Causality

Randomization, machine learning, and everything in between

David Puelz
UT Austin

February 11, 2024

Causality?

What is the effect of kale intake on cholesterol?

Were COVID-19 lockdowns effective at reducing the spread?

Do workplace wellness programs make employees healthier?

Does increased policing reduce crime?

Causality?

Observational ... **data is just observed**

What is the effect of kale intake on cholesterol?

Were COVID-19 lockdowns effective at reducing the spread?

Do workplace wellness programs make employees healthier?

Does increased policing reduce crime?

Causality?

Randomized ... data is experimentally generated

What is the effect of kale intake on cholesterol?

Were COVID-19 lockdowns effective at reducing the spread?

Do workplace wellness programs make employees healthier?

Does increased policing reduce crime? $\leftarrow$

All share a common theme

Cause **Effect**

What is the effect of kale intake on cholesterol?

Were COVID-19 lockdowns effective at reducing the spread?

Do workplace wellness programs make employees healthier?

Does increased policing reduce crime?

Causal inference

My research program focuses on both applied and methodological problems in causal inference, ie, understanding **cause** and **effect** relationships. *I love teaching these topics, too!*

$\rightarrow$ Applied: Use existing empirical technique to investigate problems in social science, economics, and policy.

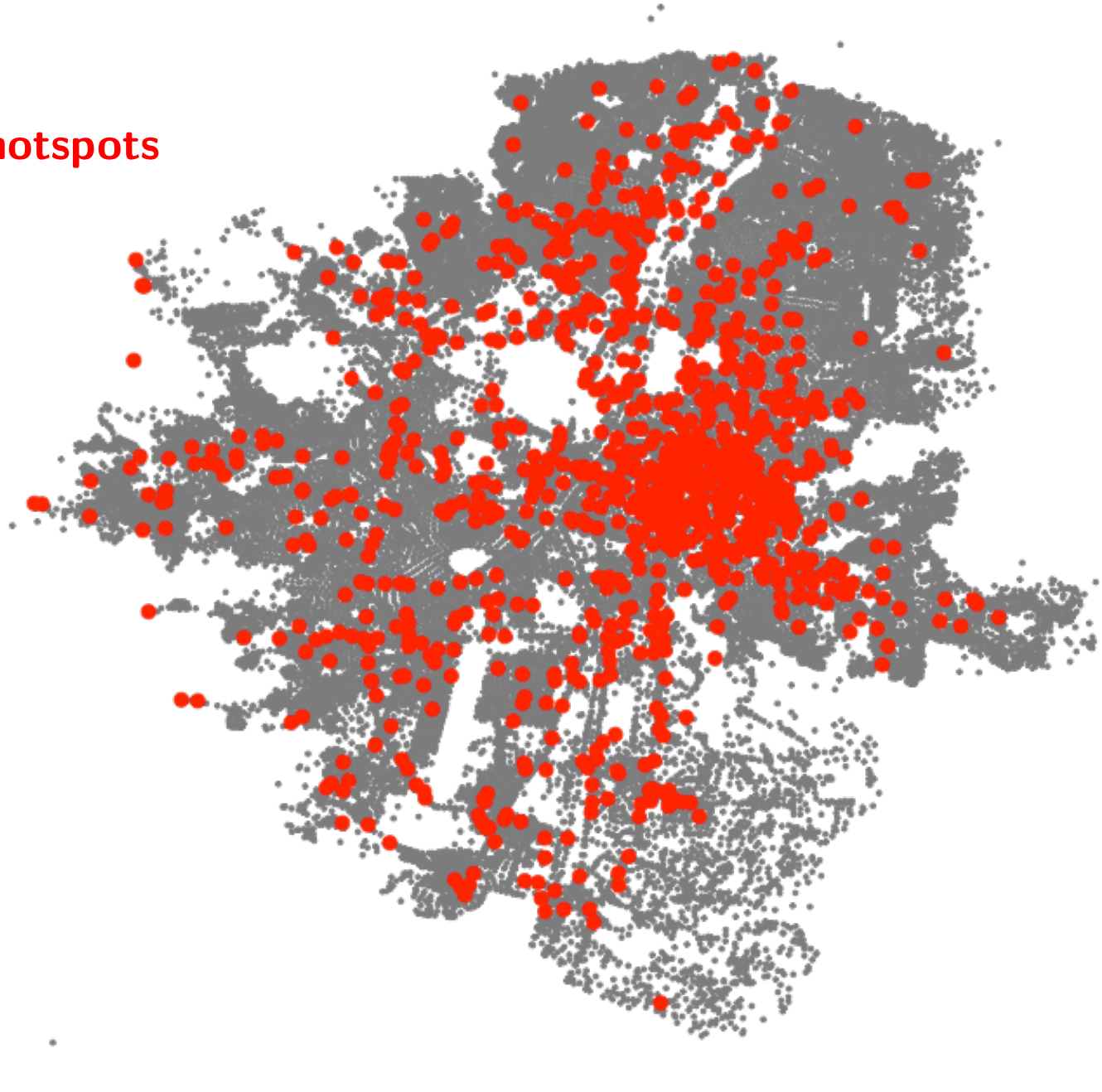
$\rightarrow$ Methodological: Design new statistical methods to understand "deeper" effects in systems, ie, peer/interference/network effects. Bring advanced ML methods to these messy problems.

