NBER WORKING PAPER SERIES

LINEAR ESTIMATION OF GLOBAL AVERAGE TREATMENT EFFECTS

Stefan Faridani
Paul Niehaus

ABSTRACT

We study the problem of estimating the average causal effect of treating every member of a population, as opposed to none, using an experiment that treats only some. We consider settings where spillovers have global support and decay slowly with (a generalized notion of) distance. We derive the minimax rate over both estimators and designs, and show that it increases with the spatial rate of spillover decay. Estimators based on OLS regressions like those used to analyze recent large-scale experiments are consistent (though only after de-weighting), achieve the minimax rate when the DGP is linear, and converge faster than IPW-based alternatives when treatment clusters are small, providing one justification for OLS's ubiquity. When the DGP is nonlinear they remain consistent but converge slowly. We further address inference and bandwidth selection. Applied to the cash transfer experiment studied by Egger et al (2022) these methods yield a 20% larger estimated effect on consumption.

Stefan Faridani
Georgia Institute of Technology
sfaridani6@gatech.edu

Paul Niehaus
Department of Economics
University of California, San Diego
9500 Gilman Drive #0508
La Jolla, CA 92093
and NBER
pniehaus@ucsd.edu

1   Introduction

Economists often study situations in which some units received a given treatment in order to estimate the consequences of treating more of them. For example, Miguel and Kremer's (2004) evaluation of deworming medication in a sample of 2,300 schoolchildren shaped subsequent debate over whether to administer similar medication to all schoolchildren in Kenya and beyond. A key quantity for policy-making in such settings is the "global average treatment effect" (GATE), the average causal effect of treating all units within some population of interest as opposed to treating none of them.

Learning the GATE is challenging, even when treatment is randomly assigned, because it compares two counterfactual states which are never directly observed for any unit. The econometrican must extrapolate, for example, from the observed outcome for a unit with many treated neighbors to its counterfactual outcome had everyone been treated. This extrapolation requires an assumption controlling interference or "spillovers" between units.

The predominant approach has been to restrict the support of the spillovers via exposure mappings (Manski, 2013) which imply that treating one unit affects only a small number of neighbors (e.g., those within the same classroom or village).[1] But such assumptions are sometimes in uncomfortable tension with economic logic. Economic models often imply that everyone's behavior affects everyone else to some degree through prices, strategic interactions, and other mechanisms. Recent, large-scale field experiments have illustrated these kinds of far-reaching general-equilibrium effects (Muralidharan and Niehaus, 2017). Reforming a public workfare program in India, for example, affected market wages and employment, land rents, and firm entry, with significant spillovers across sub-district boundaries (Muralidharan et al., 2023). Causal channels like these are difficult to reconcile with strictly local spillovers. The result is that empirical practice in this area has moved ahead of its theoretical underpinnings. Researchers have specified linear regressions that include the mean treatment status of other units within a given radius as a regressor (e.g., Miguel and Kremer, 2004; Egger et al., 2022; Muralidharan et al., 2023) without corresponding theoretical results establishing whether, when, and for what estimands this allows for consistent or efficient estimation.

In this paper we study the estimation of the GATE, and in particular whether consistent and rate-efficient estimation can be achieved using estimators like those predominant in practice. The general approach we adopt is to bound the magnitude of spillovers, rather than restricting their support. The specific assumption we work with is that interference decays with at least a power $\eta > 0$ of distance. This yields a statistical framework better-suited to accommodate many economic ones: endogenous economic interactions decay geometrically with distance in the gravity (Allen et al., 2020) and market access (Donaldson and Hornbeck, 2016) traditions, for example. We employ a flexible notion of "distance" in order to make transparent the range of applications that fit the assumptions, including settings where spillovers are asymmetric (e.g. if pollution travels from upwind to downwind sites), do not exactly satisfy the triangle inequality (e.g. trade flows in a gravity model), or where the researcher is simply uncertain what notion of distance is most relevant and so uses the minimum of several proper metrics. Finally, we study

---

[1] The "Stable Unit Treatment Value Assumption" (SUTVA) is a very strong version, implying that *no* other units are affected.

the class of estimators that are linear functions of the observed outcomes, which includes all of those commonly used in both the applied and theoretical literatures.

In this setting we show that no linear estimator, regardless of the experimental design, can be guaranteed to converge to the GATE faster than $n^{-\frac{1}{2+1/\eta}}$. The fact that this is slower than the parametric rate $n^{-\frac{1}{2}}$ illustrates the intrinsic difficulty of estimating the GATE, while the fact that it is faster for large η illustrates the intuitive idea that the problem becomes less difficult when the population is economically less interconnected. The optimal rate can be achieved using inverse probability weighting (IPW) when the experimental design uses treatment clusters that grow with the population at the right rate. The key idea is to prevent the probability of any one unit having *all* of its neighbors (within some growing radius) treated or untreated from decaying to zero. These initial results build on and extend an important recent contribution by Leung (2022b), who demonstrates that when using a Horvitz-Thompson estimator in $\mathbb{R}^2$ the optimal rate is achievable using clusters that take the form of growing squares if $\eta > 1$. We extend this result to non-Euclidean spaces, optimize over all linear estimators as well as all experimental designs, and allow $\eta \in (0, \infty)$. The key technical step is to derive a lower bound on the rate of convergence in terms of the covering number of the population by demonstrating that spillovers with unbounded support allow every outcome in the population to depend on a small number of treatment assignments.

In practice, analyses of large economic experiments have reported not IPW estimators but rather linear regressions on measures of neighborhood treatment intensity. One might expect estimators based on this approach to perform well if potential outcomes are themselves linear functions of treatment assignments. We show that, in this case, the optimal rate is unchanged and the coefficient from an OLS regression of a specific form does converge to the GATE at the optimal rate. This form—a regression of outcomes on the ratio of the number of treated neighboring clusters to the mean number of neighboring clusters—is similar, but not equivalent, to the common approach of regressing outcomes on the share of treated neighbors. The latter specification concentrates around a weighted average of the spillover effects, rather than the unweighted GATE.[2] While there is no theoretical limit on how different these estimands can be, in our empirical application the two estimates differ by 0-20%.

The fact that linear regressions converge no faster than IPW even when the DGP is itself linear may seem somewhat surprising. Our next result shows that this property hinges on the researcher's ability to choose any experimental design. If instead the researcher is constrained to run an experiment with "small" clusters (i.e. clusters that grow slowly) then OLS can converge faster than IPW. Intuitively, linearity lets us learn from the observed outcome of a unit with *mostly* treated neighbors about its potential outcome were *all* of its neighbors treated. Such extrapolation is useful when fully saturated neighborhoods are rare due to small cluster sizes. This case is relevant to situations in which researchers must use pre-existing administrative units as their clusters (as did each of the field experiments cited above), or choose to prioritize comparisons between treated and untreated *individuals* (which require small clusters for precision) but still wish to learn something about the GATE.

---

[2]The weights are non-negative and convex, and in this sense the problem is less dire than that recently identified in two-way fixed effect designs (de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021). However, since the weights depend on both the treated unit and the affected unit, they have no welfarist interpretation.

The true data-generating process may of course be non-linear; many economic models imply interactions between peer effects that violate linearity. The OLS-based estimator described above remains consistent, however, under weak conditions on the experimental design. The main drawback is that without linearity OLS converges at a slower rate when the true DGP is highly nonlinear. Overall, the results suggest a simple tradeoff for practitioners: OLS-based approaches will be attractive when the underlying DGP is not too non-linear, and when cluster sizes must be small. They also admit statistical inference: we confirm the OLS estimator's asymptotic normality as the bandwidth and clusters grow, propose a conservative estimator for the variance, and find good coverage in simulations.[3]

There remains the practical problem of choosing the "bandwidth" of an OLS estimator in a given, finite population. Should the researcher regress outcomes on the mean treatment status of clusters within 1km, 10km, or some other distance? Choosing the bandwidth that minimizes mean square error (MSE) is infeasible, since bias is always unknown. We propose a minimax procedure that is asymptotically guaranteed to select the bandwidth that minimizes the worst-case MSE over a set of potential outcomes that is guaranteed to include the true potential outcomes. The few tuning parameters have natural default values, so that this approach does not require the use of outcome data or the exercise of researcher degrees of freedom.

We illustrate these methods in Monte Carlo simulations and using data from the large-scale cash transfer experiment in rural Kenya studied by Egger et al. (2022) (henceforth, EHMNW). That experiment randomly assigned half of 653 study villages to a treatment in which eligible households were offered an unconditional transfer of \$1,871 PPP. Many outcomes were affected; we focus for simplicity on estimating the GATE on annualized household consumption a year later. We estimate a GATE of \$445 PPP and reject the null of no effect at the 1% level. If instead we construct a GATE estimator from the regression coefficients in EHMNW, we obtain an estimate of \$371 and cannot reject the null of no effect. This difference is largely attributable to the fact that our bandwidth selection rule recommends a 400m bandwidth, substantially smaller than the pre-specified 2km one used by EHMNW, which in turn yields less volatile estimates.

The primary theoretical contribution of the paper is to show that, broadly speaking, estimating the GATE is more realistic than had previously been understood. Theorists have cast the GATE as an important estimand because it is interpretable even when the researcher is uncertain about the exact channels over which spillovers propagate. For example, in response to Sävje's (2024) characterization of estimands available when exposure mappings are somewhat misspecified, Auerbach et al. (2024a) point out that these estimands need not have causal interpretations, to which Leung (2024) responds that the GATE is one of the few that does, but explains that it is difficult to estimate in practice. Our results show that the GATE can in fact be rate-optimally estimated under weaker conditions than those in Leung (2022b), that OLS-based approaches can be utilized and may achieve faster rates than alternatives, and consequently that the analysis of several recent large-scale field experiments in economics has

---

[3]Another potential virtue of the OLS-based approach is that it is amenable to the introduction of prior information the researcher may have about the pattern of economic interactions, which can be incorporated in a TSLS setup. We omit this approach here for brevity but study it in an earlier working paper version (Faridani and Niehaus, 2024).

already come quite close to consistent estimation of the GATE.

This approach complements recent theoretical work that has focused on consistency or on optimal experimental design, leaving open the question of how a researcher should jointly choose a design and estimator. Auerbach and Tabord-Meehan (2023) provide consistency results for a class of estimands including the GATE under very general assumptions, for example, but do not address how choices of design and estimator map into the rate of convergence and what rates are possible. Leung (2023) and Viviano et al. (2024) find optimal experimental designs to estimate the GATE assuming an IPW or difference-in-means estimator will be used. We emphasize the joint choice of estimator and design to achieve rate-optimality, the ways in which this is affected by the linearity assumptions on potential outcomes that are implicit in applied work, and the design constraints that researchers often face.

An adjacent line of work has focused on estimands other than the GATE, such as the average causal effect of receiving some specific level of neighborhood exposure to treatment (Aronow and Samii, 2017; Leung, 2022a; Auerbach et al., 2024b). The estimands studied in these papers can be consistently estimated and interpreted even when the exposure mapping is mildly misspecified (Zhang, 2023; Sävje, 2024), but do not include the GATE, since consistent estimation is achieved using designs in which the probability of universal exposure does not vanish as the population grows. More generally, Basse and Airoldi (2018) show that there is no consistent estimator of the GATE under the assumptions made in these papers.[4]

Finally, we provide a theoretical foundation for and reinterpretation of the applied literature on program evaluation in the presence of spillovers. At least since Miguel and Kremer (2004) it has been understood that capturing spillover effects is potentially crucial to getting policy inferences right, and that the decay of spillover effects with distance can be exploited to construct estimators that capture them (e.g. Baird et al., 2016; Egger et al., 2022; Muralidharan et al., 2023). But it has been unknown whether these estimators are consistent for any estimand of interest; if so, under what conditions; and whether they are efficient in any sense. We address each of these questions, providing conditions under which the familiar approach of regressing on an appropriately weighted average of neighbors' treatment statuses yields an estimation procedure that converges at the optimal rate to the GATE. In doing so our broader aim is to provide a bridge between the experimental approach to economics—which, up until now, has rested on the assumption that a "pure control group" exists—and economic theory, which typically implies that one does not.

2   Setup

Notational conventions are as follows. We denote vectors in bold Latin letters $\mathbf{Y}$, fixed deterministic scalars as lowercase Latin or Greek letters n, p, and matrices as capital letters A. A subscripted vector b_i denotes the ith element while a subscripted matrix A_i denotes the ith row. A letter with round brackets denotes a scalar-valued function $\theta(\cdot)$. A script letter $\mathcal{N}$ denotes

---

[4]Work in this genre grew out of an earlier literature on linear models of peer effects in which outcomes are linear in exogenous and endogenous variables and either the coefficients are known up to scale (Manski, 1993; Lee, 2007; Goldsmith-Pinkham and Imbens, 2013; Bramoullé et al., 2014) or exposure mappings rule out interference from faraway units (Jagadeesan et al., 2020; Sussman and Airoldi, 2017). Our results under linearity show how one can consistently and rate-optimally estimate global average treatment effects even without such assumptions.

a set. Fixed universal constants that do not change across populations are denoted with the letter K subscripted by the number of the assumption that introduced them, e.g. Assumption 3 introduces K_3. (A K subscripted by a letter, on the other hand, is not linked to an assumption.) Where there is little risk of confusion, we suppress subscripts noting dependence on n.

A researcher observes a finite population $\mathcal{N}_n = \{1, 2, \cdots, n\}$ of n units indexed by i. The researcher assigns an $n \times 1$ random vector of treatments $\{d_i\}_{i=1}^n \in \{0, 1\}^n$ and observes realized unit outcomes $\{Y_i\}_{i=1}^n \in \mathbb{R}^n$. The potential outcome for unit i given the entire treatment assignment is denoted $Y_i(\mathbf{d})$ where the argument is the full $n \times 1$ vector of realized treatment assignments $\mathbf{d}$. Since our estimand is the causal effect of treating an entire population and not a random sample, we model potential outcomes as deterministic functions and the treatment assignment $\mathbf{d}$ as random (and correspondingly will study asymptotics with respect to a sequence of growing finite populations).[5] $F(\mathbf{d})$ denotes the CDF of $\mathbf{d}$. We remain largely agnostic about the origins of the potential outcomes $Y_i(\cdot)$, which could be the outcome of nearly any spatial process, and in particular we impose no assumptions on the support of the spillover effects.

Our estimand of interest is the average causal effect of assigning *every* unit in the entire population to treatment vs assigning *no* units at all to treatment. This is called the *global average treatment effect* (GATE) because it sums both the own-effects and spillover effects of treating all units (within some relevant class).

Definition 1. Global Average Treatment Effect (GATE)

$$\theta_n \equiv \frac{1}{n} \sum_{i=1}^n Y_i(\mathbf{1}) - Y_i(\mathbf{0})$$

The GATE is often the estimand most relevant for policy decisions; a policymaker using RCT data to decide whether to scale up a program to the control group, for example, would base this decision on the GATE rather than the direct treatment-on-treated effect. Note that we assume the researcher observes outcomes for the entire population for whom she wishes to estimate the GATE.[6] Results concerning the limiting properties of our estimators will require a notion of a growing population, and our estimand θ_n can change as the population grows. Later assumptions will characterize the conditions this sequence of populations must follow.

To solve the core identification challenge, we consider restrictions on the magnitude of spillovers, as opposed to less economically plausible restrictions on their support. Specifically, we consider a restriction on the (maximum) impact that faraway treatments have on each unit. We will ultimately express this restriction and other assumptions solely in terms of neighborhoods around each unit, but for the sake of intuition it may be helpful to think of these neighborhoods as being induced by a function $\rho_n : \mathcal{N}_n \times \mathcal{N}_n \to \mathbb{R}$ that measures the "distance" between any two units. Henceforth we will frequently suppress the n subscript of ρ. We require $\rho(i, i) = 0$ and $\rho(i, j) > 0 \,\forall i \neq j$, but place no further restrictions on ρ_n: we do not require symmetry ($\rho(i, j) = \rho(j, i)$) or that the triangle inequality hold exactly even after a monotonic transformation, for example. In mathematical terms, ρ_n is a "premetric" but need not be a

---

[5]Fixing potential outcomes is standard in this literature, see for example Aronow and Samii (2017); Sussman and Airoldi (2017); Leung (2022a,b).

[6]Appendix F extends to settings where only a subset of the entire population is eligible for treatment.

metric or semimetric. As we discuss below, this will allow us to capture the structure of a broader class of economic models.

The distance function induces a natural notion of neighborhoods as s-balls:

$$\mathcal{N}_n(i,s) \equiv \{j \in \mathcal{N}_n : \rho(i,j) \leq s\} \tag{1}$$

Here $s > 0$ parameterizes the radius of the neighborhood. This will allow us to study estimators that make use of increasingly large neighborhoods as n grows. Note that every unit is a member of its own neighborhoods ($i \in \mathcal{N}_n(i,s)$) and that for any given n, i there is an $s > 0$ such that $\mathcal{N}_n(i,s) = \{i\}$. Our results will hold under *any* definition of neighborhoods that satisfies these two properties, whether they were induced by a distance function or not.

Our first assumption captures the idea of spatial decay in the spillover effects. Heuristically, it says that when s is large and when all of the units in $\mathcal{N}_n(i,s)$ are (non)treated, then unit i's outcome is close to its potential outcome under *universal* (non)treatment. Formally, call a neighborhood "saturated" when all of its members are treated and "dissaturated" when none of its members are treated. Define the random variable $\tilde{s}_i$ as the largest s for which i's s-neighborhood is fully saturated or fully dissaturated.

Definition 2. Largest (Dis)Saturated Neighborhood

$$\tilde{s}_i \equiv \max\{s > 0\} \ s.t. \ d_j = 1 \, \forall \, j \in \mathcal{N}_n(i,s) \ or \ d_j = 0 \, \forall \, j \in \mathcal{N}_n(i,s)$$

Assumption 1 states that the difference between unit i's realized outcome and its potential outcome under universal (non)treatment decays with a power of $\tilde{s}_i$.

Assumption 1. Decay of Interference

There exist universal constants $\eta, K_1 > 0$ such that:

$$\left|Y_i(\mathbf{d}) - d_i Y_i(\mathbf{1}) - (1 - d_i)Y_i(\mathbf{0})\right| \leq K_1 \min\left\{1, \tilde{s}_i^{-\eta}\right\}, \qquad \forall \, n \in \mathbb{N}, \, i \in \mathcal{N}_n, \, \mathbf{d} \in \{0,1\}^n$$

The assumption that spillovers decay in a power of distance is consistent in a broad sense with applied general equilibrium modeling in the "gravity" and "market access" traditions, and mild relative to alternatives such as exponential decay.[7] Power-of-distance decay will also simplify the statement of results and (we will see) help to make salient the relationship between the rate at which spillovers decay spatially and the rate at which estimators of the GATE converge asymptotically. We do not require any particular power η, however.[8]

Remark 1. Assumption 1 is closely related to Assumption 3 of Leung (2022b). It is weaker in two ways. First, we bound only interference outside of (dis)saturated neighborhoods while Leung bounds distant interference even when local neighborhoods are not (dis)saturated. Second, we do not require that $\eta > 1$ (in Leung's notation, $\gamma > 2$). We are able to relax this rate bound by showing that (while the pairwise spillover effects between any two units must indeed decay

---

[7]Faster rates of convergence might of course be achievable under even faster rates of spatial decay (e.g., exponential).

[8]Assumption 1 implies Near Epoch Dependence (NED). In the language of Definition 1 of Jenish and Prucha (2012), we could say that Assumption 1 implies that "Y_i is uniformly L_1-NED with respect to d_i with size η."

faster than the square of s) the sum of all outside-neighborhood spillover effects can be allowed to decay at any geometric rate.

Remark 2. An alternative way of embodying the idea that spillovers decay, which might at first seem more conservative, is to assume that the (absolute value of the) effect on unit i of changing a *single* treatment assignment outside unit i's s-neighborhood is less than $s^{-\gamma}$ for some $\gamma > 0$. This assumption turns out to be *stronger* than Assumption 1, however, provided the density of the population is well-behaved. Formally, suppose that for any $s > 0$ and two treatment vectors $\mathbf{d}, \mathbf{d}'$ that differ only at component j where $j \notin \mathcal{N}_n(i, s)$, we have $|Y_i(\mathbf{d}) - Y_i(\mathbf{d}')| < K_0 s^{-\gamma}$. If the maximum number of units inside any $\mathcal{N}_n(i, s) \setminus \mathcal{N}_n(i, s-1)$ is upper bounded by a constant times $s^{\omega-1} + 1$ and $\gamma > \omega > 0$, then Assumption 1 holds with $\eta = \gamma - \omega$. Proof: Section A.1. The spatial moving average model in Equation (3) of Leung (2022b) satisfies these conditions with his γ equal to our γ and $\omega = 2$. He requires $\gamma > 2$ which corresponds to our requirement in this remark that $\gamma > \omega$. In summary, Assumption 1 bounds spillovers in a significantly more general way by bounding only the total influence of all spillovers outside a given distance, rather than restricting each spillover effect individually.

Assumption 1 ensures that spillover effects decay with space. Next we rule out the pathological possibility that the space between units itself increases too rapidly as the population grows, in which case spillovers would become asymptotically irrelevant to the GATE, which would converge rapidly to the ATE. To focus attention on cases where spillovers are a meaningful challenge, we rule out such unbounded sparsity. We use the concept of an *s-covering* of a set of units $Q \subseteq \mathcal{N}_n$: this is a subset $Q' \subseteq Q$ such that for each $i \in Q$, there is a unit $q \in Q'$ such that $i \in \mathcal{N}_n(q, s)$. The *s-covering number* of Q is then the number of units in the smallest possible s-covering of Q. This conception of the s-covering number always exists. Using it we can state Assumption 2.

Assumption 2. Covering Number

There exists a universal constant $K_2 > 0$ such that for all $s > 0$ and all $n \in \mathbb{N}$, the s-covering number of $\mathcal{N}_n$ is no greater than $K_2 \frac{n}{s}$.

Assumption 2 says that it takes no more than $K_2 \frac{n}{s}$ neighborhoods to cover the population. Example 1 below illustrates that it holds in familiar Euclidean applications.

The experimenter uses information about units' positions in space to design the experiment. Define an experimental design as a mapping $\mathcal{D} : \mathcal{N}_n \times \rho_n(\cdot, \cdot) \to F(\mathbf{d})$ from any pair of a population and distance function to a distribution function for the vector of treatments $\mathbf{d}$. We will impose restrictions on the design later on, but for now let $\mathcal{D}$ be unrestricted.

The last assumption we maintain throughout is a regularity condition that uniformly bounds potential outcomes. While less restrictive assumptions could be sufficient, Assumption 3 eases exposition and is standard (e.g. Aronow and Samii, 2017; Manta et al., 2022; Leung, 2022b).

Assumption 3. Bounded Outcomes

There is a constant $\overline{Y} > 0$ such that: $|Y_i(\mathbf{d})| < \overline{Y} < \infty \quad \forall n \in \mathbb{N}, i \in \mathcal{N}_n, \mathbf{d} \in \{0, 1\}^n$

3 Rate-Optimal Linear Estimation

We study estimators that are linear, i.e. that take the form of weighted averages of observed outcomes Y_i where the weights $\omega_{in}(\mathbf{d})$ are arbitrary functions of the full vector of treatment assignments $\mathbf{d}$. Weights can be specific to units i and may change with n, though we will sometimes suppress the latter subscript for brevity. Such estimators can thus be written as:

$$(2) \qquad \widehat{\theta}_n \equiv \sum_{i=1}^{n} Y_i \omega_{in}(\mathbf{d})$$

All of the estimators studied in this paper take this form, including the IPW estimators commonly used in theoretical analysis (Aronow and Samii, 2017; Leung, 2022b) as well as (after conditioning on covariates) the regression-based estimators common in applied work.

Theorem 1. Optimal Rate

Fix any sequence of populations $\mathcal{N}_n$ equipped with neighborhoods that satisfy Assumption 2. Let $\widehat{\theta}_n$ be any estimator of the form of Equation (2). Then for all $\delta > 0$ and any sequence $b_n \to \infty$,

$$\limsup_{n\to\infty} \sup_{\mathbf{Y}\in\mathcal{Y}_n} \mathbb{P}\left(b_n n^{\frac{1}{2+1/\eta}} \left|\widehat{\theta}_n - \theta_n\right| > \delta \right) > 0$$

where $\mathcal{Y}_n$ is the set of all potential outcomes $\mathbf{Y}$ that satisfy Assumptions 1 and 3.

Proof: Section A.2

Theorem 1 states that no estimator taking the form of a weighted mean of outcomes can be guaranteed to converge to θ_n at a rate strictly faster than $n^{-\frac{1}{2+1/\eta}}$ for all potential outcomes satisfying Assumptions 1 and 3. The intuition is that spillover effects with unbounded support allow the entire set of outcomes to potentially depend on the treatment statuses of only a vanishing fraction of the population as it grows. The slower spillover effects decay, the faster this fraction can vanish as the population grows while still determining all of the outcomes. This result is not driven by any pathological configuration of units in space (as it applies to *any* set of neighborhoods satisfying Assumption 2); by a pathological series of potential outcomes (those used in the proof are simply indicators of the treatment status of one nearby neighbor each); or by restrictions on the available experimental designs (as we have imposed none).

The rates allowed by Theorem 1 are slower than the parametric rate $n^{-\frac{1}{2}}$ because $\eta \in (0, \infty)$. The result thus illustrates a sense in which GATE estimation in the presence of spillovers with global support is a more difficult problem than estimation with localized spillovers, where the parametric rate can be guaranteed (Aronow and Samii, 2017). Theorem 1 also links the spatial rate of decay of the spillover effects with the rate of convergence of linear estimators. The economic principle this suggests is that estimators based on Assumption 1 will tend to converge faster in less integrated economies, where distance is an important constraint on economic interactions, than in highly integrated ones where all units are "close" to each other.

