1] - \frac{T-3}{2}\mathbb{E}[\lambda_0^{K_{j,j}-K_{i,j}}(\lambda_3\lambda_4)^{K_{i,j}}|K_{i,j}\geq 1] \\
<&\frac{T-1}{2}\mathbb{E}[\lambda_0^{K_{j,j}}|K_{i,j}\geq 1] - \frac{T-1}{2}\mathbb{E}[\lambda_0^{K_{j,j}}|K_{i,j}\geq 1] = 0
\end{aligned}$$

where the inequality holds because $\dfrac{\lambda_3\lambda_4}{\lambda_0} = \dfrac{(T-1)(T-2)}{(T-3)(T-4)} > \dfrac{(T-1)}{(T-3)}$. Putting two scenarios together by the tower rule, we prove that $\theta^3_{i,j}(K,T) \leq 0$.

Now we are ready to derive an upper bound for the variance. First of all, unlike the single-unit case where the variance of treatment effect is definitely non-negative, we can not guarantee the covariance of treatment effect $S^{ct}_{i,j}$ between units i and j is non-negative. This could be an assumption if one finds it reasonable in the business context. For example, for both units, the revenue gains of a new algorithm might be high in peak hours and low in non-peak hours. We impose this assumption first to give a variance estimate and get back later to discuss the case when it does not hold.

For the remaining terms with positive coefficients, we use the basic inequality to obtain an upper bound, e.g.

$$Y_{i,t}(\mathbf{1})Y_{j,t}(\mathbf{1}) \leq \frac{Y^2_{i,t}(\mathbf{1}) + Y^2_{j,t}(\mathbf{1})}{2}.$$

For the terms with negative coefficients, we use the non-negativity to obtain an upper bound, e.g.

$$-Y_{i,t}(\mathbf{1})Y_{j,t}(\mathbf{1}) \leq 0.$$

Finally, we obtain the following upper bound of the variance:

$$\begin{aligned}
&\mathrm{Var}_{\eta_d^\dagger}(\widehat{\tau}|\mathbb{Y}) \\
&\leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\theta_{i,j}^0(K,T)(S_{i,j}^t + S_{i,j}^c) + \frac{\theta_{i,j}^2(K,T) + 2\theta_{i,j}^5(K,T)}{T-1}\sum_{t=2}^{T}\frac{Y_{i,t}^2(\mathbf{1}) + Y_{i,t}^2(\mathbf{0}) + Y_{j,t}^2(\mathbf{1}) + Y_{j,t}^2(\mathbf{0})}{2}\right).
\end{aligned}$$

Using the same technique in the single-unit case, we construct the following unbiased estimate:

$$\begin{aligned}
\hat{\sigma}_{\mathrm{U}}^2 = &\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\theta_{i,j}^0(K,T)(\widehat{S}_{i,j}^t + \widehat{S}_{i,j}^c) \\
&+ \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\theta_{i,j}^2(K,T) + 2\theta_{i,j}^5(K,T)}{2(T-1)}\sum_{t=2}^{T}\left(\frac{Y_{i,t}^2\mathbb{1}\{\boldsymbol{w}_{\mathcal{N}(i),t-1:t} \in \{\mathbf{1},\mathbf{0}\}\}}{\mathbb{P}(\boldsymbol{W}_{\mathcal{N}(i),t-1:t} = \mathbf{1})} + \frac{Y_{j,t}^2\mathbb{1}\{\boldsymbol{w}_{\mathcal{N}(j),t-1:t} \in \{\mathbf{1},\mathbf{0}\}\}}{\mathbb{P}(\boldsymbol{W}_{\mathcal{N}(j),t-1:t} = \mathbf{1})}\right)
\end{aligned}$$