Medellín, Colombia

Medellín, Colombia

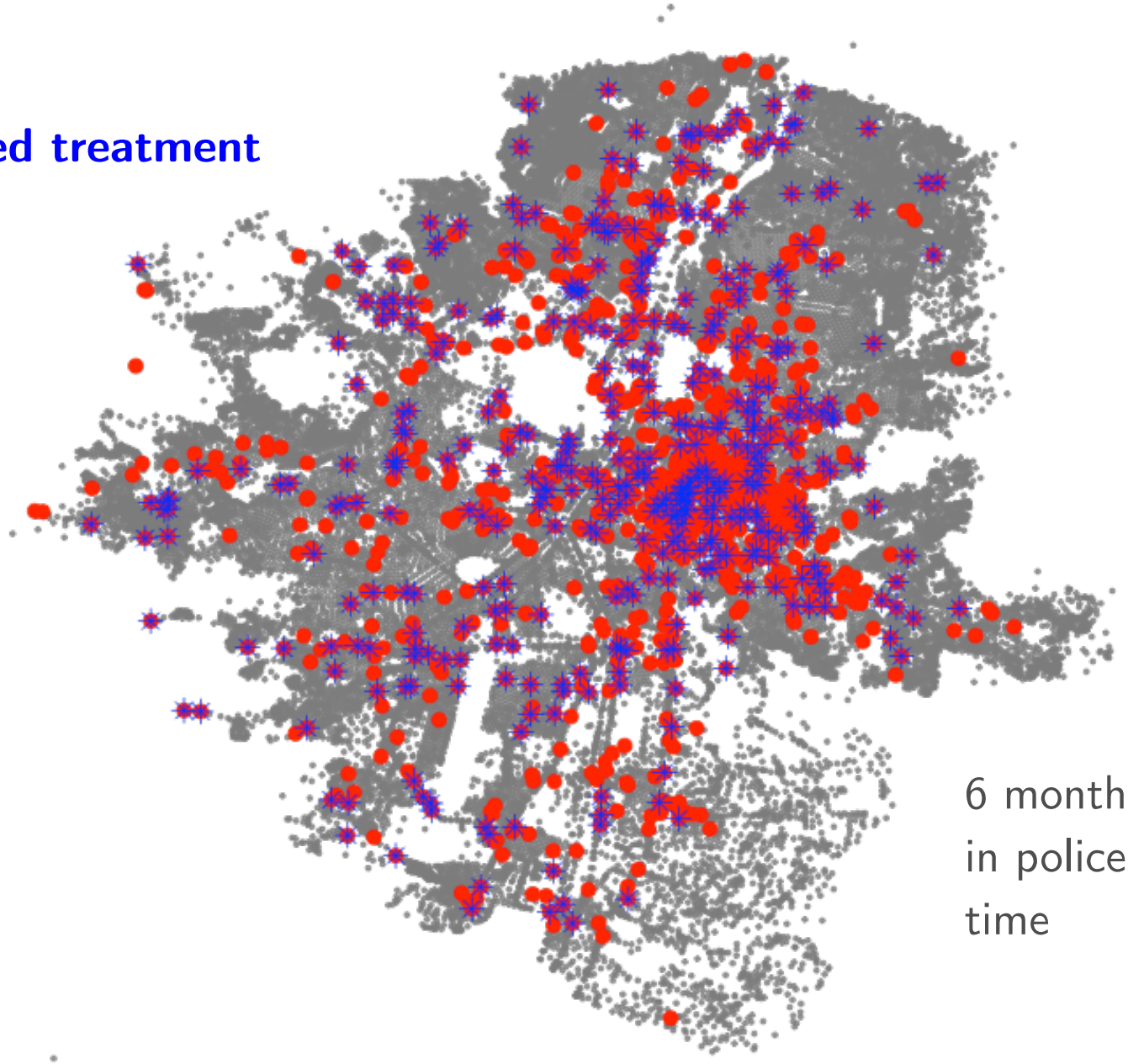
Medellín, Colombia

crime hotspots



Medellín, Colombia

observed treatment

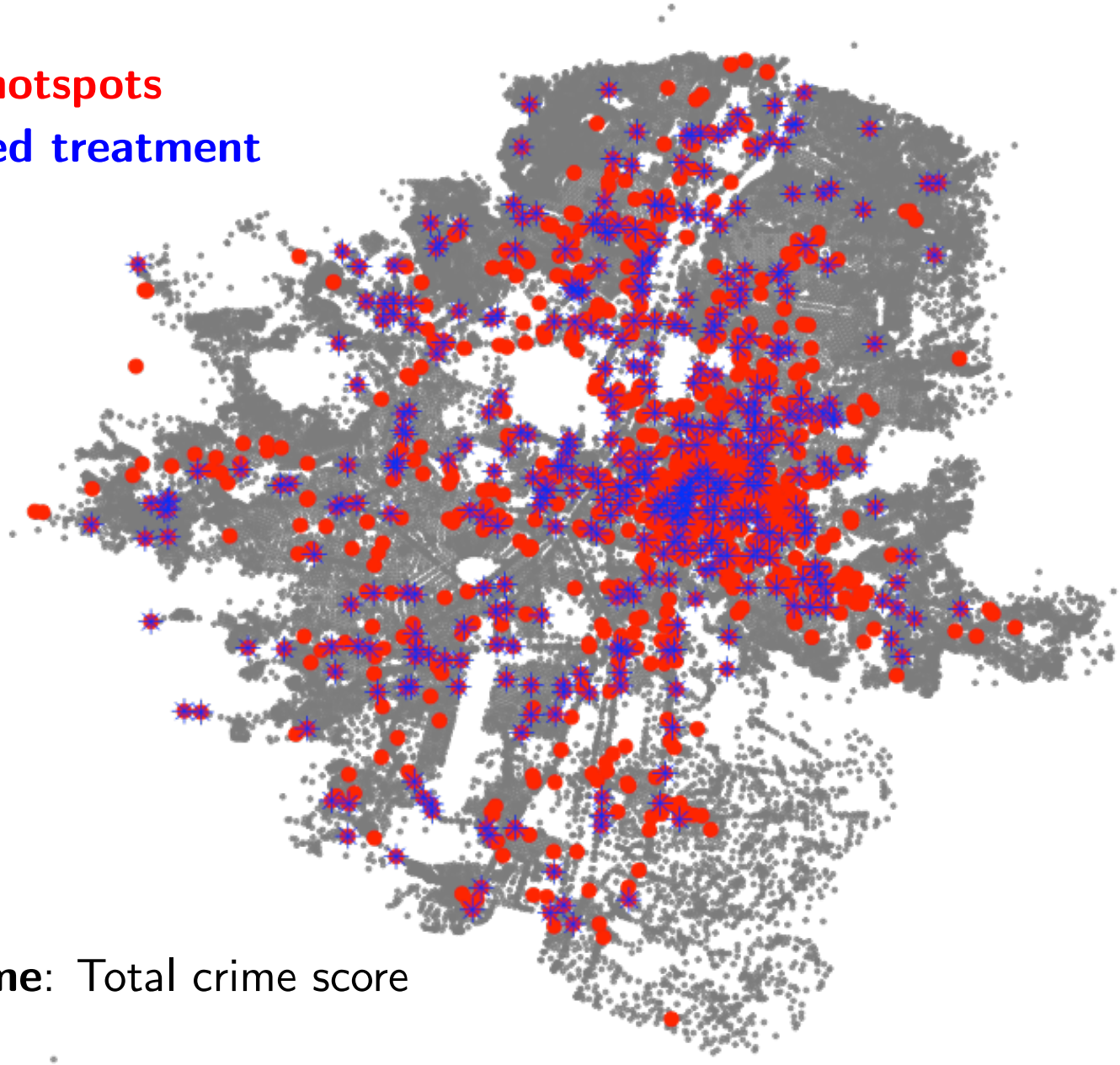


6 month increase
in police patrolling
time

Medellín, Colombia

crime hotspots

observed treatment

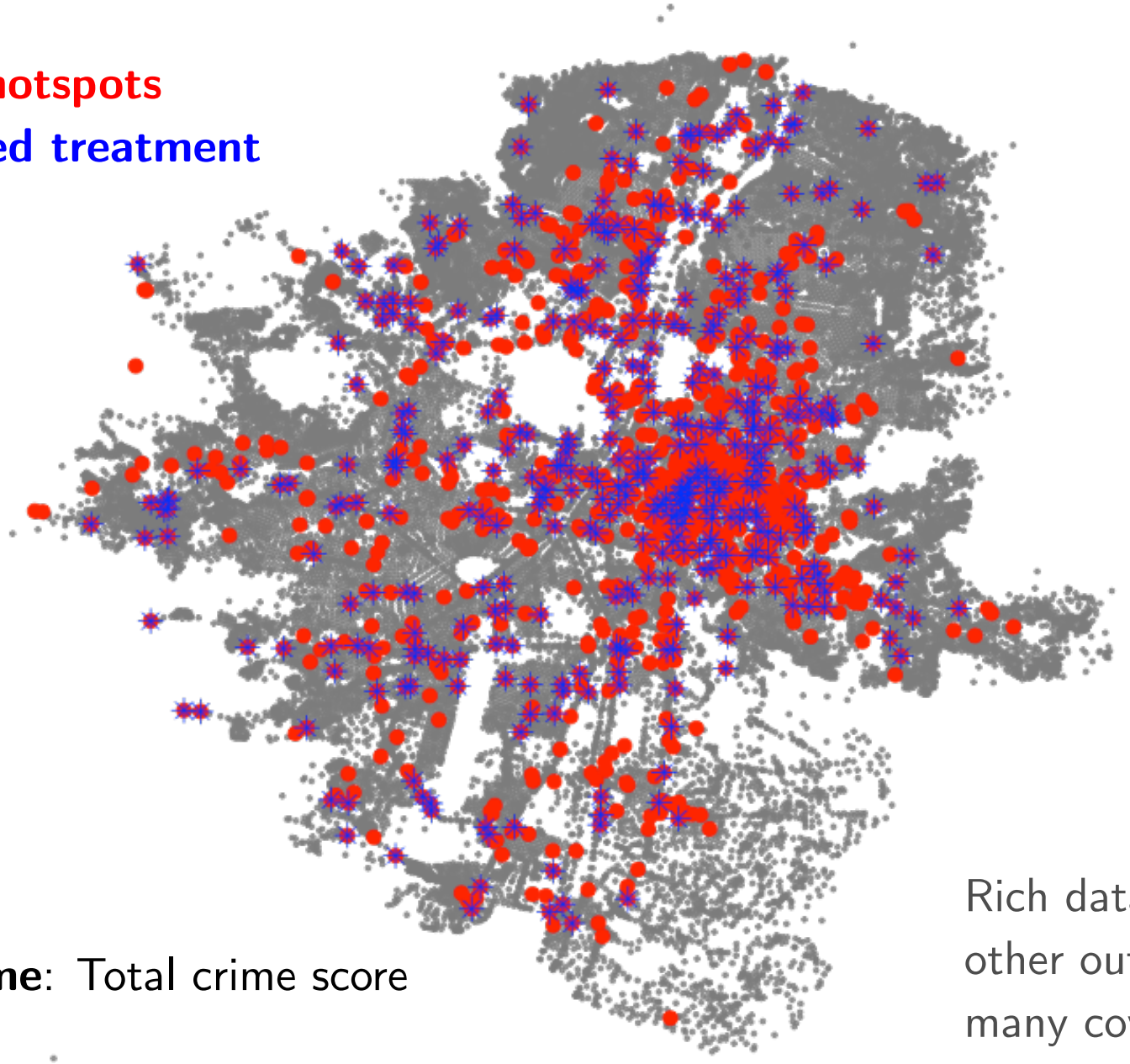


Outcome: Total crime score

Medellín, Colombia

crime hotspots

observed treatment



Outcome: Total crime score

Rich data set with other outcomes, many covariates

Questions we aim to answer

How does the intervention affect crime?

$\rightarrow$ direct effect?

$\rightarrow$ spillovers to adjacent streets?

Questions we aim to answer

How does the intervention affect crime?

→ direct effect?

→ spillovers to adjacent streets?

We will answer these through hypothesis testing.

We prefer a model-free approach, so we will use the randomization method of inference.

Notation

Define observed data:

$Z = (Z_1, \ldots, Z_N)$ as binary treatment assignment;

$Y = (Y_1, \ldots, Y_N)$ as vector of observed outcomes.

$\hookrightarrow$ *Z', Y': "counterfactuals", pr(Z'): design.*

The potential outcome of unit i under assignment z: $Y_i(z)$

i.e., total crime score

First, let's assume no interference: $Y_i(z)$ depends only on z_i.

$\hookrightarrow$ *Only two potential outcomes, $Y_i(0)$, $Y_i(1)$, for every i.*

Classical question of randomization inference

Does treatment have an effect at all?

$$\mathbf{H_0}: \quad Y_i(z_i = 0) = Y_i(z_i = 1) \text{ for every } i.$$

Key implication of $\mathbf{H_0}$ is that Y is fixed across all possible randomizations, a.k.a. treatment assignments.

Fisher randomization test (1935)

$\mathbf{H_0}$: $Y_i(Z_i = 0) = Y_i(Z_i = 1)$ for every i.

The procedure:

Choose test statistic $T = T(y, z)$ (e.g., difference in means).

1. $T_{\text{obs}} = T(Y, Z)$.
2. Sample $Z' \sim \text{pr}(Z')$, store $T_r = T(Y', Z') \overset{H_0}{=} T(Y, Z')$.
3. p-value $= \mathbb{E}[\mathbb{1}\{T_r \geq T_{\text{obs}}\}]$.

Advantages of Fisherian randomization

- Exact. The test is valid in finite samples.
- Minimal assumptions. No model for Y.
- Robust. Same answer under many transformations of Y.

Advantages of Fisherian randomization

- Exact. The test is valid in finite samples.
- Minimal assumptions. No model for Y.
- Robust. Same answer under many transformations of Y.

Disadvantages

- Can only test "strong," uninteresting nulls.
- Difficult to generalize out of sample.

Our goal is to use Fisherian randomization under interference.

No interference assumption is too strong ...

Assume: $Y_i(z)$ depends only on z_i (no interference)

$\hookrightarrow$ *not very realistic for our application*.

In reality, $Y_i(z)$ is **exposed** to (depends on) multiple parts of z.

No interference assumption is too strong ...

Assume: $Y_i(z)$ depends only on z_i (no interference)

$\hookrightarrow$ *not very realistic for our application*.

In reality, $Y_i(z)$ is **exposed** to (depends on) multiple parts of z.

$\hookrightarrow$ ***one extreme***: *no interference – 2 potential outcomes*

$\hookrightarrow$ ***another extreme***: *max interference –* 2^N *potential outcomes!*

No interference assumption is too strong ...

Assume: $Y_i(z)$ depends only on z_i (no interference)

$\hookrightarrow$ *not very realistic for our application*.

In reality, $Y_i(z)$ is **exposed** to (depends on) multiple parts of z.

$\hookrightarrow$ ***one extreme***: *no interference – 2 potential outcomes*

$\hookrightarrow$ ***another extreme***: *max interference –* 2^N *potential outcomes!*

To make progress, we express more interesting hypotheses with **exposure functions**. These functions are just lower dimensional summaries of the treatment vector.

Harder question: Is there a short-range spillover effect?

$\mathbf{H_0}$: $Y_i(Z) = Y_i(Z')$ for every i, Z, Z',

such that $f_i(Z), f_i(Z') \in \{\texttt{short}, \texttt{control}\}$.

$$f_i(Z) := \begin{cases} \texttt{short} & Z_i = 0, \text{dist}_i < 125\text{m} \\ \texttt{control} & Z_i = 0, \text{dist}_i > 500\text{m} \\ \texttt{neither} & \text{else} \end{cases}$$

$\text{dist}_i :=$ distance to closest treated street.

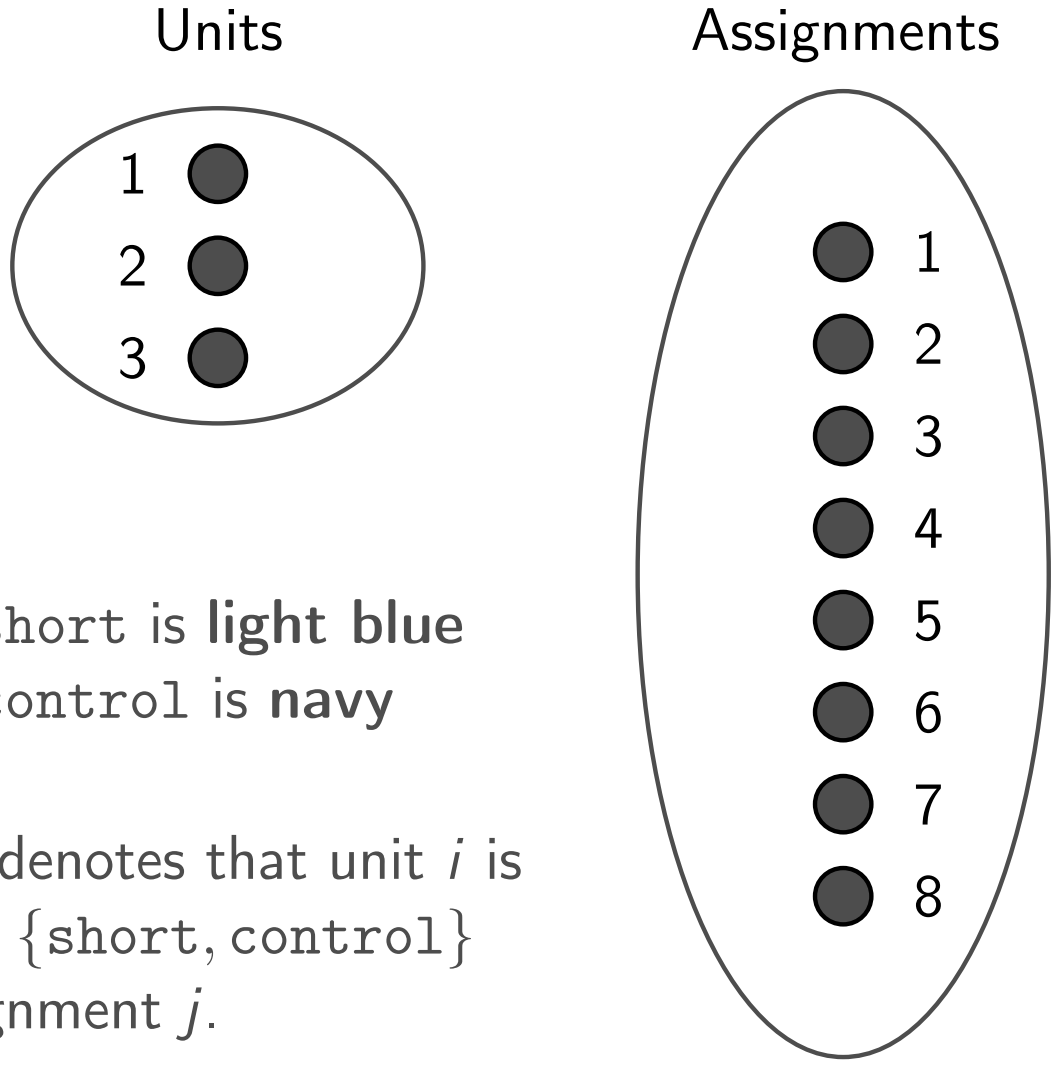
Testing $Y_i(\texttt{short}) = Y_i(\texttt{control}), \forall i$

Given a null hypothesis and assignment from pr(Z), we know which units are exposed to `short` or `control` using $f_i(\cdot)$.

This is a binary relationship!

How can we visualize?