The result extends Theorem 2 of Leung (2022b) in three ways. First, it relaxes the requirement that spillovers decay faster than the geometric rate $\eta \geq 1$ and instead accommodates any $\eta > 0$. Second, it generalizes from the case in which the population is contained in a circle in $\mathbb{R}^2$ to the larger set of spaces defined by Assumption 2, which nests $\mathbb{R}^q$ among others. Third,

it finds the optimal rate jointly over all linear estimators and all experimental designs. This is relevant given that applied work has used linear estimators other than IPW.

3.1 Achieving the optimal rate

We next show that the optimal rate is achievable provided that neighborhoods and population structure are sufficiently "well-behaved." This requires selecting a suitable experimental design; we will in fact define a class of "Scaling Clusters" designs all of which will permit rate-optimal estimation and which encompasses many standard designs used in real-world experiments.

First consider the larger set of *cluster-randomized* designs:

Assumption 4. Cluster-Randomized Design

If $\mathbf{d} \sim \mathcal{D}(\mathcal{N}_n, \rho_n(\cdot, \cdot))$, *then* $d_i \sim Bern(p)$ *for each* $i \in \mathcal{N}_n$. *In addition, there exists a partition of* $\mathcal{N}_n$ *consisting of* $c_n \in \mathbb{N}$ *clusters called* $\{\mathcal{P}_c\}_{c=1}^{c_n}$ *such that:*

$$d_i = d_j \text{ if } i, j \in \mathcal{P}_c \text{ for some } c$$
$$d_i, d_j \text{ independent otherwise}$$

Let $c(i)$ *denote the cluster that contains unit* i.

Assumption 4 allows for many designs used in practice, including those in recent large-scale experimental work. It technically includes the familiar Bernoulli design, in which each unit is its own cluster, but this design is unlikely to perform well since (by the law of large numbers) it tends to generate few large neighborhoods in which most units are treated (or untreated), and which thus approximate the counterfactual scenarios of interest.[9] Assumption 4 rules out stratified designs for ease of exposition; under stratification many estimators require additional weights, which are provided in the extension in Appendix E.

To consistently estimate the GATE we will need treatment clusters to be of a shape and size such that the number of clusters within any s-neighborhood does not grow too rapidly with s. To achieve this, consider a class of *Scaling Clusters* designs where the treatment clusters are shaped like neighborhoods and grow with n. The rate of cluster growth is governed by an increasing sequence of positive numbers $\{g_n\}_{n=1}^{\infty} \subseteq \mathbb{R}^+$ as follows:

Definition 3. Scaling Clusters Designs

Consider any increasing sequence $\{g_n\}_{n=1}^{\infty} \subseteq \mathbb{R}^+$. *A cluster-randomized design is called a "Scaling Clusters" design when there is a* $K_D > 1$ *such that for every cluster* $\mathcal{P}_c$ *there is a corresponding unit* $q_c \in \mathcal{N}_n$ *such that (i)* $q_c \in \mathcal{P}_c \subseteq \mathcal{N}_n(q_c, g_n K_D)$ *and (ii) the set of* $\{q_c\}$ *form a* g_n*-packing of the population.*

---

[9]Notice the absence of an overlap (or "positivity") assumption bounding away from zero the probability that any given unit experiences full exposure to treatment or to control conditions in some relevant neighborhood. Such restrictions on the design are often used when the estimand compares the effects of "local" exposures. In these settings it (helpfully) ensures that the probability that the researcher observes both units whose relevant neighborhoods are fully treated and units whose relevant neighborhoods are fully untreated goes to one, and that the researcher can construct consistent estimators by comparing these unit. In our setting, however, the relevant neighborhood is the entire population, so we can *never* observe both such units at once: if one unit's is exposed to universal treatment, then there cannot be a comparison unit that was exposed to universal non-treatment. As a result a positivity assumption would hurt rather than help us.

A Scaling Clusters design is therefore any cluster-randomized setup where treatment clusters both contain and are contained by neighborhoods that grow in proportion to g_n.[10] In order for such a design to control the number of clusters per neighborhood, we require that the neighborhoods share two primitive properties with typical Euclidean spatial settings which the following two assumptions ensure. Assumption 5 requires that a loosened version of the triangle inequality holds. It states that if there is a neighborhood that contains both i and j, then the distance between them is bounded by an amount that does not depend on i, j, n, or the absolute size of the neighborhoods.

Assumption 5. Weakened Triangle Inequality

There is a universal constant $K_5 \geq 1$, such that

$$i, j \in \mathcal{N}_n(k, s) \implies i \in \mathcal{N}_n(j, K_5 s), \qquad \forall\, i, j, k \in \mathcal{N}_n, \, \forall\, n \in \mathbb{N}, \, \forall s > 0$$

Assumption 5 ensures that a Scaling Clusters design exists; Algorithm 1 in Appendix C demonstrates how to find one. Slight weakenings of this assumption (as well as Assumption 6 below) are possible, but as the weaker versions become quite technical, we state here versions which are relatively easy to check in application. Assumption 5 is satisfied, for example, in any metric space (due to symmetry and the triangle inequality) with $K_5 = 2$. Allowing for a wider range of K_5 lets us accommodate settings where distances are not metrics, while maintaining a link between coverings and packings. On the other hand, Assumption 5 rules out totally random neighborhoods: it will be violated as the population grows if each distance $\rho(i, j)$ is an i.i.d. random number, for example, or if the population is connected in a "small world" fashion (Watts and Strogatz, 1998).

The final assumption we use to achieve the optimal rate controls population dispersion as n grows, ensuring that the membership of neighborhoods scales (approximately) linearly in s.

Assumption 6. Bounded Density

There exist universal constants $\overline{K}_6 \geq \underline{K}_6 > 0$ such that:

$$\underline{K}_6 s + 1 \leq |\mathcal{N}_n(i, s)| \leq \overline{K}_6 s + 1 \qquad \forall\, s > 0, \ n \in \mathbb{N}, \ i, j \in \mathcal{N}_n$$

One can of course always pick constants such that this holds given a fixed n; what gives it meaning is the restrictions it imposes as n grows. Intuitively, the upper bound says that increasing the population means "growing the map" and not solely adding units to a map of a fixed size, in which case it would not be possible to have both saturated and dissaturated neighborhoods of increasing size. The lower bound controls the number of clusters that a Scaling Clusters design can produce by ruling out large numbers of "isolated" units.

The last step is to define estimators. We consider first the Horvitz-Thompson (HT) and Hajek (HJ) inverse probability weighting (IPW) estimators; while not common in practice, these are conveniently tractable and standard in theoretical work on network effects (see for example Aronow and Samii, 2017; Sussman and Airoldi, 2017; Leung, 2022a,b, 2023). They will

---

[10]Condition (ii) of Definition 3 refers to the standard concept of an *s-packing*, i.e. a subset of the population such that for all i, j in the packing, $\rho(i, j), \rho(j, i) \geq s$, and is straightforward to verify in practice.

also serve below as a benchmark for the performance of OLS estimators. To define them, fix a sequence of neighborhood radii $\kappa_n \subset \mathbb{R}^+$, which will grow to infinity with the sample size. The researcher can choose this sequence, and we will study later how to optimize this choice. Define T_{i1}, T_{i0} as indicators that all of i's neighbors within distance κ_n are treated or untreated:

$$T_{i1} \equiv \prod_{j \in \mathcal{N}_n(i,\kappa_n)} d_j, \qquad T_{i0} \equiv \prod_{j \in \mathcal{N}_n(i,\kappa_n)} (1 - d_j)$$

The HT estimator is the weighted difference in outcomes between units with fully treated v.s. fully untreated neighborhoods.

$$(3) \qquad \widehat{\theta}_{n,\kappa_n}^{HT} \equiv \frac{1}{n} \sum_{i=1}^n Y_i \left(\frac{T_{i1}}{\mathbb{P}(T_{i1}=1)} - \frac{T_{i0}}{\mathbb{P}(T_{i0}=1)} \right)$$

HT may perform poorly because the probabilities in the denominators can be small, and is sensitive to adding a constant to all outcomes. The HJ estimator makes an adjustment to (partially) address such issues, ensuring that the sum of weights given to units with fully saturated neighborhoods equals the sum of weights given to units with fully dissaturated neighborhoods.

$$(4) \qquad \widehat{\theta}_{n,\kappa_n}^{HJ} \equiv \sum_{i=1}^n Y_i \left(\frac{T_{i1}}{\mathbb{P}(T_{i1}=1) \sum_{k=1}^n \frac{T_{k1}}{\mathbb{P}(T_{k1}=1)}} - \frac{T_{i0}}{\mathbb{P}(T_{i0}=1) \sum_{k=1}^n \frac{T_{k0}}{\mathbb{P}(T_{k0}=1)}} \right)$$

For these estimators to converge, it is necessary to lower bound the probabilities in the denominators of the weights. This is equivalent to upper bounding the number of clusters $\phi_n(i,\kappa_n)$ that each neighborhood $\mathcal{N}_n(i,\kappa_n)$ intersects. Lemma 1 in Appendix B shows that Assumptions 5 and 6 combined with a Scaling Clusters Design make $\phi_n(i,\kappa_n)$ scale with κ_n/g_n. Theorem 2 then shows that either estimator paired with a Scaling Clusters Design achieves the optimal rate from Theorem 1.

Theorem 2. Inverse Probability Weighting

If Assumptions 1–6 hold and the researcher uses a Scaling Clusters Design with $g_n \propto \kappa_n$, then

$$\mathbb{E}\left[\left|\widehat{\theta}_{n,\kappa_n}^{HT} - \theta_n\right|\right] = \mathcal{O}\left(\sqrt{\frac{\kappa_n}{n}} + \kappa_n^{-\eta}\right), \qquad \mathbb{E}\left[\left|\widehat{\theta}_{n,\kappa_n}^{HJ} - \theta_n\right|\right] = \mathcal{O}\left(\sqrt{\frac{\kappa_n}{n}} + \kappa_n^{-\eta}\right)$$

When $\kappa_n \propto n^{\frac{1}{2\eta+1}}$, the optimal rate is achieved:

$$\sqrt{\frac{\kappa_n}{n}} + \kappa_n^{-\eta} = \mathcal{O}\left(n^{-\frac{1}{2+\frac{1}{\eta}}}\right)$$

Proof: Section A.3.

In contrast to the HJ estimator for local effects studied in Aronow and Samii (2017), the HJ estimator of the GATE here converges at a subparametric rate. Extending results for the HT estimator in Leung (2022b), our result applies beyond $\mathbb{R}^2$, allows for $\eta \in (0,1)$, and (in conjunction with Theorem 1) shows that the researcher cannot improve on the IPW + Scaling Clusters rate with another *pair* of linear estimator and cluster-randomized design.

3.2 Discussion: notions of space

With the main spatial assumptions now stated, we pause to take stock of the applications they encompass. Example 1 first verifies they are consistent with standard Euclidean space:

Example 1. Spatial Designs Suppose that units are all located in m-dimensional Euclidean space where the $m \times 1$ coordinate vector of unit i is $\mathbf{x}_i$. Let $\rho(i,j) = ||\mathbf{x}_i - \mathbf{x}_j||_2^m$. Assumption 5 is met with $K_5 = 2^m$ because the Euclidean distance satisfies the triangle inequality. If no two units can be closer than ϵ together and the entire population is contained within a 2-ball of radius $n^{1/m}$, Assumption 6 is also met. Finally, Assumption 2 is met with $K_2 = 16^m$.[11]

In other applications researchers may wish to explicitly model an economically relevant notion of "distance" that is not a proper metric. It may be asymmetric; for example, downwind locations may experience more pollution from upwind ones than vice versa, as in Rangel and Vogl (2019). Or it may not exactly satisfy the triangle inequality, even up to monotonic transformations. In market access models, for example, two small towns may be heavily affected by the economy of a large nearby city, but have little influence on each other, so that—from the point of view of spillover effects—the "distance" between them is greater than the sum of the distances between them and a common neighbor (Donaldson and Hornbeck, 2016).[12] Applications like these *may* still fit within a purely Euclidean model of space, but checking this requires working out whether some bound η on the rate of spillover decay with respect to Euclidean distance is defensible given that said distance is, in a sense, misspecified. Using more flexible notions of distance may be simpler. The following example illustrates this, showing that if we modify the distance from i to j in Example 1 by some bounded positive factor $h(i,j)$ then our spatial assumptions continue to hold:

Example 2. Spatial Designs (continued) Let $\rho(i,j) = h(i,j)||i - j||_2^m$ for some function $h(\cdot,\cdot)$ satisfying $0 < \underline{h} \le h(i,j) \le \overline{h} < \infty$ for constants $\overline{h} > \underline{h} > 0$, and let the other conditions in Example 1 hold. Then Assumption 2 is met with $K_2 = \underline{h}^{-1}16^m$, and Assumptions 5–6 are similarly met as in Example 1.

Another scenario in which flexibility is useful is where the researcher is uncertain about which of several potential channels for spillovers matter. For example, training small business owners might lead to knowledge spillovers for which the social network defines the relevant notion of distance, businesses stealing for which physical distance and product substitutability between firms define distance, input demand effects for which the buyer-supplier network defines distance, and so on.[13] The next example shows that in this scenario the researcher need not take a stand on any one notion of distance, but can instead work with the minimum of them, which preserves the essential properties they need:

---

[11]To see this final point, note that the packing number of a Euclidean ball of radius $n^{1/m}$ in Euclidean balls of radius $s^{1/m}$ is upper bounded by $\left(1 + 4\left(\frac{n}{s}\right)^{1/m}\right)^m$. The δ-covering number of a subset of a 2-ball is no greater than its $\delta/2$-packing number. So the covering number is upper bounded by $\left(1 + 8\left(\frac{n}{s}\right)^{1/m}\right)^m$. If we restrict $s \le 2n$, then $8\left(\frac{n}{s}\right)^{1/m} \ge 16$. So the covering number is upper bounded by $16^m\left(\frac{n}{s}\right)$ and so $K_2 = 16^m$. This constant can be sharpened. See p. 126 of Wainwright (2019).

[12]Appendix C of an earlier version of the paper provides a formal illustration (Faridani and Niehaus, 2024).

[13]We thank David McKenzie for suggesting this example. See Sävje (2024), Auerbach et al. (2024a), and Leung (2024) for further discussion of the desirability of robustness to misspecification.

Example 3. Suppose that we have two distance functions ρ_1, ρ_2 and a universal constant C such that $\rho_1(i,j) \leq C\rho_2(i,j)$ and vice versa. If ρ_1, ρ_2 satisfy Assumptions 2, 5, and 6, then $\rho_3(i,j) = \min\{\rho_1(i,j), \rho_2(i,j)\}$ does as well (possibly with different values for the constants $K_2, K_5,$ and K_6). Notice that ρ_3 does not necessarily follow the triangle inequality, even if ρ_1 and ρ_2 do. Proof: Section A.4.

4   Estimation under linearity

We now turn to defining and studying a class of estimators based on OLS regressions like those used in the applied literature. These estimators regress unit-level outcomes on a weighted average of the treatment statuses of nearby clusters. Formally, let $\mathbf{b}$ denote the $c_n \times 1$ vector of i.i.d. Bernoulli(p) cluster treatment assignments. Let $\mathbf{v}$ denote the centered $c_n \times 1$ vector $\mathbf{v} \equiv \mathbf{b} - \mathbb{E}[\mathbf{b}]$. Consider a sequence of $n \times c_n$ matrices B_n of real numbers that satisfy two properties: First, the absolute row sums of B must be uniformly bounded, i.e. $\sup_n ||B_n||_\infty < K$ for some constant K. This is to rule out pathological regressors which diverge to infinity. Second, we require that, for the sequence $\kappa_n \subseteq \mathbb{R}^+$ chosen by the researcher, the i, c element $B_{n,ic} = 0$ if cluster c is not intersected by $\mathcal{N}_n(i, \kappa_n)$. These two conditions mean that $B_n \mathbf{v}$ is some weighted average of the centered treatment statuses of clusters within distance κ_n.

Given such a sequence of B_n (and suppressing the n subscript from now on) we define the estimator obtained from a regression of the outcomes on $B\mathbf{v}$ as follows:

$$(5) \qquad \widehat{\theta}^{OLS}_{n,\kappa_n} \equiv \frac{\widehat{\text{Cov}}(B\mathbf{v}, \mathbf{Y})}{\widehat{\text{Cov}}(B\mathbf{v}, B\mathbf{v})} \equiv \frac{\frac{1}{n}\mathbf{v}'B'\mathbf{Y} - (\mathbf{1}'B\mathbf{v})(\mathbf{1}'\mathbf{Y})/n^2}{\frac{1}{n}\mathbf{v}'B'B\mathbf{v} - (\mathbf{1}'B\mathbf{v}/n)^2}$$

While it is defined in terms of weighted averages across *clusters*, this class includes estimators in which, as is common practice, outcomes are regressed on a weighted average of treatment status across *units*. Specifically, setting B to the "unit-weighted" matrix B^U defined by

$$(6) \qquad B^U_{ic} = \frac{\sum_{j=1}^n \mathbb{1}\{j \in \mathcal{N}_n(i, \kappa_n) \cap \mathcal{P}_c\}}{|\mathcal{N}_n(i, \kappa_n)|}$$

yields a regressor similar to those used in Miguel and Kremer (2004), Muralidharan et al. (2023), and Egger et al. (2022).[14] The neighborhoods over which treatment status is averaged are interpreted as growing with n (since κ_n grows), an interpretation not explicitly discussed in the original papers but which admits the possibility of consistency for the GATE.[15]

Remark 3. One can view the regression in (5) as one of a class of estimators that regress outcomes on average treatment status within one or more "rings" and then sum the coefficients. For instance, EHMNW regress outcomes on an indicator for own-village treatment and the

---

[14]These papers regressed outcomes on $B^U\mathbf{b}$ rather than $B^U\mathbf{v}$, but these are numerically equivalent because (a) the rows of B^U all sum to 1, and (b) the designs in those papers were such that all units had equal probability of being treated. This equivalence holds only for B^U, because its row sums are uniform.

[15]Technically one can read the specification in (5) as nesting some IPW estimators in the sense that if κ_n is kept small enough that each unit's neighborhood $\mathcal{N}_n(i, \kappa_n)$ intersects only one treatment cluster then it is possible for $\widehat{\theta}^{OLS}_{n,\kappa_n}$ to agree with an IPW estimator. If $\kappa_n/g_n \to \infty$, however, they cannot agree, and this will be the case of most interest to us as it is the one in which OLS will prove to have a potential rate advantage.

mean treatment status of neighbors within a ring, and then calculate a weighted sum of the two coefficients. We conjecture that there may be scope for such "multi-ring" approaches to improve finite-sample performance in some cases, but in our application to the EHMNW data we find that the point estimates are virtually unchanged whether we regress on one ring or two (see Appendix G). We therefore focus for simplicity on the one-ring case.

4.1 Linear DGPs

One might expect estimators of the form in (5) to perform well if potential outcomes are in fact linear functions of the regressors. Assumption 7 formalizes this idea, stipulating that the effect of treating unit j on unit i does not depend on any of the other treatment assignments.

Assumption 7. Linear Potential Outcomes

If $\mathbf{d}, \mathbf{d}' \in \{0,1\}^n$ are disjoint treatment assignment vectors, then:

$$Y_i(\mathbf{d} + \mathbf{d}') - Y_i(\mathbf{0}) = Y_i(\mathbf{d}) + Y_i(\mathbf{d}') - 2Y_i(\mathbf{0})$$

As an immediate consequence of the Riesz Representation Theorem, for every population $\mathcal{N}_n$, there exists an $n \times n$ matrix A of real numbers with rows denoted A_i and $\beta_0 > 0$ such that

$$(7) \qquad Y_i(\mathbf{d}) = \beta_0 + A_i\mathbf{d} + \epsilon_i, \qquad \sum_{i=1}^n \epsilon_i = 0$$

The residual term ϵ_i is non-random and sums to exactly zero in all populations by construction; since treatment is randomly assigned, $\mathbb{E}[\mathbf{d}_i] = p$ for all i regardless of ϵ_i. The convergence of the OLS estimator will not depend on how ϵ is related to the neighborhoods.

It will be convenient to do a change of variables and write the potential outcomes in terms of the centered cluster treatment assignments $\mathbf{v}$. Let P be the $n \times c_n$ matrix where P_{ic} indicates that unit i is a member of cluster c, so $P\mathbf{v} = \mathbf{d} - \mathbb{E}[\mathbf{d}]$. Equation 8 below expresses potential outcomes in terms of $\mathbf{v}$. Both Equations 7 and 8 are implied by Assumption 7. Here $(AP)_i$ denotes the ith row of the matrix product of A and P. As before, the residual term μ_i is nonrandom and sums to zero.

$$(8) \qquad Y_i(\mathbf{v}) = \alpha_0 + (AP)_i\mathbf{v} + \mu_i, \qquad \sum_{i=1}^n \mu_i = 0$$

In economic terms, linearity of the potential outcomes will often be a strong assumption.[16] Yet introducing Assumption 7 does not change the minimax rate of convergence in Theorem 1; it is still the case that no linear estimator can guarantee a rate of convergence to θ_n faster than $n^{-\frac{1}{2+\frac{1}{\eta}}}$, regardless of the experimental design.

Theorem 1′. Optimal Rate for Linear DGPs

Requiring that potential outcomes also satisfy Assumption 7 does not change the conclusions of Theorem 1. Proof: Section A.2

---

[16]In the case of households that are the monopolist providers of goods to their neighbors, for example, one can show that linearity of profits in transfers requires that the absolute markup be invariant to demand shocks.

The reason that the researcher cannot guarantee a faster rate of convergence with linearity than without it is that the sub-parametric rate at which Theorem 1 bounds the speed of convergence is governed by the possibility that, when spillovers have unbounded support, all of the outcomes depend on the treatment status of a vanishingly small fraction of the population. Uncertainty about functional forms is second order to this.

With Assumption 7 in hand, we can analyze the convergence of OLS for different choices of the $n \times c_n$ matrix B. Theorem 3 provides a convergence result that applies for any matrix B satisfying the two conditions on its row sums and support above.

Theorem 3. Convergence of OLS

Let Assumptions 1-7 hold and let the design be a Scaling Clusters design with cluster radius g_n. If $\frac{\max\{\kappa_n, g_n\}\sqrt{n}}{\sqrt{g_n} tr(B'B)} \to 0$, then

$$\mathbb{E}\left[\left|\frac{\widehat{\mathrm{Cov}}\left(\mathbf{Y}, B\mathbf{v}\right)}{\widehat{\mathrm{Cov}}\left(B\mathbf{v}, B\mathbf{v}\right)} - \frac{tr(B'AP)}{tr(B'B)}\right|\right] = \mathcal{O}\left(\frac{\max\{\kappa_n, g_n\}\sqrt{n}}{tr(B'B)\sqrt{g_n}}\right)$$

Proof: Section A.5.

Theorem 3 provides two insights. First, the rate of convergence of OLS depends principally on the column sums of B, which measure how influential each individual's treatment status is on the rest of the population. Second, OLS concentrates around the ratio of traces $\frac{tr(B'AP)}{tr(B'B)}$. This lets us see that the workhorse regressor in the applied literature (i.e. that using B^U) yields a weighted average of treatment effects with weights that are positive but sensitive to the exact specification of the clusters. We call this estimand θ_n^U. Specifically, some arithmetic yields

$$(9) \qquad \frac{tr(B^{U'}AP)}{tr(B^{U'}B^U)} = \sum_{i=1}^{n}\sum_{k=1}^{n}\underbrace{A_{ik}}_{\text{Effects}}\underbrace{\left(\frac{\frac{\sum_{j=1}^{n} \mathbb{1}\left\{j \in \mathcal{N}_n(i, \kappa_n) \cap \mathcal{P}_{c(k)}\right\}}{|\mathcal{N}_n(i, \kappa_n)|}}{\sum_{m=1}^{n}\sum_{c=1}^{c_n}\left(\frac{\sum_{j=1}^{n} \mathbb{1}\left\{j \in \mathcal{N}_n(m, \kappa_n) \cap \mathcal{P}_c\right\}}{|\mathcal{N}_n(m, \kappa_n)|}\right)^2}\right)}_{\text{Weights}} \equiv \theta_n^U$$

Here the causal effect on unit i of treating unit k receives a weight that depends on both i and k. This means that it is not generically feasible to obtain a rate-efficient estimator of the unweighted GATE by first re-weighting the observations and then running a regression using B^U. An exception is when randomization is i.i.d. *and* neighborhoods are of uniform size, in which case the weights are uniform. In the empirical application in Section 8, the unwanted weights in Equation 9 will shift the point estimate by 0-20%—but there is no theoretical limit on how large the difference can be.

A consistent OLS estimate of the GATE can be obtained by regressing on the (normalized) number of treated *clusters* within the neighborhood. Let

$$(10) \qquad B_{ic}^* = \overline{\phi}_n^{-1}\mathbb{1}\left\{\mathcal{N}_n(i, \kappa_n) \cap \mathcal{P}_c \neq \emptyset\right\}$$

where $\phi_n(i, \kappa_n) \equiv \sum_{c=1}^{c_n} \mathbb{1}\left\{\mathcal{P}_{n,c} \cap \mathcal{N}_n(i, s) \neq \emptyset\right\}$ denotes the number of treatment clusters within κ_n of unit i, and $\overline{\phi}_n \equiv \frac{1}{n}\sum_{i=1}^{n}\phi_n(i, \kappa_n)$ denotes its mean. Using B^* guarantees that the

difference between the ratio of traces and the GATE converges to zero at the appropriate rate.

$$(11) \qquad \frac{tr(B^{*\prime}AP)}{tr(B^{*\prime}B^*)} - \theta_n = \mathcal{O}\left(\kappa_n^{-\eta}\right)$$

Theorem 4 characterizes the estimand and rate of convergence obtained when using B^*, showing that it achieves rate-optimal consistency for the GATE.

