Title

Rate-optimal cluster-randomized designs for spatial interference

Permalink

https://escholarship.org/uc/item/8t44s021

Journal

The Annals of Statistics, 50(5)

ISSN

0090-5364

Author

Leung, Michael P

Publication Date

2022

DOI

10.1214/22-aos2224

Copyright Information

Peer reviewed

RATE-OPTIMAL CLUSTER-RANDOMIZED DESIGNS FOR SPATIAL INTERFERENCE

BY MICHAEL P. LEUNG[1]

[1]*Department of Economics, UC Santa Cruz, leungm@ucsc.edu*

We consider a potential outcomes model in which interference may be present between any two units but the extent of interference diminishes with spatial distance. The causal estimand is the global average treatment effect, which compares outcomes under the counterfactuals that all or no units are treated. We study a class of designs in which space is partitioned into clusters that are randomized into treatment and control. For each design, we estimate the treatment effect using a Horvitz-Thompson estimator that compares the average outcomes of units with all or no neighbors treated, where the neighborhood radius is of the same order as the cluster size dictated by the design. We derive the estimator's rate of convergence as a function of the design and degree of interference and use this to obtain estimator-design pairs that achieve near-optimal rates of convergence under relatively minimal assumptions on interference. We prove that the estimators are asymptotically normal and provide a variance estimator. For practical implementation of the designs, we suggest partitioning space using clustering algorithms.

1. Introduction. Consider a population of n experimental units. Denote by $Y_i(\boldsymbol{d})$ the potential outcome of unit i under the counterfactual that the population is assigned treatments according to the vector $\boldsymbol{d} = (d_i)_{i=1}^n \in \{0,1\}^n$, where $d_i = 1$ ($d_i = 0$) implies unit i is assigned to treatment (control). Treatments assigned to alters can influence the ego since $Y_i(\boldsymbol{d})$ is a function of d_j for $j \neq i$, what is known as *interference*.

An important estimand of practical interest is the *global average treatment effect*

$$\theta_n = \frac{1}{n} \sum_{i \in \mathcal{N}_n} \left(Y_i(\mathbf{1}_n) - Y_i(\mathbf{0}_n) \right)$$

where $\mathbf{1}_n$ ($\mathbf{0}_n$) is the n-dimensional vector of ones (zeros). This compares average outcomes under the counterfactuals that all or no units are treated. Each average can only be directly observed in the data under an extreme design that assigns all units to the same treatment arm, which would necessarily preclude observation of the other counterfactual. Common designs used in the literature, including those studied here, assign different units to different treatment arms, so neither average is directly observed in the data. Nonetheless, we show that asymptotic inference on θ_n is possible for a class of cluster-randomized designs under spatial interference where the degree of interference diminishes with distance.

Many phenomena diffuse primarily through physical interaction. The government of a large city may wish to compare the effect of two different policing strategies on crime, but more intensive policing in one neighborhood may displace crime to adjacent neighborhoods [8, 49]. A rideshare company may wish to compare the performance of two different pricing algorithms, but these may induce behavior that generates spatial externalities, such as congestion. Other phenomena exhibiting spatial interference include infectious diseases in animal [14] and human [35] populations, pollution [18], and environmental conservation programs [36].
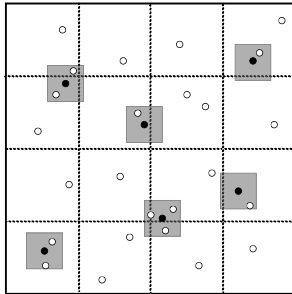

FIG 1. *White and black dots depict units. White squares correspond to clusters and gray squares to neighborhoods of black units used to construct the estimator.*

Much of the existing literature assumes that interference is summarized by a low-dimensional exposure mapping and that units are individually randomized into treatment or control either via Bernoulli or complete randomization [e.g. 3, 7, 16, 34, 45]. Jagadeesan et al. [23] and Ugander et al. [46] also utilize exposure mappings but depart from unit-level randomization. They propose new designs that introduce cluster dependence in unit-level assignments in order to improve estimator precision. We build on this literature by (1) studying rate-optimal choices of both cluster-randomized designs and Horvitz-Thompson estimators, (2) avoiding exposure mapping restrictions on interference, which can be quite strong [15], and (3) developing a distributional theory for the estimator and a variance estimator.

Regarding (2), most exposure mappings used in the literature imply that only units within a small, known distance from the ego can interfere with the ego's outcome. We instead study a weaker restriction on interference similar to [31], which states that the degree of interference decreases with distance but does not necessarily zero out at any given distance. This is analogous to the widespread use of mixing-type conditions in the time series and spatial literature instead of m-dependence because the latter rules out interesting forms of autocorrelation, including models as basic as AR(1).

Regarding (1), we study cluster-randomized designs in which units are partitioned into spatial clusters, clusters are independently randomized into treatment and control, and the assignment of a unit is dictated by the assignment of its cluster. By introducing correlation in assignments, such designs can avoid overlap problems common under Bernoulli randomization, which improves the rate of convergence. For analytical tractability, we focus on designs in which clusters are equally sized squares, each design distinguished by the number of such squares. We pair each design with a Horvitz-Thompson estimator that compares the average outcomes of units with all or no treated neighbors, where the neighborhood radius is of the same order as the cluster size dictated by the design. See Figure 1 for a depiction of a hypothetical design and neighborhoods used to construct the estimator.

Our results inform how the analyst should choose the number of clusters (and hence, the cluster size and neighborhood radius of the estimator) to minimize the rate of convergence of the estimator. Notably, existing work on cluster randomization with interference utilizes small clusters (those with asymptotically bounded size). We show that such designs are generally asymptotically biased under the weaker restriction on interference we impose, which motivates the large-cluster designs we study.

Finally, regarding (3), we show that the estimator is asymptotically normal and provide a variance estimator. These results appear to be novel, as no existing central limit theorems seem to apply to our setup in which treatments exhibit cluster dependence, clusters can be large, and units in different clusters are spatially dependent due to interference. As usual, the variance estimator is biased due to heterogeneity in unit-level treatment effects. However,

we show that, in a superpopulation setting in which potential outcomes are weakly spatially dependent, the bias is asymptotically negligible.

Based on our theory, we provide practical recommendations for implementing cluster-randomized designs in §3.3. Of course, rate-optimality results do not determine the choice of nonasymptotic constants that are often important in practice under smaller sample sizes. Still, they constitute an important first step toward designing practical procedures. Due to the generality of the setting, which imposes quite minimal assumptions on interference, it seems reasonable to first study rate-optimality, as finite-sample optimality appears to require substantial additional structure on the problem. We note that existing results on graph cluster randomization, which require stronger restrictions on interference than this paper, are nonetheless limited to rates, and how "best" to construct clusters in practice has been an open question.

1.1. *Related Literature.* Most of the literature supposes interference is mediated by a network. Studying optimal design in this setting is difficult because network clusters can be highly heterogeneous in topology, and their graph-theoretic properties can closely depend on the generative model of the network [32]. We study spatial interference, and to make the optimal design problem analytically tractable, we focus on a class of designs that partitions space into equally sized squares while exploring in simulations the performance of more realistic designs that partition using clustering algorithms. We discuss in §6 the (pessimistic) prospects of extending our approach to network interference.

There is relatively little work on optimal experimental design under interference. Viviano [50] proposes variance-minimizing two-wave experiments under network interference. Baird et al. [5] study the power of randomized saturation designs under partial interference.

A recent literature studies designs for interference that depart from unit-level randomization. A key paper motivating our work is [46], who propose graph cluster randomization designs under network interference. Ugander and Yin [47] study a new variant of these designs, and [19] consider related designs for bipartite experiments. These papers assume interference is summarized by exposure mappings, which enables the construction of unbiased estimators and use of designs in which clusters are small. Under our weaker restriction on interference, we show that large clusters are required to reduce bias, which creates a bias-variance trade-off.

Eckles et al. [15] show that graph cluster randomization can reduce the bias of common estimators for θ_n in the absence of correctly specified exposure mappings. Pouget-Abadie et al. [40] propose two-stage cluster-randomized designs to minimize bias under a monotonicity restriction on interference. Several papers [6, 23, 44] study linear potential outcome models and propose designs targeting the direct average treatment effect, rather than θ_n. Under a normal-sum model, [6] compute the mean-squared error of the difference-in-means estimator, which they use to suggest model-assisted designs.

The aforementioned papers on cluster randomization target global effects such as θ_n [also see 9, 10]. Much of the literature on interference considers fundamentally different estimands defined by exposure mappings. When these mappings are misspecified, the estimands are functions of assignment probabilities, in which case their interpretations can be specific to the experiments run [42, 43]. Hu et al. [21] (§5) views this as "largely unavoidable" in nonparametric settings with interference. Our results show that inference on θ_n, which avoids this issue, is possible under restrictions on interference weaker than those typically used in the literature. Additionally, papers in the literature impose an overlap assumption, which implicitly restricts the estimand [31]. We study cluster-randomized designs that directly satisfy overlap.

There is a large literature on cluster-randomized trials [e.g. 20, 37]. This literature predominantly studies partial interference, meaning that units are divided into clusters such that

those in distinct clusters do not interfere. That is, the clusters themselves impose restrictions on interference. In our setting, clusters are determined by the design and do not restrict interference.

Finally, [4], [39], and [51] study spatial interference in a different "bipartite" setting in which treatments are assigned to units or locations that are distinct from the units whose outcomes are of interest. This shares some similarities with spatial cluster randomization, where different spatial regions are randomized into treatment, so some of the ideas here may be applicable to optimal design there.

1.2. *Outline.* The next section defines the model of spatial interference and the class of designs and estimators studied. In §3, we derive the estimator's rate of convergence, discuss rate-optimal designs, and provide practical design recommendations. In §4, we prove that the estimator is asymptotically normal, propose a variance estimator, and characterize its asymptotic properties. We report results from a simulation study in §5, exploring the use of spectral clustering to implement the designs. Finally, §6 concludes.

2. Setup. Let $\mathcal{N}_n$ be a set of n units. We study experiments in which units are cluster-randomized into treatment and control, postponing to §2.2 the specifics of the design. For each $i \in \mathcal{N}_n$, let D_i be a binary random variable where $D_i = 1$ indicates that i is assigned to treatment and $D_i = 0$ indicates assignment to control. Let $\boldsymbol{D} = (D_i)_{i \in \mathcal{N}_n}$ be the vector of realized treatments and $\boldsymbol{d} = (d_i)_{i \in \mathcal{N}_n} \in \{0, 1\}^n$ denote a non-random vector of counterfactual treatments. Recall from §1 that $Y_i(\boldsymbol{d})$ is the *potential outcome* of unit i under the counterfactual that units are assigned treatments according to $\boldsymbol{d}$. Formally, for each $n \in \mathbb{N}$ and $i \in \mathcal{N}_n$, $Y_i(\cdot)$ is a non-random function from $\{0, 1\}^n$ to $\mathbb{R}$. We denote i's factual, or observed, outcome by $Y_i = Y_i(\boldsymbol{D})$ and maintain the standard assumption that potential outcomes are uniformly bounded.

ASSUMPTION 1 (Bounded Outcomes). $\sup_{n \in \mathbb{N}} \max_{i \in \mathcal{N}_n} \max_{\boldsymbol{d} \in \{0,1\}^n} |Y_i(\boldsymbol{d})| < \infty$.

2.1. *Spatial Interference.* Thus far, the model allows for unrestricted interference in the sense that $Y_i(\boldsymbol{d})$ may vary essentially arbitrarily in any component of $\boldsymbol{d}$. In order to obtain a positive result on asymptotic inference, it is necessary to impose restrictions on interference to establish some form of weak dependence across unit outcomes. The existing literature primarily focuses on restrictions captured by K-*neighborhood exposure mappings*, which imply that D_j can only interfere with $Y_i(\boldsymbol{D})$ if the distance between i, j is at most K. We will discuss how this assumption is potentially restrictive and establish results under weaker conditions.

We assume each unit is located in $\mathbb{R}^2$. Label each unit by its location, so that $\mathcal{N}_n \subset \mathbb{R}^2$, and equip this space with the sup metric $\rho(i, j) = \max_{t=1,2} |i_t - j_t|$ for $i = (i_1, i_2)$, $j = (j_1, j_2)$, and $i, j \in \mathbb{R}^2$. Let $Q(i, r) = \{j \in \mathbb{R}^2 : \rho(i, j) \leqslant r\}$, the ball of radius r centered at i. Under the sup metric, balls are squares, and the radius is half the side length of the square. Letting $\boldsymbol{0}$ denote the origin, we consider a sequence of *population regions* $\{Q(\boldsymbol{0}, R_n)\}_{n \in \mathbb{N}}$ such that

$$\mathcal{N}_n \subset Q(\boldsymbol{0}, R_n), \quad \text{where} \quad R_n n^{-1/2} \to c \in (0, \infty).$$

That is, units are located in the square $Q(\boldsymbol{0}, R_n)$ with growing radius R_n. Combined with the next increasing domain assumption, the number of units in the region is $O(n)$, but throughout, we will simply assume the number is exactly n.

ASSUMPTION 2 (Increasing Domain). There exists $\rho_0 > 0$ such that, for any $n \in \mathbb{N}$ and $i, j \in \mathcal{N}_n$, $\rho(i, j) \geqslant \rho_0$.

This allows units to be arbitrarily irregularly spaced, subject to being minimally separated by some distance ρ_0, a widely used sampling framework in the spatial literature [e.g. 25]. In contrast, "infill" asymptotic approaches that do not require minimal separation and instead assume increasingly dense sampling from a fixed region can yield nonstandard limiting behavior [28]. For some applications, the spatial distribution of units may exhibit "hotspots" with unusually high densities, perhaps making the infill approach more plausible. Some work adopts a hybrid of the two approaches [29, 30], and it may be possible to extend our results to this framework.

Let $\mathbb{R}_+$ denote the set of non-negative reals and

$$\mathcal{N}(i,K) = Q(i,K) \cap \mathcal{N}_n$$

denote the K-*neighborhood* of i. We study the following model of interference similar to that proposed by [31].