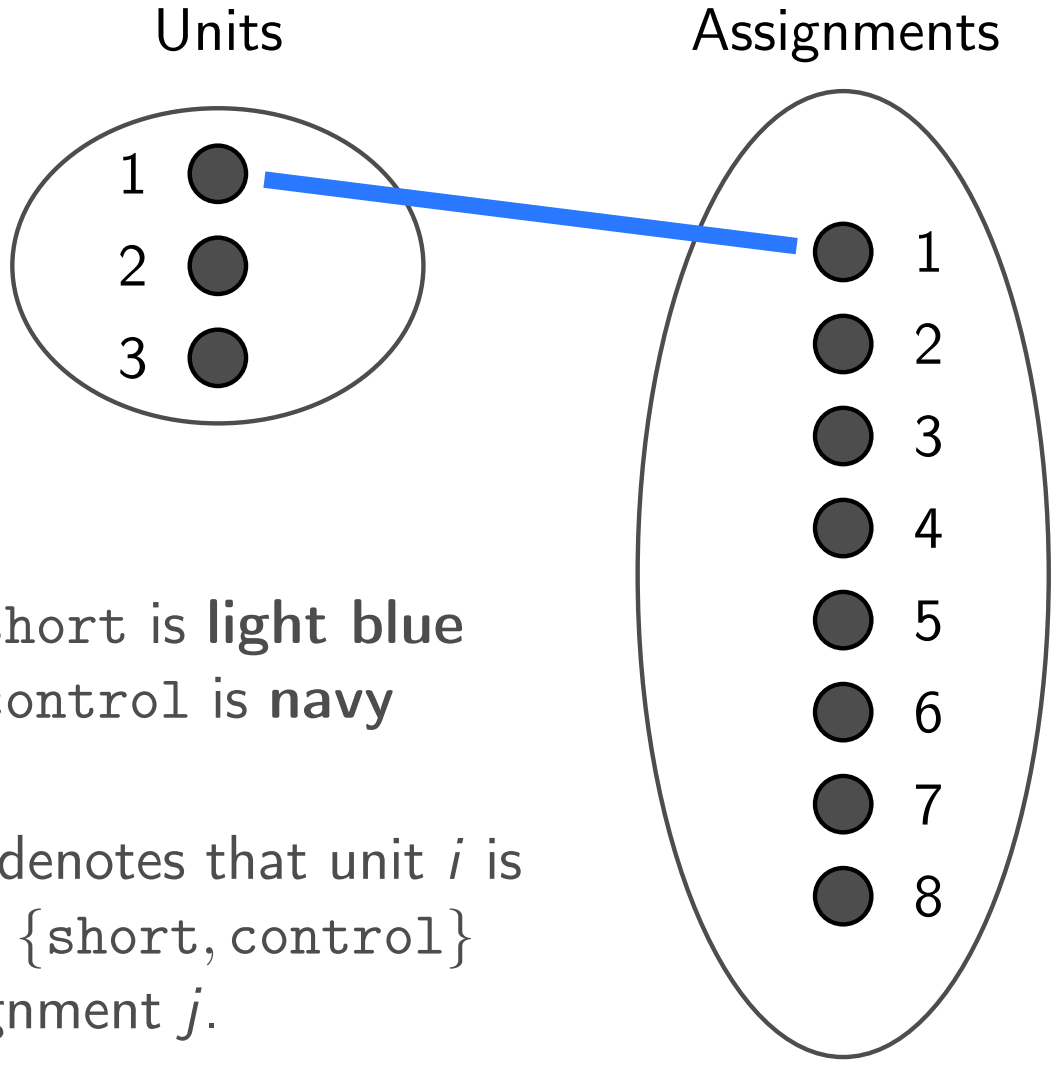
Our main contribution: The null exposure graph



Exposure `short` is **light blue**
Exposure `control` is **navy**

edge (i, j) denotes that unit i is exposed to $\{\texttt{short}, \texttt{control}\}$ under assignment j.

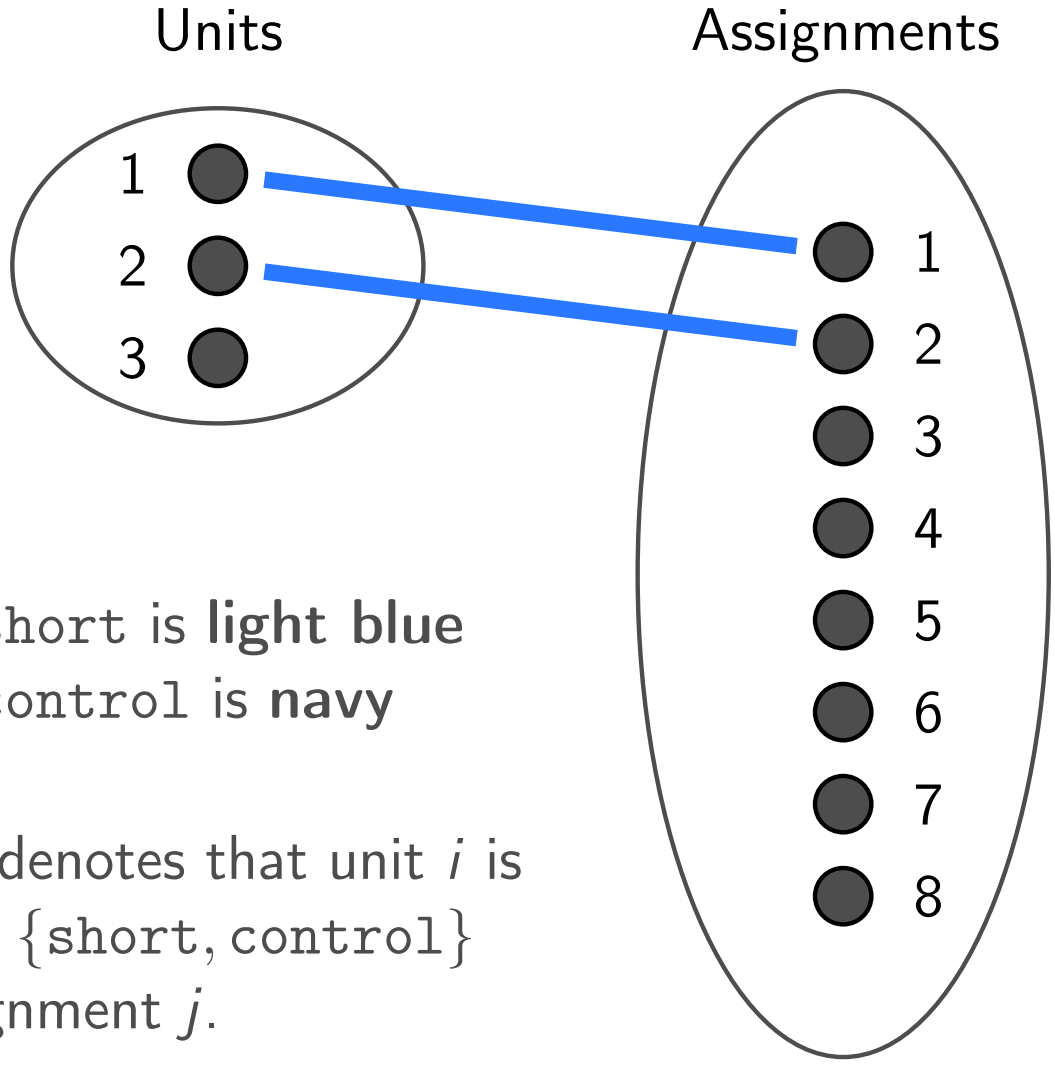
Our main contribution: The null exposure graph



Exposure `short` is **light blue**
Exposure `control` is **navy**

edge (i, j) denotes that unit i is exposed to $\{\texttt{short}, \texttt{control}\}$ under assignment j.

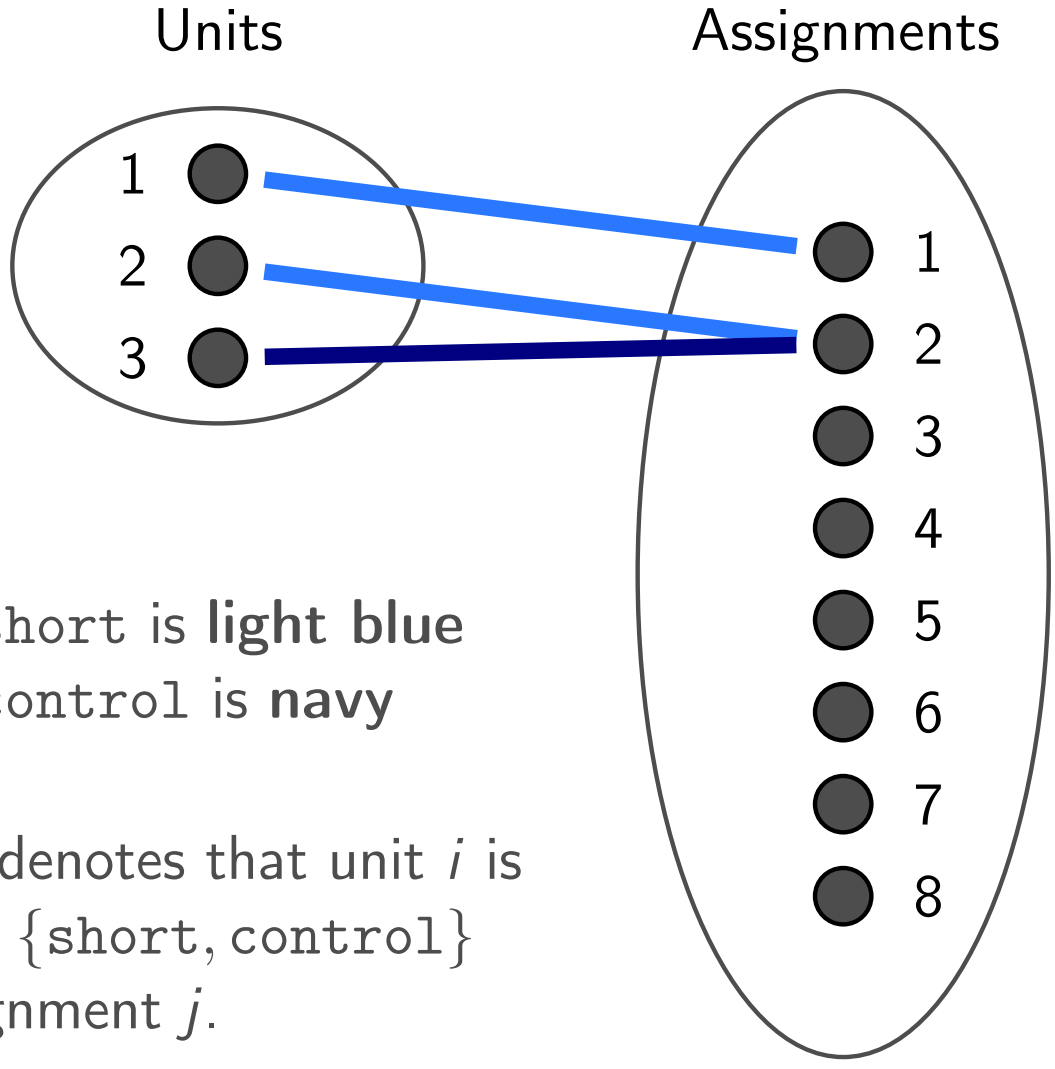
Our main contribution: The null exposure graph



Exposure `short` is **light blue**
Exposure `control` is **navy**

edge (i, j) denotes that unit i is exposed to $\{\texttt{short}, \texttt{control}\}$ under assignment j.