Theorem 4. Consistency of OLS for the GATE

If Assumptions 1-7 hold and we use the Scaling Clusters design with scaling sequence g_n and $\max\left\{\frac{\kappa_n^2}{g_n}, \kappa_n\right\} \Big/ (\sqrt{g_n}\sqrt{n}) \to 0$, then:

$$\mathbb{E}\left[\left\|\frac{\widehat{\mathrm{Cov}}\left(B^*\mathbf{v}, \mathbf{Y}\right)}{\widehat{\mathrm{Cov}}\left(B^*\mathbf{v}, B^*\mathbf{v}\right)} - \theta_n\right\|\right] = \mathcal{O}\left(\frac{\max\left\{\frac{\kappa_n^2}{g_n}, \kappa_n\right\}}{\sqrt{g_n}\sqrt{n}} + \kappa_n^{-\eta}\right)$$

Proof: Section A.6.

The rate $\max\left\{\frac{\kappa_n^2}{g_n}, \kappa_n\right\} \Big/ (\sqrt{g_n}\sqrt{n}) + \kappa_n^{-\eta}$ is optimized by setting $\kappa_n \propto g_n \propto n^{\frac{1}{2\eta+1}}$ which yields the optimal $n^{\frac{-1}{2+\frac{1}{\eta}}}$ rate of convergence from Theorem 1, which (per Theorem 1') is also the best rate attainable with the addition of Assumption 7.

Remark 4. It may seem surprising at first that estimating a regression equation of $Y_i = \beta_0 + \beta_1 B_i^*\mathbf{v} + e_i$ yields a coefficient that converges to the GATE even though the model in Assumption 7 says that actually $Y_i = \beta_0 + (AP)_i\mathbf{v} + \epsilon_i$. The OLS estimator is consistent because our assumptions guarantee that $\frac{Cov\left((AP)_i\mathbf{v}, B_i^*\mathbf{v}\right)}{Var(B_i^*\mathbf{v})} = (AP)_i\mathbf{1} + \mathcal{O}\left(\kappa_n^{-\eta}\right)$. This is a consequence of the construction of B^* and does not apply to B^U.

The result that OLS estimators converge to the GATE no faster than alternatives even when the true DGP is linear is (or was, at least to us) somewhat surprising. It turns out to hinge, however, on the assumption that the researcher can select a design with any cluster size. In practice this is often not the case. Equity considerations or the logistics of treatment delivery may require that the experimental design be based on pre-existing administrative units. This was true of all of the large-scale experiments referenced above: schools in Miguel and Kremer (2004), villages in Egger et al. (2022), and sub-districts in Muralidharan et al. (2023). Researchers may also choose to randomize in small clusters or even at the individual level because they do not have sufficient budget for many large clusters and wish to prioritize precise comparisons between treated and untreated individuals, even though they realize these may be biased to some extent as estimates of the average individual treatment effect. In practice Muralidharan and Niehaus (2017) document that program evaluation RCTs have typically used small clusters, with a median cluster size of just 26 units.

Linearity of the potential outcomes creates opportunities in such settings. Intuitively, without linearity it is essential that clusters grow quickly, because the *only* way to learn about the potential outcomes of units surrounded by large (dis)saturated neighborhoods is to actually observe such units. Linearity of the potential outcomes implies, however, that we can also learn

about these potential outcomes from observing units surrounded by large neighborhoods that are only partially (dis)saturated. This allows the researcher to obtain consistent estimates using OLS even in a scenario where the researcher cannot grow cluster sizes as fast as she would prefer, so that estimators such as HT or HJ diverge. Our next theorem illustrates this point by imposing a constraint on the rate of growth of cluster sizes, measured by g_n.

Theorem 5. Small Clusters

If Assumptions 1-7 hold and we use a Scaling Clusters design but set $g_n = n^q$, where $q < \frac{1}{2\eta+1}$, then

- For any choice of κ_n that makes $\widehat{\theta}_{HT}, \widehat{\theta}_{HJ}$ consistent, the bias dominates the standard deviation and cannot be guaranteed to decay faster than $n^{-\eta q}$.

- If we set $\kappa_n \propto n^{\frac{1+3q}{4+2\eta}}$, then the bias and standard deviation of $\widehat{\theta}_{OLS}$ both converge with $n^{-\frac{1+3q}{4/\eta+2}}$, which is faster than $n^{-\eta q}$.

Proof: Section A.7.

Another potential advantage of the linear approach is that it allows the practitioner to adapt and use other standard econometric techniques. For example, one natural question that arises in our setting is how the researcher might exploit additional information they have about the structure of economic interactions between units, in addition to the distance pre-metric ρ. The researcher might collect information on shopping patterns, for example, or locations of resident and employment. If Assumption 7 holds then it turns out to be straightforward to incorporate this information to refine the estimate of the GATE using a TSLS approach. For brevity we omit the details here, providing them in a earlier version of the paper (Faridani and Niehaus, 2024, Section 4.1).

4.2 Robustness to Nonlinear DGPs

While linearity of the DGP (i.e., Assumption 7) provides a natural justification for OLS estimation, it is not necessary for its consistency. Mild conditions on the experimental design are in fact sufficient, without any conditions on the potential outcomes. Specifically, the OLS estimator defined by the matrix B^* above will be consistent if the growth of the treatment clusters is such that potential outcomes asymptotically start to resemble functions of only their own cluster's treatment assignment. This process is slow, and so the OLS estimator includes a nonlinearity bias term that converges with $n^{-\frac{\eta}{1+\eta}q}$, slower than IPW estimator's rate of $n^{-\eta q}$. If spillovers decay slowly (η is small), however, the resulting difference in rates will be small.

Assumption 8 controls the number of units that sit near the boundaries of their own treatment clusters. It says that the fraction of the population within distance t of another treatment cluster scales with $\left(\frac{t}{g_n}\right)^\tau$ for some constant $\tau > 0$.

Assumption 8. *There are universal constants $K_8, \tau > 0$ such that*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{\mathcal{N}_n(i,t) \setminus \mathcal{P}_c(i) \neq \emptyset\right\} < K_8 \left(\frac{t}{g_n}\right)^\tau$$

Assumption 8 simply requires that not too many units are located near cluster boundaries. In practice, checking Assumption 8 may simply mean that the applied researcher checks whether treatment clusters are predominantly convex. Example 4 illustrates conditions under which it holds in the canonical case of two-dimensional Euclidean space if treatment clusters are convex and the researcher uses a Scaling Clusters Design.

Example 4. Suppose that units are located within two-dimensional Euclidean space and that $\rho(i, j)$ is the square of the Euclidean distance. Assume that a strengthened version of bounded density holds where the number of units inside any convex shape of Euclidean area a but not inside a second convex shape contained within the first of area b is upper bounded by $K(a - b)$ for some $K > 0$. Finally, assume a scaling clusters design and that all treatment clusters take the form of convex shapes. Then Assumption 8 holds with $\tau = \frac{1}{2}$. Proof: Section A.8

Assumption 8 ensures that, as treatment clusters grow, the fraction of units located near their boundaries goes to zero. For the remaining units located far from cluster boundaries, meanwhile, the spatial decay of spillover effects ensures that potential outcomes $Y_i(\mathbf{d})$ become close to $d_i Y_i(\mathbf{1}) + (1 - d_i) Y_i(\mathbf{0})$, which is linear. As clusters grow, the DGP becomes as good as linear and the intuition from Remark 4 comes into effect.

Theorem 6. *If Assumptions 1-6 and 8 hold and the researcher uses a Scaling Clusters design with $g_n \propto n^q$ where $q \leq \frac{1}{2\eta+1}$, and the researcher sets $\kappa_n = \mathcal{O}\left(n^{\frac{1+3q}{4+2\eta}}\right)$, then:*

$$\mathbb{E}\left[\left\|\frac{\widehat{\mathrm{Cov}}\left(B^* \mathbf{v}, \mathbf{Y}\right)}{\widehat{\mathrm{Cov}}\left(B^* \mathbf{v}, B^* \mathbf{v}\right)} - \theta_n\right\|\right] = \mathcal{O}\left(n^{-\frac{\eta}{1+\eta/\tau}q}\right)$$

Proof: Section A.9.

Theorem 6 implies a robustness property: if the researcher is incorrect in believing that potential outcomes are linear functions of treatment assignments (Assumption 7), but uses a design such that the treatment clusters are approximately convex shapes in Euclidean space, then the OLS estimator will still converge to the GATE, albeit more slowly than IPW. This is true whether clusters grow at the optimal rate ($q = \frac{1}{2\eta+1}$) or more slowly. Note also that the ratio of the rates of convergence of the two estimators is proportional to $1 + \eta/\tau$, and so is small when η is small.

5 Inference

This section presents a central limit theorem and variance estimator for OLS in order to construct confidence intervals. To see the main challenges, let $\mathcal{F}_i$ denote the σ-algebra generated by the treatment assignments of all units in $\mathcal{N}_n(i, \kappa_n)$. We can then decompose the difference between the GATE and any linear estimator as defined in Equation 2 (including all those studied above) in the following way:

$$\frac{1}{n}\sum_{i=1}^n \omega_i Y_i - \theta_n = \underbrace{\left(\frac{1}{n}\sum_{i=1}^n \omega_i \mathbb{E}[Y_i \mid \mathcal{F}_i] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i \omega_i]\right)}_{[A]} + \underbrace{\frac{1}{n}\sum_{i=1}^n \omega_i (Y_i - \mathbb{E}[Y_i \mid \mathcal{F}_i])}_{[B]} + \underbrace{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\omega_i Y_i] - \theta_n}_{[C]}$$

Term $[A]$ will turn out to be asymptotically normal, but with variance that is not identified. Term $[B]$ is asymptotically negligible compared to the other two, and can be ignored. Term $[C]$ is a bias term which, when cluster sizes are set to increase at the optimal rate, will converge to zero at the same rate as $[A]$ and so cannot be ignored. The two challenges for inference are therefore controlling the decay of $[C]$, which depends on the degree of interdependency across units, and constructing a variance estimator for $[A]$ despite the fact that it is unidentified.

Proposition 1 provides an upper bound on $[C]$, and thus shows that $[A]$ approximates the estimation error up to that known bound. In particular, this bound is such that when κ_n grows faster than the optimal rate given the size of the clusters, then the bias term $[C]$ decays faster than $\overline{\phi}_n C_n^{-1/2}$. We will see later that—barring degenerate cases—this will make $[C]$ asymptotically negligible for inference when the researcher chooses to grow κ_n faster than optimal rate.

Proposition 1. Deviation Bound for OLS.

If Assumptions 1-7 hold and the researcher uses a Scaling Clusters design with cluster size sequence g_n, then

$$\left| \frac{\widehat{\mathrm{Cov}}\,(B^*\mathbf{v}, \mathbf{Y})}{\widehat{\mathrm{Cov}}\,(B^*\mathbf{v}, B^*\mathbf{v})} - \theta_n - [A] \right| \leq \frac{K_1}{n} \sum_{i=1}^{n} \tilde{s}_i^{-\eta} + \mathcal{O}\left(\kappa_n^{-\eta} \frac{\max\left\{ \frac{\kappa_n^2}{g_n}, \kappa_n \right\}}{\sqrt{g_n}\sqrt{n}} \right)$$

Proof: Section A.10.

To control the asymptotic distribution of $[A]$ we then use the following central limit theorem.

Theorem 7. Central Limit Theorem

Let Assumptions 1-7 hold and suppose that the researcher uses a Scaling Clusters design and such that $\frac{g_n}{\kappa_n} = \mathcal{O}(1)$ and $\kappa_n/n \to 0$. If $\liminf_{n\to\infty} \sqrt{\mathbb{V}\left([A]\right)} \frac{n^{2/3}}{\kappa_n^{2/3}\overline{\phi}_n} > 0$, then:

$$[A]/\sqrt{\mathbb{V}\left([A]\right)} \to_d N(0,1)$$

Proof: Section A.11.

This is similar to the CLT in Leung (2022b), but generalized to allow $\kappa_n/g_n \to \infty$. It is also subject to the same caveat as his: only $[A]$ is asymptotically normal, so the estimator is asymptotically normal centered on $\theta_n + [C]$, which we will refer to for convenience as $\tilde{\theta}$. If κ_n increases faster than the optimal rate $n^{\frac{1}{2\eta+1}}$, then $[C]$ will decay faster than $[A]$ and the difference between θ_n and $\tilde{\theta}_n$ will be asymptotically negligible. However, if κ_n increases at exactly the optimal rate, then confidence intervals will cover $\tilde{\theta}_n$ and not necessarily θ_n.

If we knew $\mathbb{V}\left([A]\right)$ then Theorem 7 would allow us to conduct inference. The challenge, as noted above, is that $\mathbb{V}\left([A]\right)$ is unidentified. This is a well-known consequence of unobserved heterogeneous treatment effects in a fixed-population setting (and not of spillovers or of the choice of estimator per se). Specifically, the unidentified term in the variance of any linear estimator is

$$H_n \equiv \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Lambda_{ij} \left(\mathbb{E}[\omega_i Y_i \mid \mathcal{F}_i] - \mathbb{E}[\widehat{\theta}] \right) \left(\mathbb{E}[\omega_j Y_j \mid \mathcal{F}_i] - \mathbb{E}[\widehat{\theta}] \right)$$

where Λ denotes the adjacency matrix that connects two units if and only if their κ_n-neighborhoods intersect a common cluster. As this expression illustrates, H_n is unidentified because we cannot identify the treatment effect of any unit individually. Moreover, its sign is not known, as Λ usually has both positive and negative eigenvalues.

A few approaches to solving this problem have been suggested in the literature. One can always bound H_n using Young's Inequality (Aronow and Samii, 2017), but this can lead to very conservative confidence sets when treatment clusters are large. An alternative is to require that treatment effect heterogeneity is positively correlated over large enough neighborhoods, in which case H_n vanishes in large samples (Leung, 2022b). This is attractive when it is defensible, but in some of the settings of particular interest to us treatment effects in one large region may be systematically different from those in others in economically important ways. Muralidharan and Niehaus (2017) document, for example, that the effects of the reform studied by Muralidharan et al. (2023) varied substantially across districts in India (see Figure 2). If spillovers are potentially negatively correlated across space, then H_n could be negative.

We therefore take a new approach to bound H_n. We propose the variance estimator

$$(12) \qquad \widehat{\mathbb{V}}\left([A]\right) = \frac{\widehat{\mathbf{r}}'\Lambda\widehat{\mathbf{r}}}{n^2\left(p(1-p)\overline{\phi}_n^{-1}\right)^2} - \frac{\widehat{\mathbf{q}}'(\Lambda - Q)\widehat{\mathbf{q}}}{n^2}$$

where Q denotes the block-diagonal adjacency matrix that connects two units if and only if they are in the same cluster, and

$$\widehat{r}_i \equiv B_i\mathbf{v}(Y_i - \overline{Y} - \widehat{\theta}_{n,\kappa_n}B_i\mathbf{v})$$
$$\widehat{q}_i \equiv \frac{(2d_i - 1)}{p}Y_i - \frac{\phi_i}{\overline{\phi}}\widehat{\theta}_{n,\kappa_n}$$

The second term $\frac{\widehat{\mathbf{q}}'(\Lambda-Q)\widehat{\mathbf{q}}}{n^2}$ in (12) is an asymptotically valid bound on H_n. Theorem 8 says that this procedure is guaranteed to be asymptotically conservative for the variance of $[A]$.

Theorem 8. Variance Estimation

If Assumptions 1-7 hold $\liminf_{n\to\infty} \frac{n^{2/3}}{\phi_n\kappa_n^{2/3}}\sqrt{\mathbb{V}\left([A]\right)} > 0$ *and* $g_n = O(\kappa_n) = o(n)$ *and* $g_n \to \infty$, *then under a Scaling Clusters Design:*

$$\frac{\widehat{\mathbb{V}}\left([A]\right)}{\mathbb{V}\left([A]\right)} \geq 1 + o_p(1)$$

Proof: Section A.12.

We have found that in small samples the second term of Equation 12 can be somewhat volatile, and so to temper this we recommend that the practitioner be conservative and set the second term to zero if it turns out to be positive. This preserves the guarantee from Theorem 8 that standard errors are conservative while improving performance in small samples. All of our simulations and applications use the more conservative $\widehat{\mathbb{V}}^*\left([A]\right)$ below:

$$\widehat{\mathbb{V}}^*\left([A]\right) = \frac{\widehat{\mathbf{r}}'\Lambda\widehat{\mathbf{r}}}{n^2\left(p(1-p)\overline{\phi}_n^{-1}\right)^2} - \min\left\{0, \frac{\widehat{\mathbf{q}}'(\Lambda - Q)\widehat{\mathbf{q}}}{n^2}\right\}$$

20

Theorems 7 and 8 allow us to construct valid confidence intervals covering the expectation of the estimator (which converges to θ_n). Appendix D further shows with theorems and simulations that letting κ_n grow faster than the optimal rate guarantees that confidence intervals also offer valid coverage of θ_n. Choosing κ_n to balance coverage with mean square error is analogous to the problem of choosing a smoothing parameter in nonparametric statistics.

6   Minimax Bandwidth Selection

We turn next to the problem of selecting a bandwidth κ_n given a specific, finite study population and experimental design, and an OLS estimator. Per se the convergence results above do not select a bandwidth, as they hold provided that $\kappa_n = C_\kappa n^{\frac{1}{2+\frac{1}{\eta}}}$ for an arbitrary constant C_κ. In fact, without further structure nearly any choice of bandwidth and clustering can be justified by some DGP satisfying Assumptions 1, 3, and 7. To see this, notice that for any $\kappa_n > 0$ and any experimental design, if we set potential outcomes to be $Y_i = K_1 \kappa_n^{-\eta} B_i^* \mathbf{v}$, then $\widehat{\theta}_{n,\kappa}^{OLS} = \theta_n$ almost surely. Thus, any choice of bandwidth and design is optimal under some circumstances.

When available, data from a pilot experiment could be used to select among options, but as they typically are not we focus here on obtaining a priori performance guarantees. A researcher willing to commit to specific values of K_1, η and willing to accept bias of up to X could simply choose $\kappa_n = (X/K_1)^{-1/\eta}$. For researchers uncertain about K_1, η or unwilling to exercise the researcher degrees of freedom that this involves, we provide here a default selection method that does not require contextual knowledge.

One such approach is to minimize risk under the worst-case-scenario potential outcomes. We define risk $\mathcal{R}\left(\widehat{\theta}, \mathcal{D}_n, \mathbf{Y}\right)$ for a given estimator $\widehat{\theta}$, experimental design $\mathcal{D}$, and set of potential outcomes $\mathbf{Y}(\cdot)$ in the usual way as the expected squared deviation between the estimator and the GATE. Since potential outcomes are fixed, the expectation is taken with respect to random assignments over the distribution induced by the design $\mathcal{D}$.

$$(13) \qquad \mathcal{R}\left(\widehat{\theta}, \mathcal{D}, \mathbf{Y}\right) \equiv \mathbb{E}_{\mathcal{D}}\left[\left(\widehat{\theta} - \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Y}(\mathbf{1}) - \mathbf{Y}(\mathbf{0}))\right)^2\right]$$

We consider how to select an estimator $\widehat{\theta}$ according to the "minimax" criterion, maximizing risk over the potential outcomes allowed in some set $\mathcal{Y}_n'$ and then minimizing the result over an arbitrary finite set of available estimators and designs. With some abuse of notation,

$$(14) \qquad \widehat{\theta}^* = \arg\min_{\widehat{\theta}} \ \sup_{\mathbf{Y} \in \mathcal{Y}_n'} \mathcal{R}\left(\widehat{\theta}, \mathcal{D}_n, \mathbf{Y}\right)$$

To fully characterize this problem we must choose the set of potential outcomes $\mathcal{Y}_n'$ over which to maximize risk. For tractability, we set this to be a (large) subset of the potential outcomes allowed by Assumptions 1-7. Let $\mathcal{Y}_n'(G, \mathbf{u})$ be the set of potential outcomes satisfying:

$$\mathcal{Y}_n'(G, \mathbf{u}) \equiv \left\{\mathbf{Y}(\mathbf{d}) \ : \ Y_i(\mathbf{d}) = \beta_0 + A_i\mathbf{d} + \mu_i, \ |A_{ij}| \leq G_{ij}, \ |\mu_i| < u_i, \beta_0 < \overline{Y}\right\}$$

where $\mathbf{u}$ is an $n \times 1$ vector of positive numbers and G is an $n \times n$ matrix with positive elements

satisfying $\sup_{n,s} \max_i \sum_{j \notin \mathcal{N}_n(i,s)} s^\eta G_{ij} \leq K_1$. We discuss the selection of $\mathbf{u}$ and G below. The definition of $\mathcal{Y}'_n$ imposes, in addition to Assumptions 1 and 7, a bound on the *pairwise* spillover effects, ruling out DGPs in which one pairwise spillover effect dominates all the rest. This greatly simplifies maximization while guaranteeing that $\mathcal{Y}'_n$ is still a very large set.

To minimax risk we also strengthen slightly our assumptions on the structure of neighborhoods. Assumption 9 says that we can partition the population into two halves whose intersecting neighborhoods are few relative to the size of the population. It is violated by very connected networks (e.g. "small worlds"), but like Assumption 5 will hold in any application that is similar enough to Euclidean space.

Assumption 9. North-South

The sequence of populations can be partitioned into two halves N_1, N_2 with equal membership such that there is constant $K_9 > 0$ that guarantees:

$$\frac{1}{n} \sum_{i \in N_1} \mathbf{1}\left\{\mathcal{N}_n(i,s) \cap N_2\right\} + \frac{1}{n} \sum_{i \in N_2} \mathbf{1}\left\{\mathcal{N}_n(i,s) \cap N_1\right\} \leq K_9 s n^{-\frac{1}{2}} \quad \forall s > 0$$

We now proceed to define a deterministic and observable function $R_n(G, \mathbf{u}, \kappa)$ that, under Assumption 9 and given a bandwidth $\kappa > 0$, is an asymptotically sharp upper bound on the risk. To do so, let B^* be the $n \times n$ matrix from Equation 10 constructed using the argument κ. Let $\tilde{G} \equiv G \odot \Lambda$ where $\odot$ is the element-by-element product and Λ is the $n \times n$ matrix where Λ_{ij} indicates whether $\mathcal{N}_n(i,\kappa)$ and $\mathcal{N}_n(j,\kappa)$ intersect a common cluster. Also $\mu_2 \equiv p(1-p)$ and $Q \equiv (\tilde{G}P)'B^*$ and $\overline{U}$ is the $n \times c_n$ matrix where each column equals $\mathbf{u}$.

$$
\begin{aligned}
R_n(G, \mathbf{u}, \kappa) \equiv\ & \frac{\overline{\phi}_n^2}{n^2 \mu_2^2} \left((\mu_4 - 3\mu_2^2) \sum_{c=1}^{c_n} Q_{c,c}^2 + \mu_2^2 tr(Q(Q+Q')) \right) \\
& + \frac{\overline{\phi}_n^2}{n^2 \mu_2^2} \left(\mu_2 ||\mathbf{u}'B^*||_2^2 + 2\mu_2 tr(((B^* \odot (\tilde{G}P))'(B^* \odot \overline{U})) \right) + \left(\frac{\mathbf{1}'(G - \tilde{G})\mathbf{1}}{n} \right)^2
\end{aligned}
$$

$R_n(G, \mathbf{u}, \kappa)$ is deterministic and computable (without simulation) using information only on the geography of the population. Theorem 9 shows that R_n converges to the worst-case risk of $\widehat{\theta}_{n,\kappa_n}^{OLS}$ when treatment clusters are small.

Theorem 9. Worst-Case Risk for OLS

If Assumptions 4, 5, 6, and 9 hold and G_n is a sequence of $n \times n$ positive matrices and $\mathbf{u}_n$ is a sequence of $n \times 1$ positive vectors and $\mathcal{D}_n$ is a sequence of Scaling Clusters experimental designs and $\frac{g_n}{\kappa_n} + \frac{\kappa_n}{\sqrt{g_n n}} + \kappa_n^{-\eta} \to 0$:

$$\frac{R_n(G_n, \mathbf{u}_n, \kappa_n)}{\sup_{\mathbf{Y} \in \mathcal{Y}'_n(G_n, \mathbf{u}_n)} \mathcal{R}\left(\widehat{\theta}_{n,\kappa_n}^{OLS}, \mathcal{D}_n, \mathbf{Y}\right)} \to 1$$

Proof: Appendix A.13

Minimizing R_n will thus minimax (asymptotic) risk provided that the matrix G and the vector $\mathbf{u}$ are such that the set $\mathcal{Y}'_n(G, \mathbf{u})$ being maximized over contains the true DGP. Theorem 10 shows that this can be achieved using default values for G and $\mathbf{u}$.

