Sharp Closed-Form Bounds for Interference Bias in Linear ATT Estimators

Joao Alipio-Correa

Department of Political Science & Department of Statistics, University of Pittsburgh

Overview

Setting.

- Social-science theories constantly entertain interference: policies diffuse, information travels, interventions reshape nearby units. Comparison groups are rarely beyond a treatment's reach.
- Every linear ATT estimator: $\hat{\tau}(w) = \overline{\Delta Y}_{\mathcal{T}} - \sum_j w_j \, \Delta Y_j^{\text{obs}}$. DiD, staggered adoption, synthetic control differ only in w.
- Spillovers reach nearby controls → the counterfactual is partially treated → a bias $\mathcal{B}(w)$ built from unobserved counterfactuals.

Contribution.

A. Detect interference (doubly robust $\hat{\psi}$).

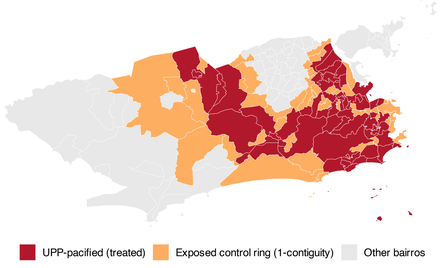
B. Bound its damage (sharp closed-form bounds, confidence intervals, sensitivity). **An effect bounded against interference, with estimator unchanged.**

C. Prove the identified mean of the missing counterfactuals is enough to sharply bound the bias of any linear estimator.

D. The width of the bounds is governed by how unevenly the estimator weights exposed controls: uniform weights point-identify; concentration widens the set.

Motivating problem: police pacification in Rio

2008–2014: Rio de Janeiro installs Pacifying Police Units (UPPs) in 55 communities. Outcome: the **left bloc's vote share** across 156 neighborhoods. A control neighborhood bordering a pacified community shares its streets, commerce, and policing. Its observed change embeds part of the intervention.



Control	E_j	ΔY_j^{obs}	$\Delta Y_j(0,0)$
connected to a UPP	1	✓	**missing**
distant	0	✓	$= \Delta Y_j^{\text{obs}}$

Each control reveals one of two potential changes:

$$\Delta Y_j^{\text{obs}} = E_j \, \Delta Y_j(0,1) + (1 - E_j) \, \Delta Y_j(0,0)$$

Theorem 1.2 (Contamination decomposition).

$$\hat{\tau}(w) = \tau^{\text{SUTVA}}(w) - \mathcal{B}(w)$$
$$\mathcal{B}(w) = \sum_{j \in \mathcal{C}_1} w_j \big[\Delta Y_j^{\text{obs}} - \Delta Y_j(0,0) \big]$$

Only exposed controls enter. Nonnegative spillovers ⇒ $\mathcal{B}(w) \geq 0$: the effect is *understated*.

How large is the interference bias $\mathcal{B}(w)$, and how tightly can it be bounded from the data?

▸ *The proposal: identify and recover the mean of the missing vector $\{\Delta Y_j(0,0)\}_{j \in \mathcal{C}_1}$.*

Assumptions

Notation guide

$\Delta Y_j(a,e)$	outcome change under own treatment a, exposure e
E_j	$= 1$ if a treated unit lies within distance r of j — computed from the data, never assumed
$\mathcal{C}_0$, $\mathcal{C}_1$; n_1; ρ	unexposed / exposed controls; $\lvert\mathcal{C}_1\rvert$; $n_1/\lvert\mathcal{C}\rvert$
$\delta_j = \Delta Y_j(0,1) - \Delta Y_j(0,0)$	unit-level spillover
$\psi = \mathbb{E}[\delta_j \mid A_j = 0]$	average spillover on controls
$d_j = \Delta Y_j^{\text{obs}} - a$	spillover capacity (a: support floor)

Assumptions

A1 — Exposure ignorability. $\Delta Y_j(0,0) \perp\!\!\!\perp E_j \mid X_j, \{X_k\}_{k \in \mathcal{N}_j}, A_j = 0$.
Sitting near a treated unit is as good as random, given own and neighbors' covariates.