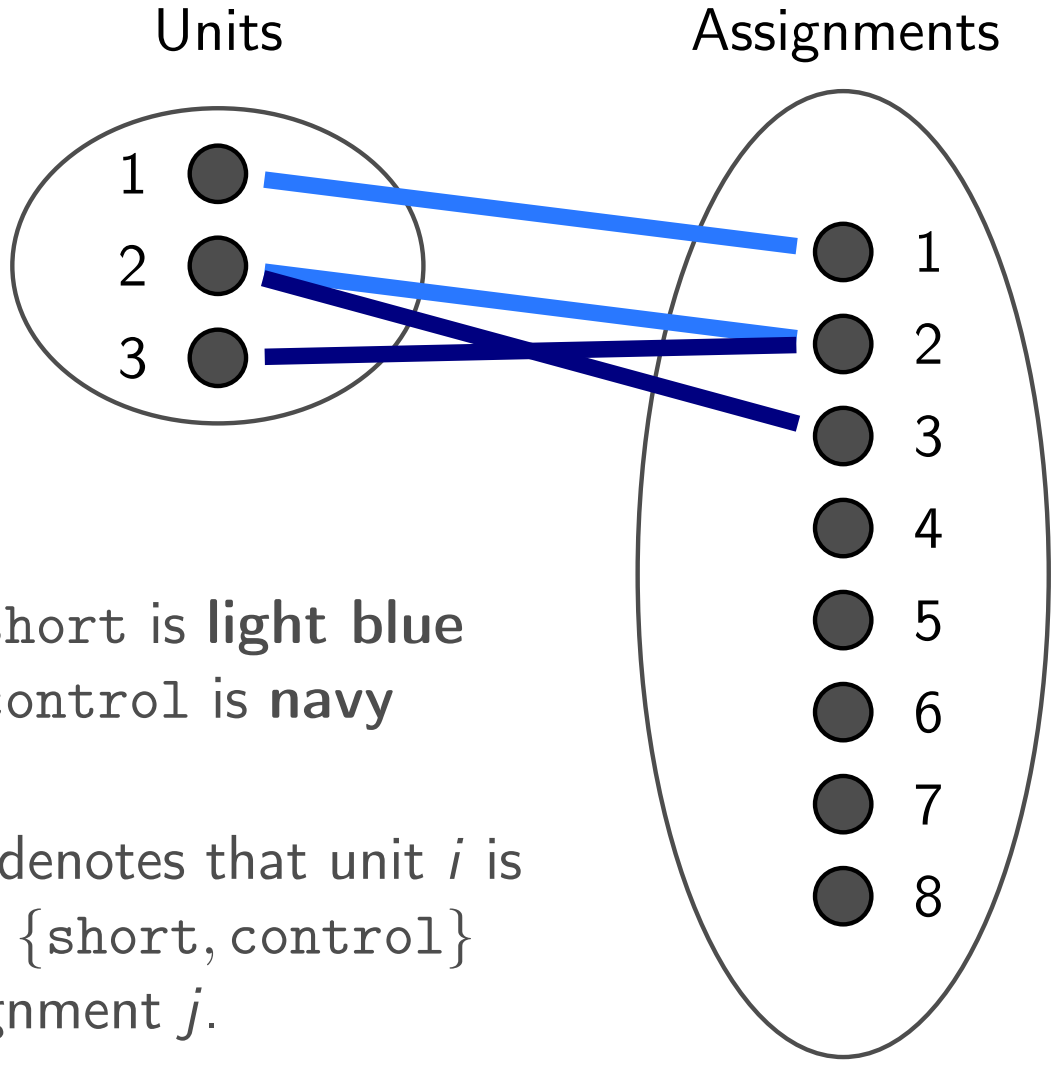
Our main contribution: The null exposure graph



Exposure `short` is **light blue**
Exposure `control` is **navy**

edge (i, j) denotes that unit i is exposed to $\{\texttt{short}, \texttt{control}\}$ under assignment j.

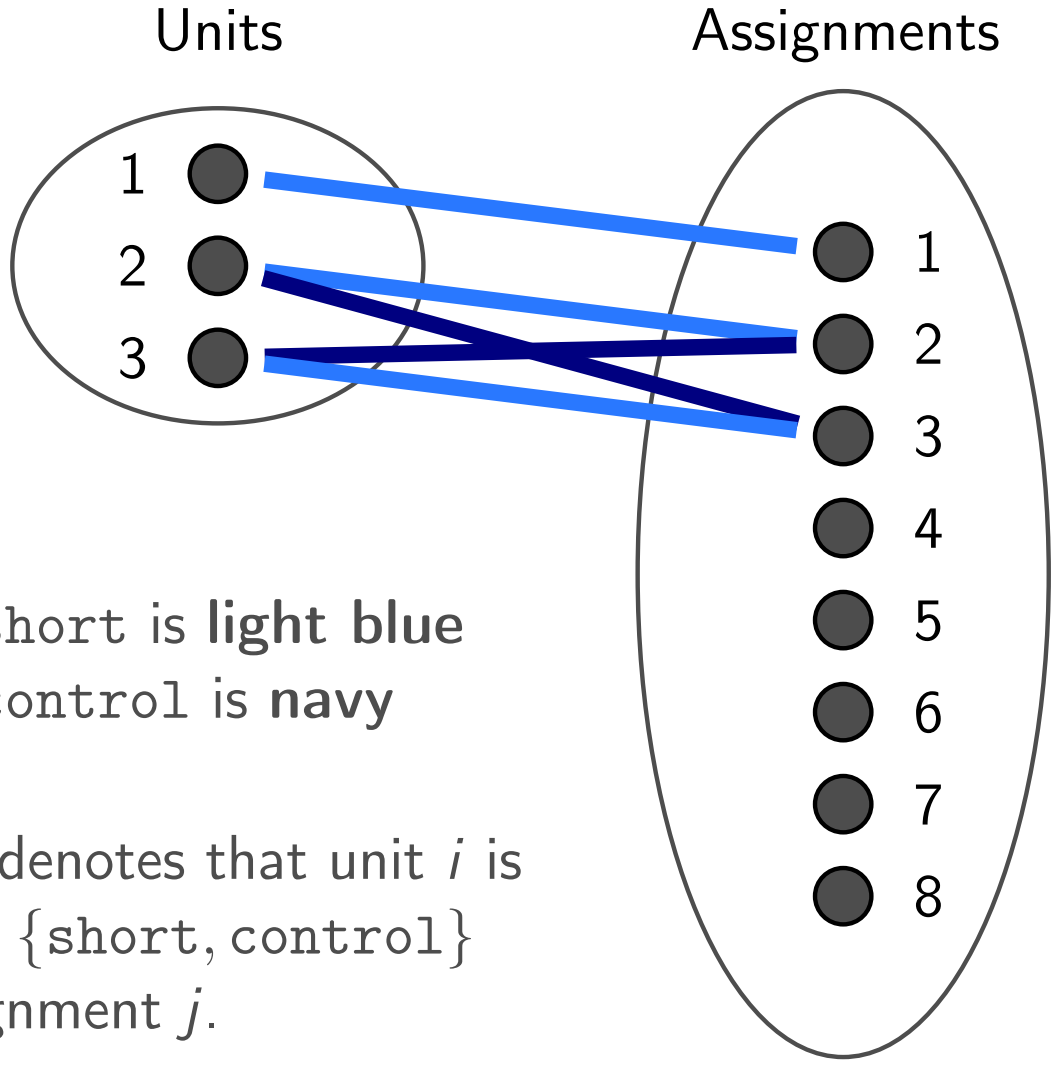
Our main contribution: The null exposure graph



Exposure `short` is **light blue**
Exposure `control` is **navy**

edge (i, j) denotes that unit i is exposed to $\{\texttt{short}, \texttt{control}\}$ under assignment j.

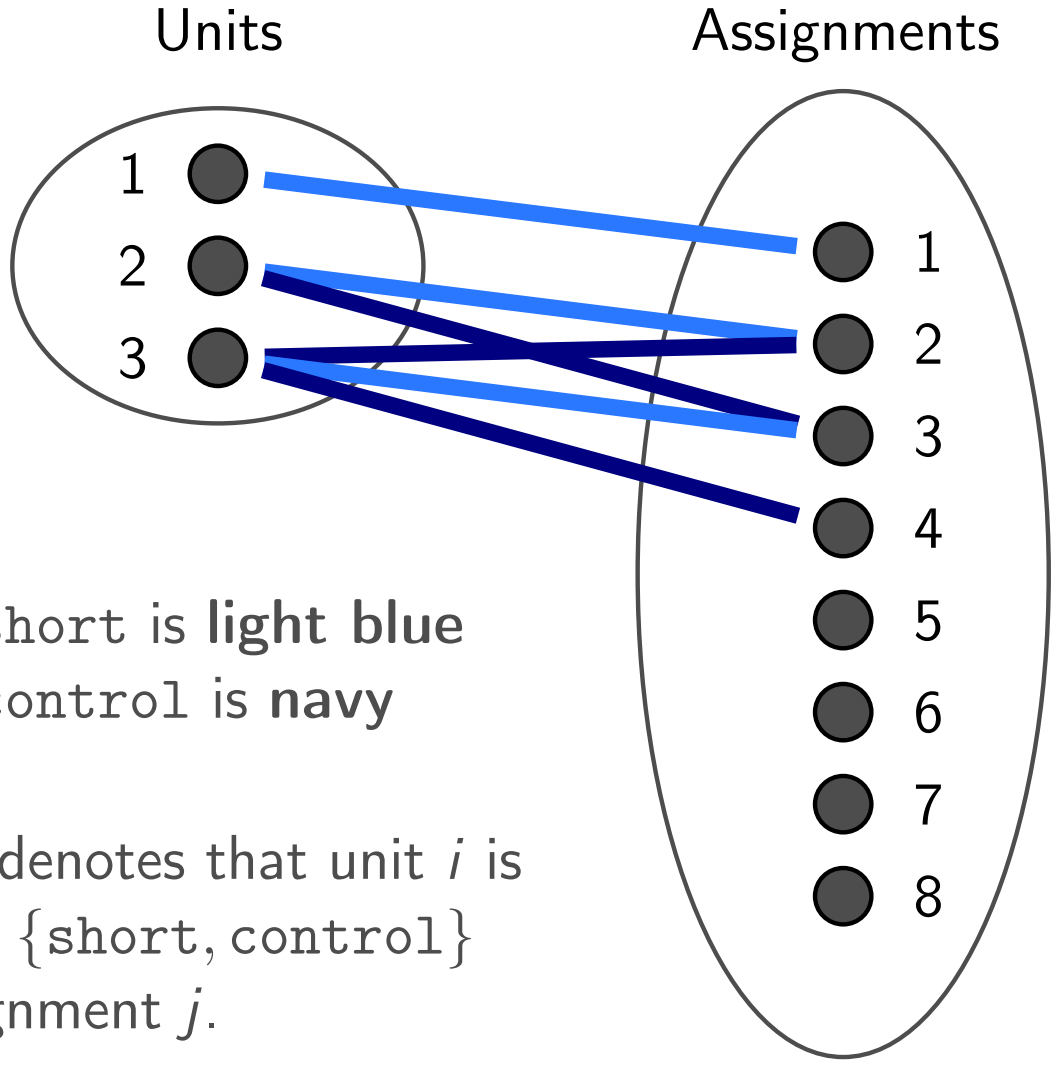
Our main contribution: The null exposure graph



Exposure `short` is **light blue**
Exposure `control` is **navy**

edge (i, j) denotes that unit i is exposed to $\{\texttt{short}, \texttt{control}\}$ under assignment j.

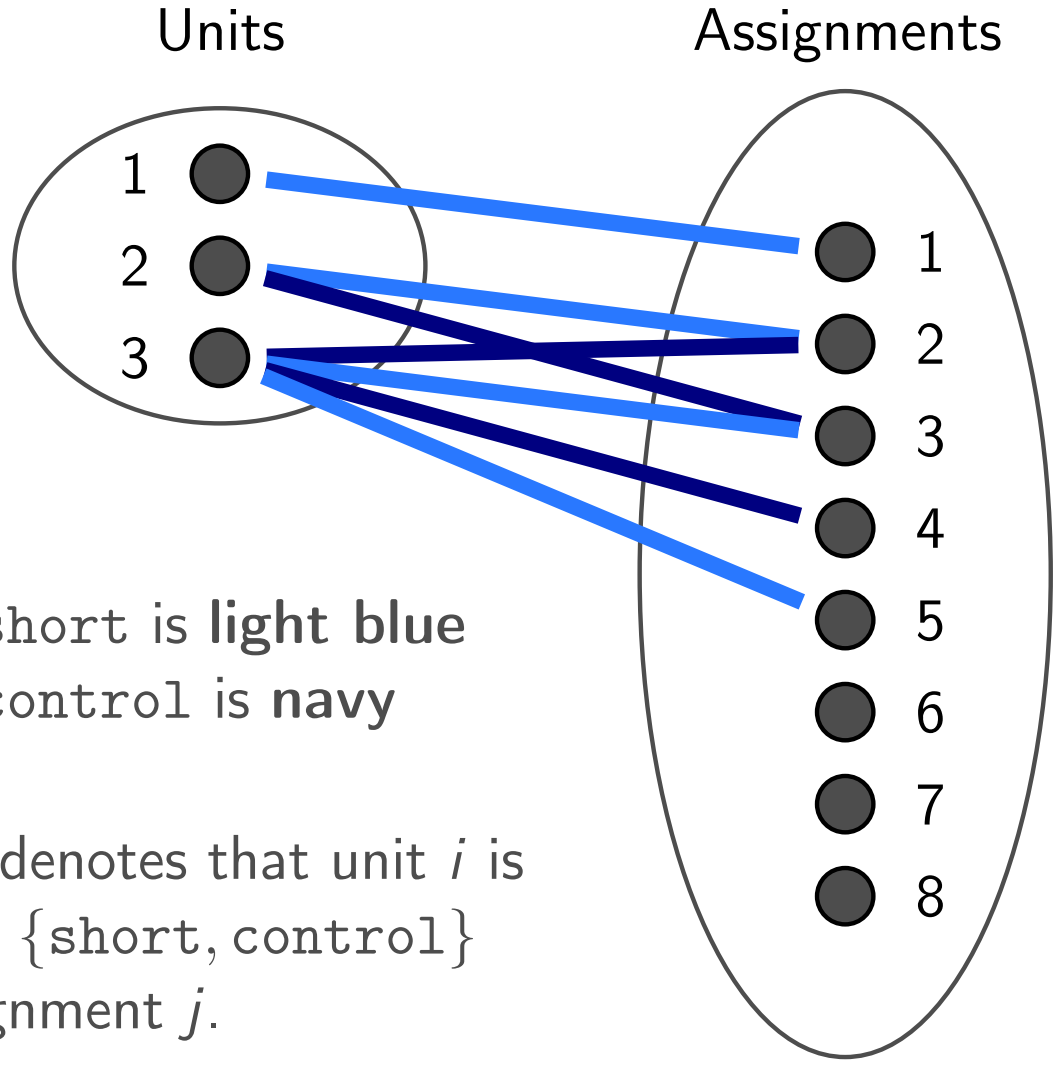
Our main contribution: The null exposure graph



Exposure `short` is **light blue**
Exposure `control` is **navy**

edge (i, j) denotes that unit i is exposed to $\{\texttt{short}, \texttt{control}\}$ under assignment j.