Theorem 10. Bandwidth Selection

Let $\{\mathbf{Y}_n^*\}_{n=1}^\infty$ be any sequence of potential outcomes where Assumptions *1-9* hold. Let the experimental designs $\mathcal{D}_n$ be Scaling Clusters designs with cluster sizes g_n. Let $\mathcal{K}_n \subseteq \mathbb{R}$ be a sequence of sets of real numbers such that for any sequence of elements $\kappa_n \in \mathcal{K}_n$, we have $\frac{g_n}{\kappa_n} + \frac{\kappa_n}{\sqrt{g_n n}} + \kappa_n^{-\eta} \to 0$. Let G_n be any sequence of $n \times n$ matrices of positive numbers. Let $\mathbf{1}_n$ denote the $n \times 1$ vector of ones. Let $\kappa_n^* = \arg\min_\kappa R_n(G_n, \mathbf{1}_n, \kappa)$.[17] Then there exists a sequence of sets of vectors of potential outcomes $\mathcal{Y}_n'$ such that for all n, $\mathbf{Y}_n^* \in \mathcal{Y}_n'$ and:

$$\frac{\sup_{\mathbf{Y} \in \mathcal{Y}_n'} \mathcal{R}\left(\widehat{\theta}_{n,\kappa_n^*}^{OLS}, \mathcal{D}_n, \mathbf{Y}\right)}{\min_{\kappa_n \in \mathcal{K}} \sup_{\mathbf{Y} \in \mathcal{Y}_n'} \mathcal{R}\left(\widehat{\theta}_{n,\kappa_n}^{OLS}, \mathcal{D}_n, \mathbf{Y}\right)} \to 1$$

If, in addition, the matrices G_n are chosen such that Assumptions *1, 3,* and *7* all hold for the potential outcomes $\mathbf{Y}_n = G_n \mathbf{v}$, then these assumptions also hold for every member of every set of potential outcomes $\mathcal{Y}_n'$.

Proof: Appendix A.14

While somewhat involved to state, this result lets us provide the researcher with a simple algorithm to follow involving minimal discretion. It says that (asymptotically) the bandwidth κ that minimizes the known function $R_n(G, \mathbf{1}, \kappa)$ is the minimax bandwidth over a set of potential outcomes that is guaranteed to contain the true potential outcomes. In Section 8 we suggest a default G for Euclidean spaces that does not require the use of outcome or pilot data.[18] The researcher can thus select a minimax bandwidth without exercising degrees of freedom by choosing a finite set of candidate bandwidths, select the one κ_n^* that minimizes $R_n(G, \mathbf{1}, \kappa)$, and then using the estimator $\widehat{\theta}_{n,\kappa_n^*}^{OLS}$.

7 Simulations

This section uses simulations in which the true values of θ and η are known to (a) check that the finite-population performance of the procedures defined above is in line with the asymptotic guarantees established by the theorems, and (b) compare the finite-population performance of the Hajek and OLS estimators.[19] We specify functional forms to be as simple and easy to reproduce as possible, and calibrate constants to produce observable moments similar to those in the empirical application in the next section.

We specify the geography as follows. We first assign units uniformly to equidistant grid positions within a two-dimensional square of area n and then add a small amount of noise to each unit's position. We define the distance function $\rho(i,j) = d(i,j)^2$ where $d(i,j)$ is the Euclidean distance. This satisfies Assumptions 2, 5, and 6. For a given cluster radius g_n, we assign units to clusters by imposing a tesselation of squares of area g_n onto the population and assigning all units within each square to their own cluster. This satisfies Definition 3 of a Scaling

---

[17]If $R_n(G_n, \mathbf{1}_n, \kappa)$ has more than one minimizer, then DEFINE κ_n^* As the largest (it does not matter).

[18]A researcher who wishes to deviate from this default should think of G_{ij} as an upper bound on $|nA_{ij}|/\sum_{m,k}|A_{mk}|$, the effect of treating unit j on outcome Y_i as a fraction of the sum of all absolute effects.

[19]We focus henceforth on the Hajek estimator as its MSE is almost always dramatically less than that of the Horvitz-Thompson estimator.

Table 1: Simulations with growing clusters

				Hajek		OLS				
n	c_n	$\overline{\phi}_n$	$\tilde{\theta}/\theta$	Mean	SD	Mean	SD	Mean SE	Cover θ	Cover $\tilde{\theta}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(9)	(10)	(11)	(12)
500	210	1.29	0.46	226	172	226	171	198	0.74	0.96
1000	310	1.65	0.56	269	142	269	139	157	0.70	0.96
2500	484	2.39	0.72	355	120	356	106	113	0.78	0.96
5000	676	2.67	0.78	390	94	389	78	82	0.75	0.96
10000	961	2.70	0.81	408	70	408	57	57	0.63	0.95

Table showing simulation performance of $\hat{\theta}_n^{OLS}$ when units are on a 2-d grid and clusters are squares. The true θ is 500 and $\eta = 0.5$ and we ran 2000 simulation repetitions per parameterization. Each row is a parameterization. Column 1 is the population size, column 2 is the number of clusters, column 3 is the mean number of clusters per κ_n-neighborhood, column 4 is the average effect of treating units within κ_n of each i divided by the true effect θ (the ratio of small-sample target to the asymptotic target), columns 5 and 6 are the mean and standard deviation of the Hajek estimates, columns 7 and 8 are the mean and standard deviation of the OLS estimates, column 9 is the mean of the standard error estimates for OLS, and columns 11 and 12 are the coverage rates for the true θ and $\tilde{\theta}$. Since we are not undersmoothing, coverage of θ is low; see Table 5 for a choice of κ_n that asymptotically covers θ.

Clusters design without necessitating the use of Algorithm 1, and thus (arguably) keeping the design as simple as possible). Overall this setup aims not for realism (which we prioritize in the application below) but for simplicity and reproducibility, and it is a natural default.

We generate outcomes using a model chosen to be as parsimonious as possible while satisfying Assumptions 1 and 7 and implying meaningful spillover effects. Given a GATE θ (which we fix with respect to n, and thus refer to in this section without a subscript) and rate of spillover decay η, we create the matrix $A_{ij}^{pre} = \min\{1, (d(i,j)/\overline{m})^{-2(\eta+1)}\}$, normalizing by the median distance $\overline{m}$ to the nearest neighbor, and use it to calculate the spillover matrix $A = \frac{\theta}{\mathbf{1}'A^{pre}\mathbf{1}/n}$. For each simulation repetition we assign each cluster to treatment independently with probability 0.5, yielding the $n \times 1$ unit treatment assignment vector $\mathbf{d}$, and then compute simulated outcomes Y_i as $Y_i = A_i\mathbf{d} + \epsilon_i$, where $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ is the error term.

We parameterize this model to yield observable moments roughly in line with those in our application below. We set $\theta = 500$ to yield point estimates of the GATE close to those in our application. We set $\sigma = 1708$ to match the standard deviation of the outcome variable among untreated households. We consider sample sizes $n \in \{500, 1000, 2500, 5000\}$ and (when computationally feasible) $n = 10000$. The size $n = 5000$ is close to the (relatively large) sample size of 5,428 households in our application, while the other values let us examine results in the smaller samples that are more typical of experimental program evaluations. We set $\eta = 0.5$ to guarantee that there is large bias at every sample size so that we can see bias decay, and then set $\kappa_n = 0.03n^{\frac{1}{2\eta+1}}$, choosing the proportionality constant so that $\overline{\phi}$ is similar to that which we will see in our empirical application below when the radius is 400 meters.

Table 1 examines results when cluster sizes scale at the theoretically suggested rate; specifically, we set $g_n = 0.11n^{\frac{1}{2\eta+1}}$, choosing the proportionality constant 0.11 so that the number of clusters c_n is close to the number in our application when $n = 5000$. We highlight four

Table 2: Simulations with small clusters

n	c_n	$\overline{\phi}_n$	SD OLS	SD Hajek
(1)	(2)	(3)	(4)	(5)
500	500	1.62	197	202
1000	1000	2.51	168	194
2500	2500	4.70	149	276
5000	5000	6.93	127	414

Table showing how $\hat{\theta}_{n\kappa_n}^{OLS}$ converges when clusters are small while Hajek does not. All parameters are the same as Table 1 except that $g_n = 0$. Each row is a parameterization. Column 1 is the sample size, column 2 is the number of clusters, column 3 is the mean number of clusters per κ_n-neighborhood, and columns 4 and 5 compare the standard deviation of OLS and Hajek. For computational reasons we omit $n = 10000$.

main conclusions. First, the Hajek and OLS estimators (which in this case are very similar on average) both approach the true value of the GATE (500) as sample size nears the larger end of the range we examine here. For small values of n, both substantially under-estimate the GATE. Second, OLS has meaningfully lower variance than Hajek, particularly in larger samples—in a sample of 5,000, approximately that available in the application below, the standard deviation of OLS is 17% lower than that of Hajek. Third, with respect to inference, standard errors for OLS constructed using the variance estimator developed in Section 5 are conservative on average relative to its actual design-based standard deviation, as predicted asymptotically by Theorem 8. And fourth, confidence intervals constructed using these standard errors provide appropriate nominal coverage of $\tilde{\theta}$, as predicted asymptotically by Theorem 7.

Remark 5. The high bias and low coverage of the true θ in Table 1 Column 8 are in line with our theorems since our confidence intervals are properly covering $\tilde{\theta}$ which converges to θ by Theorem 7. This convergence is evident in Column 4. As explained in Section 5, a researcher wishing to prioritize coverage of θ rather than the MSE of the point estimate should increase κ_n at a faster rate with n, which causes coverage of θ to converge to 0.95 at the cost of slowing down the rate of convergence of $\widehat{\theta}$ to θ. Table 5 in Appendix D illustrates this tradeoff.

Table 2 presents a second simulation exercise that illustrates the advantage of the OLS estimator over Hajek guaranteed by Theorem 5. For this exercise we force the researcher to randomize with only one unit per cluster, setting $g_n = 0$. Here the variance of OLS decreases with n while the variance of Hajek eventually increases rapidly with n. This demonstrates that in settings where spillovers decay slowly, Assumption 7 is plausible, and treatment clusters are small, OLS significantly outperforms Hajek as n grows.

8 Empirical Application

This section illustrates the application of the methods above to an actual research problem in which θ and η are unknown, so that choosing a bandwidth κ_n is a central consideration. Specifically, we reanalyze data from EHMNW, a cluster RCT that evaluated the impact of a

Table 3: Estimates of the GATE of the Give Directly Cash Transfer

Radius (meters)	0m	100m	200m	300m	400m	500m	2000m
$\widehat{\theta}$ with B^*	306	337	384	429	445	421	345
	(56)	(61)	(69)	(88)	(95)	(96)	(206)
$\widehat{\theta}$ with B^U	306	327	332	354	386	398	272
$\overline{\phi}$	1	1.21	1.59	2.07	2.66	3.32	18.69
Degree Λ	94	104	119	138	164	200	1295

Estimates of the GATE on annual household consumption expenditure in PPP US dollars of 5,419 eligible households in 653 villages. Columns indicate the radius in meters (κ_n is the square of the radius in meters). The top row shows the estimate of the GATE using our recommended regression matrix B^*. The third row shows that the choice of B^* matters by displaying the (biased) GATE estimate using a regression of outcomes on fraction of treated neighbors. Both estimates use the same sampling weights as Egger et al. (2022). The bottom two rows show $\overline{\phi}$, the mean number of villages within the radius and the mean degree of the $n \times n$ adjacency matrix Λ, which is the mean number of households in saturation groups within the radius. We account for the saturation groups in the experimental design using Appendix E, but this changes point estimates and standard errors by less than 1% (except for the 2000m radius, where saturation groups matter but the 95% confidence interval contains zero regardless of whether saturation groups are accounted for). Data: Egger et al. (2022).

randomly assigned cash transfer worth USD 1,871 PPP in a sample of 5,419 treatment-eligible households in 653 villages in rural Kenya.[20] They estimated total effects, including general equilibrium effects, of these transfers on several outcomes; for simplicity we will focus on one of these, annual household consumption expenditure (denominated in PPP USD). They used several empirical strategies, the most relevant of which for our purposes was in effect to regress outcomes for transfer-eligible households on both an indicator for living in a treated village and on the share of neighbors within 2000m that received transfers. This yielded an estimate of the average effect of the intervention on the treated of $339 (their Table 1, row 1, column 2). If instead we use their coefficients but calculate the implied average spillover effect if *all* neighbors were treated—yielding what one could interpret as an estimate of the GATE—we obtain a somewhat larger estimate of $371 but with a standard error of $191 and thus a 95% confidence interval of $[-3, 746]$ which does not exclude zero.[21]

In comparison, Table 3 reports estimates of the GATE obtained using the methods developed in this paper. We show point estimates and standard errors for six radii, ranging from zero meters to the two-kilometer radius used in the original paper; we will return to the question of bandwidth selection shortly. All estimates use the same inverse-probability sampling weights as in the original study. The row labeled B^* uses our preferred OLS estimator that regresses outcomes on the (normalized) share of treated nearby clusters, while that labeled B^U uses the naive alternative that regresses outcomes on the fraction of treated nearby units. Comparing these

---

[20]The details of treatment assignment were as follows. The 653 villages were spatially sorted into 68 saturation groups, half of which were randomly assigned to "high saturation" and half to "low saturation." In high (low) saturation groups, 2/3 (1/3) of villages were randomly assigned to treatment. We account for this stratified structure using the modified regression described in Appendix E although row 3 of Table 6 shows that this does not change any point estimate by more than 1% except that for the 2km radius which is imprecise in any case.

[21]We compute standard errors of the sum of coeffients from EHMNW using the variance covariance matrix of the original regression without any modification.

Table 4: Minimax Radii

η	Radius (meters)	Estimate (USD PPP)	95% CI
≈ 0	400	445	[257, 633]
0.25	300	429	[257, 601]
0.50	200	384	[249, 519]
0.75	200	384	[249, 519]
1.00	200	384	[249, 519]
1.25	200	384	[249, 519]
1.50	200	384	[249, 519]

Minimax radius for the GATE calculated with Theorem 9 over bandwidths {0m, 100m, 200m, 300m, 400m, 500m, 2000m}. Each row corresponds to a different G matrix where spillovers decay with a different η. We emphasize the bandwidth chosen in the top row because it yields the largest radius and is therefore the most conservative with respect to coverage. Data: Egger et al. (2022).

two rows we see that the differences are small for small or larger bandwidths, but meaningful (on the order of 10%) for intermediate ones.

The table also illustrates that the choice of a bandwidth can be quite consequential. Estimates range from \$262 at a radius of 2000m (the value that EHMNW pre-specified based on prior intuition) to \$445 at a radius of 400m. Which estimate is best? We address this, in a minimax risk sense, using the bandwidth selection methods from Section 6. Recall that the key step is to construct the matrix G, which (per Theorem 10) need only have elements that are positive and decay with distance. To make this choice less arbitrary we propose a G can be used whenever units are located in two-dimensional Euclidean space. Let $G_{ij}^{pre} = \min\{1, (d(i,j)/\overline{m})^{-2(\eta+1)}\}$, where as before $d(i,j)$ as the two-dimensional Euclidean distance (in meters) and $\overline{m}$ is the median Euclidean distance to the closest neighbor, and let $G = \frac{1}{\mathbf{1}'G^{pre}\mathbf{1}/n}$.[22] This determines G up to the choice of η. Theorem 10 guarantees that the set over which we maximize risk will contain the true DGP for any choice of η; in practice we recommend using $\eta \approx 0$ as this selects the widest bandwidth, leading to better coverage, but for the sake of illustration we compute results for a range of values.

Table 4 displays these results. For each value of η indicated in the first column, the second reports the minimax bandwidth selected from among the options {0m, 100m, 200m, 300m, 400m, 500m, 2000m}, while the third and fourth report the corresponding point estimate and confidence interval (reading off Table 3). Note that having extended the theoretical results to $\eta < 1$ is what allows us to consider several of these cases. At our recommended parametrization of $\eta \approx 0$ we select 400m as our preferred bandwidth, yielding a point estimate of \$445, which is 29% larger than the one we obtain using the hard-coded bandwidth of 2000m pre-specified in EHMNW. Even under less conservative values of η we obtain estimates substantially larger than for a 2000m bandwidth. Taken together, these results illustrate how consequential bandwidth selection can be.

---

[22]For spaces that are not two-dimensional Euclidean space, we recommend the generalization $G_{ij}^{pre} = \min\{1, (\rho(i,j)/\overline{m})^{-\eta-1}\}$ where here $\overline{m}$ is the median distance to the closest neighbor in the distance ρ.

9   Conclusion

Our analysis of the problem of GATE estimation in a general class of spaces has found that it is feasible—i.e., that consistent and rate-optimal estimator-design pairs exist—under more general conditions than was previously understood; that jointly optimizing over designs as well as linear estimators does not fundamentally change the rate bound; that there are reasonable justifications for using estimators based on OLS regressions like those used in practice to study large-scale field experiments, particularly when clusters must be "small," albeit with the tradeoff that convergence may slow down when the underlying DGP is non-linear; and that disciplined bandwidth selection for and valid inference on such procedures is feasible. Applying these methods to data from one such large-scale experiment, we obtain estimates 20% larger than those we obtain if we construct a GATE estimate from the original coefficients, with the difference largely due to differences in the bandwidth selected by our minimax procedure. Overall, the results demonstrate the viability of GATE estimation without reliance on assumptions about strictly local spillovers, which are often in uncomfortable tension with general equilibrium economic reasoning.

References

Allen, Treb, Costas Arkolakis, and Yuta Takahashi, "Universal Gravity," *Journal of Political Economy*, 2020, *128* (2), 393–433.

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

Auerbach, Eric and Max Tabord-Meehan, "The Local Approach to Causal Inference under Network Interference," 2023.

_ , **Jonathan Auerbach, and Max Tabord-Meehan**, "Discussion of 'Causal inference with misspecified exposure mappings: separating definitions and assumptions'," *Biometrika*, 02 2024, *111* (1), 21–24.

_ , **Yong Cai, and Ahnaf Rafi**, "Regression Discontinuity Design with Spillovers," 2024.

Baird, Sarah, Joan Hamory Hicks, Michael Kremer, and Edward Miguel, "Worms at Work: Long-run Impacts of a Child Health Investment*," *The Quarterly Journal of Economics*, 07 2016, *131* (4), 1637–1680.

Basse, Guillaume W. and Edoardo M. Airoldi, "Limitations of Design-based Causal Inference and A/B Testing under Arbitrary and Network Interference," *Sociological Methodology*, 2018, *48* (1), 136–151.

Bramoullé, Yann, Rachel Kranton, and Martin D'Amours, "Strategic Interaction and Networks," *American Economic Review*, March 2014, *104* (3), 898–930.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, September 2020, *110* (9), 2964–96.

Donaldson, Dave and Richard Hornbeck, " Railroads and American Economic Growth: A "Market Access" Approach *," *The Quarterly Journal of Economics*, 02 2016, *131* (2), 799–858.

Egger, Dennis, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael Walker, "General Equilibrium Effects of Cash Transfers: Experimental Evidence From Kenya," *Econometrica*, 2022, *90* (6), 2603–2643.

Faridani, Stefan and Paul Niehaus, "Linear estimation of average global effects [v4]]," 2024.

Goldsmith-Pinkham, Paul and Guido Imbens, "Social Networks and the Identification of Peer Effects," *Journal of Business, Economics and Statistics*, 2013.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 2021, *225* (2), 254–277. Themed Issue: Treatment Effect 1.

Jagadeesan, Ravi, Natesh S. Pillai, and Alexander Volfovsky, "Designs for estimating the treatment effect in networks with interference," *The Annals of Statistics*, 2020, *48* (2), 679 – 712.

Jenish, Nazgul and Ingmar R. Prucha, "On spatial processes and asymptotic inference under near-epoch dependence," *Journal of Econometrics*, 2012, *170* (1), 178–190.

Khoshnevisan, Davar, "Lecture notes on Quadratic Forms," 2011.

Lee, Lung-fei, "GMM and 2SLS estimation of mixed regressive, spatial autoregressive models," *Journal of Econometrics*, 2007, *137* (2), 489–514.

Leung, Michael P., "Causal Inference Under Approximate Neighborhood Interference," *Econometrica*, 2022, *90* (1), 267–293.

_ , "Rate-optimal cluster-randomized designs for spatial interference," *The Annals of Statistics*, 2022, *50* (5), 3064 – 3087.

_ , "Design of Cluster-Randomized Trials with Cross-Cluster Interference," 2023.

Leung, Michael P, "Discussion of 'Causal inference with misspecified exposure mappings: separating definitions and assumptions'," *Biometrika*, 02 2024, *111* (1), 17–20.

Manski, Charles F., "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, 1993, *60* (3), 531–542.

_ , "Identification of treatment response with social interactions," *The Econometrics Journal*, 2013, *16* (1), S1–S23.

Manta, Alexandra, Anson T.Y. Ho, Kim P. Huynh, and David T. Jacho-Chávez, "Estimating social effects in a multilayered Linear-in-Means model with network data," *Statistics and Probability Letters*, 2022, *183*, 109331.

Miguel, Edward and Michael Kremer, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 2004, *72* (1), 159–217.

Muralidharan, Karthik and Paul Niehaus, "Experimentation at Scale," *Journal of Economic Perspectives*, 2017, *31* (4), 103–24.

_ , _ , **and Sandip Sukhtankar**, "General Equilibrium Effects of (Improving) Public Employment Programs: Experimental Evidence From India," *Econometrica*, 2023, *91* (4), 1261–1295.

Donaldson, Dave and Richard Hornbeck, " Railroads and American Economic Growth: A "Market Access" Approach *," *The Quarterly Journal of Economics*, 02 2016, *131* (2), 799–858.

Egger, Dennis, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael Walker, "General Equilibrium Effects of Cash Transfers: Experimental Evidence From Kenya," *Econometrica*, 2022, *90* (6), 2603–2643.

Faridani, Stefan and Paul Niehaus, "Linear estimation of average global effects [v4]]," 2024.

Goldsmith-Pinkham, Paul and Guido Imbens, "Social Networks and the Identification of Peer Effects," *Journal of Business, Economics and Statistics*, 2013.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 2021, *225* (2), 254–277. Themed Issue: Treatment Effect 1.

Jagadeesan, Ravi, Natesh S. Pillai, and Alexander Volfovsky, "Designs for estimating the treatment effect in networks with interference," *The Annals of Statistics*, 2020, *48* (2), 679 – 712.

Jenish, Nazgul and Ingmar R. Prucha, "On spatial processes and asymptotic inference under near-epoch dependence," *Journal of Econometrics*, 2012, *170* (1), 178–190.

Khoshnevisan, Davar, "Lecture notes on Quadratic Forms," 2011.

Lee, Lung-fei, "GMM and 2SLS estimation of mixed regressive, spatial autoregressive models," *Journal of Econometrics*, 2007, *137* (2), 489–514.

Leung, Michael P., "Causal Inference Under Approximate Neighborhood Interference," *Econometrica*, 2022, *90* (1), 267–293.

_ , "Rate-optimal cluster-randomized designs for spatial interference," *The Annals of Statistics*, 2022, *50* (5), 3064 – 3087.

_ , "Design of Cluster-Randomized Trials with Cross-Cluster Interference," 2023.

Leung, Michael P, "Discussion of 'Causal inference with misspecified exposure mappings: separating definitions and assumptions'," *Biometrika*, 02 2024, *111* (1), 17–20.

Manski, Charles F., "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, 1993, *60* (3), 531–542.

_ , "Identification of treatment response with social interactions," *The Econometrics Journal*, 2013, *16* (1), S1–S23.

Manta, Alexandra, Anson T.Y. Ho, Kim P. Huynh, and David T. Jacho-Chávez, "Estimating social effects in a multilayered Linear-in-Means model with network data," *Statistics and Probability Letters*, 2022, *183*, 109331.

Miguel, Edward and Michael Kremer, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 2004, *72* (1), 159–217.

Muralidharan, Karthik and Paul Niehaus, "Experimentation at Scale," *Journal of Economic Perspectives*, 2017, *31* (4), 103–24.

_ , _ , **and Sandip Sukhtankar**, "General Equilibrium Effects of (Improving) Public Employment Programs: Experimental Evidence From India," *Econometrica*, 2023, *91* (4), 1261–1295.

Rangel, Marcos A. and Tom S. Vogl, "Agricultural Fires and Health at Birth," *The Review of Economics and Statistics*, 10 2019, *101* (4), 616–630.

Sussman, Daniel L. and Edoardo M. Airoldi, "Elements of estimation theory for causal effects in the presence of network interference," 2017.

Sävje, F, "Causal inference with misspecified exposure mappings: separating definitions and assumptions," *Biometrika*, 03 2024, *111* (1), 1–15.

Viviano, Davide, Lihua Lei, Guido Imbens, Brian Karrer, Okke Schrijvers, and Liang Shi, "Causal clustering: design of cluster experiments under network interference," 2024.

Wainwright, Martin J., *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019.

Watts, Duncan J. and Steven H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, 1998, *393* (6684), 440–442.

Zhang, Lina, "Spillovers of Program Benefits with Mismeasured Networks," 2023.

A   Proofs

Define the following notation: Let $\phi_n(i, s)$ denote the number of clusters that intersect the s-neighborhood of unit i. Let $\gamma_n(c, s)$ denote the number of s-neighborhoods that intersect cluster c. Formally, $\phi_n(i, s) \equiv \sum_{c=1}^{c_n} \mathbb{1}\{\mathcal{P}_{n,c} \cap \mathcal{N}_n(i, s) \neq \emptyset\}$ and $\gamma_n(c, s) \equiv \sum_{i=1}^{n} \mathbb{1}\{\mathcal{P}_{n,c} \cap \mathcal{N}_n(i, s) \neq \emptyset\}$. We omit the i or c argument when we wish to denote the maximum over a population: $\phi_n(s) \equiv \max_{i \in \mathcal{N}_n} \phi_n(i, s)$ and $\gamma_n(s) \equiv \max_{c \in \mathcal{P}_n} \gamma_n(i, s)$.