A2 — Treatment ignorability. $A_i \perp\!\!\!\perp \Delta Y_i(a,e) \mid X_i$. *Adoption unconfounded given covariates.*

A3 — Independent assignments. $A_i \perp\!\!\!\perp A_j \mid X_i, X_j$. *With A2, delivers A1.*

A4 — Positivity. $\eta < \Pr(E_j = 1 \mid \cdot) < 1 - \eta$. *Every control could sit near or far.*

A5 — Overlap. *Exposed controls have unexposed counterparts with the same covariates.*

A6 — Feasibility. $0 \leq \psi \leq \bar{d}$. *Average spillover within average capacity; checked from data.*

▸ *A1–A5: unexposed neighborhoods reveal, on average, what exposed ones would have done.*

From detection to partial identification

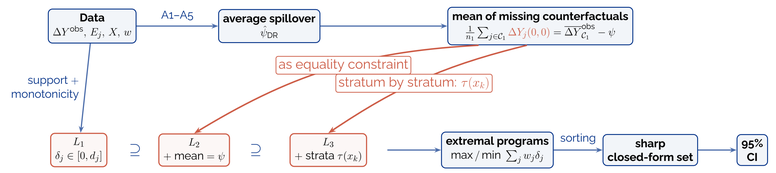
Detection estimand: $\psi = \mathbb{E}[\Delta Y_j(0,1) - \Delta Y_j(0,0) \mid A_j = 0]$, the ATT of exposure on controls, identified under A1–A5. Doubly robust estimation:

$$\hat{\psi}_{\text{DR}} = \frac{1}{n_1} \underbrace{\sum_{j \in \mathcal{C}_1} \big[\Delta Y_j - \hat{\mu}_0(X_j)\big]}_{\text{outcome-model residuals (exposed)}} - \frac{1}{n_1} \underbrace{\sum_{j \in \mathcal{C}_0} \frac{\hat{\pi}_j}{1 - \hat{\pi}_j} \big[\Delta Y_j - \hat{\mu}_0(X_j)\big]}_{\text{propensity-reweighted correction (unexposed)}}$$

$\hat{\mu}_0(x)$: outcome model on unexposed controls. $\pi_j = 1 - \prod_{k \in \mathcal{N}_j}(1 - e(X_k))$: exposure propensity, from neighbors' adoption probabilities.

Properties: doubly robust — either nuisance correct suffices; $\sqrt{n}$-normal; semiparametric efficient.

$\hat{\psi}_{\text{DR}}$ identifies the mean of the missing counterfactuals: $\frac{1}{n_1} \sum_{j \in \mathcal{C}_1} \Delta Y_j(0,0) = \overline{\Delta Y}_{\mathcal{C}_1}^{\text{obs}} - \psi$.
Not the end product: the detected average enters the bounding problem as an equality constraint on the missing vector.



▸ *Nested requirements, nested sets: $\mathcal{F}_3 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_1$. Each addition shrinks the bounds.*