ASSUMPTION 3 (Interference). There exists a non-increasing function $\psi\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that $\psi(0) \in (0,\infty)$, $\sum_{s=1}^{\infty} s\,\psi(s) < \infty$, and, for all $s \in \mathbb{R}_+$,

$$\sup_{n\in\mathbb{N}} \max_{i\in\mathcal{N}_n} \max\left\{|Y_i(\boldsymbol{d}) - Y_i(\boldsymbol{d}')|\colon \boldsymbol{d},\boldsymbol{d}' \in \{0,1\}^n, d_j = d_j'\ \forall j \in \mathcal{N}(i,s)\right\} \leqslant \psi(s).$$

To interpret this, observe that $\max\left\{|Y_i(\boldsymbol{d}) - Y_i(\boldsymbol{d}')|\colon \boldsymbol{d},\boldsymbol{d}' \in \{0,1\}^n, d_j = d_j'\ \forall j \in \mathcal{N}(i,s)\right\}$ maximizes over pairs of treatment assignment vectors that fix the assignments of units in i's s-neighborhood but allow assignments to freely vary outside of this neighborhood. It therefore measures the degree of spatial interference in terms of the maximum change to i's potential outcome caused by manipulating treatments assigned to units k "distant" from i in the sense that $\rho(i,k) > s$. The assumption requires the degree of interference to vanish with the neighborhood radius s so that treatments assigned to more distant alters interfere less with the ego. The rate at which interference vanishes is controlled by $\psi(s)$, which is required to decay at a rate faster than s^{-2}.

REMARK 1 (Necessity of rate condition). Assumption 3(b) of [25] and Assumption 4(c) of [27] impose the same minimum rate of decay as Assumption 3 on various measures of spatial dependence (mixing or near-epoch dependence coefficients) to establish central limit theorems. If the rate is slower, then the variance can be infinite. For example, consider a spatial process $\{Z_i\}_{i\in\mathcal{N}_n}$ such that units are positioned on the integer lattice $\mathbb{Z}^2$ and $\mathrm{Cov}(Z_i, Z_j) = f(\rho(i,j))$ for some function $f(\cdot)$ for any i,j. Then

$$\mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{s=0}^{\infty}\sum_{j:\rho(i,j)=s} \mathrm{Cov}(Z_i, Z_j) = \sum_{s=0}^{\infty} f(s)\frac{1}{n}\sum_{i=1}^{n}|\{j\colon \rho(i,j)=s\}|.$$

Note that $|\{j\colon \rho(i,j)=s\}| \leqslant 8s$, with equality achieved for all i not near the boundary of the population region. Thus, for large n, a finite variance requires that $f(s)$ decay with s faster than s^{-2}.

We next discuss two models of interference satisfying Assumption 3. The first is the standard approach of specifying a K-neighborhood exposure mapping. Such a mapping is given by $T_i = T(i,\boldsymbol{D},\mathcal{N}_n)$ with the crucial property that its dimension does not depend on n, unlike that of $\boldsymbol{D}$. The approach is to assume that the low-dimensional T_i summarizes interference by reparameterizing potential outcomes as

$$(1) \qquad\qquad Y_i(\boldsymbol{D}) = \tilde{Y}_i(T_i).$$

That is, once we fix i's exposure mapping T_i, its potential outcome is fully determined. No less important, it is also typically assumed that exposure mappings are restricted to a unit's K-neighborhood, where K is small, meaning fixed with respect to n. Formally, $T(i, \boldsymbol{d}, \mathcal{N}_n) = T(i, \boldsymbol{d}', \mathcal{N}_n)$ for any $\boldsymbol{d}, \boldsymbol{d}'$ such that $d_j = d'_j$ for all $j \in \mathcal{N}(i, K)$, which implies that the treatment assigned to a unit j only interferes with i if $\rho(i, j) \leqslant K$. In practice, choices with $K = 1$ are most common, for example $T_i = (D_i, S_i)$ for $S_i = \mathbf{1}\{\sum_{j \in \mathcal{N}_n} G_{ij} T_j > 0\}$ or $S_i = \sum_{j \in \mathcal{N}_n} G_{ij} T_j$ where $G_{ij} = \mathbf{1}\{\rho(i, j) \leqslant 1\}$. In these examples, D_i captures the direct effect of the treatment, and S_i captures interference from units near i.

Crucially, T_i and K must be known to the analyst in this approach, which is often a strong requirement. In contrast, Assumption 3 enables the analyst to impose (1) *while requiring neither to be known*. Indeed, if there exists a K-neighborhood exposure mapping satisfying (1), then Assumption 3 holds with $\psi(s) = c\,\mathbf{1}\{s \leqslant K\}$ for some c sufficiently large.

Furthermore, Assumption 3 allows for more complex forms of interference ruled out by (1) in which interference decays more smoothly with distance, rather than being truncated at some distance K. The former is analogous to mixing conditions, which are widespread in the time series and spatial literature, while the latter is analogous to m-dependence, which rules out interesting forms of autocorrelation, including models as basic as AR(1).

In the spatial context, our assumption accommodates, for example, the Cliff-Ord autoregressive model [11, 12], which is a workhorse model of spatial autocorrelation used in a variety of fields, including geography [17], ecology [48], and economics [2]. A typical formulation of the model is

$$Y_i = \alpha + \lambda \sum_{j \in \mathcal{N}_n} W_{ij} Y_j + D_i \beta + \varepsilon_i,$$

where we assume ε_i is uniformly bounded to satisfy Assumption 1. Let $\boldsymbol{W}$ be the $n \times n$ spatial weight matrix whose ijth entry is W_{ij}. These weights typically decay with distance $\rho(i, j)$ in a sense to be made precise below. While this model is highly stylized, the important aspect it captures is autocorrelation through the spatial autoregressive parameter λ. If this is nonzero, then there is no K-neighborhood exposure mapping for which (1) holds, a point previously noted by [15] in the context of network interference.

To see this, first note that coherency of the model requires nonsingularity of $\boldsymbol{I} - \lambda \boldsymbol{W}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix. Let $\boldsymbol{V}$ be the inverse of this matrix and V_{ij} its entry corresponding to units (i, j). Then the reduced form of the model is

$$(2) \qquad Y_i(\boldsymbol{D}) \equiv Y_i = \sum_{j \in \mathcal{N}_n} V_{ij}(\alpha + D_j \beta + \varepsilon_j),$$

a spatial "moving average" model with spatial weight matrix $\boldsymbol{V}$. (See §5 for some examples of $\boldsymbol{V}$.) Noticeably, $Y_i(\boldsymbol{D})$ can potentially depend on D_j for any $j \in \mathcal{N}_n$, which is ruled out if one imposes a K-neighborhood exposure mapping.

Outcomes satisfying (2) are near-epoch dependent, a notion of weak spatial dependence, when the weights decay with spatial distance in the following sense:

$$(3) \qquad \sup_{n \in \mathbb{N}} \max_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} |V_{ij}| \rho(i, j)^{\gamma} < \infty$$

for some $\gamma > 0$ [see Proposition 5 and eq. (13) of 26]. The next result shows that this condition is sufficient for verifying Assumption 3 if $\gamma > 2$.

PROPOSITION 1. *Suppose potential outcomes are given by* (2) *and spatial weights satisfy* (3) *for some $\gamma > 2$. Then Assumption 3 holds with $\psi(s) = c \min\{s^{-\gamma}, 1\}$ for some $c \in (0, \infty)$ that does not depend on s.*

PROOF. Fix $s \geqslant 0$ and any $\boldsymbol{d}, \boldsymbol{d}' \in \{0,1\}^n$ such that $d_j = d'_j$ for all $j \in \mathcal{N}(i, s)$. For $s < 1$, $|Y_i(\boldsymbol{d}) - Y_i(\boldsymbol{d}')| \leqslant c_1 \equiv 2 \sup_n \max_i \max_{\boldsymbol{d}} |Y_i(\boldsymbol{d})|$. For $s \geqslant 1$,

$$|Y_i(\boldsymbol{d}) - Y_i(\boldsymbol{d}')| \leqslant |\beta| \sum_{j \in \mathcal{N}_n} |V_{ij}||d_j - d'_j| \mathbf{1}\{\rho(i,j) > s\} \leqslant s^{-\gamma} |\beta| \sum_{j \in \mathcal{N}_n} |V_{ij}| \rho(i,j)^\gamma.$$

Defining $c_2 = \sup_n \max_i \sum_{j \in \mathcal{N}_n} |\beta||V_{ij}|\rho(i,j)^\gamma$ and $c = \max\{c_1, c_2\} + 1 > 0$, the inequality in Assumption 3 holds with $\psi(s) = c \min\{s^{-\gamma}, 1\}$ by construction. Furthermore, $c < \infty$ by (3) and uniform boundedness of $\{\varepsilon_i\}_{i \in \mathcal{N}_n}$, and $\psi(s)$ satisfies $\sum_{s=1}^\infty s\, \psi(s) < \infty$ because $\gamma > 2$. $\qquad\square$

This result shows that, unlike the standard approach of imposing a K-neighborhood exposure mapping, Assumption 3 can allow for richer forms of interference in which alters that are arbitrarily distant from the ego can interfere with the ego's response.

REMARK 2 (Literature). Assumption 3 and Proposition 1 are spatial analogs of Assumptions 4 and 6 and Proposition 1 of [31] who studies interference mediated by an unweighted network, Bernoulli designs, and a different class of estimands defined by exposure mappings satisfying overlap. We study the global average treatment effect and cluster-randomized designs that induce overlap by introducing dependence in treatment assignments, and we further derive rate-optimal designs. These differences require an entirely distinct asymptotic theory.

2.2. *Design and Estimator.* Much of the literature on interference considers designs in which units are individually randomized into treatment and control, either via Bernoulli or complete randomization. A common problem faced by such designs is limited overlap, meaning that some realizations of the exposure mapping occur with low probability. For example, suppose that (1) holds with exposure mapping $T_i = \sum_{j \in \mathcal{N}_n} \mathbf{1}\{\rho(i,j) \leqslant K\}D_j$, the number of treated units in i's K-neighborhood. Then in a Bernoulli design, for large values of K, $\mathbf{P}(T_i = 0)$ is small, tending to zero with K at an exponential rate. This is problematic for a Horvitz-Thompson estimator such as $n^{-1} \sum_{i \in \mathcal{N}_n} (p_i(t)^{-1} \mathbf{1}\{T_i = t\} - p_i(t')^{-1} \mathbf{1}\{T_i = t'\})Y_i$ where $p_i(t) = \mathbf{P}(T_i = t)$ since its variance grows rapidly with K if either t or t' is zero. Ugander et al. [46] propose cluster-randomized designs, which reduce this problem by deliberately introducing dependence in treatment assignments across certain units.

We consider the following class of such designs. We assign units to mutually exclusive clusters by partitioning the population region $Q(\boldsymbol{0}, R_n)$ into $m_n \leqslant n$ equally sized squares, assuming for simplicity that $m_n \in \{4^s : s \in \mathbb{N}\}$. That is, to obtain increasingly more clusters, we first divide the population region into four squares, then divide each of these squares into four squares, and so on, as in Figure 2. Label the m_n squares $Q_1, \ldots, Q_{m_n}$, and call

$$C_k = Q_k \cap \mathcal{N}_n$$

cluster k. Then the number of units in each cluster is uniformly $O(n/m_n)$ under Assumption 2, and the radius of each cluster is

$$r_n \equiv R_n/\sqrt{m_n} = \sqrt{n/m_n} + o(1),$$

which we assume is greater than 1. We also assume there are no units on the common boundaries of different squares, so the squares partition $\mathcal{N}_n$.

A *cluster-randomized design* first independently assigns each cluster to treatment and control with some probability

$$p \in (0, 1)$$

that is fixed with respect to n. Then within a treated (control) cluster C_k, all $i \in C_k$ are assigned $D_i = 1$ ($D_i = 0$). In order to emphasize that we use this design in later theorems, we state it as a separate assumption.
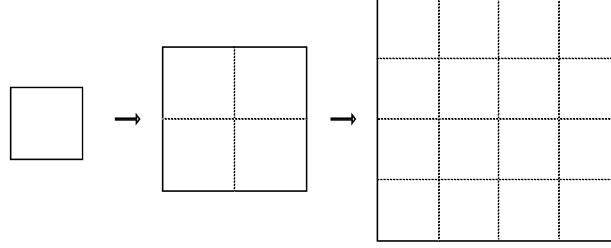
FIG 2. *Sequence of population regions and clusters as $n \to \infty$.*

ASSUMPTION 4 (Design). For any n, $\boldsymbol{D}$ is realized according to a cluster-randomized design with m_n clusters constructed as above.

Note that m_n will be required to diverge with n since a large number of clusters is needed for the estimator to concentrate. If m_n is order n, then $r_n = O(1)$, so clusters are asymptotically bounded in size, the usual case studied in the literature, which includes unit-level Bernoulli randomization as a special case. If m_n is of smaller order, then cluster sizes grow with n.

To construct the estimator, define the *neighborhood exposure indicator*

$$T_{ti} = \prod_{j \in \mathcal{N}(i, \kappa_n)} \mathbf{1}\{D_j = t\} \quad \text{for} \quad t \in \{0, 1\}, \quad \kappa_n = r_n/2.$$

This is an indicator for whether i's κ_n-neighborhood is entirely treated ($t = 1$) or untreated ($t = 0$). Unlike K-neighborhood exposure mappings, the radius κ_n will be allowed to diverge. Let $p_{ti} = \mathbf{E}[T_{ti}]$. We study the Horvitz-Thompson estimator

$$\hat{\theta} = \frac{1}{n} \sum_{i \in \mathcal{N}_n} Z_i \quad \text{for} \quad Z_i = \left(\frac{T_{1i}}{p_{1i}} - \frac{T_{0i}}{p_{0i}} \right) Y_i.$$

Intuitively, $n^{-1} \sum_{i=1}^{n} Y_i T_{1i} p_{1i}^{-1}$ estimates $n^{-1} \sum_{i \in \mathcal{N}_n} Y_i(\mathbf{1}_n)$ using the outcomes of units whose neighbors within radius κ_n are all treated. Since the radius depends on m_n through r_n, $\hat{\theta}$ is a function of the number of clusters dictated by the design. Figure 1 depicts the relationship between the clusters and the κ_n-neighborhoods that determine exposure.

Since nontrivial designs will include both treated and untreated units, $\hat{\theta}$ is biased for the global average treatment effect. The choice of design can trade off the size of the bias against that of the variance. In particular, small choices of m_n (few clusters, large radii) induce lower bias and higher variance. In §3, we discuss nearly rate-optimal choices of m_n for which the bias is asymptotically negligible.

REMARK 3 (Overlap). Under Bernoulli randomization, overlap needs to be imposed as a separate assumption for asymptotic inference. By overlap we mean the probability weights p_{1i} and p_{0i} are uniformly bounded away from zero and one, which imposes potentially strong restrictions on the types of exposure mappings the analyst can use, as previously illustrated. In our setup, however, overlap is directly satisfied because $p_{1i} = p^k$ and $p_{0i} = (1 - p)^k$, where k is the number of clusters that intersect i's κ_n-neighborhood. Our choice of κ_n implies $k \in [1, 4]$ for all i, so overlap holds.