A.1   Proof of claim in Remark 2

Under the remark's hypotheses, the quantity $|Y_i(\mathbf{d}) - Y_i(\mathbf{d}')|$ when $\mathbf{d}, \mathbf{d}'$ differ only off of the $\mathcal{N}_n(i, s)$ is upper bounded by the sum over all $x > s$ of $x^{-\gamma}$ times the number of possible members of the x neighborhood but not the $x - 1$ neighborhood. Regardless of the population size, this is upper bounded by a constant times $\sum_{x=s}^{\infty} x^{-\gamma} x^{\omega-1}$. This sum is a lower Riemann sum and is upper bounded by the integral: $\int_s^{\infty} x^{\omega-\gamma-1} dx = \frac{s^{\omega-\gamma}}{\gamma-\omega}$ which implies Assumption 1 with $\eta = \gamma - \omega$ when $\gamma > \omega > 0$.

A.2   Proof of Theorem 1

Fix an sequence of positive numbers a_n. Suppose that there is a sequence of experimental designs such that for all $\delta > 0$,

$$\sup_{\mathbf{Y} \in \mathcal{Y}_n} \mathbb{P}\left(a_n \left|\widehat{\theta}_n - \theta_n\right| > \delta\right) \to 0$$

Let $\mathcal{K}_n \subseteq \mathcal{N}_n$ be a minimal $a_n^{1/\eta}$ cover of $\mathcal{N}_n$. Assumption 2 guarantees that $|\mathcal{K}_n| \leq K_2 \frac{n}{a_n^{1/\eta}}$. Let $k(i) \equiv \arg\min_{k \in \mathcal{K}_n} \rho(i, k)$. By construction $\rho(i, k(i)) \leq a_n^{1/\eta}$. Define $w_k = \sum_{i=1}^{n} \mathbf{1}\{k(i) = k\} \omega_i$.

Without loss of generality, let $\overline{Y} = K_1 = 1$. Consider the following potential outcomes: $Y_i = d_{k(i)}a_n$ which makes $\theta_n = a_n^{-1}$. These potential outcomes satisfy all of our assumptions about potential outcomes including Assumptions 1, 3, and 7. By hypothesis, $\mathbb{P}\left(a_n \left|\frac{1}{n}\sum_k a_n^{-1}d_k\omega_k - a_n^{-1}\right| > \delta\right) \to 1$. Therefore, $\mathbb{P}\left(\frac{1}{n}\sum_k d_k\omega_k > 1 - \delta\right) \to 1$. By an identical argument, $\mathbb{P}\left(-\frac{1}{n}\sum_k(1 - d_k)\omega_k > 1 - \delta\right) \to 1$ and $\mathbb{P}\left(\frac{1}{|\mathcal{K}_n|}\sum_k \frac{|\mathcal{K}_n|}{n}|\omega_k| > 1 - \delta\right) \to 1$.

Next consider a case where we randomly select a set of potential outcomes where $Y_i = R_{k(i)}$ where the R_k are $|\mathcal{K}_n|$ iid Rademacher random variables, which makes $\theta_n = 0$. This does not violate Assumption 1 because these potential outcomes do not depend on $\mathbf{d}$ at all. Under these potential outcomes $\widehat{\theta}_n = \frac{1}{n}\sum_k R_k\omega_k$ has the same distribution as $\frac{1}{|\mathcal{K}_n|}\sum_k \frac{|\mathcal{K}_n|}{n}|\omega_k|R_k$. Since the probability that a weighted sum of iid Rademachers will exceed $\delta > 0$ is larger than if we replace the weights (independent of R_k) with their mean:

$$\mathbb{P}\left(\frac{\sqrt{|\mathcal{K}_n|}}{|\mathcal{K}_n|}\sum_k \frac{|\mathcal{K}_n|}{n}|\omega_k|R_k > \delta\right) \geq \mathbb{P}\left(\left(\frac{1}{|\mathcal{K}_n|}\sum_k \frac{|\mathcal{K}_n|}{n}|\omega_k|\right)\left(\frac{\sqrt{|\mathcal{K}_n|}}{|\mathcal{K}_n|}\sum_k R_k\right) > \delta\right)$$

$$\geq \mathbb{P}\left(\frac{(1-\delta)\sqrt{|\mathcal{K}_n|}}{|\mathcal{K}_n|}\sum_k R_k > \delta\right) - \mathbb{P}\left(\frac{1}{|\mathcal{K}_n|}\sum_k \frac{|\mathcal{K}_n|}{n}|\omega_k| < 1 - \delta\right)$$

$$= \mathbb{P}\left(\frac{\sqrt{|\mathcal{K}_n|}}{|\mathcal{K}_n|}\sum_k R_k > \frac{\delta}{1-\delta}\right) + o(1)$$

Since the R_k are iid Rademacher we can invoke the central limit theorem to see that: $\liminf_{n\to\infty} \mathbb{P}\left(\frac{\sqrt{|\mathcal{K}_n|}}{|\mathcal{K}_n|}\sum_k R_k > \frac{\delta}{1-\delta}\right) > 0$. Combining this with the lower bound above yields:

$$\liminf_{n\to\infty} \mathbb{P}\left(\frac{\sqrt{|\mathcal{K}_n|}}{|\mathcal{K}_n|}\sum_k \frac{|\mathcal{K}_n|}{n}|\omega_k|R_k > \delta\right) > 0$$

Since this statement holds in expectation over a randomly chosen set of potential outcomes, it also must hold conditional on the supremum over all $\mathbf{Y} \in \mathcal{Y}_n$. So we have two guarantees:

$$\sup_{\mathbf{Y}\in\mathcal{Y}_n} \mathbb{P}\left(a_n\left|\widehat{\theta}_n - \theta_n\right| > \delta\right) \to 0$$

$$\liminf_{n\to\infty} \sup_{\mathbf{Y}\in\mathcal{Y}_n} \mathbb{P}\left(\sqrt{|\mathcal{K}_n|}\left|\widehat{\theta}_n - \theta_n\right| > \delta\right) > 0$$

The only way for these both to hold would be for: $\liminf_{n\to\infty} \sqrt{|\mathcal{K}_n|}/a_n = \infty$. Assumption 2 guarantees that $|\mathcal{K}_n| \leq K_2\frac{n}{a_n^{1/\eta}}$. So we must have $\frac{\sqrt{n}}{a_n^{1/2\eta+1}} \to \infty$. So $\frac{n}{a_n^{2+1/\eta}} \to \infty$ or $a_n = o\left(n^{\frac{1}{2+1/\eta}}\right)$ which guarantees that $\sqrt{|\mathcal{K}_n|} = o\left(n^{\frac{1}{2+1/\eta}}\right)$. So if $b_n \to \infty$, then

$$\limsup_{n\to\infty} \sup_{\mathbf{Y}\in\mathcal{Y}_n} \mathbb{P}\left(b_n n^{\frac{1}{2+1/\eta}}\left|\widehat{\theta}_n - \theta_n\right| > \delta\right) > 0$$

A.3 Proof of Theorem 2

Define: $\tilde{\theta}_{n,\kappa_n}^{HT} \equiv \frac{1}{n}\sum_{i=1}^n \left(\frac{Y_i(\mathbf{1})T_{i1}}{\mathbb{P}(T_{i1})} - \frac{Y_i(\mathbf{0})T_{i0}}{\mathbb{P}(T_{i0})}\right)$. By Assumption 1, $\left|\tilde{\theta}_{n,\kappa_n}^{HT} - \widehat{\theta}_{n,\kappa_n}^{HT}\right| \leq \frac{K_1\kappa_n^{-\eta}}{\min_i\{\mathbb{P}(T_{i1})\mathbb{P}(T_{i0})\}}$ almost surely. So the bias and variance of $\tilde{\theta}_{n,\kappa_n}^{HT}$ control the bias and variance of $\widehat{\theta}_{n,\kappa_n}^{HT}$. Since $\mathbb{E}[\tilde{\theta}_{n,\kappa_n}^{HT}] = \theta_n$ by construction and Assumption 3 guarantees that the $|Y_i|$ are uniformly bounded,

$\left| \mathbb{E} \left[\widehat{\theta}_{n,\kappa_n}^{HT} \right] - \theta_n \right| \leq \frac{K_1 \kappa_n^{-\eta}}{\min_i \{\mathbb{P}(T_{i1}) \mathbb{P}(T_{i0})\}}$. If clusters are treated independently, then $\mathbb{P}(T_{i1}) \mathbb{P}(T_{i0}) = (p - p^2)^{\phi_n(i,\kappa_n)}$. By Lemma 1, $\phi_n(i, \kappa_n)$ is uniformly bounded above, so $\mathbb{P}(T_{i1}) \mathbb{P}(T_{i0})$ is uniformly bounded away from zero. So $\mathbb{E} \left[\widehat{\theta}_{n,\kappa_n} \right] - \theta_n = O(\kappa_n^{-\eta})$.

Next we bound the variance. For ease of notation, define $Z_{i,\kappa_n} \equiv \frac{Y_i(\mathbf{1}) T_{i1}}{\mathbb{P}(T_{i1})} - \frac{Y_i(\mathbf{0}) T_{i0}}{\mathbb{P}(T_{i0})}$. We already showed that $\min_i \{\mathbb{P}(T_{i1}) \mathbb{P}(T_{i0})\}$ is bounded away from zero, so $\sup_{i,n} \mathbb{V}(Z_{i,n}) < \infty$. If clusters are treated independently, then $\text{Cov}(Z_{i,n}, Z_{j,n}) = 0$ if $\mathcal{N}_n(i, \kappa_n), \mathcal{N}_n(j, \kappa_n)$ do not intersect any treatment clusters in common. By Lemma 2 each cluster can be intersected by at most $\overline{K}_6 K_5^2 \max\{\kappa_n, K_D g_n\}$ neighborhoods of radius κ_n. So since $g_n = c_0 \kappa_n$ for a constant $c_0 > 0$, the number of pairs of neighborhoods which can intersect a common cluster is upper bounded by $\overline{K}_6 K_5^2 \max\{1, K_D c_0 \kappa_n\} \phi_n(\kappa_n) n \kappa_n$. By Lemma 1, $\phi_n(\kappa_n)$ is bounded. So $\mathbb{V}\left(\tilde{\theta}_{n,\kappa}^{HT} \right) = \mathcal{O}\left(\frac{\kappa_n}{n} \right)$. Since $\left| \tilde{\theta}_{n,\kappa_n}^{HT} - \widehat{\theta}_{n,\kappa}^{HT} \right| \leq \frac{K_1 \kappa_n^{-\eta}}{\min_i \{\mathbb{P}(T_{i1}) \mathbb{P}(T_{i0})\}}$, then $\mathbb{V}\left(\widehat{\theta}_{n,\kappa} \right) = \mathcal{O}\left(\frac{\kappa_n}{n} + \kappa_n^{-2\eta} \right)$.

Next we show that the HJ and HT differ by an amount that converges at $\mathcal{O}_p\left(\sqrt{\frac{\kappa_n}{n}} + \kappa_n^{-\eta} \right)$. Using a first-order Taylor expansion: $\frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{T_{i1}}{\mathbb{P}(T_{i1})}} = 1 - \left(\frac{1}{n} \sum_{i=1}^n \frac{T_{i1}}{\mathbb{P}(T_{i1})} - 1 \right) + \mathcal{O}_p\left(\left(\frac{1}{n} \sum_{i=1}^n \frac{T_{i1}}{\mathbb{P}(T_{i1})} - 1 \right)^2 \right)$. We know that $\mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^n \frac{T_{i1}}{\mathbb{P}(T_{i1})} - 1 \right)^2 \right] = \mathcal{O}\left(\frac{\kappa_n}{n} + \kappa_n^{-2\eta} \right)$ because this is a HT estimate for the DGP $Y_i = d_i$. So $\widehat{\theta}_{n,\kappa_n}^{HT} - \widehat{\theta}_{n,\kappa_n}^{HJ} = \mathcal{O}_p\left(\sqrt{\frac{\kappa_n}{n}} + \kappa_n^{-\eta} \right)$.

A.4 Proof of Example 3

Assumption 2: The neighborhood under the minimum distance is the union of the neighborhoods under the ρ_1, ρ_2: $\mathcal{N}_{n,3}(i, s) = \mathcal{N}_{n,1}(i, s) \cup \mathcal{N}_{n,2}(i, s)$. So any covering under ρ_1 is a covering under ρ_3. So if ρ_1 satisfies Assumption 2, then ρ_3 does as well.

Assumption 5: Since $\mathcal{N}_{n,3}(k, s) \subseteq \mathcal{N}_{n,1}(k, Cs)$, if $i, j \in \mathcal{N}_{n,3}(k, s)$, then $\rho_1(i, j) \leq K_5 Cs$ and thus $\rho_3(i, j) \leq K_5 Cs$.

Assumption 6: Using the union bound: $|\mathcal{N}_{n,3}(i, s)| \leq |\mathcal{N}_{n,2}(i, s)| + |\mathcal{N}_{n,1}(i, s)| \lesssim s$. Since $\mathcal{N}_{n,1}(i, s) \subseteq |\mathcal{N}_{n,3}(i, s)|$, we have $s \lesssim |\mathcal{N}_{n,1}(i, s)| \leq |\mathcal{N}_{n,3}(i, s)|$

A.5 Proof of Theorem 3

First we show the convergence of the denominator $\widehat{\text{Cov}}(B\mathbf{v}, B\mathbf{v})$. By Lemma 3:

$$\mathbb{E}\left[\left(\widehat{\text{Cov}}(B\mathbf{v}, B\mathbf{v}) - \frac{p(1-p) tr(B'B)}{n} \right)^2 \right] = \mathcal{O}\left(\frac{||B'\mathbf{1}||_2^4}{n^4} \right) + \mathcal{O}\left(\frac{\sum_{c=1}^{C_n} (B'B)_{cc}^2 + tr(B'BB'B)}{n^2} \right)$$

We can analyze each of these terms. First consider the term $\sum_{c=1}^{c_n} (B'B)_{c,c}^2$. Since we assume that $||B||_\infty < K$, it must be that $\max |B_{i,c}| < K$ and therefore $\sum_{c=1}^{c_n} (B'B)_{c,c}^2 \leq K^2 ||B'\mathbf{1}||_2^2$. We assumed that B_{ic} is always zero when cluster c is not intersected by the κ_n neighborhood of unit i. So $\frac{||B'\mathbf{1}||_2}{n} \leq K \frac{\sqrt{\sum_{c=1}^{c_n} \gamma_n(c, \kappa_n)^2}}{n}$. By Lemma 2, $\frac{\sqrt{\sum_{c=1}^{c_n} \gamma_n(c, \kappa_n)^2}}{n} = \mathcal{O}\left(\frac{\max\{\kappa_n, g_n\}}{\sqrt{g_n}\sqrt{n}} \right)$. Now consider the trace. $tr(B'BB'B) = \sum_{i=1}^n \sum_{j=1}^n ((BB')_{ij})^2$. By assumption, B_{ic} equals zero if cluster c is not within distance κ_n of unit i. So $(BB')_{ij}$ equals zero if i, j do not have κ_n-neighborhoods that intersect a common cluster. In order for i, j to have κ_n-neighborhoods that intersect a common cluster under a Scaling Clusters design, by Assumption 5, $\rho(i, j), \rho(j, i) \leq$

$K_5^2(g_n + 2\kappa_n)$. By Assumption 6, for each unit i, there can be at most $K_6 K_5^2(g_n + 2\kappa_n)$ units that satisfy this. Since we assumed that $||B||_\infty < K$, then $\max_i \sum_{c=1}^{c_n} |B_{ic}| \in [0, K]$, and so $tr(B'BB'B) \leq K^2 n K_6 K_5^2(g_n + 2\kappa_n)$. Thus, we have:

$$(15) \qquad \mathbb{E}\left[\left(\widehat{\text{Cov}}\,(B\mathbf{v}, B\mathbf{v}) - \frac{p(1-p)tr(B'B)}{n}\right)^2\right] = \mathcal{O}\left(\frac{\max\{g_n^2, \kappa_n^2\}}{g_n n}\right)$$

Next we bound the numerator. Let P be the $n \times C$ matrix where $P_{i,c}$ indicates that unit i is in cluster c. Assumption 7 lets us rewrite Y_i as:

$$Y_i(\mathbf{v}) = \beta_0 + (AP)_i \mathbf{v} + \mu_i, \quad \sum_{i=1}^n \mu_i = 0$$

So we have $\widehat{\text{Cov}}\,(\mathbf{Y}, B\mathbf{v}) = \widehat{\text{Cov}}\,(AP\mathbf{v}, B\mathbf{v}) + \widehat{\text{Cov}}\,(\mu, B\mathbf{v})$. By construction μ sums to zero so $\widehat{\text{Cov}}\,(\mu, B\mathbf{v}) = \frac{\mu' B\mathbf{v}}{n}$. This has expectation zero. We bound the variance using Assumption 3: $\mathbb{E}\left[\left(\frac{\mu' B\mathbf{v}}{n}\right)^2\right] = \frac{p(1-p)\mu' BB'\mu}{n^2} \leq p(1-p)\overline{Y}^2 \frac{||B'\mathbf{1}||_2^2}{n^2} = \mathcal{O}\left(\frac{\max\{g_n^2, \kappa_n^2\}}{ng_n}\right)$. By Lemma 3:

$$\mathbb{E}\left[\left(\widehat{\text{Cov}}\,(AP\mathbf{v}, B\mathbf{v}) - \frac{p(1-p)tr(B'AP)}{n}\right)^2\right] = \mathcal{O}\left(\frac{||B'\mathbf{1}||_2^2 ||(AP)'\mathbf{1}||_2^2}{n^4}\right)$$
$$+ \mathcal{O}\left(\frac{\sum_{c=1}^{C_n}(B'AP)_{cc}^2 + tr(B'APB'AP) + tr(P'A'BB'AP)}{n^2}\right)$$

By Assumption 3 and Lemma 2, $||(AP)'\mathbf{1}||_2^2 \leq \overline{Y}^2 \sum_{c=1}^{c_n} \gamma(c, \kappa_n)^2$. Next we bound $\sum_{c=1}^n (B'AP)_{c,c}^2$,

$$\sum_{c=1}^n (B'AP)_{c,c}^2 = \sum_{c=1}^n \left(\sum_{i=1}^n B_{i,c}(AP)_{i,c}\right)^2 \leq \overline{Y}^2 \sum_{c=1}^n \left(\sum_{i=1}^n B_{i,c}\right)^2 = \overline{Y}^2 ||B'\mathbf{1}||_2^2$$

Next we bound the sum of the traces.

$$tr(B'APB'AP) + tr(B'APP'A'B) = \sum_{c=1}^{c_n}\sum_{k=1}^{c_n} \left(B'AP_{c,k}\right)^2 + \sum_{c=1}^{c_n}\sum_{k=1}^{c_n} B'AP_{c,k}B'AP_{k,c}$$

An application of Cauchy-Schwartz yields: $tr(B'APB'AP) + tr(B'APP'A'B) \leq \sum_{c=1}^{c_n}\sum_{k=1}^{c_n} \left(B'AP_{c,k}\right)^2$. We can now bound this sum using the fact that by Assumptions 3 and 7, the absolute row sums of AP are no greater than $\overline{Y}$.

$$\sum_{c=1}^{c_n}\sum_{k=1}^{c_n} \left(B'AP_{c,k}\right)^2 \leq \sum_{c=1}^{c_n} \left(\sum_{k=1}^{c_n} |B'AP_{c,k}|\right)^2 \leq \sum_{c=1}^{c_n} \left(\sum_{i=1}^n \sum_{k=1}^{c_n} B_{i,c}|(AP)_{i,k}|\right)^2 \leq \overline{Y}^2 ||B'\mathbf{1}||_2^2$$

Putting all of these results together:

$$(16) \qquad \mathbb{E}\left[\left(\widehat{\text{Cov}}\,(\mathbf{Y}, B\mathbf{v}) - \frac{p(1-p)tr(B'AP)}{n}\right)^2\right] = \mathcal{O}\left(\frac{||B'\mathbf{1}||_2^2}{n^2}\right) = \mathcal{O}\left(\frac{\max\{\kappa_n^2, g_n^2\}}{ng_n}\right)$$

Next use a Taylor expansion of $\frac{1}{\widehat{\text{Cov}}(B\mathbf{v}, B\mathbf{v})}$ about $\frac{1}{\frac{p(1-p)tr(B'B)}{n}}$. Define $X_n \equiv \frac{\widehat{\text{Cov}}(\mathbf{Y}, B\mathbf{v})}{p(1-p)tr(B'B)/n} + \frac{\widehat{\text{Cov}}(\mathbf{Y}, B\mathbf{v})}{p(1-p)tr(B'B)/n} \left(\frac{p(1-p)tr(B'B)}{n} - \widehat{\text{Cov}}(B\mathbf{v}, B\mathbf{v}) \right)$. By Taylor's Theorem:

$$\mathbb{E}\left[\left(\frac{\widehat{\text{Cov}}(\mathbf{Y}, B\mathbf{v})}{\widehat{\text{Cov}}(B\mathbf{v}, B\mathbf{v})} - X_n \right)^2 \right] = \mathcal{O}_p \left(\frac{\max\{g_n^2, \kappa_n^2\}}{g_n n} \frac{n^2}{tr(B'B)^2} \right)$$

By Equations 15 and 16:

$$\mathbb{E}\left[\left(X_n - \frac{tr(B'AP)}{tr(B'B)} \right)^2 \right] = \mathcal{O}\left(\frac{\max\{\kappa_n^2, g_n^2\}}{ng_n} \frac{n^2}{tr(B'B)^2} \right)$$

Combining the two previous equations with the triangle inequality, taking the square root of both sides, and using Jensen's Inequality yields the conclusion of the theorem:

$$\mathbb{E}\left[\left| \frac{\widehat{\text{Cov}}(\mathbf{Y}, B\mathbf{v})}{\widehat{\text{Cov}}(B\mathbf{v}, B\mathbf{v})} - \frac{tr(B'AP)}{tr(B'B)} \right| \right] = \mathcal{O}\left(\frac{\max\{\kappa_n, g_n\}\sqrt{n}}{tr(B'B)\sqrt{g_n}} \right)$$

A.6   Proof of Theorem 4

First we check that B^U, B^* satisfy the hypotheses of Theorem 3. By inspection of their definitions, both matrices clearly have elements that are non negative, no greater than unity, and equal to zero when unit j is not in a cluster that is intersected by $\mathcal{N}_n(i, \kappa_n)$. Next we show that $B\mathbf{1} \leq K\mathbf{1}$ for some K. This is true for B^U because the row sums are always equal to unit. To show that this is true for B^*, use Lemma 1.

$$B_i^* \mathbf{1} \leq \frac{\max_i \phi_n(i, \kappa_n)}{\overline{\phi}_n(\kappa_n)} \leq \frac{\max_i \phi_n(i, \kappa_n)}{\min_i \phi_n(i, \kappa_n)} \leq \frac{\left(\overline{K}_6 g_n + 1 \right) \left(\overline{K}_6 K_5^2 (\kappa_n + 2g_n) + 1 \right)}{\left(\underline{K}_6 \kappa_n + 1 \right) \left(\underline{K}_6 g_n + 1 \right)}$$

The ratio $\frac{\overline{K}_6 g_n + 1}{\underline{K}_6 g_n + 1}$ is bounded because $g_n \geq 0$. If $\frac{g_n}{\kappa_n}$ is bounded, then $\frac{\overline{K}_6 K_5^2 (\kappa_n + 2g_n) + 1}{\underline{K}_6 \kappa_n + 1}$ must be bounded as well. So $B_i^* \mathbf{1}$ is bounded. So B^*, B^U satisfy all the conditions of Theorem 3.