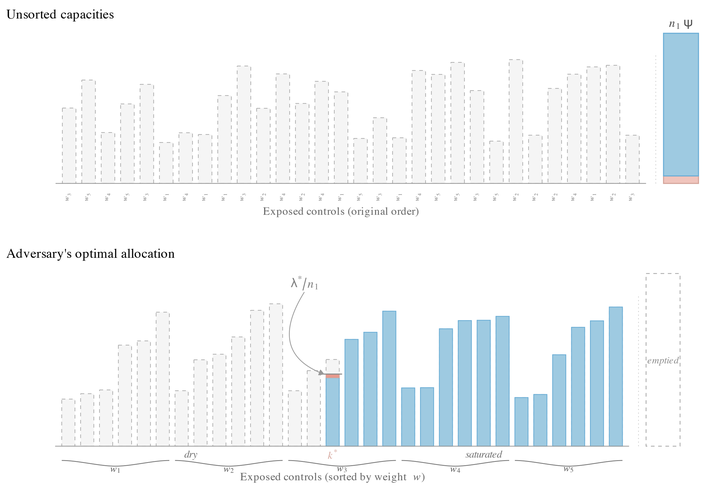
Sharp bounds by sorting

The L_2 extremes solve

$$\mathcal{B}_2^{\text{U}}(w) = \max_{\delta} \sum_{j \in \mathcal{C}_1} w_j \delta_j \quad \text{s.t.} \quad \underbrace{0 \leq \delta_j \leq d_j}_{\text{capacity}}, \qquad \underbrace{\tfrac{1}{n_1} \textstyle\sum_j \delta_j = \psi}_{\text{detection}}.$$

Theorem 3.8 (Sorting). Sort weights ascending; the maximizing allocation fills capacities from the highest weight down until the budget $n_1\psi$ is spent; one threshold unit k^* absorbs the residual. **Theorem 3.9** evaluates it in closed form:

$$\mathcal{B}_2^{\text{U}}(w) = \sum_{j > k^*} w_{(j)} d_{(j)} + w_{(k^*)} \big[n_1\psi - C(k^*{+}1)\big], \qquad C(k^*{+}1) < n_1\psi \leq C(k^*).$$



Low-weight units stay *dry*, high-weight units *saturate*, the threshold unit takes the residual. The lower bound mirrors the allocation from the lowest weight up.

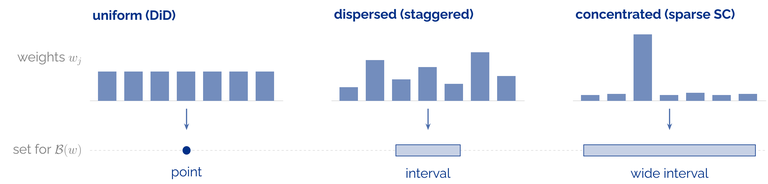
Sharp: no spillover configuration consistent with the data can escape these bounds. LP duality certifies the bounds and an explicit data-generating process attains them: sharpness is certified twice (computation: one sort, $O(n_1 \log n_1)$).

Width of the identification region

Theorem 3.10.

$$\mathcal{W}(w) = \mathcal{B}_2^{\text{U}} - \mathcal{B}_2^{\text{L}} = \sum_{j=1}^{n_1} \tilde{w}_{(j)} \Delta_{(j)} \;\leq\; \underbrace{2 \max_j |\tilde{w}_{(j)}|}_{\text{weight dispersion}} \cdot \underbrace{n_1 \min(\psi,\, \bar{d} - \psi)}_{\text{budget flexibility}}$$

with $\tilde{w}_{(j)} = w_{(j)} - \bar{w}$ the centered sorted weights and $\Delta_{(j)}$ the gap between the two extreme allocations.
Zero width $\iff$ uniform weights: then $\mathcal{B} = \rho\hat{\psi}$ and detection alone de-contaminates. Concentrating weight widens the set.

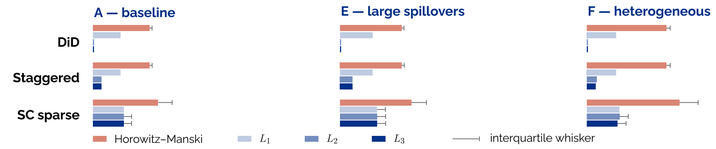


Same data, same $\hat{\psi}$: the weight profile alone sets the width.

▸ *The ranking of widths is set by the weighting strategy before any data arrive. The evidence below tests it.*

Simulation evidence

11 DGP configurations × 4 weight schemes (HHI $0.0002 \to 0.028$) × 5,000 replications; median widths at $N = 5000$, interquartile whiskers.



- **Coverage** — the probability that the reported interval contains the true contamination $\mathcal{B}(w)$. The Imbens–Manski interval covers it, and every value of the identified set, with probability $\geq 95\%$: observed ≥ 0.95 in every cell.
  - Ignore interference and the naive 95% CI covers the true ATT in **0% of replications**.
- Median contamination: 25% of the ATT at baseline, 50% under large spillovers. Stratification (Config F): a further 8–12%.