REMARK 4 (Neighborhood radius). Let $\boldsymbol{c}(C_k) \in \mathbb{R}^2$ be the centroid of cluster C_k. The choice of $\kappa_n = r_n/2$ ensures that, for any unit i in the "interior" of a cluster in the sense that $i \in Q(\boldsymbol{c}(C_k), r_n/2)$, i's κ_n-neighborhood also lies within that cluster, in which case the

exposure probabilities are simply given by the cluster assignment probability: $p_{1i} = p$ and $p_{0i} = 1 - p$. If we had instead chosen, say, $\kappa_n = r_n$, then this would be true only for the centroid, while for the remaining units, p_{1i} and p_{0i} could be as small as p^4, which means less overlap and a more variable estimate. For the purposes of the asymptotic theory, the main requirement is that κ_n has the same asymptotic order as r_n. If κ_n were of smaller order, then results in §3 show that the bias of $\hat{\theta}$ could be non-negligible, whereas if κ_n were of larger order, then k in Remark 3 would grow with n, overlap would be limited, and $\mathrm{Var}(\hat{\theta})$ could be large.

3. Rate-Optimal Designs. We next derive the rate of convergence of $\hat{\theta}$ as a function of n, m_n, and $\psi(\cdot)$, which we use to obtain rate-optimal choices of m_n. Recall that designs are parameterized by m_n, which determines the number and sizes of clusters, and also that $\hat{\theta}$ depends on m_n through the neighborhood exposure radius κ_n, so we will be optimizing over both the design and the radius that determines the estimator.

3.1. *Rate of Convergence.* We first provide asymptotic upper bounds on the bias and variance of $\hat{\theta}$. For two sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, we write $a_n \lesssim b_n$ to mean $a_n/b_n = O(1)$ and $a_n \gtrsim b_n$ to mean $b_n/a_n = O(1)$.

THEOREM 1. *Under Assumptions 1–4, $|\mathbf{E}[\hat{\theta}] - \theta_n| \lesssim \psi(r_n/2)$ and $\mathrm{Var}(\hat{\theta}) \lesssim m_n^{-1}$.*

PROOF. First, we bound the bias. If $T_{1i} = 1$, then all units in $Q(i, \kappa_n)$ are treated, so by Assumption 3,

$$|\mathbf{E}[Y_i T_{1i} p_{1i}^{-1}] - Y_i(\mathbf{1}_n)| = |\mathbf{E}[Y_i(\mathbf{D}) \,|\, T_{1i} = 1] - Y_i(\mathbf{1}_n)| \leqslant \psi(\kappa_n).$$

The same argument applies to $|\mathbf{E}[Y_i T_{0i} p_{0i}^{-1}] - Y_i(\mathbf{0}_n)|$, so combining these results, we obtain the rate for the bias.

Next, we bound the variance. The following is an important object in our analysis and also for later constructing the variance estimator:

(4) $$\Lambda_i = \{j \in \mathcal{N}_n : \exists\, k \text{ s.t. } C_k \cap \mathcal{N}(i, \kappa_n) \neq \varnothing, C_k \cap \mathcal{N}(j, \kappa_n) \neq \varnothing\}.$$

This is the set of units j whose κ_n-neighborhoods intersect a cluster C_k that also intersects i's κ_n-neighborhood. We have

(5) $$\mathrm{Var}\left(\frac{1}{n}\sum_{i\in\mathcal{N}_n} Z_i\right) = \frac{1}{n^2}\sum_{i\in\mathcal{N}_n}\sum_{j\in\Lambda_i}\mathrm{Cov}(Z_i, Z_j) + \frac{1}{n^2}\sum_{i\in\mathcal{N}_n}\sum_{j\notin\Lambda_i}\mathrm{Cov}(Z_i, Z_j) \equiv [A] + [B].$$

Note that Λ_i contains units from at most 16 clusters (the worst case is when $Q(i, \kappa_n)$ intersects four clusters), and clusters contain uniformly $O(n/m_n)$ units by Lemma A.1 of [25]. By Assumption 1 and Remark 3, $\{Z_i\}_{i\in\mathcal{N}_n}$ is uniformly bounded, so

$$[A] \lesssim \frac{1}{n^2} \cdot n \cdot \frac{n}{m_n} = \frac{1}{m_n}.$$

The difficult part of the argument is obtaining a tight enough rate for $[B]$, which requires control over the dependence between outcomes of units i, j that are "distant" in the sense that $j \notin \Lambda_i$. Lemma B.1 in §B proves that $[B] \lesssim (nm_n)^{-1/2}$. Therefore, $\mathrm{Var}(\hat{\theta}) \lesssim m_n^{-1} + (nm_n)^{-1/2}$, which is at most m_n^{-1} since $m_n \leqslant n$. □

Our second result provides asymptotic lower bounds.

THEOREM 2. *Suppose $\psi(s) = \min\{s^{-\gamma}, 0.5\}$ for some $\gamma > 2$. Under Assumption 4, there exists a sequence of units and potential outcomes $\{\{Y_i(\cdot)\}_{i \in \mathcal{N}_n} : n \in \mathbb{N}\}$ satisfying Assumptions 1–3 such that $|\mathbf{E}[\hat{\theta}] - \theta_n| \gtrsim \psi(1.5r_n)$ and $Var(\hat{\theta}) \gtrsim m_n^{-1}$.*

PROOF. See supplemental material [33].   □

The result shows that we can construct potential outcomes satisfying the assumptions of Theorem 1 such that the bias is at least order $\psi(1.5r_n)$ and the variance at least m_n^{-1}. As discussed in §1, existing work on cluster randomization under interference assumes clusters have asymptotically bounded size, which, in our setting, implies $r_n = O(1)$. Theorem 2 implies that the bias of the Horvitz-Thompson estimator can then be bounded away from zero, showing that existing results strongly rely on the exposure mapping assumption to obtain unbiased estimates. In the absence of this assumption, it is necessary to consider designs in which cluster sizes are large to ensure the bias vanishes with n.

3.2. *Design Examples.* Theorem 1 implies the mean-squared error of $\hat{\theta}$ is at most of order $\psi(r_n/2)^2 + m_n^{-1}$, and Theorem 2 provides a similar asymptotic lower bound. Under either bound, the bias increases with m_n while the variance decreases, so there exists a bias-variance trade-off in the choice of design. We next derive rates for m_n that minimize or nearly minimize the upper bound under different assumptions on $\psi(\cdot)$. Based on these results, we make recommendations for practical implementation in the next subsection.

Oracle design. Suppose $\psi(s)$ is known to decay with s at some rate $\phi(s)$. Then by definition of r_n, a rate-optimal design chooses m_n to minimize $\phi(0.5R_n m_n^{-1/2})^2 + m_n^{-1}$.

Exposure mappings. If we assume (1) holds for some K-neighborhood exposure mapping, then $\psi(s) = 0$ for all $s > K$. If K is known, then by choosing $m_n = R_n^2(2K)^{-2}$, we have $\kappa_n = K$ and zero bias. In this case, clusters are asymptotically bounded in size, the estimator converges at rate $n^{-1/2}$, and both the design and estimator qualitatively coincide with those of [46].

On the other hand, if K is unknown, then for a nearly rate-optimal design, we can choose κ_n to grow at a slow rate so that it eventually exceeds any fixed K. This may be achieved by choosing m_n to grow slightly slower than n, say $n/\log(n)$. Then for large enough n, the bias is zero, and the rate of convergence is $\sqrt{\log(n)/n}$.

Exponential decay. Common specifications of the spatial weight matrix W in the Cliff-Ord model imply that $\psi(s)$ decays exponentially with s, for example, the row-normalized matrix

$$(6) \qquad W_{ij} = \frac{\mathbf{1}\{\rho(i,j) \leqslant 1\}}{\sum_{k \in \mathcal{N}_n} \mathbf{1}\{\rho(i,k) \leqslant 1\}}.$$

If $\psi(s)$ is known to decay at some exponential rate but the exponent is unknown, then we may choose $m_n = n^{1-\epsilon}$ for any small $\epsilon > 0$ for a nearly rate-optimal design, which yields a rate of convergence of $n^{-0.5(1-\epsilon)}$, close to an $n^{-1/2}$-rate. This shows that rates close to $n^{-1/2}$ are attainable in the absence of exposure mapping assumptions, despite targeting the global average treatment effect.

Worst-case decay. In practice, we may have little prior knowledge about $\psi(s)$. Recall that Assumption 3 requires the rate of decay to be no slower than $s^{-2(1+\epsilon)}$ for $\epsilon > 0$. As discussed in Remark 1, this is the slowest rate for spatial dependence that ensures a finite variance. For this rate, since R_n is order $\sqrt{n}$, the bias is order $(n/m_n)^{-(1+\epsilon)}$. Without knowledge of ϵ, we can settle for a nearly rate-optimal design by setting $\epsilon = 0$ and choosing m_n to minimize $(n/m_n)^{-2} + m_n^{-1}$, which yields $m_n = n^{2/3}$ and an $n^{-1/3}$-rate of convergence. Under this design, cluster sizes grow at the rate $r_n^2 = n^{1/3}$.

In the last three designs, the bias is $o(m_n^{-1/2})$, which is of smaller order than the variance. This makes the bias negligible from an asymptotic standpoint, but it would be useful to develop bias-reduction methods. We also reiterate that, while this analysis only provides rates, it is apparently by necessity at this level of generality. A finite-sample optimal design seems to require substantially more knowledge of the functional form of $\psi(\cdot)$.

3.3. *Practical Recommendations.* The designs in the previous section rely on varying degrees of knowledge of $\psi(\cdot)$, the rate at which interference vanishes with distance. In practice, this is likely unknown, so we recommend operating under the worst-case rate of decay discussed in the previous subsection. The *default conservative choice* we recommend using in practice is the near-optimal rate described there, namely

$$(7) \qquad\qquad m_n = n^{2/3}.$$

To construct the clusters, we recommend partitioning space into m_n clusters using a clustering algorithm, such as spectral clustering. A confidence interval (CI) for θ_n is given in (11). In §5, we explore in simulations the performance of the CI when clusters are constructed according to these recommendations.

Our large-sample theory assumes space is subdivided into evenly sized squares in order to avoid the difficult problem of optimizing over arbitrary shapes. However, since units are typically irregularly distributed in practice, division into equally sized squares may be inefficient, which is why we recommend the use of clustering algorithms. We suggest spectral clustering because it recovers, under weak conditions, low-conductance clusters [38], and low conductance is the key property of clusters utilized in our proofs, as discussed in §6.

4. Inference. We next state results for asymptotic inference on θ_n. Define $\sigma_n^2 = \text{Var}(\sqrt{m_n}\hat{\theta})$.

ASSUMPTION 5 (Non-degeneracy). $\liminf_{n\to\infty} \sigma_n^2 > 0$.

This is a standard condition and reasonable to impose in light of the lower bound on the variance derived in Theorem 2.

THEOREM 3. *Suppose $m_n \to \infty$ and $m_n = o(n)$. Under Assumptions 1–5,*

$$\sigma_n^{-1}\sqrt{m_n}(\hat{\theta} - \mathbf{E}[\hat{\theta}]) \xrightarrow{d} \mathcal{N}(0,1).$$

PROOF. See §B. □

The result centers $\hat{\theta}$ at its expectation, not the estimand θ_n. However, designs discussed in §3.2 result in small bias, meaning $|\mathbf{E}[\hat{\theta}] - \theta_n| = o(m_n^{-1/2})$, so we can replace $\mathbf{E}[\hat{\theta}]$ with θ_n on the left-hand side. Also note that the assumption $m_n = o(n)$ implies that cluster sizes grow with n. If instead m_n were of order n, then $r_n = O(1)$, so by Theorem 2, we would additionally need to assume that there exists a K-neighborhood exposure mapping in the sense of (1) in order to guarantee that the bias vanishes at all. In this case, it is straightforward to establish a normal approximation using existing results.

12

4.1. *Proof Sketch.* To our knowledge, there is no off-the-shelf central limit theorem that we can apply to $\hat{\theta}$. Under Assumption 3, the outcomes appear to be near-epoch dependent on the input process $\{D_i\}_{i\in\mathcal{N}_n}$, but the treatments are cluster-dependent with growing cluster sizes, rather than α-mixing, as required by [27]. To prove a central limit theorem, they split the average into two parts: its expectation conditional on the dependent input process $\{D_i\}_{i\in\mathcal{N}_n}$, and a remainder that they show is small. Rather than conditioning on all treatments, we find that the following unit-specific conditioning event is more useful for proving our result.

Let $\mathcal{C}_i$ be the cluster containing unit i, and $\mathcal{F}_i = \{D_j : j \in \mathcal{C}_i \cup \mathcal{N}(i, \kappa_n)\}$. Rewrite the estimator as

$$(8) \qquad \hat{\theta} = \frac{1}{n} \sum_{i\in\mathcal{N}_n} Z_i = \frac{1}{n} \sum_{i\in\mathcal{N}_n} \mathbf{E}[Z_i \mid \mathcal{F}_i] + \frac{1}{n} \sum_{i\in\mathcal{N}_n} (Z_i - \mathbf{E}[Z_i \mid \mathcal{F}_i]).$$

We first show that the last term is relatively small, $o_p(m_n^{-1/2})$ to be precise, which means that, on average, Z_i is primarily determined by $\mathcal{F}_i$. The proof of this claim is somewhat complicated [and different from that of 27], but it is similar to the argument showing $[B] \lesssim (nm_n)^{-1/2}$ in (5). To then establish a central limit theorem for $n^{-1}\sum_{i\in\mathcal{N}_n} \mathbf{E}[Z_i \mid \mathcal{F}_i]$, we observe that the dependence between "observations" $\{\mathbf{E}[Z_i \mid \mathcal{F}_i]\}_{i\in\mathcal{N}_n}$ is characterized by the following dependency graph $\boldsymbol{A}$, which, roughly speaking, links two units only if they are dependent. Recalling the definition of Λ_i from (4), we connect units i, j in $\boldsymbol{A}$ if and only if $j \in \Lambda_i$ (or equivalently $i \in \Lambda_j$). Then $\boldsymbol{A}$ is indeed a dependency graph because, under Assumption 4, $j \notin \Lambda_i$ implies that the treatment assignments that determine $\mathcal{F}_i$ are independent of those that determine $\mathcal{F}_j$. The result follows from a central limit theorem for dependency graphs.