Next we compute the estimand using B^*.

$$\frac{tr(B^{*\prime} B^*)}{n} = \frac{\sum_{i=1}^n \sum_{c=1}^{c_n} (B_{ic}^*)^2}{n} = \frac{\sum_{i=1}^n \phi_n(i, \kappa_n)\overline{\phi}^{-2}}{n} = \overline{\phi}_n^{-1}$$

$$\frac{tr((B^*)'AP)}{n} = \frac{\sum_{i=1}^n \sum_{c=1}^{c_n} B_{ic}^* (AP)_{ic}}{n}$$

$$= \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^n A_{ik} \overline{\phi}^{-1} \sum_{c=1}^{c_n} \mathbb{1}\{k \in \mathcal{P}_c\} \mathbb{1}\{\mathcal{N}_n(i, \kappa_n) \cap \mathcal{P}_c \neq \emptyset\}$$

$$\frac{tr((B^*)'AP)}{tr((B^*)'B^*)} = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^n A_{ik} \left(\sum_{c=1}^{c_n} \mathbb{1}\{k \in \mathcal{P}_c\} \mathbb{1}\{\mathcal{N}_n(i, \kappa_n) \cap \mathcal{P}_c \neq \emptyset\} \right)$$

$$= \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^n A_{ik} \mathbb{1}\{\mathcal{N}_n(i, \kappa_n) \cap \mathcal{P}_{c(k)} \neq \emptyset\}$$

By Lemma 1, $\overline{\phi} = O\left(\frac{\kappa_n}{g_n}\right)$. By Assumption 1, $\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} A_{ik}\mathbb{1}\left\{\mathcal{N}_n(i,\kappa_n) \cap \mathcal{P}_{c(k)} \neq \emptyset\right\} = \theta_n + \mathcal{O}\left(\kappa_n^{-\eta}\right)$. So the estimand when we use B^* is: $\frac{tr((B^*)'AP)}{tr((B^*)'B^*)} = \theta_n + \mathcal{O}\left(\kappa_n^{-\eta}\right)$

Next we compute the estimand when we use B^U. Computing traces:

$$\frac{tr((B^U)'AP)}{n} = \frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{c_n} \frac{\sum_{j=1}^{n}\mathbb{1}\left\{j \in \mathcal{N}_n(i,\kappa_n) \cap \mathcal{P}_c\right\}}{|\mathcal{N}_n(i,\kappa_n)|}\sum_{k=1}^{n} A_{ik}\mathbb{1}\left\{k \in \mathcal{P}_c\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} A_{ik}\sum_{c=1}^{c_n}\mathbb{1}\left\{k \in \mathcal{P}_c\right\}\frac{\sum_{j=1}^{n}\mathbb{1}\left\{j \in \mathcal{N}_n(i,\kappa_n) \cap \mathcal{P}_c\right\}}{|\mathcal{N}_n(i,\kappa_n)|}$$

$$\frac{tr((B^U)'B^U)}{n} = \frac{\sum_{i=1}^{n}\sum_{c=1}^{c_n}(B_{ci}^U)^2}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{c_n}\left(\frac{\sum_{j=1}^{n}\mathbb{1}\left\{j \in \mathcal{N}_n(i,\kappa_n) \cap \mathcal{P}_c\right\}}{|\mathcal{N}_n(i,\kappa_n)|}\right)^2$$

So the estimand when OLS is run with B^U is:

$$\frac{tr((B^U)'AP)}{tr((B^U)'B^U)} = \sum_{i=1}^{n}\sum_{k=1}^{n} A_{ik}\left(\frac{\sum_{c=1}^{c_n}\mathbb{1}\left\{k \in \mathcal{P}_c\right\}\frac{\sum_{j=1}^{n}\mathbb{1}\{j\in\mathcal{N}_n(i,\kappa_n)\cap\mathcal{P}_c\}}{|\mathcal{N}_n(i,\kappa_n)|}}{\sum_{m=1}^{n}\sum_{c=1}^{c_n}\left(\frac{\sum_{j=1}^{n}\mathbb{1}\{j\in\mathcal{N}_n(m,\kappa_n)\cap\mathcal{P}_c\}}{|\mathcal{N}_n(m,\kappa_n)|}\right)^2}\right)$$

A.7 Proof of Theorem 5

If there is an $\epsilon > 0$ such that $n^{-\epsilon}\frac{\kappa_n}{g_n} \to \infty$, then $n^{-\epsilon}\phi_n \to \infty$ because Assumption 6 guarantees that there are at least $\underline{K}_6\kappa_n$ members of each $\mathcal{N}_n(i,\kappa_n)$ but at most $K_D\overline{K}_6 g_n$ members of each cluster. If $n^{-\epsilon}\phi_n \to \infty$ then for each unit i, having $\mathbb{P}(T_{1i} = 1)$ is for large enough n less than p^{n^ϵ}. By the Bonferroni inequality $\mathbb{P}\left(\max_i T_{1i}\right) \leq np^{n^\epsilon} \to 0$. So if $n^{-\epsilon}\frac{\kappa_n}{g_n} \to \infty$ then the HT and HJ estimators converge in probability to zero and are inconsistent. So in order to use an IPW estimator, the researcher must set κ_n to be proportional to g_n and neither HT nor HJ can converge at a geometric rate faster than the maximum bias which is $\mathcal{O}\left(g_n^{-\eta}\right) = \mathcal{O}_p\left(n^{-q\eta}\right)$. Moreover the standard deviation of these estimators is of the order of the square root of the number of clusters or $\mathcal{O}\left(\sqrt{n/g_n}\right) = \mathcal{O}\left(n^{-(1-q)/2}\right)$. Since $q < \frac{1}{2\eta+1}$, $n^{-q\eta}$ is slower than $n^{-(1-q)/2}$. So the bias dominates the variance.

If we set $\kappa_n \propto n^{\frac{1+3q}{4+2\eta}}$ then the bias of any of our estimators decays with $\kappa_n^{-\eta}$ or $n^{-\frac{1+3q}{4/\eta+2}}$. To see that OLS converges with $\mathcal{O}_p\left(\frac{1+3q}{2+4/\eta}\right)$, plug $g_n = n^q$ and $\kappa_n = n^{\frac{1+3q}{4+2\eta}}$ into the result of Theorem 4. To show that this is faster than ηq, solve the inequality $\eta\frac{1+3q}{4+2\eta} \geq \eta q$ for q and see that this is equivalent to $q \leq \frac{1}{2\eta+1}$, which is the hypothesis of this theorem.

A.8 Proof of Example 4

We have assumed that all clusters are convex shapes and are contained within circles of area $K_D g_n$. The number of units within distance t of the boundary of their cluster is the number of units within Euclidean distance $\sqrt{t}$ of the boundary of their cluster. So these units lie within a zone with area less than the perimeter of the cluster times $\sqrt{t}$. Since clusters are convex shapes contained within circles, their perimeters cannot exceed the perimeter of the circle containing them. The containing circles have area at most $K_D g_n$ so they have perimeter at most

$2\sqrt{K_D g_n / \pi}$. Multiplying the perimeter by $\sqrt{t}$ we compute that area of the zone containing all these units is less than $2\sqrt{K_D g_n t / \pi}$. By the strengthened bounded density assumption in the example there can be at most $2K\sqrt{K_D g_n t / \pi}$ units in this zone. As a proportion of the cluster this is $2K\sqrt{(K_D / \pi) t / g_n}$ which is $\mathcal{O}\left(\sqrt{\frac{t}{g_n}}\right)$ which satisfies Assumption 8 with $\tau = \frac{1}{2}$.

A.9 Proof of Theorem 6

Let $d < q$. By Assumption 1, if $\mathbf{1}\left\{\mathcal{N}_n\left(i, n^{-d}\right) \setminus \mathcal{P}_c(i) \neq \emptyset\right\}$, then $|Y_i - d_i Y_i(\mathbf{1}) - (1-d_i)Y_i(\mathbf{0})| < K_1 n^{-d\eta}$. By Assumption 8, the fraction of units for which this does not hold is $\mathcal{O}\left(n^{\tau d - \tau q}\right)$. So (regardless of κ_n) the difference between OLS run on the actual Y_i and OLS run on the DGP $d_i Y_i(\mathbf{1}) + (1 - d_i)Y_i(\mathbf{0})$ is $\mathcal{O}\left(n^{\tau d - \tau q} + n^{-d\eta}\right)$. Since this holds for all $d < q$, we maximize the rate by setting $d = \frac{q\tau}{\tau + \eta}$. Thus this bias term converges with $n^{-\frac{\tau\eta}{\tau+\eta}q}$. If the researcher sets $\kappa_n = \mathcal{O}\left(n^{\frac{1+3q}{4+2\eta}}\right)$ then by Theorem 5, the standard deviation converges faster than $n^{-\eta q}$ which is slower. So the new bias term dominates the variance and OLS converges with $n^{-\frac{\tau\eta}{\tau+\eta}q}$.

A.10 Proof of Proposition 1

First we show that $[B]$ vanishes. Let $\tilde{A}$ be the $n \times n$ matrix where $\tilde{A}_{ij} = A_{ij}$ whenever j is a member of a cluster that intersects $\mathcal{N}_n(i, \kappa_n)$ and $\tilde{A}_{ij} = 0$ otherwise. Then, $[B] = \frac{1}{n}\sum_{i=1}^n \omega_i(Y_i - \mathbb{E}[Y_i \mid \mathcal{F}_i]) = \frac{1}{n}\sum_{i=1}^n \omega_i(A_i\mathbf{v} - \tilde{A}_i\mathbf{v})$. Since ω_i is symmetric about zero and independent of $A_i\mathbf{v} - \tilde{A}_i\mathbf{v}$, $\omega_i(A_i\mathbf{v} - \tilde{A}_i\mathbf{v})$ has the same distribution as $R_i\omega_i(A_i\mathbf{v} - \tilde{A}_i\mathbf{v})$ where R_i is a Rademacher independent of every other variable. By Assumption 1, this is upper bounded by $K_1 R_i \omega_i \kappa_n^{-\eta}$. By Theorem 3, $\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n R_i\omega_i\right|\right] = \mathcal{O}\left(\max\left\{\frac{\kappa_n^2}{g_n}, \kappa_n\right\} / \sqrt{g_n}\sqrt{n}\right)$. So $\mathbb{E}[\|[B]\|] = \mathcal{O}\left(\kappa_n^{-\eta}\max\left\{\frac{\kappa_n^2}{g_n}, \kappa_n\right\} / \sqrt{g_n}\sqrt{n}\right)$. By Assumption 1, $\|[C]\| \leq K_1 \kappa_n^{-\eta}$. So by the triangle inequality $\|[B] + [C]\| \leq K_1 \frac{1}{n}\sum_{i=1}^n \tilde{s}_i^{-\eta} + \mathcal{O}\left(\kappa_n^{-\eta}\max\left\{\frac{\kappa_n^2}{g_n}, \kappa_n\right\} / \sqrt{g_n}\sqrt{n}\right)$

A.11 Proof of Theorem 7

Recall that OLS can be written $\frac{1}{n}\sum_{i=1}^n Y_i\omega_i$. Let $V_i \equiv \frac{1}{n}\omega_i\mathbb{E}[Y_i \mid \mathcal{F}_i] - \frac{1}{n}\mathbb{E}[\omega_i Y_i]$. The terms $\mathbb{E}[V_i]$ have expectation zero and dependency graph Λ (where Λ_{ij} is the indictor that $\mathcal{N}_n(i, \kappa_n), \mathcal{N}_n(j, \kappa_n)$ intersect a common cluster). Since the design is Scaling Clusters, by Assumptions 5 and 6 maximum degree of Λ is $\lesssim \kappa_n + g_n$. So the degree of the dependency graph Ψ_n satisfies: $\Psi_n \lesssim \kappa_n + g_n$.

Since the outcomes are uniformly bounded by Assumption 1 and the weights are bounded by construction, all of the moments exist and can be upper bounded: $\mathbb{E}[|V_i|^3] \lesssim \frac{\overline{\phi}_n^3}{n^3}$ and $\mathbb{E}[V_i^4] \lesssim \frac{\overline{\phi}_n^4}{n^4}$. So we can use Lemma B.2 from Leung (2022b) (a.k.a. Theorem 3.6 from Ross (2011)) to bound the Wasserschein distance between a standardized version of the distribution of $[A]$ and the standard normal distribution. (Our invocation of this theorem requires extra care because the OLS weights need not be uniformly bounded.) Define $\sigma_n^2 \equiv \mathbb{V}([A])$.

$$d\left([A]/\sigma_n, \mathcal{Z}\right) \leq \frac{\Psi_n^2}{\sigma_n^3}\sum_{i=1}^n \mathbb{E}[|V_i|^3] + \frac{\sqrt{28}\Psi_n^{3/2}}{\sqrt{\pi}\sigma_n^2}\sqrt{\sum_{i=1}^n \mathbb{E}[V_i^4]}$$

$$\lesssim \frac{\kappa_n^2}{\sigma_n^3}\frac{\overline{\phi}_n^3}{n^2} + \frac{\kappa_n^{3/2}}{\sigma_n^2}\frac{\overline{\phi}_n^2}{n^{3/2}}$$

The lim inf of the right hand side converges to zero whenever $\liminf_{n\to\infty} \frac{\sigma_n n^{2/3}}{\kappa_n^{2/3}\overline{\phi}_n} > 0$ and $\liminf_{n\to\infty} \frac{\sigma_n n^{3/4}}{\kappa_n^{3/4}\overline{\phi}_n} > 0$. Since $\kappa_n/n = \mathcal{O}(1)$ by assumption, the first limit implies the second. The first limit is assumed by the hypotheses of the theorem.

A.12 Proof of Theorem 8

First define the following terms for ease of notation:

$$x_i \equiv \overline{\phi}^{-1}\left(B_i\mathbf{v} - \overline{B\mathbf{v}}/n\right), \qquad y_i \equiv \mathbb{E}[Y_i \mid \mathcal{F}_i]$$

$$\widehat{\theta}_{n,\kappa_n}^{OLS} = \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i}{\frac{1}{n}\sum_{i=1}^n x_i^2}, \qquad \tilde{\theta} \equiv \frac{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[x_i y_i]}{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i^2\right]}$$

Using a first-order Taylor approximation of $\frac{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[x_i^2]}{\frac{1}{n}\sum_{i=1}^n x_i^2}$ about $\frac{1}{n}\sum_{i=1}^n x_i^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i^2\right]$:

$$\frac{\frac{1}{n}\sum_{i=1}^n x_i y_i}{\frac{1}{n}\sum_{i=1}^n x_i^2} - \frac{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[x_i y_i]}{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i^2\right]} \approx \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \mathbb{E}[x_i y_i]}{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i^2\right]} - \frac{\left(\frac{1}{n}\sum_{i=1}^n x_i y_i\right)\left(\frac{1}{n}\sum_{i=1}^n x_i^2 - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i^2\right]\right)}{\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i^2\right]\right)^2}$$

$$\approx \frac{\frac{1}{n}\sum_{i=1}^n x_i(y_i - \tilde{\theta}x_i) - \mathbb{E}[x_i(y_i - \tilde{\theta}x_i)]}{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i^2\right]}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^n x_i(y_i - \tilde{\theta}x_i)}{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i^2\right]}$$

Next we approximate the numerator.

$$\frac{1}{n}\sum_{i=1}^n x_i y_i = \frac{1}{n}\sum_{i=1}^n (B_i\mathbf{v} - \overline{B\mathbf{v}})y_i = \frac{1}{n}\sum_{i=1}^n B_i\mathbf{v}\left(y_i - \frac{1}{n}\sum_{j=1}^n y_j\right)$$

$$= \frac{1}{n}\sum_{i=1}^n B_i\mathbf{v}\left(y_i - \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n y_j\right]\right) + \left(\frac{1}{n}\sum_{i=1}^n B_i\mathbf{v}\right)\left(\frac{1}{n}\sum_{j=1}^n y_j - \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n y_j\right]\right)$$

$$= \frac{1}{n}\sum_{i=1}^n B_i\mathbf{v}\left(y_i - \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n y_j\right]\right) + o_p\left(\frac{\max\{\kappa_n, g_n\}\kappa_n}{g_n^{3/2}\sqrt{n}}\right)$$

Let $r_i \equiv B_i\mathbf{v}(y_i - \mathbb{E}[\overline{y}] - \tilde{\theta}B_i\mathbf{v}) = B_i\mathbf{v}(\tilde{A}_i\mathbf{v} + \mu_i - \mathbb{E}[\overline{y}] - \tilde{\theta}B_i\mathbf{v})$. So we have: $\widehat{\theta}_{n,\kappa_n}^{OLS} - \tilde{\theta}_{n,\kappa_n} \approx \frac{\frac{1}{n}\sum_{i=1}^n r_i}{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[x_i^2]}$ Which yields:

$$Var\left([A]\right) \approx \frac{\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n Cov(r_j, r_i)}{\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i^2\right]\right)^2} = \frac{\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \Lambda_{ij}Cov(r_j, r_i)}{\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i^2\right]\right)^2}$$

$$= \frac{\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \Lambda_{ij}\mathbb{E}[r_j r_i] - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \Lambda_{ij}\mathbb{E}[r_j]\mathbb{E}[r_i]}{\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i^2\right]\right)^2}$$

$$= \frac{\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \Lambda_{ij}\mathbb{E}[r_j r_i]}{\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i^2\right]\right)^2} - \frac{\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n (\Lambda_{ij} - Q_{ij})\mathbb{E}[r_j]\mathbb{E}[r_i]}{\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i^2\right]\right)^2} - \frac{\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n Q_{ij}\mathbb{E}[r_j]\mathbb{E}[r_i]}{\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i^2\right]\right)^2}$$

$$= [D] - [E] - [F]$$

Where Q is the $n \times n$ matrix where $Q_{ij} = 1$ if i, j are in the same cluster. So Q is positive semidefinite and $[F] \geq 0$. $[F]$ is unidentified but non-negative so it can be ignored without affecting the validity of the confidence intervals. $[D]$ can be estimated easily. So we need only bound $[E]$. If $Q_{ij} = 0$, then d_i, d_j are independent. We can compute: $\mathbb{E}[r_i] = p(1-p)\frac{\phi_i}{\overline{\phi}^2}\left(\frac{\overline{\phi}}{\phi_i}\tilde{\theta}_{i,n} - \tilde{\theta}_n\right)$. So for $\Lambda_{ij} - Q_{ij} = 1$,

$$\mathbb{E}\left[\left(\frac{\overline{\phi}}{\phi_i}\left(\frac{d_i}{p} - \frac{1-d_i}{1-p}\right)\mathbb{E}[Y_i|T_{i1} + T_{i0} = 1] - \tilde{\theta}_n\right)\frac{\phi_j}{\overline{\phi}^2}\left(\frac{\overline{\phi}}{\phi_i}\left(\frac{d_j}{p} - \frac{1-d_j}{1-p}\right)\mathbb{E}[Y_j|T_{j1} + T_{j0} = 1] - \tilde{\theta}_n\right)\right]$$
$$= \mathbb{E}[r_i]\mathbb{E}[r_j]$$

The researcher never observes $\mathbb{E}[r_j]$. Nevertheless we wish to estimate $[E]$ by replacing $\mathbb{E}[r_j]$ with $\widehat{q}_i$ defined below.

$$\widehat{q}_i = p(1-p)\frac{\phi_i}{\overline{\phi}^2}\left(\frac{\overline{\phi}}{\phi_i}\left(\frac{d_i}{p} - \frac{1-d_i}{1-p}\right)Y_i - \widehat{\theta}_{n,\kappa_n}^{OLS}\right)$$

We now bound $|\widehat{q}_i(\Lambda_{ij} - Q_{ij})\widehat{q}_j - \mathbb{E}[r_i](\Lambda_{ij} - Q_{ij})\mathbb{E}[r_j]|$. By Assumptions 1 and 3, $|Y_i - \mathbb{E}[Y_i|T_{i1} + T_{i0} = 1]| \leq K_1\left(\min\left\{\overline{Y}, \tilde{s}_i^{-\eta}\right\} + \kappa_n^{-\eta}\right)$. By Assumptions 5 and 6, the Scaling Clusters design guarantees that the maximal degree of Λ is $\mathcal{O}(\kappa_n + g_n)$. So $|\widehat{q}_i(\Lambda_{ij} - Q_{ij})\widehat{q}_j - \mathbb{E}[r_i](\Lambda_{ij} - Q_{ij})\mathbb{E}[r_j]| \lesssim \overline{\phi}_n^{-1}\left(\min\left\{\overline{Y}, \tilde{s}_i^{-\eta}\right\} + \kappa_n^{-\eta}\right)$. So if $\frac{1}{n}\sum_{i=1}^n \min\left\{\tilde{s}_i^{-\eta}, 1\right\} = o_p(1)$, then $\frac{\widehat{\mathbf{q}}'(\Lambda - Q)\widehat{\mathbf{q}}}{n^2\left(p(1-p)\overline{\phi}_n^{-1}\right)^2} - [E] = o_p\left(\frac{\kappa_n + g_n}{n}\right)$.

By Assumption 1, the law of large numbers, and Theorem 3,

$$\max_i |\widehat{r}_i - r_i| = \max_i \left|B_i\mathbf{v}((Y_i - \mathbb{E}[Y_i|\mathcal{F}_i]) + (\overline{Y} - \mathbb{E}[\overline{Y}]) + B_i\mathbf{v}(\widehat{\theta}_n - \tilde{\theta}_n))\right| = O_p\left(\kappa_n^{-\eta} + \frac{\kappa_n^2}{\sqrt{n}g_n^{3/2}}\right)$$

. Since Λ has maximum degree $\lesssim \kappa_n + g_n \lesssim \kappa_n$, we can conclude:

$$\frac{\widehat{\mathbf{r}}'\Lambda\widehat{\mathbf{r}}}{n^2} - \frac{\mathbf{r}'\Lambda\mathbf{r}}{n^2} = O_p\left(\frac{\kappa_n^{1-\eta} + \frac{\kappa_n^3}{\sqrt{n}g_n^{3/2}}}{n}\right)$$

So, since $[F]$ is positive, $\mathbb{E}\left[\frac{\mathbf{r}'\Lambda\mathbf{r} - \widehat{\mathbf{q}}'(\Lambda - Q)\widehat{\mathbf{q}}}{n^2\left(p(1-p)\overline{\phi}_n^{-1}\right)^2}\right] \geq [D] - [E] - [F] + o\left(\frac{\kappa_n + g_n}{n}\right) + O\left(\frac{\kappa_n^{1-\eta} + \frac{\kappa_n^3}{\sqrt{n}g_n^{3/2}}}{n}\right)$.

We need now only show that the remainder terms decay faster than the variance itself. Since the maximum degree of Λ is $O(\kappa_n + g_n)$ and $\Lambda_i\widehat{\mathbf{r}}$ and $\Lambda_i\widehat{\mathbf{q}}$ are correlated with at most $O(\kappa_n + g_n)$ elements j and r_i, q_i are uniformly bounded, the squared sums converge with:

$\frac{1}{n^4}\sum_i\sum_j\sum_k\sum_l \Lambda_{ij}\Lambda_{kl}Cov(q_iq_j, q_kq_l) = O\left((\kappa_n + g_n)^2/n^2\right) = O\left(\kappa_n^2/n^2\right)$. So,

$$\phi_n^{-2}\left(\widehat{\mathbb{V}}([A]) - \mathbb{V}([A])\right) \geq O_p\left(\frac{\kappa_n^{1-\eta} + \frac{\kappa_n^3}{\sqrt{n}g_n^{3/2}}}{n}\right) + o_p\left(\frac{\kappa_n}{n}\right)$$

So if $\liminf_{n\to\infty}\frac{n^{2/3}}{\phi_n\kappa_n^{2/3}}\sqrt{\mathbb{V}([A])} > 0$, then: $\frac{\widehat{\mathbb{V}}([A])}{\mathbb{V}([A])} \geq 1 + o_p(1)$

A.13 Proof of Theorem 9

Step 1: Taylor Approximation of the OLS Estimator

By assumption, $\kappa_n/g_n \to \infty$. In the proof of Theorem 4 we showed that $\overline{\phi}_n\widehat{\mathrm{Cov}}\left(B^*\mathbf{v}, B^*\mathbf{v}\right) - p(1-p) = O_p\left(\frac{\kappa_n^2}{g_n^{3/2}\sqrt{n}}\right)$. Using a Taylor expansion of $\widehat{\mathrm{Cov}}\left(B^*\mathbf{v}, B^*\mathbf{v}\right)$ about $p(1-p)\overline{\phi}_n^{-1}$, the estimation error is:

$$
\frac{\widehat{\mathrm{Cov}}\left(\mathbf{Y}, B^*\mathbf{v}\right)}{\widehat{\mathrm{Cov}}\left(B^*\mathbf{v}, B^*\mathbf{v}\right)} - \theta_n = \frac{\widehat{\mathrm{Cov}}\left(\mathbf{Y} - \theta_n B^*\mathbf{v}, B^*\mathbf{v}\right)}{\widehat{\mathrm{Cov}}\left(B^*\mathbf{v}, B^*\mathbf{v}\right)} = \frac{\widehat{\mathrm{Cov}}\left(\mathbf{Y} - \theta_n B^*\mathbf{v}, B^*\mathbf{v}\right)}{p(1-p)\overline{\phi}_n^{-1}} \frac{p(1-p)\overline{\phi}_n^{-1}}{\widehat{\mathrm{Cov}}\left(B^*\mathbf{v}, B^*\mathbf{v}\right)}
$$

$$
= \frac{\widehat{\mathrm{Cov}}\left(\mathbf{Y} - \theta_n B^*\mathbf{v}, B^*\mathbf{v}\right)}{p(1-p)\overline{\phi}_n^{-1}}\left(1 - \frac{1}{p(1-p)}\left(\overline{\phi}_n\widehat{\mathrm{Cov}}\left(B^*\mathbf{v}, B^*\mathbf{v}\right) - p(1-p)\right)\right) + O_p\left(\frac{\kappa_n^4}{g_n^3 n}\right)
$$

But since $\frac{\widehat{\mathrm{Cov}}(\mathbf{Y}-\theta_n B^*\mathbf{v}, B^*\mathbf{v})}{p(1-p)\overline{\phi}_n} = O_p\left(\frac{\kappa_n^2}{g_n^{3/2}\sqrt{n}} + \kappa_n^{-\eta}\right)$, the first term of the estimation error dominates:

$$
\frac{\widehat{\mathrm{Cov}}\left(\mathbf{Y}, B^*\mathbf{v}\right)}{\widehat{\mathrm{Cov}}\left(B^*\mathbf{v}, B^*\mathbf{v}\right)} - \theta_n = \frac{\widehat{\mathrm{Cov}}\left(\mathbf{Y} - \theta_n B^*\mathbf{v}, B^*\mathbf{v}\right)}{p(1-p)\overline{\phi}_n^{-1}} + O_p\left(\left(\frac{\kappa_n^2}{g_n^{3/2}\sqrt{n}}\right)\left(\frac{\kappa_n^2}{g_n^{3/2}\sqrt{n}} + \kappa_n^{-\eta}\right)\right)
$$

Without loss of generality, let $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i(\mathbf{d})] = 0$ (subtracting a constant from all the Y_i has numerically no impact on the OLS estimator). Since $\frac{\overline{\phi}_n}{n}\sum_{i=1}^n B_i^*\mathbf{v} = \mathcal{O}_p\left(\sqrt{\frac{\kappa_n}{n}\phi_n}\right)$ and $\frac{1}{n}\sum_{i=1}^n Y_i = \mathcal{O}_p\left(\sqrt{\frac{\kappa_n}{n}\phi_n}\right)$, the leading term can be decomposed:

$$
\frac{\widehat{\mathrm{Cov}}\left(\mathbf{Y} - \theta_n B^*\mathbf{v}, B^*\mathbf{v}\right)}{p(1-p)\overline{\phi}_n^{-1}} = \frac{\overline{\phi}_n}{p(1-p)n}\sum_{i=1}^n Y_i B_i^*\mathbf{v} - \theta_n\frac{\overline{\phi}_n}{p(1-p)n}\sum_{i=1}^n (B_i^*\mathbf{v})^2 + \mathcal{O}_p\left(\frac{\kappa_n}{n}\phi_n\right)
$$

Since the v_c are mean zero and iid, $\overline{\phi}_n^{1/2}B_i^*\mathbf{v} = O_p(1)$. So $\frac{\phi_n}{np(1-p)}\sum_{i=1}^n (B_i^*\mathbf{v})^2 - 1 = O_p\left(\frac{\sqrt{\kappa_n}}{\sqrt{n}}\right)$. This lets us simplify further:

$$
\frac{\widehat{\mathrm{Cov}}\left(\mathbf{Y}, B^*\mathbf{v}\right)}{\widehat{\mathrm{Cov}}\left(B^*\mathbf{v}, B^*\mathbf{v}\right)} - \theta_n = \frac{\overline{\phi}_n}{p(1-p)n}\sum_{i=1}^n Y_i B_i^*\mathbf{v} - \theta_n + \mathcal{O}_p\left(\frac{\sqrt{\kappa_n}}{\sqrt{n}}\right)
$$

Notice that $\sup_{\mathbf{Y}\in\mathcal{Y}_n}\sqrt{\mathbb{V}\left[\frac{\overline{\phi}_n}{p(1-p)n}\sum_{i=1}^n Y_i B_i^*\mathbf{v}\right]} \geq \sqrt{\mathbb{V}\left(\overline{\phi}_n\frac{1}{n}\sum_{i=1}^n B_i^*\mathbf{v}\right)} \approx \frac{\sqrt{\kappa_n}}{\sqrt{n}}\overline{\phi}_n^{1/2}$. By the assumptions of Theorem 9, $\kappa_n/g_n \to \infty$ so by Lemma 1, $\overline{\phi}_n \to \infty$. So the $\mathcal{O}_p\left(\frac{\sqrt{\kappa_n}}{\sqrt{n}}\right)$ term is dominated and can be ignored from now on.