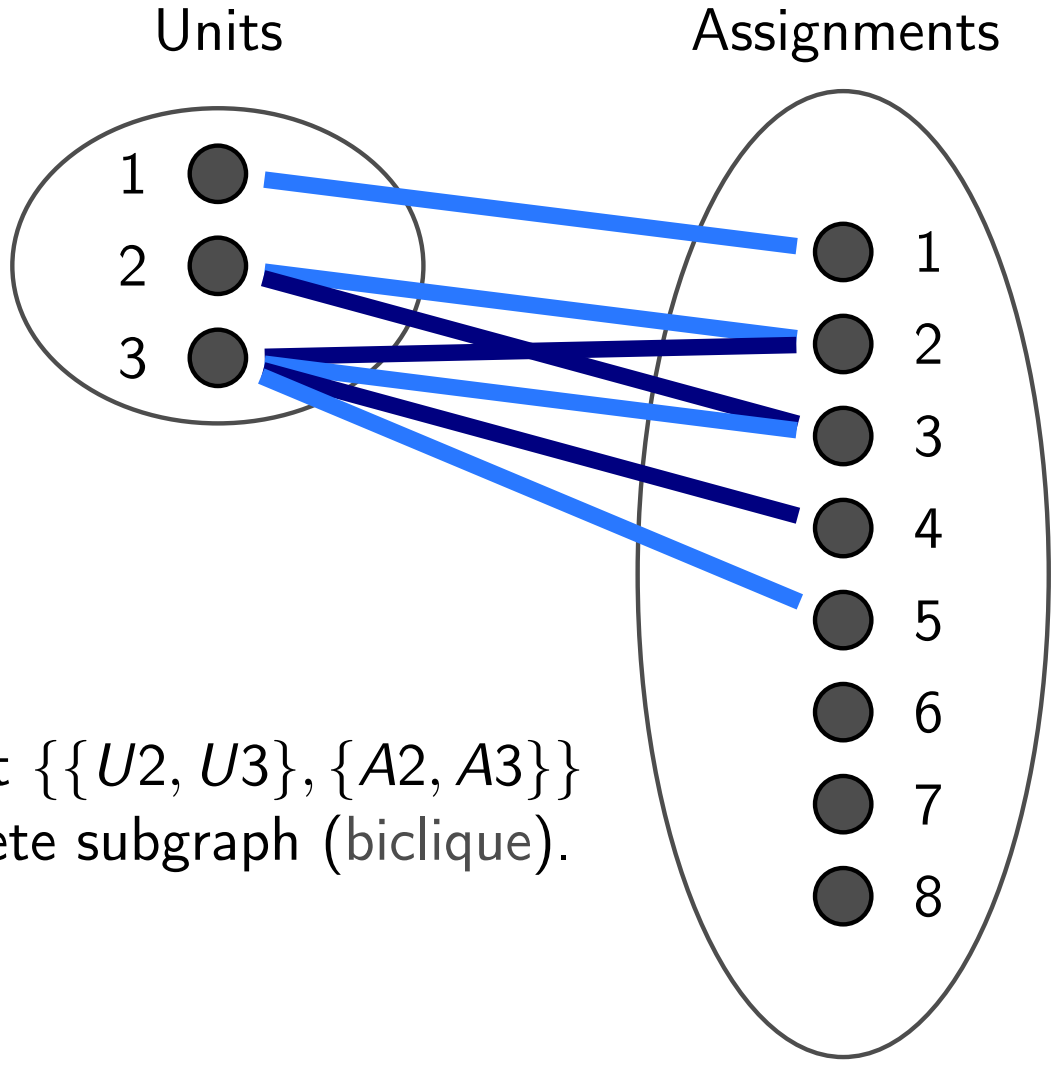
Our main contribution: The null exposure graph



Exposure `short` is **light blue**
Exposure `control` is **navy**

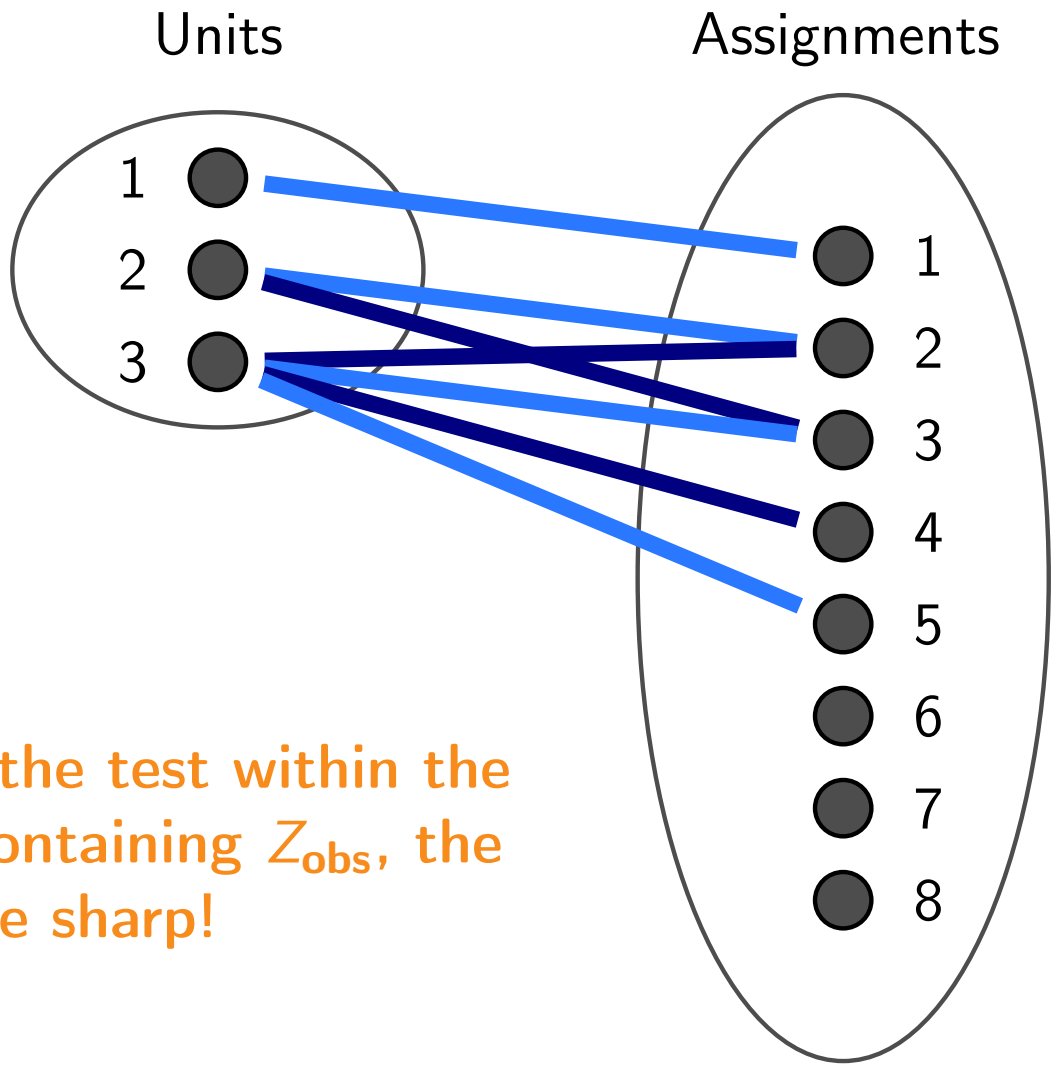
edge (i, j) denotes that unit i is exposed to $\{\texttt{short}, \texttt{control}\}$ under assignment j.

Our main contribution: The null exposure graph



Notice that $\{\{U2, U3\}, \{A2, A3\}\}$ is a complete subgraph (biclique).

Our main contribution: The null exposure graph



If we run the test within the biclique containing Z_{obs}, the null will be sharp!

Conditioning in this way gives a valid method!

Clique test statistics: $T_C = T(Y_C, Z_C)$

*T is defined only in C by **condition** step in method

For every Z, Z', we need to show $T(Y', Z') \stackrel{d}{=} T(Y, Z) \mid C$

Proof:

$$T(Y', Z') \stackrel{*}{=} T(Y'_C, Z'_C) \stackrel{H_0}{=} T(Y_C, Z'_C) \stackrel{d}{=} T(Y_C, Z_C) \stackrel{*}{=} T(Y, Z)$$

(a distribution)

Related work

We can also use our framework to describe related work:

- Aronow (2012) and Athey et al (2018) effectively propose to randomly sample *focal units* on one side, and then find the *maximum induced* clique to condition on.

  ↪ *General procedure but the random selection of focals does not exploit the problem structure — Loss of power.*

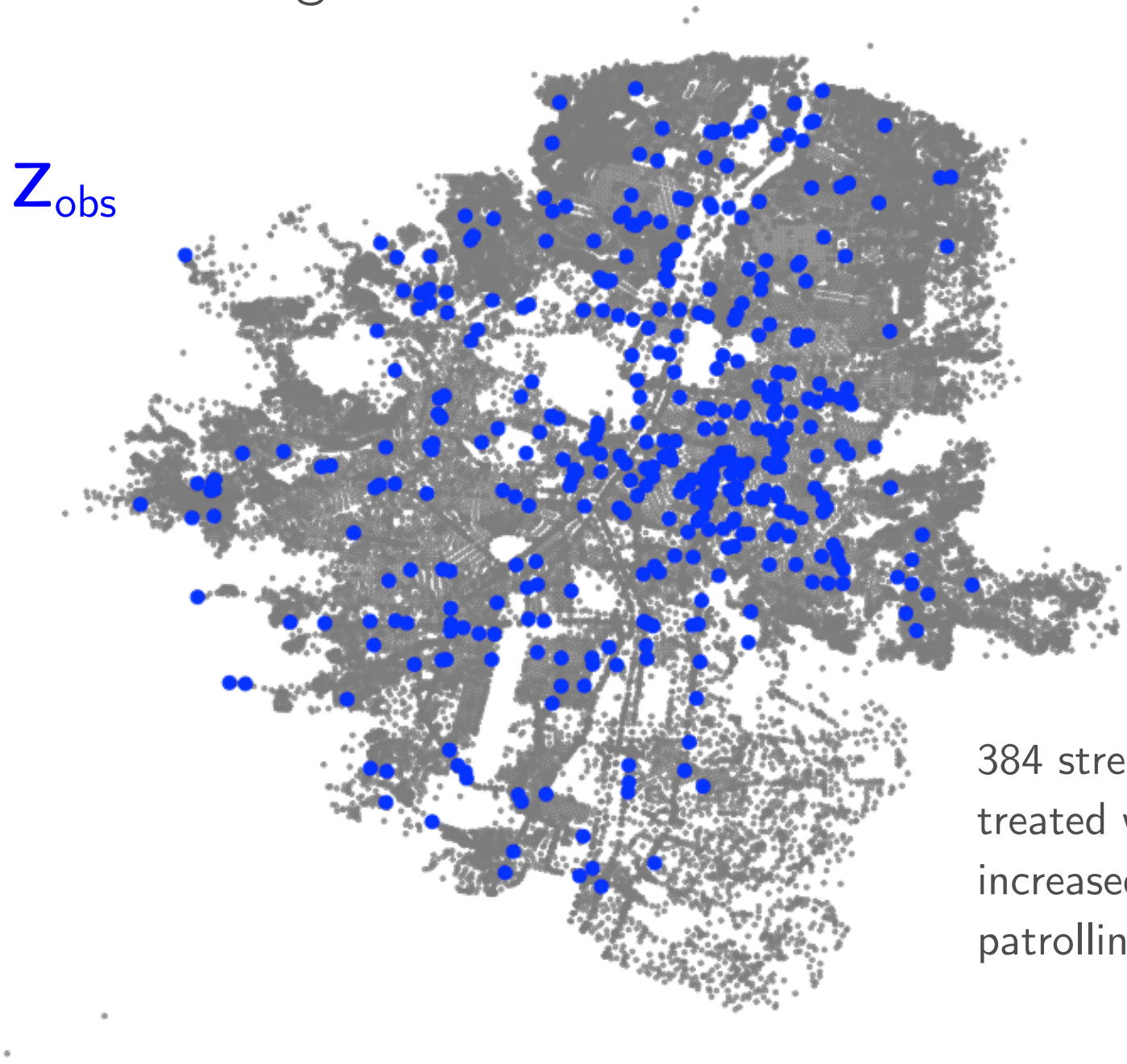
- Basse et al (2019) develop a clique decomposition that provably leads to permutation test under a setting with clustered interference.