The proof highlights two sources of dependence. The first-order source is the first term on the right-hand side of (8). Dependence in this average is due to cluster randomization, which induces correlation in the treatments determining $\mathcal{F}_i$ across i. The second-order source is the second term on the right-hand side of (8). Dependence in this average is due to interference, which decays with distance due to Assumption 3. Sävje [42] derives a similar decomposition in a different context with misspecified exposure mappings. The previous arguments show that the second-order source of dependence is small relative to the first-order source because, with large clusters, dependence induced by cluster randomization dominates dependence induced by interference. This is generally untrue with small clusters.

4.2. *Variance Estimator.* The proof sketch suggests that, to estimate σ_n^2, it suffices to account for dependence induced by cluster randomization. Define $A_{ij} = \mathbf{1}\{j \in \Lambda_i\}$, where Λ_i is defined in (4), and note that $A_{ii} = 1$ and $A_{ij} = A_{ji}$. Let $\bar{Z} = n^{-1}\sum_{i\in\mathcal{N}_n} Z_i$, which is equivalent to $\hat{\theta}$. Our proposed variance estimator is

$$(9) \qquad \hat{\sigma}^2 = \frac{m_n}{n^2} \sum_{i\in\mathcal{N}_n} \sum_{j\in\mathcal{N}_n} (Z_i - \bar{Z})(Z_j - \bar{Z}) A_{ij}.$$

THEOREM 4. *Suppose* $m_n \to \infty$ *and* $m_n = o(n)$. *Under Assumptions 1–5,* $(\hat{\sigma}^2 - \mathcal{R}_n)/\sigma_n^2 \overset{p}{\longrightarrow} 1$, *where*

$$\mathcal{R}_n = \frac{m_n}{n} \hat{\hat{\sigma}}^2 \quad and \quad \hat{\hat{\sigma}}^2 = \frac{1}{n} \sum_{i\in\mathcal{N}_n} \sum_{j\in\mathcal{N}_n} (\mathbf{E}[Z_i] - \mathbf{E}[\bar{Z}])(\mathbf{E}[Z_j] - \mathbf{E}[\bar{Z}]) A_{ij}.$$

PROOF. See supplemental material [33]. $\qquad\qquad\square$

The bias term $\mathcal{R}_n$ is typically nonzero due to the unit-level heterogeneity. That is, $|\mathbf{E}[Z_i] - \mathbf{E}[\bar{Z}]|$ does not approach zero asymptotically, except in the special case of homogeneous treatment effects where $Y_i(\mathbf{1}_n) - Y_i(\mathbf{0}_n)$ does not vary across i. In the no-interference setting, it is well-known that the variance of the difference-in-means estimator is biased for the same reason and that consistent estimation of the variance is impossible. However, due to the term $m_n/n = o(1)$ in $\mathcal{R}_n$, we will argue that typically $\mathcal{R}_n = o_p(1)$, meaning that $\hat{\sigma}^2$ is asymptotically exact.

Let us first compare $\mathcal{R}_n$ to its formulation under no interference. In this case, $Y_i(\mathbf{D}) = Y_i(D_i)$, and we replace T_{1i} with D_i and T_{0i} with $1 - D_i$ to estimate the usual average treatment effect. Furthermore, we set $A_{ij} = 0$ for all $i \neq j$ because units are independent and set $m_n = n$ since there is no longer a need to cluster units. With these changes, $Z_i = (D_i/p - (1 - D_i)/(1 - p))Y_i$, and

$$\mathcal{R}_n = \frac{1}{n} \sum_{i \in \mathcal{N}_n} (\tau_i - \bar{\tau})^2 \tag{10}$$

for $\tau_i = Y_i(1) - Y_i(0)$ and $\bar{\tau} = n^{-1} \sum_{i \in \mathcal{N}_n} \tau_i$. This is the well-known expression for the bias in the absence of interference [e.g. 22, Theorem 6.2].

In our setting, we have additional "covariance" terms included in $\mathcal{R}_n$ due to the nonzero off-diagonals of the dependency graph A_{ij}. These would be problematic if they were negative and larger in magnitude than the main variance terms since that would make $\hat{\sigma}^2$ anti-conservative. We show that this occurs with small probability, and in fact, that $\mathcal{R}_n$ is $o_p(1)$. Observe that $m_n/n = o(1)$ and $\hat{\tilde{\sigma}}^2$ has the form of a HAC (heteroskedasticity and autocorrelation consistent) variance estimator [1, 13]. Hence, under conventional regularity conditions, $\hat{\tilde{\sigma}}^2$ is consistent for a variance term $\tilde{\sigma}^2 \geqslant 0$, in which case $\mathcal{R}_n$ is nonnegative in large samples, and furthermore, $o_p(1)$. To formalize this intuition, we need to specify conditions on the superpopulation from which potential outcomes are drawn. In §A, we show that, if potential outcomes are α-mixing, then $\hat{\tilde{\sigma}}^2$ is asymptotically unbiased for $\tilde{\sigma}^2 = \mathrm{Var}(n^{-1/2} \sum_{i=1}^{n} \mathbf{E}[Z_i \mid \{Y_i(\mathbf{d})\}_{\mathbf{d} \in \{0,1\}^n}])$, and furthermore, $\mathrm{Var}(\hat{\tilde{\sigma}}^2) = O(n^2/m_n^3)$. Consequently, $\mathrm{Var}(\mathcal{R}_n) = O(m_n^{-1})$ due to the m_n/n term in its expression.

REMARK 5 (Confidence interval). As previously discussed, the bias of $\hat{\theta}_n$ is $o(m_n^{-1/2})$ for the near-optimal designs discussed at the end of §3. Thus, for such designs, the preceding discussion justifies the use of

$$\hat{\theta} \pm 1.96 \cdot \hat{\sigma} m_n^{-1/2} \tag{11}$$

as an asymptotic 95-percent CI for θ_n, where $\hat{\sigma}^2$ is defined in (9).

REMARK 6 (Literature). Leung [31] proves a result similar to Theorem 4 but for a different variance estimator under a different design and variety of interference. Due to the lack of an analogous m_n/n term, in his setting, weak dependence conditions would only ensure $\mathcal{R}_n \xrightarrow{p} \tilde{\sigma}^2 \geqslant 0$, in which case the estimator would be asymptotically conservative, whereas ours is asymptotically exact. He does not provide a formal result on the limit of $\mathcal{R}_n$.

5. Monte Carlo. We next present results from a simulation study illustrating the quality of the normal approximation in Theorem 3 and coverage of the CI (11) when constructing clusters using spectral clustering. To generate spatial locations, let $\{\tilde{\rho}_i\}_{i \in \mathcal{N}_n}$ be i.i.d. draws from $\mathcal{U}([-1, 1]^2)$. Unit locations in $\mathbb{R}^2$ are given by $\{\rho_i\}_{i \in \mathcal{N}_n}$ for $\rho_i = R_n \tilde{\rho}_i$ with $R_n = \sqrt{n}$. We let $\rho(i, j) = \|\rho_i - \rho_j\|$ where $\|\cdot\|$ is the Euclidean norm.

We set the number of clusters according to (7), rounded to the nearest integer, which corresponds to the near-optimal design under the worst-case decay discussed in §3.2. To construct
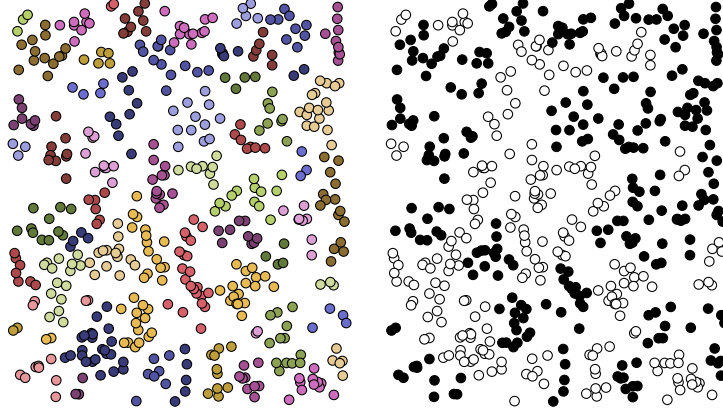
14

FIG 3. *The left figure colors units by cluster memberships obtained from spectral clustering (some colors are reused for different clusters). The right figure colors units by treatment assignment.*

clusters, we apply spectral clustering to $\{\rho_i\}_{i\in\mathcal{N}_n}$ with the standard Gaussian affinity matrix whose ijth entry is $\exp\{-\rho(i,j)^2\}$. Clusters are randomized into treatment with probability $p = 0.5$. Figure 3 displays the clusters and treatment assignments for a typical simulation draw.

We generate outcomes from three different models. Let $\{\varepsilon_i\}_{i\in\mathcal{N}_n} \overset{iid}{\sim} \mathcal{N}(0,1)$ be drawn independently of the other primitives. The first model is Cliff-Ord:

$$Y_i = \alpha + \lambda \sum_{j\in\mathcal{N}_n} W_{ij}Y_j + \delta \sum_{j\in\mathcal{N}_n} W_{ij}D_j + D_i\beta + \varepsilon_i$$

with $(\alpha,\lambda,\delta,\beta) = (-1, 0.8, 1, 1)$ and spatial weight matrix given by the row-normalized adjacency matrix (6). As discussed in §3, this model features exponentially decaying $\psi(s)$, in fact of order λ^s [31, Proposition 1].

We construct the second and third models to explore how our methods break down when Assumption 3 is violated or close to violated. For this purpose, we use the "moving average" model (2) with $(\alpha,\beta) = (-1, 1)$ and $V_{ij} = \rho(i,j)^{-\eta}$ for $\eta = 4, 5$ for the two respective models, so that $\psi(s)$ decays at a polynomial rate. Notably, the choice of $\eta = 4$ implies that the rate of decay is slow enough that Assumption 3 can fail to hold. This is because

$$(3) = \sum_{s=1}^{\infty} \sum_{j\in\mathcal{N}_n} \mathbf{1}\{\rho(i,j) \in [s-1,s)\}\rho(i,j)^{\gamma-4} \leqslant c \sum_{s=0}^{\infty} s^{\gamma-3}$$

for some $c > 0$ by Lemma A.1(iii) of [25]. The right-hand side does not converge for some $\gamma > 2$, as required by Proposition 1. On the other hand, the choice of $\eta = 5$ is large enough for Assumption 3 to be satisfied since we now replace the 3 on the right-hand side of the previous display with 4. However, in smaller samples, $\eta = 4$ or 5 may not be substantially different, so our methods may still break down from the assumption being "close to" violated.

Table 1 displays the results of 5000 simulation draws. Row "Bias($\hat{\theta}$)" displays $|\mathbf{E}[\hat{\theta} - \theta_n]|$, estimated by taking the average over the draws, while "Var($\hat{\theta}$)" is the variance of $\hat{\theta}$ across the draws. The next rows display the coverage of three different confidence intervals. "Our CI" corresponds to the empirical coverage of (11). "Naive CI" corresponds to (11) but replaces $\hat{\sigma}m_n^{-1/2}$ with the i.i.d. standard error, so the extent to which its coverage deviates from 95 percent illustrates the degree of spatial dependence. "Oracle CI" corresponds to (11)

but replaces $\hat{\sigma}m_n^{-1/2}$ with the "oracle" SE, which is the standard deviation of $\hat{\theta}$ across the draws. Note that the oracle SE approximates $\sigma_n^2 + \mathcal{R}_n$ because the variance is taken over the randomness of the design as well as of the potential outcomes. Lastly, "SE" displays our standard error $\hat{\sigma}m_n^{-1/2}$.

TABLE 1
Simulation Results

	Moving Avg, $\eta = 4$			Moving Avg, $\eta = 5$			Cliff-Ord		
n	250	500	1000	250	500	1000	250	500	1000
m_n	40	63	100	40	63	100	40	63	100
Our CI	0.909	0.925	0.924	0.922	0.937	0.940	0.979	0.983	0.982
Naive CI	0.530	0.489	0.447	0.575	0.539	0.494	0.918	0.913	0.916
Oracle CI	0.943	0.943	0.936	0.950	0.952	0.949	0.983	0.982	0.982
Bias($\hat{\theta}$)	0.108	0.093	0.083	0.033	0.027	0.024	0.009	0.004	0.003
Var($\hat{\theta}$)	0.143	0.087	0.054	0.108	0.065	0.039	0.341	0.177	0.087
SE	0.364	0.292	0.232	0.317	0.252	0.199	0.601	0.432	0.309
$\hat{\theta}$	1.432	1.467	1.492	1.289	1.307	1.319	5.804	5.822	5.851

5k simulations. The "CI" rows show the empirical coverage of 95% CIs. "Naive" and "Oracle" respectively correspond to i.i.d. and true standard errors.

There are at most 100 clusters in all designs, and the rate of convergence is quite slow at $n^{-1/3}$ for our choice of m_n. Nonetheless, across all designs, the coverage of the oracle CI is close to 95 percent or above, which illustrates the quality of the normal approximation. For the Cliff-Ord model, our CI attains at least 95 percent coverage even for small sample sizes, despite m_n being chosen suboptimally for the worst-case decay. For the moving average model with $\eta = 5$, we see some under-coverage in smaller samples due to the larger bias, which is unsurprising from the above discussion, but coverage is close to the nominal level for larger n. The results for $\eta = 4$, as expected, are worse since it is deliberately constructed to violate our main assumption. Once again, our CI exhibits under-coverage due to the larger bias, but coverage improves and bias decreases as n grows.

6. Conclusion. This paper studies the design of cluster-randomized experiments targeting the global average treatment effect under spatial interference. Each design is characterized by a parameter m_n that determines the number and sizes of clusters. We propose a Horvitz-Thompson estimator that compares units with different neighborhood exposures to treatment, where the neighborhood radius is of the same order as clusters' sizes given by the design. We asymptotically bound the estimator's bias and variance as a function of m_n and the degree of interference and derive rate-optimal choices of m_n. Our lower asymptotic bound shows that designs using small clusters (those with asymptotically bounded sizes) generally result in a non-negligible asymptotic bias. On the other hand, constructing large clusters reduces the total number of clusters, resulting in a bias-variance trade-off that we seek to optimize in terms of rates through the choice of design.

In the worst case where the degree of interference is substantial, the estimator has an $n^{-1/3}$-rate of convergence under a nearly rate-optimal design, whereas in the best case where interference is characterized by a K-neighborhood exposure mapping, the rate is $n^{-1/2}$ under a rate-optimal design. We derive the asymptotic distribution of the estimator and provide an estimate of the variance.