Step 2: Asymptotically Approximating the Risk

Let $\mathcal{F}_i$ denote the sigma-algebra generated by the treatment statuses of all units in a cluster intersected by $\mathcal{N}_n(i, \kappa_n)$. We can now decompose the sum:

$$\frac{\overline{\phi}_n}{p(1-p)n}\sum_{i=1}^n Y_i B_i^* \mathbf{v} - \theta_n = \frac{\overline{\phi}_n}{p(1-p)n}\sum_{i=1}^n \mathbb{E}[Y_i|\mathcal{F}_i]B_i^*\mathbf{v} - \frac{\overline{\phi}_n}{p(1-p)n}\sum_{i=1}^n \mathbb{E}[Y_i B_i \mathbf{v}]$$
$$+ \frac{\overline{\phi}_n}{p(1-p)n}\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i|\mathcal{F}_i])B_i\mathbf{v} + \frac{\overline{\phi}_n}{p(1-p)n}\sum_{i=1}^n \mathbb{E}[Y_i B_i \mathbf{v}] - \theta_n$$

By an identical argument to the proof of Proposition 1 in Section A.10, $\kappa_n^\eta \frac{\overline{\phi}_n}{n}\sum_{i=1}^n B_i^*\mathbf{v}(Y_i - \mathbb{E}[Y_i|\mathcal{F}_i]) \to_p 0$. Moreover, by the law of total probability, the expectation of the cross-terms vanish when we take the expectation of the square of the right hand side of the equation above. So we have:

$$\kappa_n^{2\eta}\mathbb{E}\left[\left(\frac{\overline{\phi}_n}{p(1-p)n}\sum_{i=1}^n Y_i B_i^*\mathbf{v} - \theta_n\right)^2\right] = \kappa_n^{2\eta}\mathbb{V}\left(\frac{\overline{\phi}_n}{p(1-p)n}\sum_{i=1}^n \mathbb{E}[Y_i|\mathcal{F}_i]B_i^*\mathbf{v}\right)$$
$$+ \kappa_n^{2\eta}\left(\frac{\overline{\phi}_n}{p(1-p)n}\sum_{i=1}^n \mathbb{E}[Y_i B_i^*\mathbf{v}] - \theta_n\right)^2 + o_p(1)$$

Step 3: Upper-Bounding the Bias and Variance

Recall that P is the $n \times C_n$ matrix where P_{ic} indicated whether unit i belongs to cluster c. Assumption 7 guarantees that $\mathbb{E}[Y_i|\mathcal{F}_i] = (\tilde{A}P)_i\mathbf{v} + \mu_i$ and $\mathbb{E}[Y_i B_i^*\mathbf{v}] = \frac{p(1-p)}{\overline{\phi}}(\tilde{A}P)_i\mathbf{1}$. Substituting these in:

$$\kappa_n^{2\eta}\mathbb{E}\left[\left(\frac{\overline{\phi}_n}{p(1-p)n}\sum_{i=1}^n Y_i B_i^*\mathbf{v} - \theta_n\right)^2\right] = \kappa_n^{2\eta}\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^n \left((\tilde{A}P)_i\mathbf{v} + \mu_i\right)B_i^*\mathbf{v}\right)$$
$$+ \kappa_n^{2\eta}\left(\frac{1}{n}\sum_{c=1}^{C_n}\sum_{i=1}^n (AP - \tilde{A}P)_{ic}\right)^2 + o_p(1)$$

Upper-bounding the bias term using the upper bound on the elements of A:

$$\left|(AP - \tilde{A}P)_{ic}\right| \leq G_{ic} - \tilde{G}_{ic}$$

Upper-bounding the variance term:

$$\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^n \left((\tilde{A}P)_i\mathbf{v} + \mu_i\right)B_i^*\mathbf{v}\right) \leq \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\left[B_i^*\mathbf{v}B_j^*\mathbf{v}(\tilde{G}P)_i\mathbf{v}(\tilde{G}P)_j\mathbf{v}\right]$$
$$+ \frac{2}{n^2}\sum_{i=1}^n\sum_{j=1}^n u_j\left|\mathbb{E}\left[B_i^*\mathbf{v}B_j^*\mathbf{v}(\tilde{G}P)_i\mathbf{v}\right]\right|$$
$$+ \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n u_i u_j\mathbb{E}\left[B_i^*\mathbf{v}B_j^*\mathbf{v}\right]$$

Step 4: Achieving the upper bound

Consider the following potential outcomes, where $\mathcal{N}_1$ is defined by Assumption 9

$$Y_i^* = \left(2\mathbf{1}\left\{i \in \mathcal{N}_1\right\} - 1\right)\left(\tilde{G}\mathbf{v} + M\right) + (G - \tilde{G})\mathbf{v}$$

Y_i^* achieves the upper bound on the bias by inspection. To see that it achieves the upper bound on the variance term,

$$\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Y_i^*|\mathcal{F}_i\right]B_i^*\mathbf{v}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[B_i^*\mathbf{v}B_j^*\mathbf{v}(\tilde{G}P)_i\mathbf{v}(\tilde{G}P)_j\mathbf{v}\right]\left(2\mathbf{1}\left\{i \in \mathcal{N}_1\right\} - 1\right)\left(2\mathbf{1}\left\{j \in \mathcal{N}_1\right\} - 1\right)$$

$$+ \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}u_i\mathbb{E}\left[B_i^*\mathbf{v}B_j^*\mathbf{v}(\tilde{G}P)_i\mathbf{v}\right]\left(2\mathbf{1}\left\{i \in \mathcal{N}_1\right\} - 1\right)\left(2\mathbf{1}\left\{j \in \mathcal{N}_1\right\} - 1\right)$$

$$+ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}u_i\overline{u}_j\mathbb{E}\left[B_i^*\mathbf{v}B_j^*\mathbf{v}\right]\left(2\mathbf{1}\left\{i \in \mathcal{N}_1\right\} - 1\right)\left(2\mathbf{1}\left\{j \in \mathcal{N}_1\right\} - 1\right)$$

For each sum, the expectation inside is zero if $\mathcal{N}_n(i, \kappa_n)$ and $\mathcal{N}_n(j, \kappa_n)$ intersect no common clusters. By Assumption 9, the number of such pairs for which $\left(2\mathbf{1}\left\{i \in \mathcal{N}_1\right\} - 1\right)\left(2\mathbf{1}\left\{j \in \mathcal{N}_1\right\} - 1\right) \neq 1$ is $o(n(\kappa_n + g_n))$, so the sum is $o\left(\frac{\kappa_n + g_n}{n}\right)$ which is of smaller order than the other terms. So the upper bound on the variance term is achieved asymptotically.

$$\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Y_i^*|\mathcal{F}_i\right]B_i^*\mathbf{v}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[B_i^*\mathbf{v}B_j^*\mathbf{v}(\tilde{G}P)_i\mathbf{v}(\tilde{G}P)_j\mathbf{v}\right]$$

$$+ \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}u_i\mathbb{E}\left[B_i^*\mathbf{v}B_j^*\mathbf{v}(\tilde{G}P)_i\mathbf{v}\right]$$

$$+ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}u_iu_j\mathbb{E}\left[B_i^*\mathbf{v}B_j^*\mathbf{v}\right] + o\left(\frac{\kappa_n + g_n}{n}\right)$$

Step 5: Analytic Expression

Next we simplify the expressions for maximized bias and variance so that the researcher can compute them analytically. For the first term we can use Lemma 3. Let Q be the $c_n \times c_n$ matrix where $Q \equiv (\tilde{G}P)'B^*$.

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[B_i^*\mathbf{v}B_j^*\mathbf{v}(\tilde{G}P)_i\mathbf{v}(\tilde{G}P)_j\mathbf{v}\right] = \frac{1}{n^2}\mathbb{V}\left(\mathbf{v}'(B^{*\prime}(\tilde{G}P))\mathbf{v}\right)$$

$$= \frac{(\mu_4 - 3\mu_2^2)\sum_{i=1}^{C}Q_{i,i}^2 + \mu_2 tr(Q^2) + \mu_2 tr(Q'Q)}{n^2}$$

We express the second term analytically. Here $\odot$ is the element-wise product.

$$\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}u_i\mathbb{E}\left[B_i^*\mathbf{v}B_j^*\mathbf{v}(\tilde{G}P)_i\mathbf{v}\right] = \frac{2p(1-p)tr((B^* \odot (\tilde{G}P))(B^* \odot \overline{U})')}{n^2}$$

We express the third term analytically also using Lemma 3:

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}u_i u_j \mathbb{E}\left[B_i^* \mathbf{v} B_j^* \mathbf{v}\right] = \mathbb{E}\left[\frac{(\mathbf{u}'B^*\mathbf{v})^2}{n^2}\right] = p(1-p)\frac{||\mathbf{u}'B^*||_2^2}{n^2}$$

This makes the upper bound on $\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Y_i^*|\mathcal{F}_i\right]B_i^*\mathbf{v}\right)$ equal to:

$$\frac{1}{n^2}\left((\mu_4 - 3\mu_2^2)\sum_{c=1}^{c_n}Q_{c,c}^2 + \mu_2^2 tr(Q^2) + \mu_2^2 tr(Q'Q) + \mu_2||\mathbf{u}'B^*||_2^2 + 2\mu_2 tr(((B^* \odot (\tilde{G}P))'(B^* \odot \overline{U}))\right)$$

where $\mu_2 \equiv p(1-p)$ and $\mu_4 = p(1-p)^4 + (1-p)p^4$.

Concluding the Proof

To lower bound the risk, simply notice that the bias is the sum of $\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C_n}(G_{ic} - \tilde{G}_{ic})$ which is positive and decays with $\kappa_n^{-\eta}$ so long as g_n starts small enough. The variance of the supremum over the DGPs is lower bounded by the variance if $Y_i = (2\mathbf{1}\{i \in \mathcal{N}_1\} - 1)$, which is: $\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}B_i^*\mathbf{v}\right)$ which decays no faster than $\frac{\kappa_n^2}{g_n n}$. Since κ_n/g_n does not go to zero, the variance must decay no faster than $\frac{\kappa_n + g_n}{n}$. Notice that expectations in the cross term are all positive so the absolute value is unnecessary. So we have:

$$\sup_{\mathbf{Y}\in\mathcal{Y}_n'}\mathcal{R}\left(\widehat{\theta}_{n,\kappa_n}^{OLS},\mathcal{D},\mathbf{Y}(\cdot)\right) = \left(\frac{\mathbf{1}'(G-\tilde{G})\mathbf{1}}{n}\right)^2$$

$$+ \frac{\overline{\phi}_n^2}{n^2\mu_2^2}\left((\mu_4 - 3\mu_2^2)\sum_{c=1}^{c_n}Q_{c,c}^2 + \mu_2^2 tr(Q^2) + \mu_2^2 tr(Q'Q) + \mu_2||\mathbf{u}'B^*||_2^2 + 2\mu_2 tr(((B^* \odot (\tilde{G}P))'(B^* \odot \overline{U}))\right)$$

$$+ o\left(\kappa_n^{-2\eta} + \frac{\kappa_n + g_n}{n}\right)$$

and $\left(\sup_{\mathbf{Y}\in\mathcal{Y}_n'}\mathcal{R}\left(\widehat{\theta}_{n,\kappa_n}^{OLS},\mathcal{D},\mathbf{Y}(\cdot)\right)\right)^{-1} = \mathcal{O}\left(\kappa_n^{-2\eta} + \frac{\kappa_n + g_n}{n}\right)$. Thus, $\frac{R_n(G,\mathbf{u},\kappa_n)}{\sup_{\mathbf{Y}\in\mathcal{Y}_n'}\mathcal{R}\left(\widehat{\theta}_{n,\kappa_n}^{OLS},\mathcal{D},\mathbf{Y}(\cdot)\right)} \to 1$.

A.14 Proof of Theorem 10

Fix a number $\overline{\theta} > 0$. Claim 1: For large enough $\overline{\theta}$, the true DGP $\mathbf{Y}(\mathbf{d})$ is in the set $\mathcal{Y}_n'(\overline{\theta}G, \overline{\theta}\mathbf{1})$. To prove Claim 1, we need only point out that Assumption 3 guarantees that $|\mu_i|, |A_{ij}|$ are uniformly bounded. So when $\overline{\theta}$ is large enough, $|\mu_i| < \overline{\theta}$ and $|A_{ij}| < \overline{\theta}G_{ij}$ for all $G_{ij} > 0$.

Claim 2: For all $\overline{\theta} > 0$, $\arg\min_{\kappa_n}\mathcal{R}_n^*(\overline{\theta}G, \overline{\theta}\mathbf{1}) = \arg\min_{\kappa_n}\mathcal{R}_n^*(G, \mathbf{1})$. To prove Claim 2, notice that every term in $\mathcal{R}_n(\overline{\theta}G, \overline{\theta}\mathbf{1})$ is proportional to $\overline{\theta}^2$. Dividing by $\overline{\theta}^2$ is a monotonic transformation that does not change the minimizer. So the minimizer is not affected by $\overline{\theta}$. So $\kappa_n^* = \arg\min_{\kappa\in\mathcal{K}_n}R_n^*(\overline{\theta}G, \overline{\theta}\mathbf{1})$. Since any sequence of elements κ_n of $\mathcal{K}_n$ satisfies the conditions of Theorem 9, $\frac{R_n^*(G_n,\mathbf{1},\kappa_n)}{\sup_{\mathbf{Y}\in\mathcal{Y}_n'(\overline{\theta}G_n,\overline{\theta}\mathbf{1}_n)}\mathcal{R}\left(\widehat{\theta}_{n,\kappa_n}^{OLS},\mathcal{D},\mathbf{Y}(\cdot)\right)} \to 1$. Let a_n be the minimizers of the numerator and let b_n be the minimizers of the dominator. $\frac{R_n^*(G_n,\mathbf{1},a_n)}{\sup_{\mathbf{Y}\in\mathcal{Y}_n'(\overline{\theta}G_n,\overline{\theta}\mathbf{1}_n)}\mathcal{R}\left(\widehat{\theta}_{n,b_n}^{OLS},\mathcal{D},\mathbf{Y}(\cdot)\right)} > \frac{R_n^*(G_n,\mathbf{1},a_n)}{\sup_{\mathbf{Y}\in\mathcal{Y}_n'(\overline{\theta}G_n,\overline{\theta}\mathbf{1}_n)}\mathcal{R}\left(\widehat{\theta}_{n,a_n}^{OLS},\mathcal{D},\mathbf{Y}(\cdot)\right)} \to 1$. By an identical argument we can show that $\frac{R_n^*(G_n,\mathbf{1},a_n)}{\sup_{\mathbf{Y}\in\mathcal{Y}_n'(\overline{\theta}G_n,\overline{\theta}\mathbf{1}_n)}\mathcal{R}\left(\widehat{\theta}_{n,b_n}^{OLS},\mathcal{D},\mathbf{Y}(\cdot)\right)}$ is smaller than a sequence that converges to one. So, $\frac{R_n^*(G_n,\mathbf{1},a_n)}{\sup_{\mathbf{Y}\in\mathcal{Y}_n'(\overline{\theta}G_n,\overline{\theta}\mathbf{1}_n)}\mathcal{R}\left(\widehat{\theta}_{n,b_n}^{OLS},\mathcal{D},\mathbf{Y}(\cdot)\right)} \to 1$. Invoking Theorem 9 one more time lets us substitute the numerator and yields the result.

B Lemmas

B.1 Lemma 1

Lemma 1. Bounding ϕ

  If Assumptions 5 and 6 hold, then for a Scaling Clusters Design with sequence g_n,

$$\max\left\{\frac{\underline{K}_6 s + 1}{\overline{K}_6 g_n + 1}, 1\right\} \leq \phi_n(i,s) \leq \frac{\overline{K}_6 K_5^2(s+2g_n)+1}{\underline{K}_6 g_n + 1} \qquad \forall i, n$$

Proof: Since the design is Scaling Clusters, $\phi_n(i,s)$ is upper bounded by the number of cluster centers with $K_D g_n$-neighborhoods that intersect the s-neighborhood of i. Applying Assumption 5 twice, $\phi_n(i,s)$ is upper bounded by the number of cluster centers inside $\mathcal{N}_n(i, K_5^2(s+g_n))$. The g_n-neighborhoods of all of these cluster centers must be contained within $\mathcal{N}_n(i, K_5^2(s+2g_n))$. Assumption 6 guarantees that each of these g_n-neighborhoods must contain at least $\underline{K}_6 g_n + 1$ units. Since this is a scaling clusters design, these g_n-neighborhoods must be mutually exclusive. So there can be at most $\frac{\overline{K}_6 K_5^2(s+2g_n)+1}{\underline{K}_6 g_n + 1}$ of these g_n neighborhoods and $\phi_n(i,s) \leq \frac{\overline{K}_6 K_5^2(s+2g_n)+1}{\underline{K}_6 g_n + 1}$. Now for the lower bound. By assumption 6, each cluster contains at most $\overline{K}_6 g_n + 1$ units and $\mathcal{N}_n(i,s)$ contains at least $\underline{K}_6 s + 1$ units. So $\phi_n(i,s) \geq \frac{\underline{K}_6 s + 1}{\overline{K}_6 g_n + 1}$ units. Since each neighborhood contains at least one unit, $\phi_n(i,s) \geq 1$ as well.

B.2 Lemma 2

Recall that $\gamma_n(c,s)$ denotes the number of s-neighborhoods that intersect cluster c.

Lemma 2. Bounding Cluster-Neighborhood Intersections

  If Assumptions 5, and 6 hold, then under a Scaling Clusters design with scaling sequence g_n :

$$\max_c \gamma_n(c,s) \lesssim \max\{s, g_n\}, \qquad \forall n \in \mathbb{N} c \in \{1, \cdots c_n\}$$

$$\frac{\sqrt{\sum_{c=1}^{c_n} \gamma_n(c,s)^2}}{n} \lesssim \frac{\max\{s, g_n\}}{\sqrt{g_n}\sqrt{n}}, \qquad \forall n \in \mathbb{N}$$

Furthermore, for a sequence κ_n where $g_n \propto \kappa_n$, then: $\frac{\sqrt{\sum_{c=1}^{c_n} \gamma_n(c,\kappa_n)^2}}{n} \lesssim \sqrt{\frac{\kappa_n}{n}}, \qquad \forall n \in \mathbb{N}$

Proof: We first bound each individual $\gamma_n(c,s)$. Under a Scaling Clusters design, each cluster is contained within a $K_D g_n$-neighborhood of some unit. By Assumption 5 if unit q is contained within $\mathcal{N}_n(i, K_D g_n)$ and $\mathcal{N}_n(j,s)$, then $i \in \mathcal{N}_n(q, K_5 K_D g_n)$ and $j \in \mathcal{N}_n(q, K_5 s)$. Invoking Assumption 5 again, $j \in \mathcal{N}_n(i, K_5^2 \max\{s, K_D g_n\})$. By Assumption 6 there are at most $\overline{K}_6 K_5^2 \max\{s, K_D g_n\} + 1$ units j that meet this criterion for each i. So $\gamma_n(c,s) \leq \overline{K}_6 K_5^2 \max\{s, K_D g_n\} \lesssim \max\{s, g_n\}$.

  Now using this bound and the fact that the scaling clusters design requires that each cluster contain a g_n-neighborhood and using Assumption 6, we can conclude that the number of clusters is at most $c_n \lesssim \frac{n}{g_n}$. This yields the bound below. The final result is obtained by substituting

in κ_n.

$$\frac{\sqrt{\sum_{c=1}^{c_n} \gamma_n(c,s)^2}}{n} \lesssim \frac{1}{n}\sqrt{c_n \max\{s, g_n\}^2} \lesssim \frac{1}{n}\sqrt{\frac{n}{g_n}} \max\{s, g_n\} = \frac{\max\{s, g_n\}}{\sqrt{g_n}\sqrt{n}}$$

B.3 Proof of Lemma 3

Lemma 3. Moments of the empirical covariance

Let $\mathbf{v}$ be a $C \times 1$ vector of mean zero i.i.d. random variables. Denote their second moment as μ_2 and their fourth moment as μ_4. Let M, W be $n \times C$ matrices of real numbers. Denote $Q \equiv M'W$. Then,

$$\mathbb{E}\left[\left(\frac{\mathbf{1}'M\mathbf{v}}{n}\right)^2\right] = \frac{\mu_2 ||M'\mathbf{1}||_2^2}{n^2}$$

$$\mathbb{E}\left[\frac{\mathbf{v}'M'W\mathbf{v}}{n}\right] = \frac{\mu_2 tr(Q)}{n}$$

$$\mathbb{V}\left(\frac{\mathbf{v}'M'W\mathbf{v}}{n}\right) = \frac{(\mu_4 - 3\mu_2^2)\sum_{i=1}^C Q_{i,i}^2 + \mu_2^2 tr(Q^2) + \mu_2^2 tr(Q'Q)}{n^2}$$

Proof: To prove the first claim, $\mathbb{E}\left[\left(\frac{\mathbf{1}'M\mathbf{v}}{n}\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^n\sum_{c=1}^C\sum_{j=1}^n\sum_{d=1}^C M_{ic}M_{jd}\mathbb{E}[v_c v_d]$. Using the fact that the v_c are iid yields $\frac{1}{n^2}\sum_{i=1}^n\sum_{c=1}^C\sum_{j=1}^n M_{ic}M_{jc}\mu_2 = \frac{\mu_2}{n^2}||M'\mathbf{1}||_2^2$.