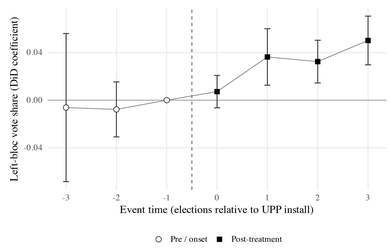
Compared with Horowitz–Manski: they use one worst-case capacity for every unit, $W_{\exp}(b - a)$, and detection plays no role. Here, unit-specific capacities $d_j = \Delta Y_j^{\text{obs}} - a$ are strictly tighter, and the identified mean has no HM analog: HM → L_2 removes 100 / 85.2 / 49.8% of the width (DiD / staggered / sparse SC, baseline).

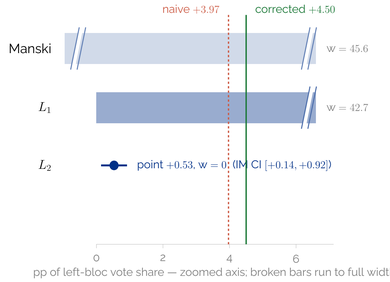
Estimation and inference

1. Fit $\hat{\mu}_0$ on unexposed controls; $\hat{\pi}_j$ from neighbors' adoption propensities (cross-fit if using ML).
2. Compute $\hat{\psi}_{\text{DR}}$; check feasibility $0 \leq \hat{\psi} \leq \bar{d}$.
3. Sort exposed controls by weight; thresholds; closed-form bounds (L_3: per stratum).
4. Influence-function SEs through the piecewise-linear bound map; Imbens–Manski interval.
5. Repeat over r; report the sensitivity curve.

Application A — UPP Rio: uniform weights, a point

Outcome: **left-bloc vote share**. 156 neighborhoods, 55 pacified, staggered event study; exposure = bordering a pacified community ($\rho = 0.40$); pre-trends flat ($p = 0.84,\ 0.51$).



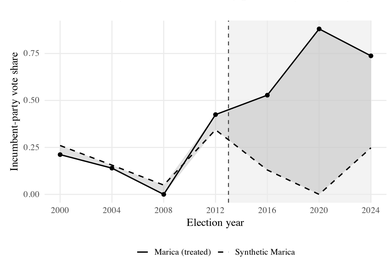


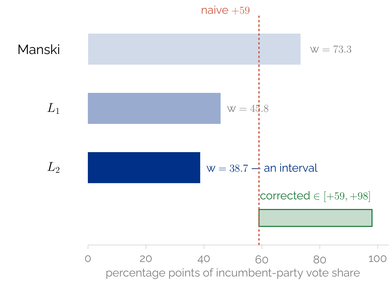
naive ATT +3.97 pp → spillover check $\hat{\psi} = +1.33$ pp (SE 0.50), $t = 2.66$ → weights uniform → sharp set point: $\rho\hat{\psi} = +0.53$ pp → corrected effect +4.50 pp

Pacification moved the left-bloc vote by 4.5 points, not 4.0: spillover into bordering neighborhoods hid 13% of the effect, and the correction is a point whose CI excludes zero.

Application B — Maricá: concentrated weights, an interval

Outcome: **incumbent-party vote share**. Synthetic control after the municipal basic income, close pre-treatment fit; donor municipalities within ~75 km exposed ($\rho = 0.24$).





naive ATT +59 pp → spillover check $\hat{\psi} = +8.9$ pp (SE 4.2), $t = 2.12$ → weights concentrated (SC) → sharp set $[0, 38.7]$ pp → corrected effect $[+59, +98]$ pp

The naive +59 points is a floor: exposed donors were lifted, deflating the synthetic counterfactual, so the true effect lies in $[+59, +98]$. A second outcome (vote concentration, $t = 3.0$) repeats the pattern: contamination follows the design, not the outcome.

Advice to applied researchers: Keep your estimator. Report the point estimate, the detected $\hat{\psi}$, the sharp de-contaminated set, and the sensitivity curve in r.