Important areas for future research include data-driven choices of m_n and κ_n and methods to reduce the bias of the estimator. However, a rigorous theory appears to require more substantive restrictions on interference than what we impose.

Our results focus on the canonical case of spatial data in $\mathbb{R}^2$. We conjecture that they can be extended to $\mathbb{R}^d$ for $d > 2$ because our proofs fundamentally rely on the following key property of Euclidean space, which is true for any dimension: it is always possible to construct many clusters with low *conductance*, or boundary-to-volume ratio, for example by partitioning space into hypercubes or by spectral clustering [32]. This appears in our proofs through the use of Lemma A.1 of [25], which, together with Assumption 3, is crucial to establish that spatially distant units have small covariance, despite dependence induced by cluster randomization and interference. In this sense, the technical idea behind this paper is to exploit a useful property of Euclidean space – the existence of many low-conductance clusters – to show that cluster-randomized designs may be fruitfully applied to the problem of spatial interference.

The story for network interference appears to be different. Existing cluster-randomized designs have theoretical guarantees under exposure mapping assumptions, but it is an open question whether such designs work under weaker restrictions on interference such as Assumption 3. In order to directly apply our idea in the previous paragraph, the network must possess many low-conductance clusters across which we can randomize. Unfortunately, this is a strong requirement in practice because, as discussed in [32], not only do some networks not possess multiple low-conductance clusters, but, of those that do, some apparently possess only a small number of such clusters. Because network "space" differs from Euclidean space in this fundamental aspect, under network interference, clusters can be strongly dependent in the absence of exposure mapping assumptions.

APPENDIX A: BIAS OF THE VARIANCE ESTIMATOR

Characterizing the asymptotic behavior of $\mathcal{R}_n$ requires conditions on the superpopulation from which units are drawn. In this section, we assume potential outcomes are random and constitute a weakly dependent spatial process (independent of $\boldsymbol{D}$). Accordingly, we rewrite $\mathbf{E}[Z_i]$ in Theorem 4 as

$$\tilde{Z}_i \equiv \mathbf{E}[Z_i \mid \{Y_i(\boldsymbol{d})\}_{\boldsymbol{d} \in \{0,1\}^n}].$$

We require the spatial process $\{\tilde{Z}_i\}_{i \in \mathcal{N}_n}$ to be α-mixing, which is a standard concept of weak spatial dependence. The results we use in fact apply to the weaker concept of near-epoch dependence, but we focus on mixing since it requires less exposition.

DEFINITION A.1. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be the probability space, $\mathcal{A}, \mathcal{B}$ be sub-σ-algebras of $\mathcal{F}$, and

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup\{|\mathbf{P}(A \cap B) - \mathbf{P}(A)\mathbf{P}(B)|; A \in \mathcal{A}, B \in \mathcal{B}\}.$$

For $U, V \subseteq \mathcal{N}_n$, let $\alpha_n(U, V) = \alpha(\sigma_n(U), \sigma_n(V))$, where $\sigma_n(U)$ is σ-algebra generated by $\{\tilde{Z}_i\}_{i \in U}$. The α-*mixing coefficient* of $\{\tilde{Z}_i\}_{i \in \mathcal{N}_n}$ is

$$\bar{\alpha}(u, v, r) = \sup_n \sup_{U,V} \{\alpha_n(U, V); |U| \leqslant u, |V| \leqslant v, \rho(U, V) \geqslant r\}$$

for $u, v \in \mathbb{N}$, $r \in \mathbb{R}_+$, and $\rho(U, V) = \min\{\rho(i, j) : i \in U, j \in V\}$.

That is, for any two sets of units $U, V \subseteq \mathcal{N}_n$ with respective sizes u, v such that the minimum distance between U, V is at least r, $\bar{\alpha}(u, v, r)$ bounds their dependence with respect to observations $\{\tilde{Z}_i\}_{i \in \mathcal{N}_n}$.

EXAMPLE A.1. Suppose that, for any n and $i \in \mathcal{N}_n$, there exists a function $f(\cdot)$ such that $Y_i(\boldsymbol{d}) = f(\varepsilon_i, \boldsymbol{d})$. If the unobserved heterogeneity $\{\varepsilon_i\}_{i \in \mathcal{N}_n}$ is α-mixing and $\tilde{Z}_i$ is a Borel-measurable function of ε_i (a mild requirement since treatments are independent of potential outcomes), then $\{\tilde{Z}_i\}_{i \in \mathcal{N}_n}$ is α-mixing.

EXAMPLE A.2. Generalizing the previous example, suppose $Y_i(\boldsymbol{d}) = f(d_i, d_{-i}, \varepsilon_i, \varepsilon_{-i})$, where $\boldsymbol{d} = (d_i, d_{-i})$ and ε_{-i} is similarly defined. Under some conditions on $f(\cdot)$, one can ensure that $\{\tilde{Z}_i\}_{i \in \mathcal{N}_n}$ is near-epoch dependent on the input $\{\varepsilon_i\}_{i \in \mathcal{N}_n}$ [e.g. 27, Proposition 1]. However, we only focus on mixing.

We next discuss the intuition behind our main result. Let $\bar{\tilde{Z}} = n^{-1} \sum_{i=1}^{n} \tilde{Z}_i$, so that

$$\mathcal{R}_n = \frac{m_n}{n} \hat{\tilde{\sigma}}^2, \quad \text{where} \quad \hat{\tilde{\sigma}}^2 = \frac{1}{n} \sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} (\tilde{Z}_i - \mathbf{E}[\bar{\tilde{Z}}])(\tilde{Z}_i - \mathbf{E}[\bar{\tilde{Z}}]) A_{ij}.$$

Observe that $m_n/n = o(1)$, and $\hat{\tilde{\sigma}}^2$ is essentially a HAC variance estimator with "kernel" A_{ij}. More precisely, A_{ij} is sandwiched between two uniform kernels:

$$(A.1) \qquad \mathbf{1}\{\rho(i,j) \leqslant \kappa_n\} \leqslant A_{ij} \leqslant \mathbf{1}\{\rho(i,j) \leqslant 2r_n + \kappa_n\}.$$

This is a consequence of the construction of clusters. The lower bound holds because Λ_i must include i's κ_n-neighborhood. The upper bound is achieved if i is located at a corner shared by four clusters, and each such cluster intersects some κ_n-neighborhood that is maximally distant from the cluster. Notably, the bandwidths of the two kernels are of the same asymptotic order (recalling that κ_n has the same order as r_n). Hence, $\hat{\tilde{\sigma}}^2$ should behave as a HAC variance estimator. This has two implications.

1. $\hat{\tilde{\sigma}}^2$ should be consistent for a non-negative variance term under standard regularity conditions, so $\mathcal{R}_n = o_p(1)$.
2. If, in the formula for $\hat{\sigma}^2$, we replace A_{ij} with the uniform kernel in the upper bound (A.1), then we would have a positive semidefinite HAC estimator, implying $\mathcal{R}_n \geqslant 0$ a.s. [1, 13, §3.3.1]. Then in the finite-population setting of our main results, $\hat{\sigma}^2$ would be at worst asymptotically conservative.

   We choose not use a spatial HAC for $\hat{\sigma}^2$ because they have a reputation for being anti-conservative in smaller samples [see references in e.g. 32]. In our estimator, A_{ij} functions as a sort of heterogeneous bandwidth determined by the design that allows different units to have different neighborhood radii in the variance estimator, whereas HAC kernels imply homogeneous radii determined by the bandwidth. The hope is that heterogeneous radii could translate to better finite-sample properties since they directly capture the first-order dependency neighborhood.

We next state regularity conditions taken from [24], which we use to apply her Theorem 4 on consistency of HAC estimators.

ASSUMPTION A.1. The mixing coefficient satisfies $\bar{\alpha}(u, v, r) \leqslant (u + v)^\varsigma \hat{\alpha}(r)$ for $\varsigma \geqslant 0$ and $\sum_{r=1}^{\infty} r^{2(\varsigma+1)-1} \hat{\alpha}(r) < \infty$.

This is Assumption 2 of [24]. The substance of the condition is the requirement that the mixing coefficient decays at a sufficiently fast rate with distance r. For $\varsigma > 0$, the rate requirement is stronger than what we require of $\psi(\cdot)$ in Assumption 3.

ASSUMPTION A.2. (a) $\mathbf{E}[\tilde{Z}_i] = \mathbf{E}[\tilde{Z}_j]$ for all i, j. (b) $\text{Var}(n^{-1/2} \sum_{i=1}^{n} \tilde{Z}_i) \xrightarrow{p} \tilde{\sigma}^2 < \infty$.

This is Assumption 7(a) of [24]. Part (a) is a standard mean-homogeneity condition required for HAC to be asymptotically unbiased. Such a requirement is untenable in the finite population setting of Theorem 4 because $\tilde{Z}_i$ is a function of i's potential outcomes which are generally heterogeneous across units. Heterogeneity is responsible for the appearance of the

bias $\mathcal{R}_n$. However, in the superpopulation setting of this section, the assumption is much more tenable since we integrate over the randomness of the potential outcomes. The condition then requires that the mean be invariant to unit labels. The finiteness requirement in part (b) can be proven as a consequence of Assumption A.1 and moment conditions, so the only substance of the assumption is the existence of a limit.

THEOREM A.1. *Under Assumptions 1, 2, A.1, and A.2, if $m_n \to \infty$, then (a) $\mathbf{E}[\hat{\tilde{\sigma}}^2] \to \tilde{\sigma}^2$, and (b) $Var(\hat{\tilde{\sigma}}) = O(n^2/m_n^3)$.*

We next discuss the implications of the theorem for $\mathcal{R}_n$ (also see the discussion in §4.2) and then conclude with the proof. The designs discussed at the end of §3, other than under the worst-case decay, choose m_n to be of substantially higher order than $n^{2/3}$. In this case, Theorem A.1 yields

$$\mathcal{R}_n = \underbrace{\frac{m_n}{n}}_{o(1)} \underbrace{\hat{\tilde{\sigma}}^2}_{\xrightarrow{p}\tilde{\sigma}^2 \geqslant 0} \xrightarrow{p} 0.$$

For the worst-case decay, $m_n = n^{2/3}$, in which case $\hat{\tilde{\sigma}}^2$ remains asymptotically unbiased for $\tilde{\sigma}^2 \geqslant 0$, and we still have $\mathrm{Var}(\mathcal{R}_n) = O(m_n^{-1})$, so again $\mathcal{R}_n = o_p(1)$.

PROOF OF THEOREM A.1. We apply Theorem 4 of [24]. Note that our setting is a simple mean estimation problem, which is a special case of her semiparametric model $Y_{1in} = h(Y_{2in}, \theta_0) + g(X_{in}) + U_{in}$ with $h(Y_{2in}, \theta_0) + g(X_{in}) = 0$ and U_{in} equal to our $\tilde{Z}_i - \mathbf{E}[\tilde{Z}] = \tilde{Z}_i - \mathbf{E}[\tilde{Z}_i]$ (by Assumption A.2(a)). The moments $m_{\varepsilon_n}(W_{in}, \hat{\theta}, \hat{\tau})$ in her formula (13) for the HAC estimator corresponds to our $\tilde{Z}_i - \mathbf{E}[\tilde{Z}_i]$. Other than Assumption 10, her assumptions are either satisfied (increasing domain corresponds to our Assumption 2 and the moment conditions hold by our Assumption 1 and Remark 3) or are irrelevant in our setting.

Assumption 10 concerns the properties of the kernel function. For context, note that if, hypothetically, we replaced A_{ij} in $\hat{\tilde{\sigma}}^2$ with its upper bound in (A.1), then the kernel $K(\cdot)$ and bandwidth β_n in [24] would correspond in our setting to the uniform kernel $\mathbf{1}\{\cdot \leqslant 1\}$ and $2r_n + \kappa_n$, respectively, so that $K(x/\beta_n)$ in Jenish's notation would correspond to our $\mathbf{1}\{x \leqslant 2r_n + \kappa_n\}$.

Now, because A_{ij} is only bounded by kernel functions but cannot be written as one, we cannot directly verify Assumption 10. However, inspection of the proof reveals that the assumption is used as follows. First, to derive bounds on the variance of the HAC estimator (Step 1 of the proof), uniform boundedness of the kernel is used, but this is trivially satisfied by $A_{ij} \in \{0, 1\}$. Second, to derive bounds on the bias (Step 2 of the proof), Assumption 10 is used to show that the term $a_{r,n} = \mathrm{argmax}_{r \leqslant x \leqslant r+1}|K(x/\beta_n) - 1| \to 0$ as $n \to \infty$ for any $r > 0$. In our case, $a_{r,n}$ corresponds to $\mathrm{argmax}_{i,j \in \mathcal{N}_n \,:\, \rho(i,j) \in [r, r+1]}|A_{ij} - 1|$. But this has the desired property; it is in fact exactly zero for n sufficiently large due to (A.1).

Hence, the conclusions of the proof of Jenish's Theorem 4 apply to $\hat{\tilde{\sigma}}^2$, which we now apply to prove our claims. Part (a) of our theorem follows from Step 2 of her proof. Next, in Step 1 of her proof, $H_{2n}, H_{3n}, H_{4n} = 0$ in our setting because the data is α-mixing rather than near-epoch dependent. Accordingly, the variance bound on H_{1n} in that step implies that the variance of the HAC estimator is $O(n^{-1}\beta_n^{3d})$ where β_n is the bandwidth and d is the dimension of the space. In our case, by (A.1), the bandwidth corresponds to $\beta_n = 2r_n + \kappa_n = O(\sqrt{n/m_n})$, so $n^{-1}\beta_n^{3d} = O(n^2/m_n^3)$, and part (b) of our theorem follows. $\square$

APPENDIX B: PROOFS

The proofs use the following definitions. Let $\mathbf{c}(C_k) \in \mathbb{R}^2$ be the centroid of cluster C_k. For $s \leq r_n - 1$, define

(B.1) $$J(s, C_k) = \big\{ j \in C_k \colon \rho(\mathbf{c}(C_k), j) \in [r_n - s - 1, r_n - s) \big\}.$$

For $s = 0$, this is the "boundary" of C_k, and as we increase s, $J(s, C_k)$ moves through contour sets within C_k that are increasingly further from the boundary. Also, for any two sets $S, T \subset \mathbb{R}^2$, let $\rho(S, T) = \min\{\rho(i, j) \colon i \in S, j \in T\}$.