The following proof of the other two claims is inspired by the lecture notes by Khoshnevisan (2011). Using the fact that $\mathbf{v}$ are mean zero and i.i.d,

$$\mathbb{E}\left[\frac{\mathbf{v}'Q\mathbf{v}}{n}\right] = \frac{\sum_{j=1}^C\sum_{i=1}^C Q_{ij}\mathbf{v}_i\mathbf{v}_j}{n} = \frac{\mu_2\sum_{i=1}^C Q_{ii}}{n} = \frac{\mu_2 tr(Q)}{n}$$

Now we only need to compute $\mathbb{V}\left(\frac{\mathbf{v}'Q\mathbf{v}}{n}\right)$. Evaluating the expectation of the squared quadratic form:

$$\mathbb{E}\left[\left(\frac{\mathbf{v}'Q\mathbf{v}}{n}\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^C\sum_{j=1}^C\sum_{k=1}^C\sum_{m=1}^C Q_{i,j}Q_{k,m}\mathbb{E}\left[\mathbf{v}_i\mathbf{v}_j\mathbf{v}_k\mathbf{v}_m\right]$$

Notice that each summand $\mathbb{E}\left[\mathbf{v}_i\mathbf{v}_j\mathbf{v}_k\mathbf{v}_m\right]$ falls into one of the following nine categories: the expectation equals μ_4 if $i = j = k = m$, it equals μ_2^2 if i, j, k, m can be sorted into to pairs which are equal within each pair, and it is zero otherwise. Now we can rewrite the sum:

$$\sum_{i=1}^C\sum_{j=1}^C\sum_{k=1}^C\sum_{m=1}^C Q_{i,j}Q_{k,m}\mathbb{E}\left[\mathbf{v}_i\mathbf{v}_j\mathbf{v}_k\mathbf{v}_m\right] = \mu_4\sum_{i=1}^C Q_{i,i}^2 + \mu_2^2\sum_{i=1}^C\sum_{k\neq i}^C Q_{i,i}Q_{k,k} + \mu_2^2\sum_{i=1}^C\sum_{j\neq i}^C Q_{i,j}^2$$

$$+ \mu_2^2\sum_{i=1}^C\sum_{j\neq i}^C Q_{i,j}Q_{j,i}$$

Now we can simplify the first of the double sums:

$$\sum_{i=1}^{C}\sum_{k\neq i}^{C}Q_{i,i}Q_{k,k} = \sum_{i=1}^{C}Q_{i,i}\sum_{k=1}^{C}Q_{k,k} - \sum_{i}^{C}Q_{i,i}^2 = tr(Q)^2 - \sum_{i=1}^{C}Q_{i,i}^2$$

We can simplify the second double sum in this way

$$\sum_{i=1}^{C}\sum_{j\neq i}^{C}Q_{i,j}^2 = \sum_{i=1}^{C}\sum_{j=1}^{C}Q_{i,j}^2 - \sum_{i=1}^{C}Q_{i,i}^2 = tr(Q'Q) - \sum_{i=1}^{C}Q_{i,i}^2$$

We can simplify the third double sum in this way:

$$\sum_{i=1}^{C}\sum_{j\neq i}^{C}Q_{i,j}Q_{j,i} = \sum_{i=1}^{C}\sum_{j=1}^{n}Q_{i,j}Q_{j,i} - \sum_{i=1}^{C}Q_{i,i}^2 = \sum_{i=1}^{C}(Q)_{i,i}^2 - \sum_{i=1}^{C}Q_{i,i}^2 = tr(Q^2) - \sum_{i=1}^{C}Q_{i,i}^2$$

Putting these terms together we have:

$$\mathbb{E}\left[\left(\mathbf{v}'Q\mathbf{v}\right)^2\right] = (\mu_4 - 3\mu_2^2)\sum_{i=1}^{C}Q_{i,i}^2 + \mu_2^2\left(tr(Q)^2 + tr(Q^2) + tr(Q'Q)\right)$$

We can get rid of the $tr(Q)^2$ term in the variance by noticing that it is equal to the square of the expectation term: $\mathbb{E}\left[\frac{\mathbf{v}'Q\mathbf{v}}{n}\right]^2 = \frac{\mu_2^2 tr(Q)^2}{n^2}$. So the variance is:

$$\mathbb{V}\left[\frac{\mathbf{v}'Q\mathbf{v}}{n}\right] = \mathbb{E}\left[\left(\frac{\mathbf{v}'Q\mathbf{v}}{n}\right)^2\right] - \mathbb{E}\left[\frac{\mathbf{v}'Q\mathbf{v}}{n}\right]^2 = \frac{(\mu_4 - 3\mu_2^2)\sum_{i=1}^{C}Q_{i,i}^2 + \mu_2^2 tr(Q^2) + \mu_2^2 tr(Q'Q)}{n^2}$$

Online Appendix

C   Guaranteed existence of a scaling clusters design

Assumption 5 ensures that a Scaling Clusters design always exists no matter how abstract the space. Algorithm 1 demonstrates how to find one. The algorithm first constructs a g_n-packing of cluster centers called $\mathcal{Q}$. Then the algorithm converts the $K_D g_n$-neighborhoods of each element of $\mathcal{Q}$ into clusters until the entire population is assigned to clusters. Since Assumption 5 guarantees that the neighborhoods of the elements of $\mathcal{Q}$ are a $K_D g_n$ cover, this algorithm is guaranteed to assign every unit to exactly one cluster.

Algorithm 1 Creating a Scaling Clusters Design

Require: $K_D \geq K_5, n \geq 0$, population $\mathcal{N}_n$
    Let $\mathcal{Q} = \{x\}$ where $x \in \mathcal{N}_n$
    while $\exists q \in \mathcal{N}_n$ such that $\rho(q', q), \rho(q, q') > g_n \; \forall_{q' \in \mathcal{Q}}$ **do**
        Add q to Q
    end while $\triangleright$ $\mathcal{Q}$ is a g_n packing and a $K_D g_n$ covering
    while $\exists q \in \mathcal{Q}$ such that q is not assigned to any cluster **do**
        Create a new cluster consisting of $q \cup (\mathcal{N}_n(q, K_D g_n) \setminus \mathcal{Q})$ minus any units already in a cluster.
    end while

D   Coverage-MSE Tradeoff

Section 5 shows how to construct confidence intervals that cover $\theta_n + [A]$ (a quantity defined in Section 5 which converges to the GATE). If the researcher prioritizes coverage of θ_n itself over minimizing the mean square error of the point estimate, then they should increase κ_n at a faster than optimal rate. This forces bias to decay faster than variance, so the 95% confidence intervals cover θ_n with probability converging to 95%. The cost is that variance decays more slowly and the rate of convergence is slower. Proposition 2 below shows that increasing κ_n at a faster rate results in appropriate coverage of θ_n regardless of whether we use the variance estimator $\widehat{\mathbb{V}}([A])$ or the more conservative (and recommended) $\widehat{\mathbb{V}^*}([A])$.

Proposition 2. Inference when κ_n grows fast

Let $\Phi()$ denote the CDF of the standard normal distribution. If Assumptions 1-7 hold, the researcher uses a Scaling Clusters Design with $g_n \propto n^q$ with $q \leq \frac{1}{2\eta+1}$ and sets $\kappa_n \propto n^r$ where $r \in \left(\frac{1+3q}{4+2\eta}, \frac{1+3q}{4}\right)$ and $\liminf_{n\to\infty} \frac{g_n^{3/2} n^{1/2}}{\kappa_n^2} \sqrt{\mathbb{V}([A])} > 0$, then for any critical value cv > 0:

$$\limsup_{n\to\infty} \mathbb{P}\left(\frac{\left|\frac{\widehat{\text{Cov}}(B^*\mathbf{v},\mathbf{Y})}{\widehat{\text{Cov}}(B^*\mathbf{v},B^*\mathbf{v})} - \theta_n\right|}{\sqrt{\widehat{\mathbb{V}}([A])}} > \text{cv}\right) \leq 2(1 - \Phi(\text{cv}))$$

Proof: D.1

Table 5 below demonstrates that when we repeat the simulation exercise of Table 1 but let $\kappa_n = 0.01 n^{\frac{2}{2\eta+1}}$, the coverage of θ is much improved but variance is much higher than Table 1.

Remark 6. To obtain valid confidence intervals without growing κ_n faster than the optimal rate, i.e. $\kappa_n \propto n^r$ with $r \leq \frac{1+3q}{4+2\eta}$, then the confidence intervals can simply be appended with the upper bound on the leading bias term: $\frac{K_1}{n}\sum_{i=1}^{n}\tilde{s}_i^{-\eta}$. So we would have the following guarantee:

$$\limsup_{n\to\infty} \mathbb{P}\left(\frac{\left| \frac{\widehat{\mathrm{Cov}}(B^*\mathbf{v},\mathbf{Y})}{\widehat{\mathrm{Cov}}(B^*\mathbf{v},B^*\mathbf{v})} - \theta_n \right|}{\sqrt{\widehat{\mathbb{V}}\left([A]\right)}} + \frac{K_1}{n}\sum_{i=1}^{n}\tilde{s}_i^{-\eta} > \mathrm{cv} \right) \leq 2(1-\Phi(\mathrm{cv}))$$

We argue that since in practice it may be difficult to choose the constant K_1, the researcher should simply set $\kappa_n \propto n^r$ where $r \in \left(\frac{1+3q}{4+2\eta}, \frac{1+3q}{4} \right)$.

D.1 Proof of Proposition 2

By Proposition 1, and Markov's Inequality, $\left| \widehat{\theta}_{n,\kappa_n}^{OLS} - \theta_n - [A] \right| = \mathcal{O}_p\left(\kappa_n^{-\eta} \right)$. By assumption, $\liminf_{n\to\infty} \frac{g_n^{3/2} n^{1/2}}{\kappa_n^2}\sqrt{\mathbb{V}\left([A]\right)} > 0$. So $\frac{|\widehat{\theta}_{n,\kappa_n}^{OLS}-\theta_n-[A]|}{\sqrt{\mathbb{V}([A])}} = \frac{\frac{g_n^{3/2} n^{1/2}}{\kappa_n^2}|\widehat{\theta}_{n,\kappa_n}^{OLS}-\theta_n-[A]|}{\frac{g_n^{3/2} n^{1/2}}{\kappa_n^2}\sqrt{\mathbb{V}([A])}} = \mathcal{O}_p\left(\frac{g_n^{3/2} n^{1/2}}{\kappa_n^2}\kappa^{-\eta} \right)$.

Substituting in the rates, this is $\frac{g_n^{3/2} n^{1/2}}{\kappa_n^2}\kappa_n^{-\eta} = n^{1/2+3/2q-(2+\eta)r}$. The exponent is negative whenever $r > \frac{1+3q}{4+2\eta}$. To guarantee that OLS converges, it must be that $\frac{\kappa_n^2}{g_n^{3/2} n^{1/2}} \to 0$. This is equivalent to $\frac{n^{2r}}{n^{3q/2+1/2}} \to 0$. or $r < (3q+1)/4$. So consider the case where $r \in \left(\frac{1+3q}{4}, \frac{1+3q}{4+2\eta} \right)$, which is never empty for $\eta > 0$. For these r, OLS is consistent and $\frac{|\widehat{\theta}_{n,\kappa_n}^{OLS}-\theta_n-[A]|}{\sqrt{\mathbb{V}([A])}} \to_p 0$. So by Slutsky's Theorem, the triangle inequality, and Theorem 7, $\widehat{\theta}_{n,\kappa_n}^{OLS} - \theta_n \to_d N(0,1)$.

Table 5: Simulations where κ_n grows fast

n	Optimal Rate κ_n					Fast κ_n				
	Mean	SD	Mean SE	Cover θ	Cover $\tilde{\theta}$	Mean	SD	Mean SE	Cover θ	Cover $\tilde{\theta}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(9)	(10)	(11)	(12)
500	226	171	198	0.74	0.96	386	332	350	0.92	0.93
1000	269	139	157	0.70	0.96	397	263	264	0.92	0.92
2500	356	106	113	0.78	0.96	431	182	179	0.93	0.93
5000	389	78	82	0.75	0.96	447	135	138	0.93	0.94
10000	408	57	57	0.63	0.95	460	104	103	0.93	0.94

Table showing improved coverage when κ_n grows faster than the optimal rate. The true θ is 500. η is set to 0.5 and we ran 2000 simulation repetitions per parameterization. Each row is a parameterization. Parameterizations are identical to Table 1. Columns 2-6 are exact reproductions of Table 1 where $\kappa_n = 0.03n^{\frac{1}{2}}$ grows at the optimal rate. Columns 7-12 show simulation results for a repeated exercise where $\kappa_n = 0.06n^{0.619}$ is larger and grows at nearly the fastest rate allowed by Proposition 2.

E Stratification and Block-Randomization

Not every experimental design assigns treatment to clusters independently. Let a "block randomization" be a design where clusters are grouped into "blocks" or "strata" and exactly fraction

p of clusters within each block are treated and treatments are uncorrelated across blocks. Define $\Omega(i)$ as the variance covariance matrix of the elements of $\mathbf{v}$ on which the row B_i^* has support. Let $X(i)$ be the $c_n \times c_n$ matrix equal to $\Omega(i)^{-1}$ with zero rows and columns added for the elements of $\mathbf{v}$ on which B_i^* has no support. Theorem 11 shows that when we multiply $\mathbf{v}$ by $X(i)$ to remove the cross-cluster correlation and the blocks are large enough compared to κ_n, then the weighted regression under block randomization converges in the same fashion as the OLS would if the clusters were treated iid. We can write this regression as a Wald estimator where $B_i^* X(i)\mathbf{v}$ instruments for $B_i^*\mathbf{v}$. Comparing Row 2 of Table 3 to Row 3 of Table 6 shows that this adjustment happens to make almost no difference for our point estimates in our empirical application.

Theorem 11. *Assume that the conditions of Theorem 3 hold, except that clusters are block-randomized. Assume that there is a universal constant K such that the number of clusters per block is less than K. Assume that for each block b intersected by $\mathcal{N}_n(i, \kappa_n)$, at least one cluster in b does not intersect $\mathcal{N}_n(i, \kappa_n)$. Then,*

$$\overline{\phi}\frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i B_i^* X(i)\mathbf{v} - \theta_n = \mathcal{O}_p\left(\kappa_n^{-\eta} + \frac{\max\{\kappa_n^2/g_n, \kappa_n\}}{\sqrt{g_n}\sqrt{n}}\right)$$

In addition, the following Wald estimator converges at the same rate:

$$\frac{\frac{1}{n}\sum_{i=1}^n Y_i B_i^* X(i)\mathbf{v} - \left(\frac{1}{n}\sum_{i=1}^n Y_i\right)\frac{1}{n}\sum_{i=1}^n B_i^* X(i)\mathbf{v}}{\frac{1}{n}\sum_{i=1}^n B_i^*\mathbf{v}B_i^* X(i)\mathbf{v} - (\mathbf{1}'B^*\mathbf{v}/n)\frac{1}{n}\sum_{i=1}^n B_i^* X(i)\mathbf{v}} - \theta_n = \mathcal{O}_p\left(\kappa_n^{-\eta} + \frac{\max\{\kappa_n^2/g_n, \kappa_n\}}{\sqrt{g_n}\sqrt{n}}\right)$$

Proof: By the same arguments as in Theorem 3, $Y_i = \beta_1 + A_i P\mathbf{v} + \mu_i$ where $\sum_{i=1}^n \mu_i = 0$. Next, for each unit i, make a list of the blocks that intersect $\mathcal{N}_n(i, \kappa_n)$. For each such block, choose one cluster that does not intersect $\mathcal{N}_n(i, \kappa_n)$. This is always possible by hypothesis. Call the chosen block $z(i)$. For each row of B^*, define $\tilde{\mathbf{v}}(i)$ as $\mathbf{v}$ with all the clusters in block $z(i)$ left out. Notice that if cluster c was deleted from $\tilde{\mathbf{v}}(i)$, then $B_{ic}^* = 0$. Let $\Omega(i)$ be the variance covariance matrix of the set of clusters on which B^* has support. No co-linearity is possible so $\Omega(i)$ must be invertible. Let $X(i)$ be the inverse of $\Omega(i)$ with rows of zeros added for each cluster left out of $\tilde{\mathbf{v}}_i$. Let $\tilde{P}_i$ denote ith row of the matrix P with the elements representing the clusters in block $z(i)$ deleted. Now we can use the "trace trick" and Assumption 1 to compute:

$$\begin{aligned}
\mathbb{E}[Y_i B_i^* X(i)^{-1}\mathbf{v}] &= \mathbb{E}[Y_i B_i^* \Omega(i)^{-1}\tilde{\mathbf{v}}(i)] \\
&= \mathbb{E}[\mathbf{v}'(AP)_i' B_i^* \Omega(i)^{-1}\tilde{\mathbf{v}}(i)] \\
&= \mathbb{E}[\tilde{\mathbf{v}}(i)'(A\tilde{P})_i' B_i^* \Omega(i)^{-1}\tilde{\mathbf{v}}(i)] + \mathcal{O}\left(\kappa_n^{-\eta}\right) \\
&= \mathbb{E}\left[tr\left(\Omega(i)^{-1}\tilde{\mathbf{v}}(i)\tilde{\mathbf{v}}(i)'(A\tilde{P})_i' B_i^*\right)\right] + \mathcal{O}\left(\kappa_n^{-\eta}\right) \\
&= tr\left((A\tilde{P})_i' B_i^*\right) + \mathcal{O}\left(\kappa_n^{-\eta}\right) \\
&= tr\left((AP)_i' B_i^*\right) + \mathcal{O}\left(\kappa_n^{-\eta}\right) \\
&= \overline{\phi}^{-1} A_i \mathbf{1} + \mathcal{O}\left(\kappa_n^{-\eta}\right)
\end{aligned}$$

Therefore:

$$\overline{\phi}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}Y_iB_i^*\Omega(i)^{-1}\mathbf{v}\right]=\theta_n+\mathcal{O}\left(\kappa_n^{-\eta}\right)$$

The next step is to control the variance. Let $\tilde{Y}_i\equiv\beta_1+\tilde{A}_i\mathbf{v}+\mu_i$ where $\tilde{A}_{ij}=A_{ij}$ whenever $\mathcal{N}_n(i,\kappa_n)$ and $\mathcal{N}_n(j,\kappa_n)$ intersect a common block and zero otherwise. By Assumption 1:

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_iB_i^*\Omega(i)^{-1}\mathbf{v}-\frac{1}{n}\sum_{i=1}^{n}Y_iB_i^*\Omega(i)^{-1}\mathbf{v}=\mathcal{O}_p\left(\kappa_n^{-\eta}\right)$$

Since Y_i and the number of clusters per block is bounded, the variance of $\tilde{Y}_iB_i^*\Omega(i)^{-1}\mathbf{v}$ is bounded. If the units in $\mathcal{N}_n(i,\kappa_n)$ and $\mathcal{N}_n(j,\kappa_n)$ do not intersect any common block, then $\tilde{Y}_iB_i^*\Omega(i)^{-1}\mathbf{v}$ and $\tilde{Y}_jB_j^*\Omega(j)^{-1}\mathbf{v}$ are independent. $\mathcal{N}_n(i,\kappa_n)$ can only intersect at most ϕ_n clusters. So $\mathcal{N}_n(i,\kappa_n)$ can intersect at most ϕ_nK blocks which connects it to ϕ_nK^2 clusters. So for each i, the number of units for which $\tilde{Y}_iB_i^*\Omega(i)^{-1}\mathbf{v}$ and $\tilde{Y}_jB_j^*\Omega(j)^{-1}\mathbf{v}$ are not independent is of the same order as if there were no blocks. So the rate of convergence of the variance is the same as if there were no blocks.

We can also use a Wald estimator. The denominator of the Wald estimator is:

$$\frac{1}{n}\sum_{i=1}^{n}B_i^*\mathbf{v}B_i^*X(i)\mathbf{v}-\left(\mathbf{1}'B^*\mathbf{v}/n\right)\frac{1}{n}\sum_{i=1}^{n}B_i^*X(i)\mathbf{v}=\frac{1}{n}\sum_{i=1}^{n}\left(B_i^*\mathbf{v}B_i^*X(i)\mathbf{v}-\left(\mathbf{1}'B^*\mathbf{v}/n\right)\right)B_i^*X(i)\mathbf{v}$$

$$=\frac{1}{n}\sum_{i=1}^{n}B_i^*\mathbf{v}B_i^*X(i)\mathbf{v}+\mathcal{O}\left(\frac{\kappa_n}{n}\right)$$

The term $\mathcal{O}\left(\frac{\kappa_n}{n}\right)$ will be dominated. Taking the expectation of the rest:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}B_i^*\mathbf{v}B_i^*X(i)\mathbf{v}\right]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(B_i^*\mathbf{v})(B_i^*X(i)\mathbf{v})\right]$$

$$=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[tr\left(\Omega(i)^{-1}\tilde{\mathbf{v}}\tilde{\mathbf{v}}'(B_i^*)'B_i^*\right)\right]$$

$$=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[tr\left(I(B_i^*)'B_i^*\right)\right]$$

$$=\frac{1}{n}\sum_{i=1}^{n}\frac{\phi_n}{\overline{\phi}_n^2}$$

$$=\overline{\phi}_n^{-1}$$

The $\mathcal{O}\left(\frac{\kappa_n}{n}\right)$ term is dominated because $\overline{\phi}_n^{-1}$ decays no faster than $\frac{g_n}{\kappa_n}$. Since the number of clusters per block is assumed to be bounded, the elements of $X(i)$ are uniformly bounded. So $Var((B_i^*\mathbf{v})(B_i^*X(i)\mathbf{v}))$ is bounded. Since $X(i)$ has no more support than B^*, each summand of $\frac{1}{n}\sum_{i=1}^{n}(B_i^*\mathbf{v})(B_i^*X(i)\mathbf{v})$ cannot be correlated with more other summands than the maximum degree of Λ times K which is $\mathcal{O}\left(\frac{\kappa_n^2}{g_n}\right)$. Thus, the variance of the denominator is $\mathcal{O}\left(\frac{\kappa_n^2}{g_nn}\right)$. So $\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}B_i^*\mathbf{v}B_i^*X(i)\mathbf{v}-(\mathbf{1}'B^*\mathbf{v}/n)\frac{1}{n}\sum_{i=1}^{n}B_i^*X(i)\mathbf{v}-\overline{\phi}_n^{-1}\right|\right]=\mathcal{O}\left(\frac{\kappa_n}{\sqrt{n}g_n}\right)$.

49

F   The GATE when a Subset are Ineligible for Treatment

Consider a study population where a subset of units $\mathcal{I}_n \subset \mathcal{N}_n$ are "treatment ineligible." This means that they belong to no cluster and the probability that they are treated is zero. We call all the other units in $\mathcal{N}_n \setminus \mathcal{I}_n$ the "treatment eligible" units and assume that they are all members of a scaling clusters design with cluster size parameter g_n and treatment probability $p \in (0, 1)$.

The researcher now wishes to estimate a variant of the GATE, $\theta_n^{(1)}$, which is the average causal effect of treating all of the eligible units, as opposed to none of them, on the whole study population $\mathcal{N}_n$. Let $\mathbf{1}_{\mathcal{N}_n \setminus \mathcal{I}_n}$ be the $n \times 1$ vector which indicates membership in $\mathcal{N}_n \setminus \mathcal{I}_n$. Define this variant as:

$$\theta_n^{(1)} \equiv \frac{1}{n} \sum_{i=1}^n (Y_i(\mathbf{1}_{\mathcal{N}_n \setminus \mathcal{I}_n}) - Y_i(\mathbf{0}))$$

The estimand $\theta_n^{(1)}$ is straightforward to estimate using the methods developed earlier. To do so the researcher need only treat each unit in $\mathcal{I}_n$ as if it were a member of its respective nearest treatment cluster and use any of the estimators in this paper—ignoring the fact that some units are ineligible and that ineligible units lumped into treated clusters were not actually treated. Then all of the results in this paper will apply, but for the estimand $\theta_n^{(1)}$ instead of θ_n.

Proof: Consider an alternative DGP $(\mathcal{D}_n', \mathbf{Y}')$ with the same population and distance function as the factual one but a different experimental design $\mathcal{D}_n'$ and set of potential outcomes $\mathbf{Y}'$. Let the alternative experimental design $\mathcal{D}_n'$ be the same as the factual one $\mathcal{D}_n$, except that now each ineligible unit is instead a member of the treatment cluster nearest to it. Let the alternative set of potential outcomes be $Y_i'(\mathbf{d}) = Y_i(\mathbf{d} \odot \mathbf{1}_{\mathcal{N}_n \setminus \mathcal{I}_n}))$ where $\odot$ is the element-wise product. The GATE under the alternative DGP is now equal to $\theta_n^{(1)}$ since under alternative DGP, the treatment assignment of ineligible units has no effect on any outcome. If Y_i satisfies Assumptions 1, 3, and/or 7 then so does Y_i'. So under the alternative DGP, the conclusions of Theorems 1-6 would hold.

Since the vector of unit treatment assignments $\mathbf{d}$ is a function of the vector of cluster assignments $\mathbf{b}$ we can always write any set of potential outcomes as functions of $\mathbf{b}$ rather than $\mathbf{d}$. Notice also that when we do this, $Y_i'(\mathbf{b}) = Y_i(\mathbf{b})$ for all vectors of cluster treatment assignments $\mathbf{b}$. The vector $\mathbf{b}$ is identical under both the factual and alternative DGP. So the factual joint distribution of $\mathbf{Y}, \mathbf{b}$ under $\mathcal{D}_n$ is identical to the alternative joint distribution of $\mathbf{Y}', \mathbf{b}'$ under $\mathcal{D}_n'$. All estimators in this paper are functions of $\mathbf{Y}, \mathbf{b}$. So the researcher can always calculate based on the factual data what the value of any of the estimators in this paper would have been under the alternative DGP $(\mathcal{D}_n', \mathbf{Y}')$—and the conclusions of Theorems 1-6 hold under alternative DGP $(\mathcal{D}_n', \mathbf{Y}')$.

G   Regressing on Multiple Rings

This paper considers a procedure where we regress outcomes on $B^* \mathbf{v}$ or an average of treatments within a single ring. But EHMNW and others instead use two regressors: own treatment and

Table 6: Regressing on Own Treatment

	100m	200m	300m	400m	500m	2000m
Two Regressors	333	380	423	434	403	305
One Regressor	337	384	429	445	421	345

Table showing point estimates of the GATE for six radii. The second row shows the point estimate when we use two regressors: (own village treatment and $B^*\mathbf{v}$ with own village excluded) and sum the two coefficients. The third row shows the point estimate when we regress only on $B^*\mathbf{v}$ with own village included. The third row differs slightly from Table 3 because for comparability across columns we do not include the adjustment for the block-randomization defined in Appendix E here. As in Table 3, there are 5419 observations.

neighbors' treatment within a ring. They then take a (weighted) sum of the two coefficients. Table 6 shows that whether we on own village treatment separately in this way does not change point estimates in our empirical application by a meaningful amount.