  ↪ *Case-by-case analysis — Cannot generalize.*

Returning to the map

The observed assignment

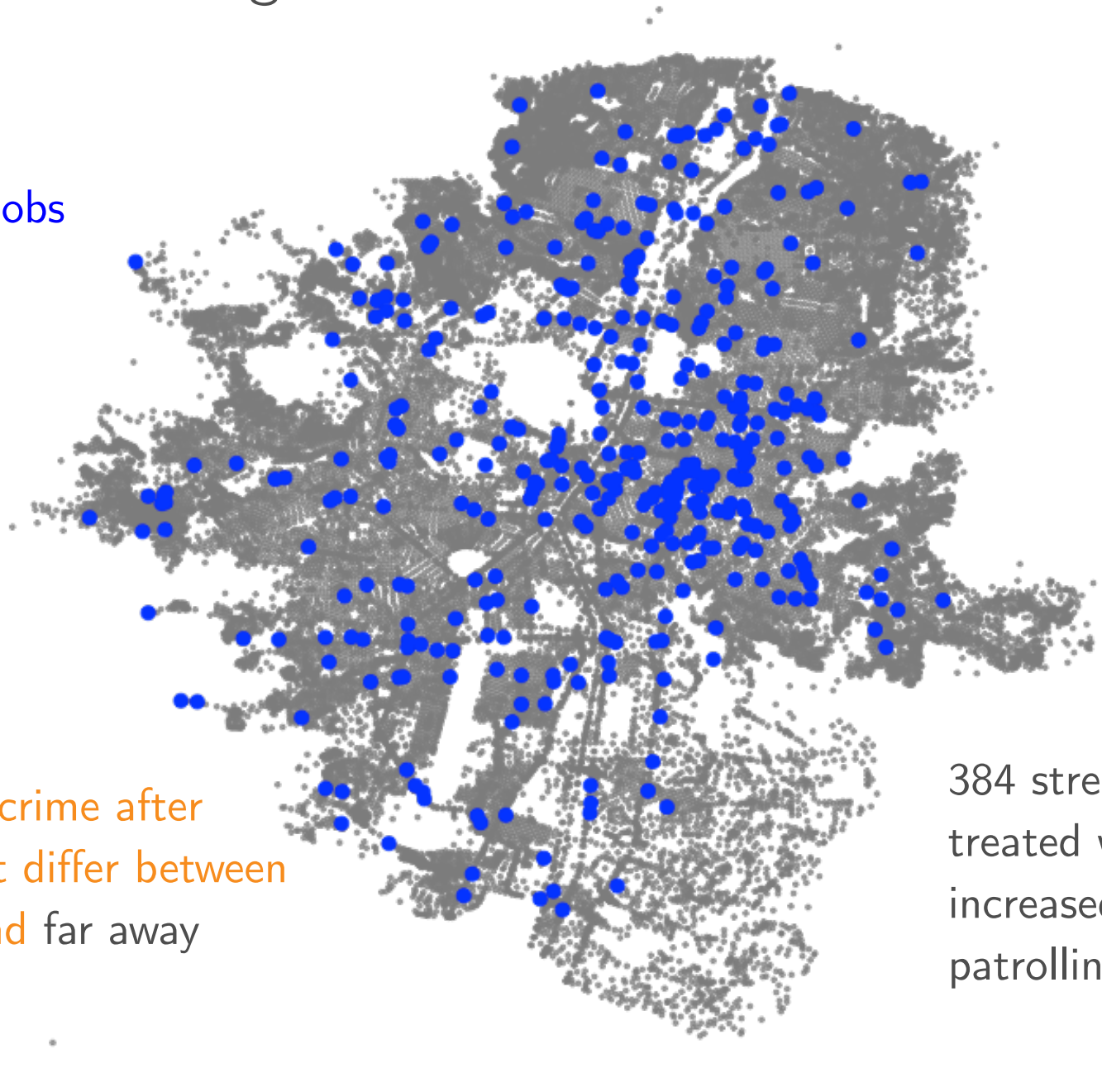
Z_{obs}



384 streets are treated with increased police patrolling

The observed assignment

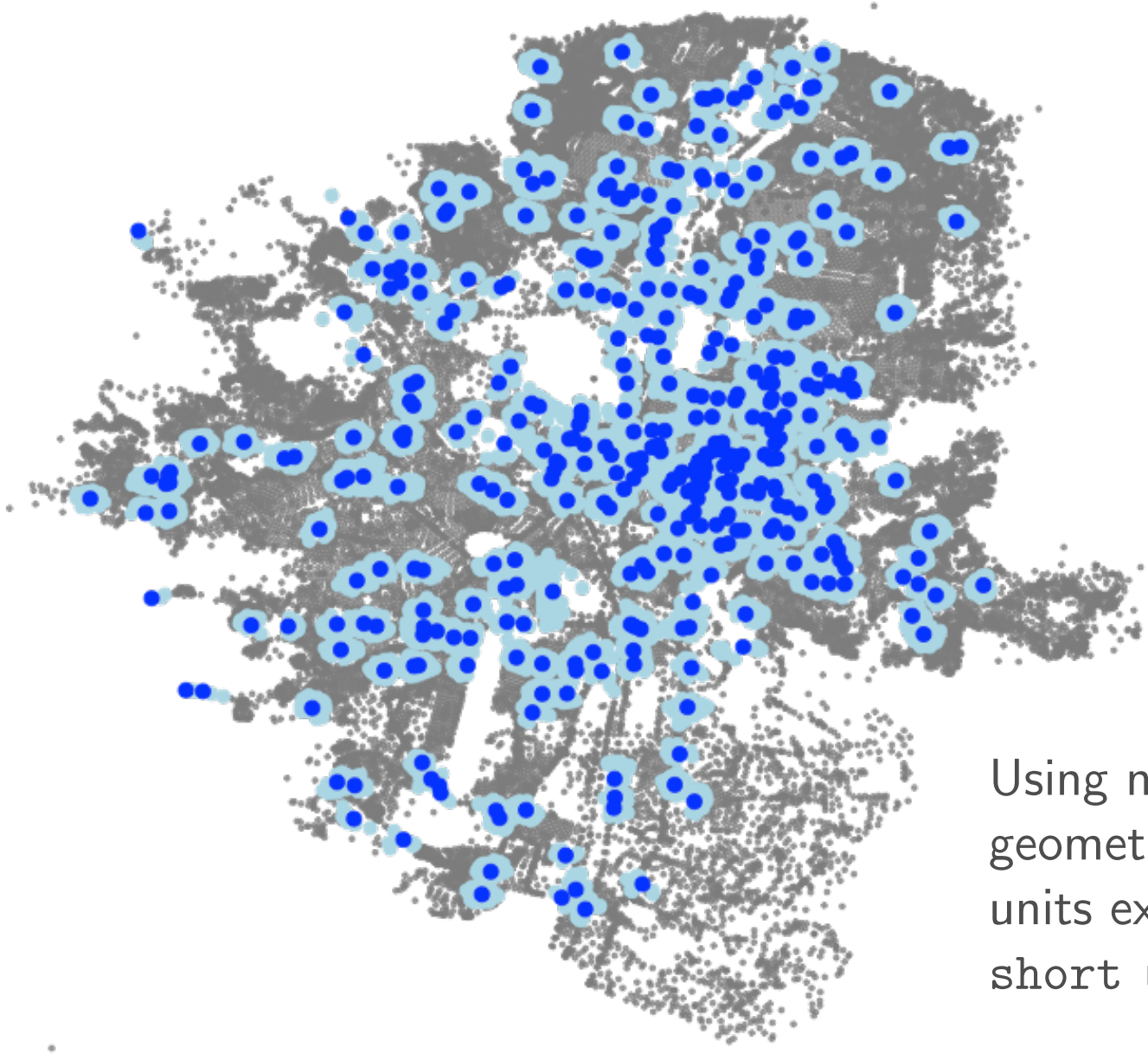
Z_{obs}



Q: Does crime after treatment differ between nearby and far away streets?

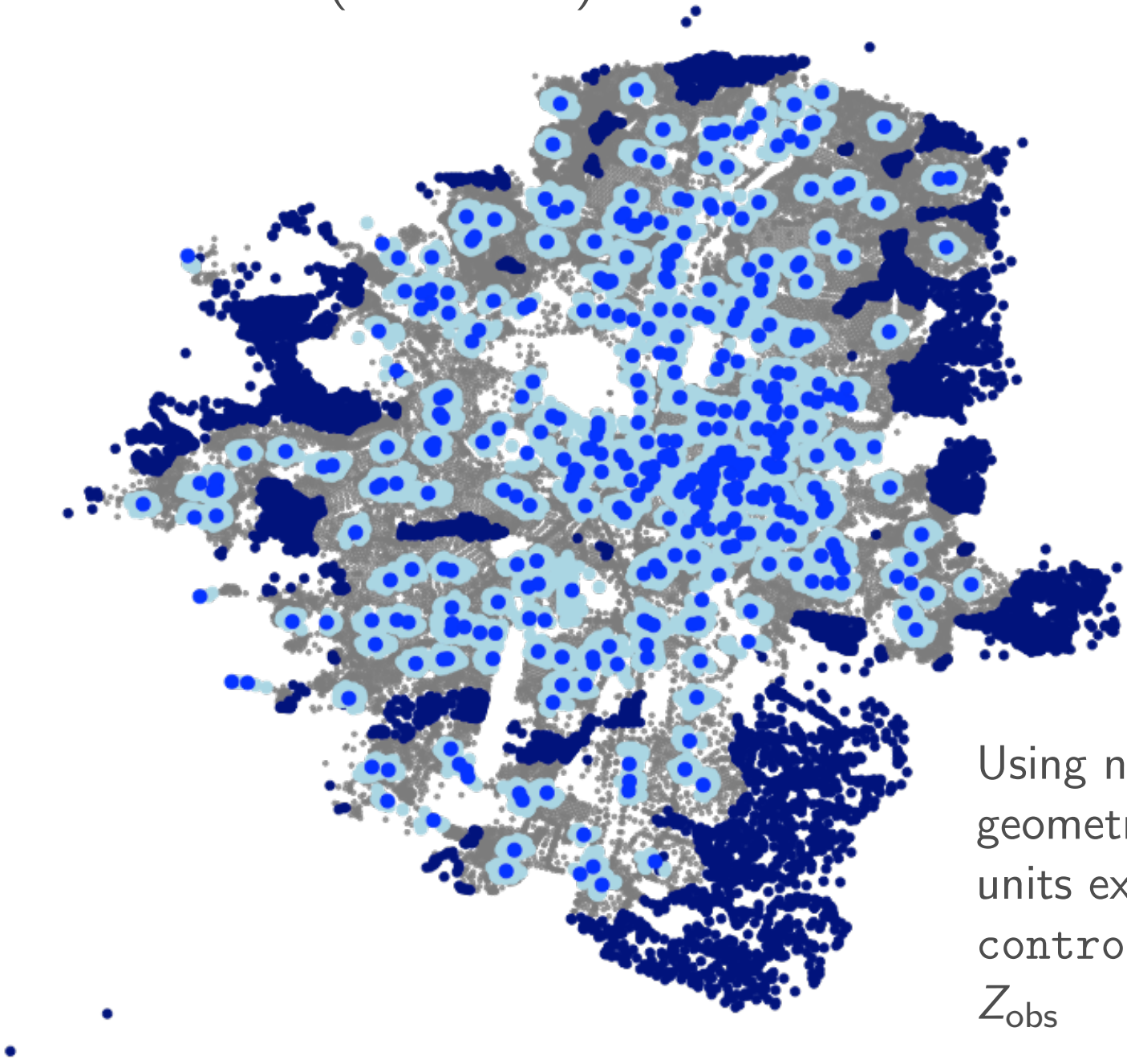
384 streets are treated with increased police patrolling