where

$$\widehat{S}_{i,j}^t = \frac{\sum_{t=2}^{T}\left(Y_{i,t-1} + Y_{i,t} - 2\frac{\sum_{t'=2}^{T} Y_{i,t}\cdot\mathbb{1}\{\boldsymbol{w}_{\mathcal{N}(i),t'-1:t'}=\mathbf{1}\}}{\sum_{t'=2}^{T}\mathbb{1}\{\boldsymbol{w}_{\mathcal{N}(i),t'-1:t'}=\mathbf{1}\}}\right)\left(Y_{j,t-1} + Y_{j,t} - 2\frac{\sum_{t'=2}^{T} Y_{j,t}\cdot\mathbb{1}\{\boldsymbol{w}_{\mathcal{N}(j),t'-1:t'}=\mathbf{1}\}}{\sum_{t'=2}^{T}\mathbb{1}\{\boldsymbol{w}_{\mathcal{N}(j),t'-1:t'}=\mathbf{1}\}}\right)\mathbb{1}\{\boldsymbol{w}_{\mathcal{N}(i)\cup\mathcal{N}(j),t-2:t}=\mathbf{1}\}}{4\left(\sum_{t=2}^{T}\mathbb{1}\{\boldsymbol{w}_{\mathcal{N}(i)\cup\mathcal{N}(j),t-2:t}=\mathbf{1}\}-1\right)}$$

and $\widehat{S}_{i,j}^c$ is defined similarly.

Suppose the assumption on non-negative covariance of treatment effect is not valid, given that $S_{i,j}^{ct}$ between two specific units is inestimable, we need some reformulation. In particular, we write

$$\begin{aligned}
S_{i,j}^{ct} =& S_{i,j}^c + S_{i,j}^t - \frac{1}{T-1}\sum_{t=2}^{T}\left(\frac{Y_{i,t}(\mathbf{1}) + Y_{i,t+1}(\mathbf{1})}{2} - \overline{Y}_i(\mathbf{1})\right)\left(\frac{Y_{j,t}(\mathbf{0}) + Y_{j,t+1}(\mathbf{0})}{2} - \overline{Y}_j(\mathbf{0})\right) \\
&- \frac{1}{T-1}\sum_{t=2}^{T}\left(\frac{Y_{j,t}(\mathbf{1}) + Y_{j,t+1}(\mathbf{1})}{2} - \overline{Y}_j(\mathbf{1})\right)\left(\frac{Y_{i,t}(\mathbf{0}) + Y_{i,t+1}(\mathbf{0})}{2} - \overline{Y}_i(\mathbf{0})\right) \\
=& S_{i,j}^c + S_{i,j}^t + \overline{Y}_i(\mathbf{1})\overline{Y}_j(\mathbf{0}) + \overline{Y}_i(\mathbf{0})\overline{Y}_j(\mathbf{1}) \\
&- \frac{1}{4(T-1)}\sum_{t=2}^{T}\left(Y_{i,t}(\mathbf{1})Y_{j,t}(\mathbf{0}) + Y_{i,t+1}(\mathbf{1})Y_{j,t}(\mathbf{0}) + Y_{i,t}(\mathbf{1})Y_{j,t+1}(\mathbf{0}) + Y_{i,t+1}(\mathbf{1})Y_{j,t+1}(\mathbf{0})\right) \\
&- \frac{1}{4(T-1)}\sum_{t=2}^{T}\left(Y_{j,t}(\mathbf{1})Y_{i,t}(\mathbf{0}) + Y_{j,t+1}(\mathbf{1})Y_{i,t}(\mathbf{0}) + Y_{j,t}(\mathbf{1})Y_{i,t+1}(\mathbf{0}) + Y_{j,t+1}(\mathbf{1})Y_{i,t+1}(\mathbf{0})\right).
\end{aligned}$$

Given that $\theta_{i,j}^1(K,T) \leq 0$, we have a new upper bound

$$\begin{aligned}
\mathrm{Var}_{\eta_d^\dagger}(\widehat{\tau}|\mathbb{Y}) \leq &\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left((\theta_{i,j}^0(K,T) + \theta_{i,j}^1(K,T))(S_{i,j}^t + S_{i,j}^c)\right) \\
&+ \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\theta_{i,j}^2(K,T) + 2\theta_{i,j}^5(K,T) - \theta_{i,j}^1(K,T)}{T-1}\sum_{t=2}^{T}\frac{Y_{i,t}^2(\mathbf{1}) + Y_{i,t}^2(\mathbf{0}) + Y_{j,t}^2(\mathbf{1}) + Y_{j,t}^2(\mathbf{0})}{2},
\end{aligned}$$

for which an unbiased estimate can be constructed as above.