The proofs make use of the following facts, which are a consequence of Lemma A.1 of [25] and use Assumption 2. Given that C_k has radius r_n, $|C_k| \leq c' r_n^2$ and $|J(0, C_k)| \leq c' r_n$ for some universal constant $c' > 0$ that does not depend on n or k. Also, since $J(s, C_k)$ is the boundary of a ball of radius $r_n - s$, $|J(s, C_k)| \leq c'(r_n - s)$.

LEMMA B.1. *Recall the definition of $[B]$ from* (5). *Under the assumptions of Theorem 1,* $[B] \lesssim (n m_n)^{-1/2}$.

PROOF. *Step 1.* We first establish covariance bounds. Fix i, j such that $j \notin \Lambda_i$, the latter defined in (4), and set $s = \rho(i, j)$. Trivially,

$$|\mathrm{Cov}(Z_i, Z_j)| = |\mathrm{Cov}(Z_i, Z_j)|\mathbf{1}\{s \leq 4 r_n\} + |\mathrm{Cov}(Z_i, Z_j)|\mathbf{1}\{s > 4 r_n\}.$$

First consider the case $s \leq 4 r_n$. Let $\mathcal{C}_i$ be the cluster containing unit i, $\mathcal{F}_i(r) = \{D_j \colon \rho(i, j) \leq r \text{ or } j \in \mathcal{C}_i\}$, and $X_i^r = \mathbf{E}[Z_i \mid \mathcal{F}_i(r)]$. As a preliminary result, we bound the discrepancy between Z_i and X_i^s.

Let $t = \rho(\{i\}, J(0, \mathcal{C}_i))$, the distance between i and the nearest unit in the boundary of $\mathcal{C}_i$. By Assumption 3, for any $q > 0$,

$$\mathbf{E}\big[|Z_i - X_i^s|^q \mid T_{1i} = 1\big] = p_{1i}^{-q} \mathbf{E}\big[|Y_i(\mathbf{D}) - \mathbf{E}[Y_i(\mathbf{D}) \mid \mathcal{F}_i(s)]|^q \mid T_{1i} = 1\big]$$
$$\leq p_{1i}^{-q} \psi(\max\{t, s\})^q.$$

The equality holds because $\mathcal{F}_i(s)$ conditions on D_i, and by Assumption 4, $T_{1i} = 1$ implies $D_i = 1$, which implies $T_{0i} = 0$. The inequality holds because $\mathcal{F}_i(s)$ fixes $\{D_j \colon j \in Q(i, t) \cup Q(i, s)\}$ at their realized values. Similarly, $\mathbf{E}[|Z_i - X_i^s|^q \mid T_{0i} = 1] \leq p_{0i}^{-q} \psi(\max\{t, s\})^q$, so by the law of total probability and Remark 3, for some universal constant $c'' > 0$,

(B.2) $$\mathbf{E}[|Z_i - X_i^s|^q]^{1/q} \leq c'' \psi(\max\{t, s\}).$$

Define $R_i = Z_i - X_i^{\kappa_n}$. Since $j \notin \Lambda_i$, $|\mathrm{Cov}(X_i^{\kappa_n}, X_j^{\kappa_n})| = 0$ by Assumption 4. Applying the Cauchy-Schwarz and Jensen's inequalities and (B.2) for $q = 2$,

$$|\mathrm{Cov}(Z_i, Z_j)| \leq |\mathrm{Cov}(X_i^{\kappa_n}, X_j^{\kappa_n})| + |\mathrm{Cov}(X_i^{\kappa_n}, R_j)| + |\mathrm{Cov}(R_i, X_j^{\kappa_n})| + |\mathrm{Cov}(R_i, R_j)|$$
$$\leq 2c''(\|Z_i\|_2 \psi(\max\{t, \kappa_n\}) + \|Z_j\|_2 \psi(\max\{t, \kappa_n\}) + \psi(\max\{t, \kappa_n\})^2)$$
$$\leq c\, \psi(\max\{t, \kappa_n\})$$

for some universal $c > 0$.

Next consider the case $s > 4 r_n$. Abbreviate $X_i = X_i^{s/2 - r_n}$, and redefine $R_i = Z_i - X_i$. Note that $\rho(Q(i, s/2 - r_n), Q(j, s/2 - r_n)) > 2 r_n$ and $2 r_n$ is the diameter of a cluster, so $X_i \perp\!\!\!\perp X_j$. Consequently, following the previous argument,

(B.3) $$|\mathrm{Cov}(Z_i, Z_j)| \leq |\mathrm{Cov}(X_i, X_j)| + |\mathrm{Cov}(X_i, R_j)| + |\mathrm{Cov}(R_i, X_j)| + |\mathrm{Cov}(R_i, R_j)|$$
$$\leq c\, \psi(\max\{t, s/2 - r_n\}).$$

Step 2. For any $c \in \mathbb{R}$, let $\lfloor c \rfloor$ denote c rounded down to the nearest integer. Using the covariance bounds derived in step 1,

(B.4)

$$\frac{1}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \notin \Lambda_i} |\mathrm{Cov}(Z_i, Z_j)| \leq \frac{c}{n^2} \sum_{k=1}^{m_n} \sum_{s=1}^{\lfloor 2R_n \rfloor} \sum_{t=0}^{\lfloor \min\{s, r_n - 1\} \rfloor} \sum_{i \in J(t, C_k)} \sum_{j \notin \Lambda_i} \mathbf{1}\{\rho(i,j) \in [s-1, s)\}$$

$$\times \left(\psi(\max\{t, r_n/2\}) \mathbf{1}\{s \leq 4r_n\} + \psi(\max\{t, s/2 - r_n\}) \mathbf{1}\{s > 4r_n\} \right)$$

$$\equiv [B1] + [B2],$$

where $[B1]$ takes the part involving $s \leq 4r_n$ and $[B2]$ the remainder. Note that t can be at most $r_n - 1$ because $J(r_n - 1, C_k)$ is the 1-ball centered at the centroid of C_k, and it can be at most s because $\rho(i,j) \in [s-1, s)$ and $j \notin C_k$ since $j \notin \Lambda_i$.

As discussed at the start of this section, $\sum_{j \notin \Lambda_i} \mathbf{1}\{\rho(i,j) \in [s-1, s)\} \leq \sum_{j \in \mathcal{N}_n} \mathbf{1}\{\rho(i,j) \in [s-1, s)\} \leq c' s$ and $|J(t, C_k)| \leq c'(r_n - t)$ for some universal $c' > 0$ by Assumption 2. Then by Assumption 3,

$$[B1] \leq c \frac{m_n}{n^2} \sum_{s=1}^{\lfloor 4r_n \rfloor} c' s \sum_{t=0}^{\lfloor \min\{s, r_n - 1\} \rfloor} c'(r_n - t)\psi(t)$$

(B.5)

$$\lesssim \frac{m_n}{n^2} \sum_{s=1}^{\lfloor 4r_n \rfloor} s \left(r_n \sum_{t=0}^{\infty} \psi(t) + \sum_{t=0}^{\infty} t \, \psi(t) \right) \lesssim \frac{m_n}{n^2} r_n^2 r_n \lesssim \frac{1}{\sqrt{n m_n}}.$$

Finally,

$$[B2] \leq c \frac{m_n}{n^2} \sum_{s=\lfloor 4r_n \rfloor}^{\lfloor 2R_n \rfloor} c' s \sum_{t=0}^{\lfloor \min\{s, r_n - 1\} \rfloor} c'(r_n - t)\psi(s/2 - r_n)$$

$$\lesssim \frac{m_n}{n^2} \sum_{t=0}^{\lfloor r_n - 1 \rfloor} (r_n - t) \sum_{s=\lfloor 4r_n \rfloor}^{\lfloor 2R_n \rfloor} s \, \psi(s/2 - r_n) \lesssim \frac{m_n}{n^2} r_n^2 \sum_{s=\lfloor 4r_n \rfloor}^{\lfloor 2R_n \rfloor} s \, \psi(s/2 - r_n)$$

(B.6)

$$\lesssim \frac{m_n}{n^2} r_n^2 r_n \lesssim \frac{1}{\sqrt{n m_n}}.$$

$\square$

PROOF OF THEOREM 3. Recall that $\mathcal{F}_i = \{D_j : j \in \mathcal{C}_i \cup \mathcal{N}(i, \kappa_n)\}$, and define $R_i = Z_i - \mathbf{E}[Z_i \mid \mathcal{F}_i]$.

Step 1. We show that

$$\mathbf{E}\left[\left(\sqrt{m_n} \frac{1}{n} \sum_{i \in \mathcal{N}_n} R_i \right)^2 \right] = o(1).$$

Recalling (4), the left-hand side equals

$$\frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \in \Lambda_i} \mathbf{E}[R_i R_j] + \frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \notin \Lambda_i} \mathbf{E}[R_i R_j] \equiv [C] + [D].$$

If $j \in \Lambda_i$, then $\rho(i,j) \leq 3r_n$. Using this fact and (B.2) with $s = r_n/2$,

$$|[C]| \leq \frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \mathbf{E}[R_i^2] + \frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{s=1}^{\lfloor 3r_n \rfloor} \sum_{j \neq i} (c'' \psi(r_n/2))^2 \mathbf{1}\{\rho(i,j) \in [s-1, s)\}$$

$$\lesssim \frac{m_n}{n} + \frac{m_n}{n} \sum_{s=1}^{\lfloor 3r_n \rfloor} s\,\psi(r_n/2)^2 \lesssim \frac{m_n}{n} + \psi(r_n/2)^2 = o(1).$$

For $|[D]|$, we first establish some covariance bounds. Fix i, j such that $j \notin \Lambda_i$, and let $s = \rho(i, j)$. Let $t = \rho(\{i\}, J(0, \mathcal{C}_i))$, the distance between i and the nearest unit in the boundary of $\mathcal{C}_i$. By Assumption 3,

$$\mathbf{E}[|R_i|^2 \mid T_{1i} = 1] = p_{1i}^{-2} \mathbf{E}\big[|Y_i(\boldsymbol{D}) - \mathbf{E}[Y_i(\boldsymbol{D}) \mid \mathcal{F}_i]|^2 \mid T_{1i} = 1\big] \leqslant p_{1i}^{-2} \psi(t)^2.$$

Similarly, $\mathbf{E}[|R_i|^2 \mid T_{0i} = 1] \leqslant p_{0i}^{-2} \psi(t)^2$, so by the law of total probability and Remark 3, for some universal constant $c > 0$,

$$\mathbf{E}[|R_i|^2] \leqslant c\,\psi(t)^2 \quad \text{and} \quad |\mathrm{Cov}(R_i, R_j)| \leqslant c\,\psi(t).$$

We derive an alternate bound for the case $s > 4r_n$. Let $\mathcal{F}_i(r) = \{D_j : \rho(i, j) \leqslant r \text{ or } j \in \mathcal{C}_i\}$ and $X_i = \mathbf{E}[Z_i \mid \mathcal{F}_i(s/2 - r_n)]$. Then

$$S_i \equiv R_i - \mathbf{E}[R_i \mid \mathcal{F}_i(s/2 - r_n)] = \tilde{R}_i \equiv Z_i - X_i$$

since $\mathcal{N}(i, s/2 - r_n) \supseteq \mathcal{N}(i, \kappa_n)$. Moreover, $X_i \perp\!\!\!\perp X_j$ since $\rho(Q(i, s/2 - r_n), Q(j, s/2 - r_n)) > 2r_n$ and $2r_n$ is the diameter of a cluster. Consequently, by (B.3),

$$|\mathrm{Cov}(R_i, R_j)| \leqslant |\mathrm{Cov}(X_i, X_j)| + |\mathrm{Cov}(X_i, S_j)| + |\mathrm{Cov}(S_i, X_j)| + |\mathrm{Cov}(S_i, S_j)|$$

$$= |\mathrm{Cov}(X_i, \tilde{R}_j)| + |\mathrm{Cov}(\tilde{R}_i, X_j)| + |\mathrm{Cov}(\tilde{R}_i, \tilde{R}_j)| \leqslant c\,\psi(s/2 - r_n).$$

Applying the covariance bounds,

$$|[D]| \leqslant \frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \notin \Lambda_i} |\mathbf{E}[R_i R_j]|$$

$$\leqslant c\frac{m_n}{n^2} \sum_{k=1}^{m_n} \sum_{s=1}^{\lfloor 2R_n \rfloor} \sum_{t=0}^{\lfloor \min\{s, r_n - 1\} \rfloor} \sum_{i \in J(t, C_k)} \sum_{j \notin \Lambda_i} \mathbf{1}\{\rho(i, j) \in [s-1, s)\}$$

$$\times \big(\psi(t)\mathbf{1}\{s \leqslant 4r_n\} + \psi(s/2 - r_n)\mathbf{1}\{s > 4r_n\}\big),$$

which is order $(m_n/n)^{1/2} = o(1)$ by an argument similar to (B.5) and (B.6).

Step 2. We show that

$$\sigma_n^{-1} \sqrt{m_n} \frac{1}{n} \sum_{i \in \mathcal{N}_n} \big(\mathbf{E}[Z_i \mid \mathcal{F}_i] - \mathbf{E}[Z_i]\big) \xrightarrow{d} \mathcal{N}(0, 1).$$

First, let $\tilde{\sigma}_n^2 = \mathrm{Var}(\sqrt{m_n} n^{-1} \sum_{i \in \mathcal{N}_n} \mathbf{E}[Z_i \mid \mathcal{F}_i])$. By Minkowski's inequality,

(B.7) $$|\sigma_n - \tilde{\sigma}_n| \leqslant \mathrm{Var}\left(\sqrt{m_n}\frac{1}{n}\sum_{i \in \mathcal{N}_n}(Z_i - \mathbf{E}[Z_i \mid \mathcal{F}_i])\right)^{1/2},$$

which is $o(1)$ by step 1. By Assumption 5, it then suffices to show

(B.8) $$\tilde{\sigma}_n^{-1} \sqrt{m_n} \frac{1}{n} \sum_{i \in \mathcal{N}_n} \big(\mathbf{E}[Z_i \mid \mathcal{F}_i] - \mathbf{E}[Z_i]\big) \xrightarrow{d} \mathcal{N}(0, 1).$$

We apply Lemma B.2 with $X_i = n^{-1} m_n^{1/2}(\mathbf{E}[Z_i \mid \mathcal{F}_i] - \mathbf{E}[Z_i])$ and dependency graph $\boldsymbol{A}$ defined after (8). The maximum degree of $\boldsymbol{A}$ is at most $16 \max_k |C_k|$, and $\max_k |C_k| = O(n/m_n)$. Therefore, by Assumptions 1 and 5,

(B.9) $$\lesssim \left(\frac{n}{m_n}\right)^2 n \left(\frac{\sqrt{m_n}}{n}\right)^3 + \left(\frac{n}{m_n}\right)^{3/2} \sqrt{n \left(\frac{\sqrt{m_n}}{n}\right)^4} \lesssim m_n^{-1/2}.$$

Since $m_n \to \infty$, (B.8) follows. $\qquad\square$

LEMMA B.2 ([41], Theorem 3.6). *Let $\{X_i\}_{i=1}^n$ be random variables with dependency graph $\mathbf{A}$ such that $\mathbf{E}[X_i^4] < \infty$ and $\mathbf{E}[X_i] = 0$. Define $\sigma^2 \equiv Var(\sum_{i=1}^n X_i)$, $\mathcal{W} = \sum_{i=1}^n X_i/\sigma^2$, and $\Psi = \max_{i=1,\dots,n} \sum_{j=1}^n A_{ij}$. For $\mathcal{Z} \sim \mathcal{N}(0,1)$,*

$$(B.9) \qquad d(\mathcal{W}, \mathcal{Z}) \leqslant \frac{\Psi^2}{\sigma^3} \sum_{i=1}^n \mathbf{E}[|X_i|^3] + \frac{\sqrt{28}\Psi^{3/2}}{\sqrt{\pi}\sigma^2} \left(\sum_{i=1}^n \mathbf{E}[X_i^4] \right)^{1/2},$$

where $d(\cdot,\cdot)$ is the Wasserstein distance.