Short-range spillover units (`short`)



Using network geometry, color units exposed to `short` under Z_{obs}

Pure control units (`control`)



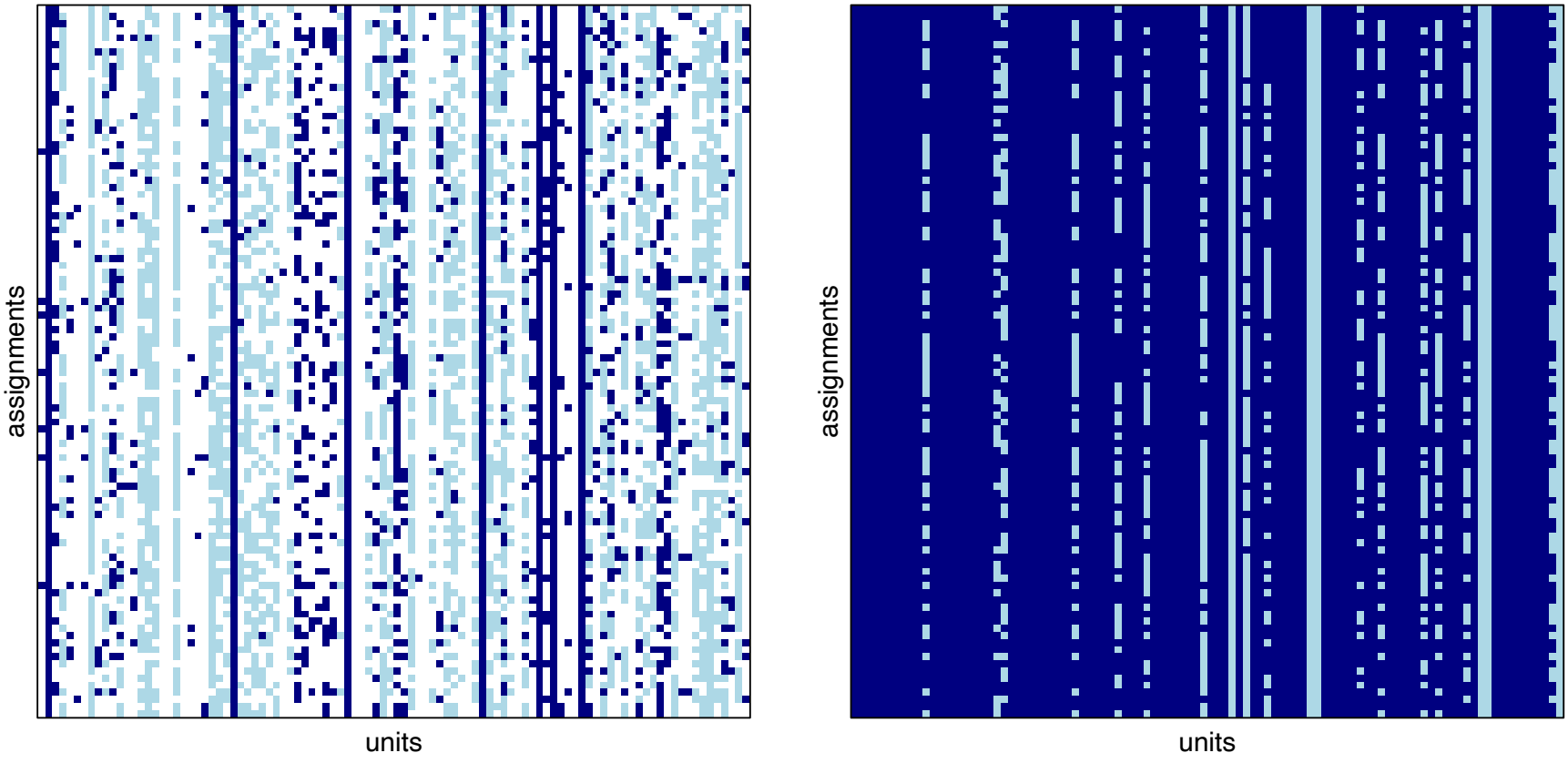
Using network geometry, color units exposed to `control` under Z_{obs}

We can remake these pictures for every assignment Z drawn from $\mathrm{pr}(Z)$ ...

We can remake these pictures for every assignment Z drawn from pr(Z) ...

$\rightarrow$ The output is our null exposure graph!

Null exposure graph (left) and biclique (right)



A statistical problem becomes a computational problem. Our goal is to find a large biclique (to minimize data loss).

Biclique-based randomization test

$\rightarrow$ A null exposure graph uniquely defined given $\mathbf{H_0}$.

$\rightarrow$ A test statistic $T = T(y, z)$.

1. **Decompose:** Compute biclique decomposition of null exposure graph. Pick out biclique with Z_{obs}, call it C.

2. **Condition:** Compute test statistic values with units and assignments only in C, T_C and T_{obs}.

3. **Summarize:** p-value $= \mathbb{E}\left[\mathbb{1}\{T_C \geq T_{\text{obs}}\} \mid C\right]$.

   $\hookrightarrow$ *Here, we sample with respect to*

$$P(Z' \mid C) \propto \underbrace{P(C \mid Z')}_{\text{conditioning mech.}} \cdot \underbrace{P(Z')}_{\text{design}}$$

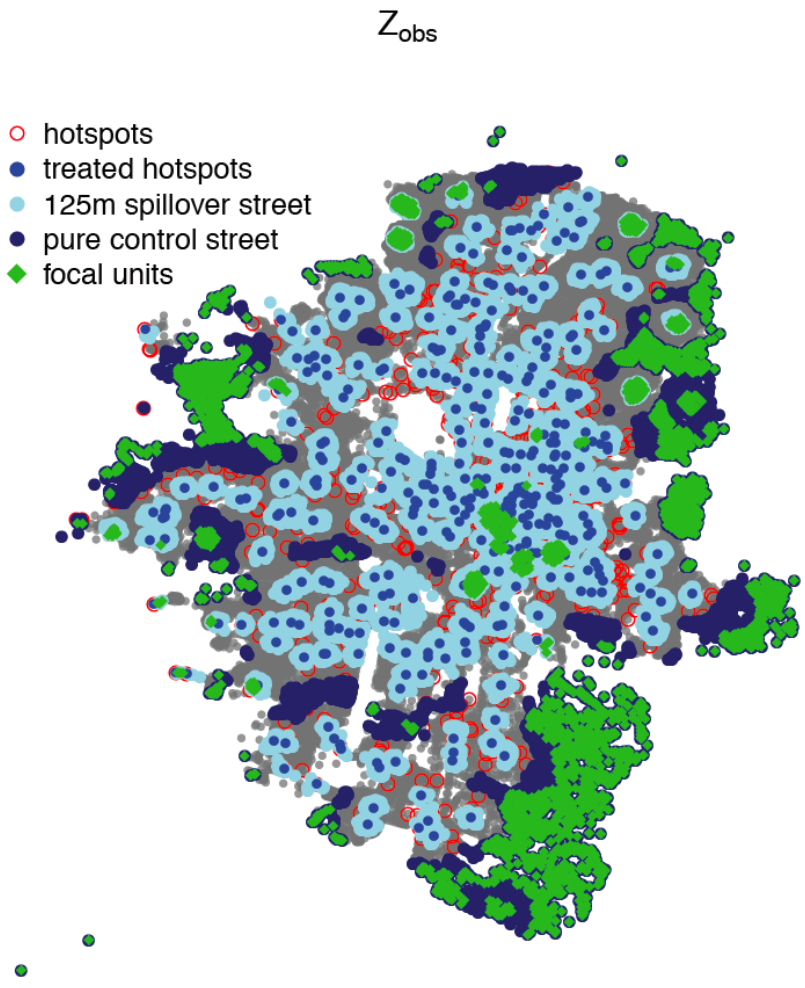
Spatial interference: Medellín data

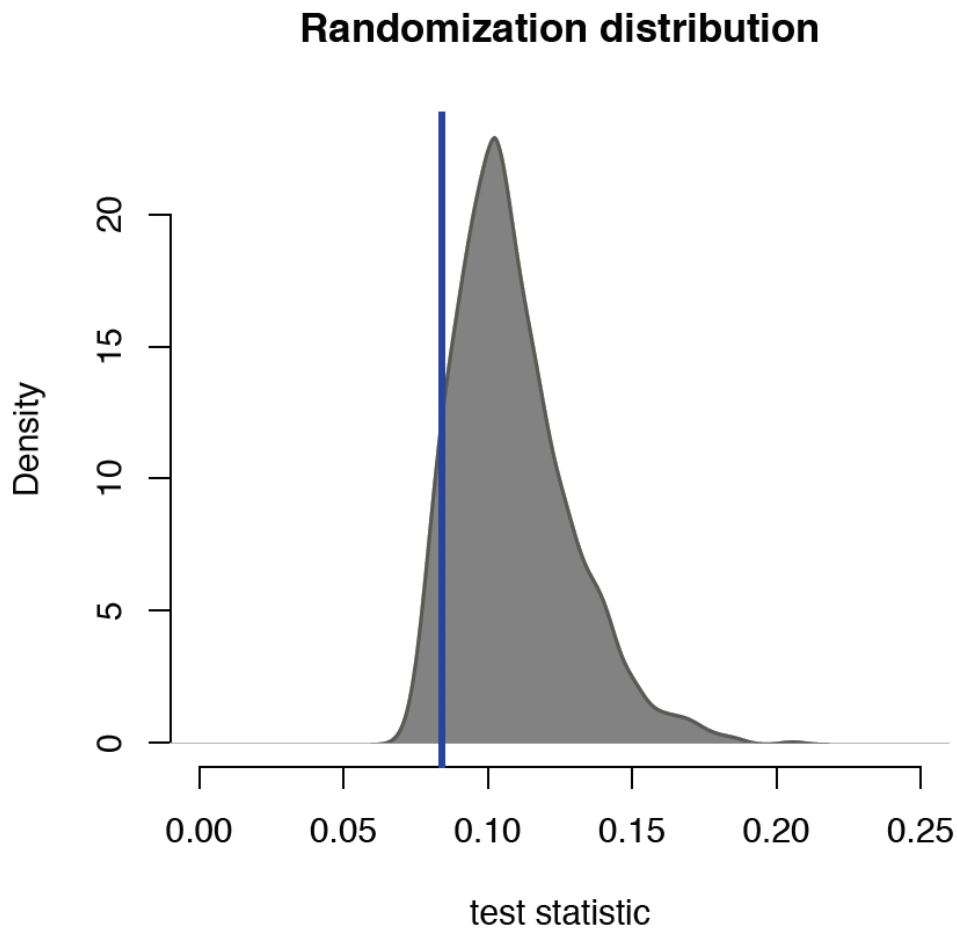
Statistics of the null-exposure graph:

- #units = 37,055.
- #assignments = 10,000.
- #edges = 163,836,445.
- density (#edges / total #of possible edges) = 44.2%

Statistics of the clique we condition on:

- #units in clique = 3,981.
- #assignments in clique $\approx 1,000$.





Focal units (in green) are in downtown and outskirts.
Clique test automatically discovers this pattern.

Concluding thoughts

- New method is presented for testing causal effects under general interference using null exposure graphs and bicliques.
- The method represents the first, general approach for testing under interference.
- Research plan for the next five years?: Next slide!

Five year research plan

These cornerstone projects and their off-shoots will carry me through the next five years. Below, I indicate whether it is machine-learning-focused (ML), randomization-focused (R), or both.

BicliqueRT: A software package for causal testing and experimental design under interference (R)

Financial literacy and financial well-being (ML)

Fisher meets BART: Integrating causal machine learning and randomization tests (ML+R)

A Bayesian classification trees approach to treatment effect variation with noncompliance (ML)

Thank You!