SUPPLEMENTARY MATERIAL

supplement.zip
This zip file contains a PDF with proofs omitted in this text and code used to produce the simulation results in §5.

REFERENCES

[1] ANDREWS, D. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica* 817–858.

[2] ANSELIN, L. (2001). Spatial Econometrics. In *A Companion to Theoretical Econometrics* (B. Baltagi, ed.) 14 Blackwell Publishing Ltd.

[3] ARONOW, P. and SAMII, C. (2017). Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment. *The Annals of Applied Statistics* **11** 1912–1947.

[4] ARONOW, P., SAMII, C. and WANG, Y. (2020). Design-Based Inference for Spatial Experiments with Interference. *arXiv preprint arXiv:2010.13599*.

[5] BAIRD, S., BOHREN, J., MCINTOSH, C. and ÖZLER, B. (2018). Optimal Design of Experiments in the Presence of Interference. *Review of Economics and Statistics* **100** 844–860.

[6] BASSE, G. and AIROLDI, E. (2018). Model-Assisted Design of Experiments in the Presence of Network-Correlated Outcomes. *Biometrika* **105** 849–858.

[7] BASSE, G., FELLER, A. and TOULIS, P. (2019). Randomization Tests of Causal Effects Under Interference. *Biometrika* **106** 487–494.

[8] BLATTMAN, C., GREEN, D., ORTEGA, D. and TOBÓN, S. (2021). Place-Based Interventions at Scale: The Direct and Spillover Effects of Policing and City Services on Crime. *Journal of the European Economic Association* **19** 2022–2051.

[9] CHIN, A. (2019). Regression Adjustments for Estimating the Global Treatment Effect in Experiments with Interference. *Journal of Causal Inference* **7**.

[10] CHOI, D. (2017). Estimation of Monotone Treatment Effects in Network Experiments. *Journal of the American Statistical Association* **112** 1147–1155.

[11] CLIFF, A. and ORD, J. (1973). *Spatial Autocorrelation*. London: Pion.

[12] CLIFF, A. and ORD, J. (1981). *Spatial Processes: Models and Applications*. London: Pion.

[13] CONLEY, T. (1999). GMM Estimation with Cross Sectional Dependence. *Journal of Econometrics* **92** 1–45.

[14] DONNELLY, C., WOODROFFE, R., COX, D., BOURNE, J., GETTINBY, G., LE FEVRE, A., MCINERNEY, J. and MORRISON, I. (2003). Impact of Localized Badger Culling on Tuberculosis Incidence in British Cattle. *Nature* **426** 834–837.

[15] ECKLES, D., KARRER, B. and UGANDER, J. (2017). Design and Analysis of Experiments in Networks: Reducing Bias from Interference. *Journal of Causal Inference* **5**.

[16] FORASTIERE, L., AIROLDI, E. and MEALLI, F. (2021). Identification and Estimation of Treatment and Interference Effects in Observational Studies on Networks. *Journal of the American Statistical Association* **116** 901–918.

[17] GETIS, A. (2008). A History of the Concept of Spatial Autocorrelation: A Geographer's Perspective. *Geographical Analysis* **40** 297–309.

[18] GIFFIN, A., REICH, B., YANG, S. and RAPPOLD, A. (2020). Generalized Propensity Score Approach to Causal Inference with Spatial Interference. *arXiv preprint arXiv:2007.00106*.

[19] HARSHAW, C., SÄVJE, F., EISENSTAT, D., MIRROKNI, V. and POUGET-ABADIE, J. (2021). Design and Analysis of Bipartite Experiments Under a Linear Exposure-Response Model. *arXiv preprint arXiv:2103.06392*.

[20] HAYES, R. and MOULTON, L. (2017). *Cluster Randomised Trials*. Chapman and Hall/CRC.

[21] HU, Y., LI, S. and WAGER, S. (2022). Average Direct and Indirect Causal Effects Under Interference. *Biometrika (forthcoming)*.

[22] IMBENS, G. and RUBIN, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

[23] JAGADEESAN, R., PILLAI, N. and VOLFOVSKY, A. (2020). Designs for Estimating the Treatment Effect in Networks with Interference. *The Annals of Statistics* **48** 679–712.

[24] JENISH, N. (2016). Spatial Semiparametric Model with Endogenous Regressors. *Econometric Theory* **32** 714–739.

[25] JENISH, N. and PRUCHA, I. (2009). Central Limit Theorems and Uniform Laws of Large Numbers for Arrays of Random Fields. *Journal of Econometrics* **150** 86–98.

[26] JENISH, N. and PRUCHA, I. (2011). On Spatial Processes and Asymptotic Inference Under Near-Epoch Dependence. *U. Maryland working paper*.

[27] JENISH, N. and PRUCHA, I. (2012). On Spatial Processes and Asymptotic Inference Under Near-Epoch Dependence. *Journal of Econometrics* **170** 178–190.

[28] LAHIRI, S. (1996). On Inconsistency of Estimators Based on Spatial Data Under Infill Asymptotics. *Sankhyā: The Indian Journal of Statistics, Series A* 403–417.

[29] LAHIRI, S. (2003). Central Limit Theorems for Weighted Sums of a Spatial Process Under a Class of Stochastic and Fixed Designs. *Sankhyā: The Indian Journal of Statistics* 356–388.

[30] LAHIRI, S. and ZHU, J. (2006). Resampling Methods for Spatial Regression Models Under a Class of Stochastic Designs. *The Annals of Statistics* **34** 1774–1813.

[31] LEUNG, M. (2022). Causal Inference Under Approximate Neighborhood Interference. *Econometrica* **90** 267-293.

[32] LEUNG, M. (2022). Network Cluster-Robust Inference. *arXiv preprint arXiv:2103.01470*.

[33] LEUNG, M. (2022). Supplement to "Rate-Optimal Cluster-Randomized Designs for Spatial Interference". *DOI: 10.1214/22-AOS2224SUPP*.

[34] MANSKI, C. (2013). Identification of Treatment Response with Social Interactions. *The Econometrics Journal* **16** S1–S23.

[35] MIGUEL, E. and KREMER, M. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica* **72** 159–217.

[36] PALER, L., SAMII, C., LISIECKI, M. and MOREL, A. (2015). Social and Environmental Impact of the Community Rangers Program in Aceh Technical Report, World Bank, Washington, DC.

[37] PARK, C. and KANG, H. (2021). Assumption-Lean Analysis of Cluster Randomized Trials in Infectious Diseases for Intent-to-Treat Effects and Spillover Effects Among a Vulnerable Subpopulation. *Journal of the American Statistical Association (forthcoming)*.

[38] PENG, R., SUN, H. and ZANETTI, L. (2017). Partitioning Well-Clustered Graphs: Spectral Clustering Works! *SIAM Journal on Computing* **46** 710–743.

[39] POLLMANN, M. (2020). Causal Inference for Spatial Treatments. *arXiv preprint arXiv:2011.00373*.

[40] POUGET-ABADIE, J., MIRROKNI, V., PARKES, D. and AIROLDI, E. (2018). Optimizing Cluster-Based Randomized Experiments Under Monotonicity. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2090–2099.

[41] ROSS, N. (2011). Fundamentals of Stein's Method. *Probability Surveys* **8** 210–293.

[42] SÄVJE, F. (2021). Causal Inference with Misspecified Exposure Mappings. *arXiv preprint arXiv:2103.06471*.

[43] SÄVJE, F., ARONOW, P. and HUDGENS, M. (2021). Average Treatment Effects in the Presence of Unknown Interference. *The Annals of Statistics* **49** 673–701.

[44] SUSSMAN, D. and AIROLDI, E. (2017). Elements of Estimation Theory for Causal Effects in the Presence of Network Interference. *arXiv preprint arXiv:1702.03578*.

[45] TOULIS, P. and KAO, E. (2013). Estimation of Causal Peer Influence Effects. In *International Conference on Machine Learning* 1489–1497.

[46] UGANDER, J., KARRER, B., BACKSTROM, L. and KLEINBERG, J. (2013). Graph Cluster Randomization: Network Exposure to Multiple Universes. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 329–337.

[47] UGANDER, J. and YIN, H. (2020). Randomized Graph Cluster Randomization. *arXiv preprint arXiv:2009.02297*.

[48] VALCU, M. and KEMPENAERS, B. (2010). Spatial Autocorrelation: an Overlooked Concept in Behavioral Ecology. *Behavioral Ecology* **21** 902–905.

[49] VERBITSKY-SAVITZ, N. and RAUDENBUSH, S. (2012). Causal Inference Under Interference in Spatial Settings: A Case Study Evaluating Community Policing Program in Chicago. *Epidemiologic Methods* **1** 107–130.

[50] VIVIANO, D. (2020). Experimental Design Under Network Interference. *arXiv preprint arXiv:2003.08421*.

[51] ZIGLER, C. and PAPADOGEORGOU, G. (2021). Bipartite Causal Inference with Interference. *Statistical Science* **36** 109.

SUPPLEMENT TO "RATE-OPTIMAL CLUSTER-RANDOMIZED DESIGNS FOR SPATIAL INTERFERENCE"

BY MICHAEL P. LEUNG[1]

[1]*Department of Economics, UC Santa Cruz, leungm@ucsc.edu*

This supplement contains proofs omitted in [1].

PROOF OF THEOREM 2. Let $\mathcal{N}_n$ equal $Q(\mathbf{0}, 0.5\sqrt{n}) \cap \mathbb{Z}^2$, and arbitrarily assign half of the units at the coincidental boundaries of two clusters to each of the clusters ($|\mathcal{N}_n|$ may not exactly equal n, but we may add additional units just outside of $Q(\mathbf{0}, 0.5\sqrt{n})$ to make up the difference). Let $\psi'(\cdot)$ denote the derivative of $\psi(\cdot)$. We construct potential outcomes as follows:

$$Y_i(\boldsymbol{d}) = \left(\prod_{j \in Q(i, r_n/2)} d_j\right)\left(\sum_{s=1}^{2R_n} |\psi'(s)| \frac{\sum_{j \in \mathcal{N}_n} d_j \mathbf{1}\{\rho(i,j) = s\}}{\sum_{j \in \mathcal{N}_n} \mathbf{1}\{\rho(i,j) = s\}}\right)$$

for $R_n = 0.5\sqrt{n}$. Note that the sum in the denominator is at always least one.

Assumption 1 holds since

$$\sum_{s=1}^{\infty} |\psi'(s)| \leqslant |\psi'(1)| + \int_1^{\infty} |\psi'(s)|\, \mathrm{d}s = |\psi'(1)| + \psi(1) < \infty.$$

To verify Assumption 3, for any $s \geqslant 0$, fix $\boldsymbol{d}, \boldsymbol{d}' \in \{0,1\}^n$ such that $d_j = d'_j$ for all $j \in \mathcal{N}(i,s)$. If it is not the case that $d_j = 1$ for all $j \in Q(i, r_n/2)$, then $|Y_i(\boldsymbol{d}) - Y_i(\boldsymbol{d}')| = 0$, so assume it is the case. Then

$$|Y_i(\boldsymbol{d}) - Y_i(\boldsymbol{d}')| = \left|\sum_{t=\max\{s, r_n/2\}}^{2R_n} |\psi'(t)| \frac{\sum_{j \in \mathcal{N}_n}(d_j - d'_j)\mathbf{1}\{\rho(i,j) = t\}}{\sum_{j \in \mathcal{N}_n} \mathbf{1}\{\rho(i,j) = t\}}\right|$$

$$\leqslant \sum_{t=s}^{2R_n} |\psi'(t)| \leqslant |\psi'(s)| + \int_s^{\infty} |\psi'(t)|\, \mathrm{d}t = |\psi'(s)| + \psi(s),$$

and $\sum_{s=1}^{\infty} s\left(|\psi'(s)| + \psi(s)\right) < \infty$, as desired.

We next derive the lower bound on the bias. For any $c \in \mathbb{R}$, let $\lceil c \rceil$ ($\lfloor c \rfloor$) denote c rounded up (down) to the nearest integer. For $\Gamma_i = \{j \in \mathcal{N}_n \colon j \in C_k \text{ for some } k \text{ s.t. } \mathcal{N}(i, \kappa_n) \cap C_k \neq \varnothing\}$,

$$\mathbf{E}\left[Y_i\left(\frac{T_{1i}}{p_{1i}} - \frac{T_{0i}}{p_{0i}}\right)\right] - \left(Y_i(\mathbf{1}_n) - Y_i(\mathbf{0}_n)\right) = \mathbf{E}[Y_i(\boldsymbol{D}) \mid T_{1i} = 1] - Y_i(\mathbf{1}_n)$$

$$= \sum_{t=1}^{2R_n} |\psi'(t)| \frac{\sum_{j \notin \Gamma_i}(p-1)\mathbf{1}\{\rho(i,j) = t\}}{\sum_{j \in \mathcal{N}_n} \mathbf{1}\{\rho(i,j) = t\}} \equiv w_i.$$

The last line follows because, by Assumption 4, $T_{1i} = 1$ implies $D_j = 1$ for all $j \in \Gamma_i$, and for all $j \notin \Gamma_i$, $\mathbf{E}[D_j \mid T_{1i} = 1] = p$. By construction of $\mathcal{N}_n$, for all $t \in [\lceil 1.5r_n \rceil + 1, \lfloor R_n \rfloor]$,

(S.1) $$\liminf_{n \to \infty} \frac{1}{n} \sum_{i \in \mathcal{N}_n} \frac{\sum_{j \notin \Gamma_i} \mathbf{1}\{\rho(i,j) = t\}}{\sum_{j \in \mathcal{N}_n} \mathbf{1}\{\rho(i,j) = t\}} > 0.$$

To see this, consider any i for which $i \in Q(\boldsymbol{c}(C_k), r_n/2)$ for some cluster k. That is, i lies in the "interior" of some cluster. For any such i, a nontrivial share of units ℓ such that $\rho(i, \ell) = t$

for $t \geqslant \lceil 1.5r_n \rceil + 1$ satisfy $j \notin \Gamma_i$ in $\mathbb{Z}^2$. For example, this share is unity if $i = \boldsymbol{c}(C_k)$. Since a nontrivial share of units i lie in the interior of some cluster, (S.1) follows. We consider $t \geqslant \lceil 1.5r_n \rceil + 1$ because for some clusters k and $i \in Q(\boldsymbol{c}(C_k), r_n/2)$, the closest $j \notin \Gamma_i$ satisfies $\rho(i, j) \geqslant \lceil 1.5r_n \rceil + 1$.

We therefore have

$$
|\mathbf{E}[\hat{\theta}] - \theta_n| = \left| \frac{1}{n} \sum_{i \in \mathcal{N}_n} w_i \right| \gtrsim \sum_{t = \lceil 1.5r_n \rceil + 1}^{\lfloor R_n \rfloor} |\psi'(t)|
$$

$$
\geqslant \int_{\lceil 1.5r_n \rceil + 1}^{\lfloor R_n \rfloor} |\psi'(t)| \, \mathrm{d}t = \psi(\lceil 1.5r_n \rceil + 1) - \psi(\lfloor R_n \rfloor).
$$

Since $r_n = R_n / \sqrt{m_n}$, the right-hand side is of order $\psi(1.5r_n)$.

Lastly, we consider the variance: $\mathrm{Cov}(Z_i, Z_j)$ equals

$$
\sum_{k \in \mathcal{N}_n} \sum_{\ell \in \mathcal{N}_n} \sum_{s=1}^{2R_n} \sum_{t=1}^{2R_n} \frac{|\psi'(s)| \mathbf{1}\{\rho(i, k) = s\}}{p_{1i} \sum_{k' \in \mathcal{N}_n} \mathbf{1}\{\rho(i, k') = s\}} \frac{|\psi'(t)| \mathbf{1}\{\rho(j, \ell) = t\}}{p_{1j} \sum_{\ell' \in \mathcal{N}_n} \mathbf{1}\{\rho(j, \ell') = t\}} \mathrm{Cov}(T_{1i} D_k, T_{1j} D_\ell).
$$

Note that $\mathrm{Cov}(T_{1i} D_k, T_{1j} D_\ell)$ is uniformly bounded away from zero if $j \in \Lambda_i$ and is always non-negative by the design. Also, p_{1i} and p_{0i} are uniformly bounded away from zero by Remark 3, so $\mathrm{Var}(n^{-1} \sum_{i \in \mathcal{N}_n} Z_i)$ is bounded below by

$$
\frac{c}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \in \Lambda_i} \sum_{s=1}^{2R_n} \sum_{t=1}^{2R_n} \sum_{k \in \mathcal{N}_n} \sum_{\ell \in \mathcal{N}_n} \frac{|\psi'(s)| \mathbf{1}\{\rho(i, k) = s\}}{\sum_{k' \in \mathcal{N}_n} \mathbf{1}\{\rho(i, k') = s\}} \frac{|\psi'(t)| \mathbf{1}\{\rho(j, \ell) = t\}}{\sum_{\ell' \in \mathcal{N}_n} \mathbf{1}\{\rho(j, \ell') = t\}}
$$

$$
= \frac{c}{n^2} \sum_{i \in \mathcal{N}_n} |\Lambda_i| \sum_{s=1}^{2R_n} \sum_{t=1}^{2R_n} |\psi'(s)| |\psi'(t)|
$$

for some universal constant $c > 0$. Since $\min_i |\Lambda_i| \gtrsim n/m_n$, $\mathrm{Var}(\hat{\theta}) \gtrsim m_n^{-1}$. $\qquad\square$

PROOF OF THEOREM 4. As shown in (B.7), $|\sigma_n - \tilde{\sigma}_n| = o(1)$, and by Assumption 4,

$$
\tilde{\sigma}_n^2 = \frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} \mathrm{Cov}(\mathbf{E}[Z_i \mid \mathcal{F}_i], \mathbf{E}[Z_j \mid \mathcal{F}_j]) A_{ij}.
$$

Then by Assumption 5, it is enough to show

(S.2) $\qquad \hat{\sigma}^2 - \mathcal{R}_n = \dfrac{m_n}{n^2} \displaystyle\sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} \mathrm{Cov}(\mathbf{E}[Z_i \mid \mathcal{F}_i], \mathbf{E}[Z_j \mid \mathcal{F}_j]) A_{ij} + o_p(1).$

The first two steps below will establish that

(S.3) $\qquad \hat{\sigma}^2 - \mathcal{R}_n = \dfrac{m_n}{n^2} \displaystyle\sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} (Z_i - \mathbf{E}[Z_i])(Z_j - \mathbf{E}[Z_j]) A_{ij} + o_p(1).$

Step 1. In the formula of $\hat{\sigma}^2$, replace $\bar{Z}$ with $\bar{Z} - \mathbf{E}[\bar{Z}] + \mathbf{E}[\bar{Z}]$ to obtain

$$
\hat{\sigma}^2 = \underbrace{\frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} (Z_i - \mathbf{E}[\bar{Z}])(Z_j - \mathbf{E}[\bar{Z}]) A_{ij}}_{\check{\sigma}^2}
$$

$$
+ (\mathbf{E}[\bar{Z}] - \bar{Z}) \frac{2m_n}{n^2} \sum_{i \in \mathcal{N}_n} (Z_i - \mathbf{E}[\bar{Z}]) \sum_{j \in \mathcal{N}_n} A_{ij} + (\mathbf{E}[\bar{Z}] - \bar{Z})^2 \frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} A_{ij}.
$$

The last two terms on the right-hand side are $o_p(1)$ since $\sum_{j \in \mathcal{N}_n} A_{ij} \leqslant 16 \max_k |C_k| \lesssim n/m_n$, and $|\bar{Z} - \mathbf{E}[\bar{Z}]| \xrightarrow{p} 0$ by Theorem 1.

Step 2. In the formula of $\check{\sigma}^2$, replace Z_i with $Z_i - \mathbf{E}[Z_i] + \mathbf{E}[Z_i]$ to obtain

$$\check{\sigma}^2 = \frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} (Z_i - \mathbf{E}[Z_i])(Z_j - \mathbf{E}[Z_j]) A_{ij}$$

$$+ \frac{2m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} (Z_i - \mathbf{E}[Z_i])(\mathbf{E}[Z_j] - \mathbf{E}[\bar{Z}]) A_{ij} + \mathcal{R}_n.$$

We show that the second term on the right-hand side is $o_p(1)$. For $W_i \equiv \sum_{j \in \mathcal{N}_n} (\mathbf{E}[Z_j] - \mathbf{E}[\bar{Z}]) A_{ij}$,

$$\mathbf{E}\left[\left| \frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} (Z_i - \mathbf{E}[Z_i])(\mathbf{E}[Z_j] - \mathbf{E}[\bar{Z}]) A_{ij} \right| \right]$$

$$\leqslant \mathbf{E}\left[\left(\frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} (Z_i - \mathbf{E}[Z_i]) W_i \right)^2 \right]^{1/2}$$

$$= \left(\frac{m_n^2}{n^4} \sum_{i \in \mathcal{N}_n} \sum_{j \in \Lambda_i} \mathrm{Cov}(Z_i, Z_j) W_i W_j + \frac{m_n^2}{n^4} \sum_{i \in \mathcal{N}_n} \sum_{j \notin \Lambda_i} \mathrm{Cov}(Z_i, Z_j) W_i W_j \right)^{1/2}$$

By Assumption 1 and Remark 3, Z_i is uniformly bounded. Since $\max_i \sum_{j \in \mathcal{N}_n} A_{ij} \lesssim n/m_n$, W_i is uniformly bounded by a constant times n/m_n. It follows from the variance bound for (5) that the last line is $o(1)$.

Step 3. Having proven (S.3) and the sufficiency of establishing (S.2), it remains to show that the following is $o_p(1)$:

$$\frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} (Z_i - \mathbf{E}[Z_i])(Z_j - \mathbf{E}[Z_j]) A_{ij}$$

$$- \frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} \mathrm{Cov}(\mathbf{E}[Z_i \mid \mathcal{F}_i], \mathbf{E}[Z_j \mid \mathcal{F}_j]) A_{ij}$$

$$= \frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} (Z_i Z_j - \mathbf{E}[\mathbf{E}[Z_i \mid \mathcal{F}_i] \mathbf{E}[Z_j \mid \mathcal{F}_j]]) A_{ij}$$

$$- 2 \frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} (Z_i - \mathbf{E}[Z_i]) \underbrace{\sum_{j \in \mathcal{N}_n} \mathbf{E}[Z_j] A_{ij}}_{\leqslant c \, n/m_n}.$$

The last term on the right-hand side is $o_p(1)$ by Theorem 1. The first term equals

$$\frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} (Z_i - \mathbf{E}[Z_i \mid \mathcal{F}_i])(Z_j - \mathbf{E}[Z_j \mid \mathcal{F}_j]) A_{ij}$$

$$+ 2 \frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} \mathbf{E}[Z_i \mid \mathcal{F}_i](Z_j - \mathbf{E}[Z_j \mid \mathcal{F}_j]) A_{ij}$$

$$+ \frac{m_n}{n^2} \sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} (\mathbf{E}[Z_i \mid \mathcal{F}_i] \mathbf{E}[Z_j \mid \mathcal{F}_j] - \mathbf{E}[\mathbf{E}[Z_i \mid \mathcal{F}_i] \mathbf{E}[Z_j \mid \mathcal{F}_j]]) A_{ij}$$

$$\equiv [E] + [F] + [G].$$

4

We claim that $[E]$ and $[F]$ are $o_p(1)$. Note that $Z_i - \mathbf{E}[Z_i \mid \mathcal{F}_i]$ equals

$$\frac{T_{1i}}{p_{1i}}(Y_i(\boldsymbol{D}) - \mathbf{E}[Y_i(\boldsymbol{D}) \mid T_{1i} = 1, \mathcal{F}_i]) - \frac{T_{0i}}{p_{0i}}(Y_i(\boldsymbol{D}) - \mathbf{E}[Y_i(\boldsymbol{D}) \mid T_{0i} = 1, \mathcal{F}_i]).$$

Furthermore, for any $t \in \{0, 1\}$,

$$T_{ti}|Y_i(\boldsymbol{D}) - \mathbf{E}[Y_i(\boldsymbol{D}) \mid T_{ti} = 1, \mathcal{F}_i]| \leqslant \psi(r_n/2)$$

by Assumption 3 since $T_{ti} = 1$ implies $D_j = t$ for all $j \in Q(i, r_n/2)$. Then because, as previously discussed, $\max_i \sum_{j \in \mathcal{N}_n} A_{ij} \lesssim n/m_n$,

$$[E] \lesssim \frac{m_n}{n^2} \cdot n \cdot \frac{n}{m_n} \cdot \psi(r_n/2)^2 = o(1) \quad \text{and} \quad [F] \lesssim \frac{m_n}{n^2} \cdot n \cdot \frac{n}{m_n} \cdot \psi(r_n/2) = o(1).$$

Since the mean of $[G]$ is zero, it remains to bound its variance, which equals

(S.4) $$\frac{m_n^2}{n^4} \sum_{i \in \mathcal{N}_n} \sum_{j \in \mathcal{N}_n} \sum_{k \in \mathcal{N}_n} \sum_{\ell \in \mathcal{N}_n} \mathrm{Cov}(\mathbf{E}[Z_i \mid \mathcal{F}_i]\mathbf{E}[Z_j \mid \mathcal{F}_j], \mathbf{E}[Z_k \mid \mathcal{F}_k]\mathbf{E}[Z_\ell \mid \mathcal{F}_\ell])A_{ij}A_{k\ell}.$$

The covariance term is zero if $k, \ell \notin \Lambda_i \cup \Lambda_j$, so by Assumption 1 and Remark 3, there exists a universal constant $c > 0$ such that

(S.5) $$\text{(S.4)} \leqslant c\frac{m_n^2}{n^4} \sum_{i \in \mathcal{N}_n} \sum_{j \in \Lambda_i} |\Xi_{ij}|$$

for $\Xi_{ij} = \{(k, \ell) \colon \ell \in \Lambda_k \text{ and either } k \in \Lambda_i \cup \Lambda_j, \ell \in \Lambda_i \cup \Lambda_j, \text{or both}\}$. Since $\max_i \Lambda_i \lesssim n/m_n$ and $\ell \in \Lambda_k$ implies $\rho(k, \ell) \leqslant 3r_n$,

$$\sup_{n \in \mathbb{N}} \max_{i,j \in \mathcal{N}_n} |\Xi_{ij}| \lesssim \underbrace{\frac{n}{m_n}r_n}_{\substack{\text{only } k \in \Lambda_i \cup \Lambda_j, \\ \ell \in \Lambda_k}} + \underbrace{r_n\frac{n}{m_n}}_{\substack{\text{only } \ell \in \Lambda_i \cup \Lambda_j, \\ \ell \in \Lambda_k}} + \underbrace{\frac{n^2}{m_n^2}}_{\substack{k, \ell \in \Lambda_i \cup \Lambda_j}} \lesssim \frac{n^2}{m_n^2}.$$

Therefore,

$$\text{(S.5)} \lesssim \frac{m_n^2}{n^4} \cdot n \cdot \frac{n}{m_n} \cdot \frac{n^2}{m_n^2} \lesssim m_n^{-1}.$$

$\square$

REFERENCES

[1] LEUNG, M. (forthcoming). Rate-Optimal Cluster-Randomized Designs for Spatial Interference. *The Annals of Statistics*.