"A Graph-Theoretic Approach to Randomization Tests of Causal Effects Under General Interference" (JRSS-B, 2022)

Athey, Eckles, Imbens,"Exact p-Values for Network Interference" (JASA, 2018)

Basse, Feller, Toulis, "Randomization tests of causal effects under interference" (Biometrika, 2019)

Aronow, "A general method for detecting interference between units in randomized experiments." (Sociol. Methods Res., 2012)

Collazos, D., Garcia, E., Mejia, D., Ortega, D., and Tobon, S., "Hot spots policing in a high crime environment: An experimental evaluation in Medellin". Documento CEDE, (2019-01).

What else can be done? Experimental design under interference

Consider a design space $(p_0, p_1) \in [0, 1]^2$ where p_0=probability of increased police in city-center, and p_1 probability of increased police in outskirts.

What is the optimal design – (p_0^*, p_1^*) – for testing the spillover hypothesis?

Experimental design under interference

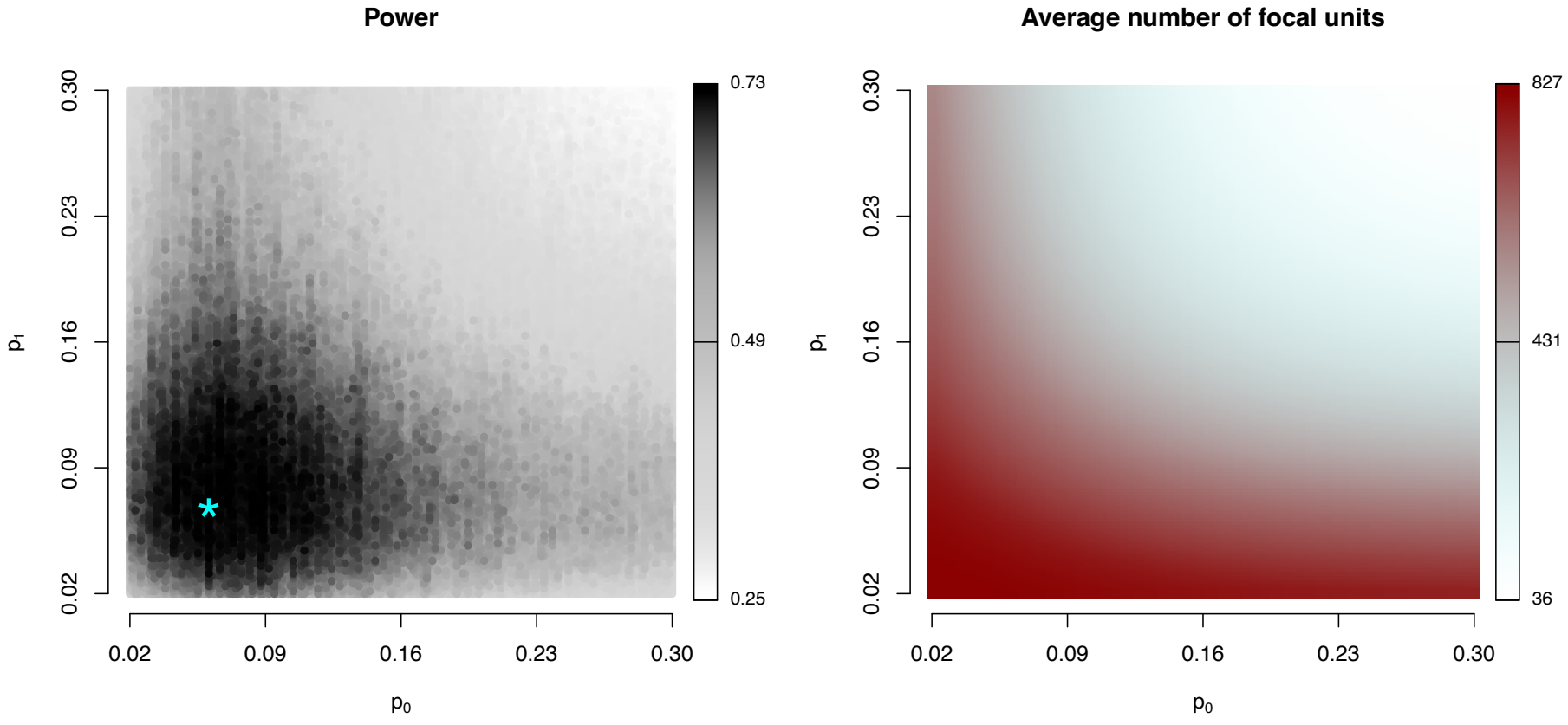


Figure: *Left:* The power of the test for different combinations of p_0, p_1 calculated via simulation. Darker colors denote larger power values, while lighter colors denote smaller power values. *Right:* Average number of focal units in clique for combinations of p_0, p_1.

Extra slides

Testing $Y_i(\texttt{short}) = Y_i(\texttt{control})$, $\forall i$.

Idea: If we focus on units only exposed to `short` or `control` then we can impute their missing outcomes in the randomization test.

$\hookrightarrow$*i.e., conditional randomization test (Aronow, 2012) (Athey et al, 2018) (Basse et al, 2019). Will discuss later.*

Basse et al (2019) formalize this approach: given Z^{obs} we condition on some event C according to $P(C|Z^{\text{obs}})$, known as the conditioning mechanism. The conditional test is valid as long as:

1. Can impute outcomes conditional on C.
2. We randomize according to the correct conditional distribution:

$$P(Z'|C) \propto \underbrace{P(C|Z')}_{\text{conditioning mech.}} \cdot \underbrace{P(Z')}_{\text{design}}.$$

$\star$ But how to construct $P(C|Z^{\text{obs}})$? ...

Answer: Cliques as conditioning mechanisms

We need to impute all potential outcomes given a null hypothesis. **Cliques** of the null exposure graph encode this imputability.

Idea: This implies a conditioning mechanism of the form:

$P(C|Z^{\text{obs}}) = \mathbb{1}\{Z^{\text{obs}} \in C\}$.

So C should be unique (full definition coming soon).

Details: Null-exposure graphs

◦ A null-exposure graph, G_f, is thus uniquely defined given $H_0, \{f_i\}$.

◦ H_0 is sharp in a clique of G_f. So, we run a conditional randomization test within a clique.

↪ *i.e.,* $P(C|Z^{obs}) = \mathbb{1}\{Z^{obs} \in C\}$.

◦ Such test requires a "clique test statistic" $t(y, z; C)$ where C is a clique in G_f such that

$$t(y, z; C) = t(y', z'; C), \text{ if } y_C = y'_C \text{ and } z_C = z'_C.$$

↪ y_C, z_C *are sub-vectors of* y, z *only with units/assignments in* C.

∗ But which clique to condition on?

A naive test (which doesn't work)

Not all approaches lead to a valid test. For example, consider:

1. Given Z^{obs} calculate maximum clique in null-exposure graph, G_f, that contains Z^{obs}, say,

$$C^* = \texttt{mc}(Z^{\text{obs}}; G_f); \quad (\texttt{mc} = \text{"}\texttt{max clique}\text{"}).$$

2. Condition the randomization test on C^*, resampling assignments according to

$$P_R(Z'|C^*) = \frac{\mathbb{1}\{Z' \in C^*\}P(Z')}{P(C^*)}.$$

Proof of invalidity:

The correct conditional distribution is:

$$P(Z'|C^*) = \frac{P(C^*|Z')P(Z')}{P(C^*)} = \frac{\mathbb{1}\{\texttt{mc}(Z'; G_f) = C^*\}P(Z')}{P(C^*)} \neq P_R.$$

Details: Clique-based randomization test

- The randomization distribution in the test is:

$$P_R(Z'|C) = \frac{\mathbb{1}\{Z' \in C\}P(Z')}{P(C)}.$$

- The correct conditional distribution is:

$$P(Z'|C) = \frac{P(C|Z')P(Z')}{P(C)} = \frac{\mathbb{1}\{C \in \mathcal{C}\}\mathbb{1}\{Z' \in C\}P(Z')}{P(C)} = P_R,$$

whenever we use only cliques from decomposition $\mathcal{C}$.

Proof of validity:

$$t(Y, Z'; C) \overset{H_0, C}{=} t(Y', Z'; C) \overset{d}{=} t(Y, Z; C)$$

"$T_R \sim T_{\text{obs}}$ (under null conditional on C)"

Biclique decomposition

- Finding cliques is NP-hard (Peeters, 2003; Zhang et al, 2014).

- We use the "Binary Inclusion-Maximal Biclustering Algorithm", which uses a "divide and conquer" method to find cliques (`Bimax`, Prelic et. al, 2006).

  ↪*works well for thousands of nodes/millions of edges.*

- Our method is **constructive**, still needs to be optimized.

  ↪*i.e., different biclique decompositions will have different power properties, but all are **valid**.*

Extensions of H_0

Athey et al (2018) consider more complex hypotheses than what can be defined based on exposures; e.g.:

$$H_0 : Y_i(z) = Y_i(z') \text{ if } z_i = z_i'.$$

This H_0 is an intersection hypothesis:

Define $f_i(z) = z_i$. Then H_0 is an intersection of:

$$H_0^0 : Y_i(z) = Y_i(z') \text{ if } f_i(z) = f_i(z') = 0. \quad (1)$$

$$H_0^1 : Y_i(z) = Y_i(z') \text{ if } f_i(z) = f_i(z') = 1. \quad (2)$$

We can still apply our method by extending the definition of the null-exposure graph:

$$\widetilde{E} = \left\{(i, z) \in \mathbb{U} \times \mathbb{Z} : f_i(z) = Z_i^{\text{obs}}\right\}. \quad (3)$$

Definition. Let $\mathbb{U}, \mathbb{Z}$ denote the units and assignments, respectively. Let $\mathrm{a}, \mathrm{b} \in \mathcal{E}$ be any two exposures and consider the hypothesis:

$$H_0^{\mathrm{a},\mathrm{b}} : Y_i(z) = Y_i(z'), \text{ for all } i, z, z' \text{ such that } f_i(z), f_i(z') \in \{\mathrm{a}, \mathrm{b}\}.$$

Define the vertex set as $V = \mathbb{U} \cup \mathbb{Z}$, and the edge set as

$$E = \{(i, z) \in \mathbb{U} \times \mathbb{Z} : f_i(z) \in \{\mathrm{a}, \mathrm{b}\}\}. \tag{4}$$

Then, $G_f = (V, E)$ is the null-exposure graph of $H_0^{\mathrm{a},\mathrm{b}}$ wrt f.

- For given $H_0^{\mathrm{a},\mathrm{b}}$ and $\{f_i\}$ the null exposure graph G_f is unique.
- Imputation is possible within the clique that contains obs. Z:

Proposition. Consider a null-exposure graph, G_f, with some clique $C = (U, \mathcal{Z})$. If $Z^{\text{obs}} \in \mathcal{Z}$, then $Y_i(z) = Y_i(Z^{\text{obs}})$ under $H_0^{\mathrm{a},\mathrm{b}}$, for all $i \in U$ and all $z \in \mathcal{Z}$.

Clique Decomposition

Let $\mathbb{U}$ be the set of units and $\mathbb{Z}$ the set of population assignments. A clique decomposition, $\mathcal{C} = \{C_1, \ldots, C_K\}$, of the null-exposure graph is a finite set of cliques, $C_k = (\mathbb{U}_k, \mathcal{Z}_k)$, $k = 1, \ldots, K$, such that

$$\bigcup_k \mathcal{Z}_k = \mathbb{Z}, \text{ and } \mathcal{Z}_k \bigcap \mathcal{Z}_{k'} = \emptyset, \text{ for any } k \neq k'.$$

$\hookrightarrow$ *The set of units does not need to be partitioned.*

Power study: clustered interference

To illustrate, we consider a **clustered interference** setting.

Suppose we have N units spread equally in K clusters.The clusters could be classrooms or households.

Experiment: Randomly treat K/2 clusters. Within each treated cluster, randomly treat 1 unit.

$\hookrightarrow$*Motivated by student absenteeism study (Basse et al, 2019).*

Power study: clustered interference

To illustrate, we consider a **clustered interference** setting.

Suppose we have N units spread equally in K clusters.The clusters could be classrooms or households.

Experiment: Randomly treat K/2 clusters. Within each treated cluster, randomly treat 1 unit.

$\hookrightarrow$*Motivated by student absenteeism study (Basse et al, 2019).*

◦ Do outcomes of a control unit in control cluster differ from outcomes of a control unit in a treated cluster?

The null and competing methods

$$H_0: \quad Y_i(\texttt{control}) = Y_i(\texttt{exposed}),\ \forall i,$$

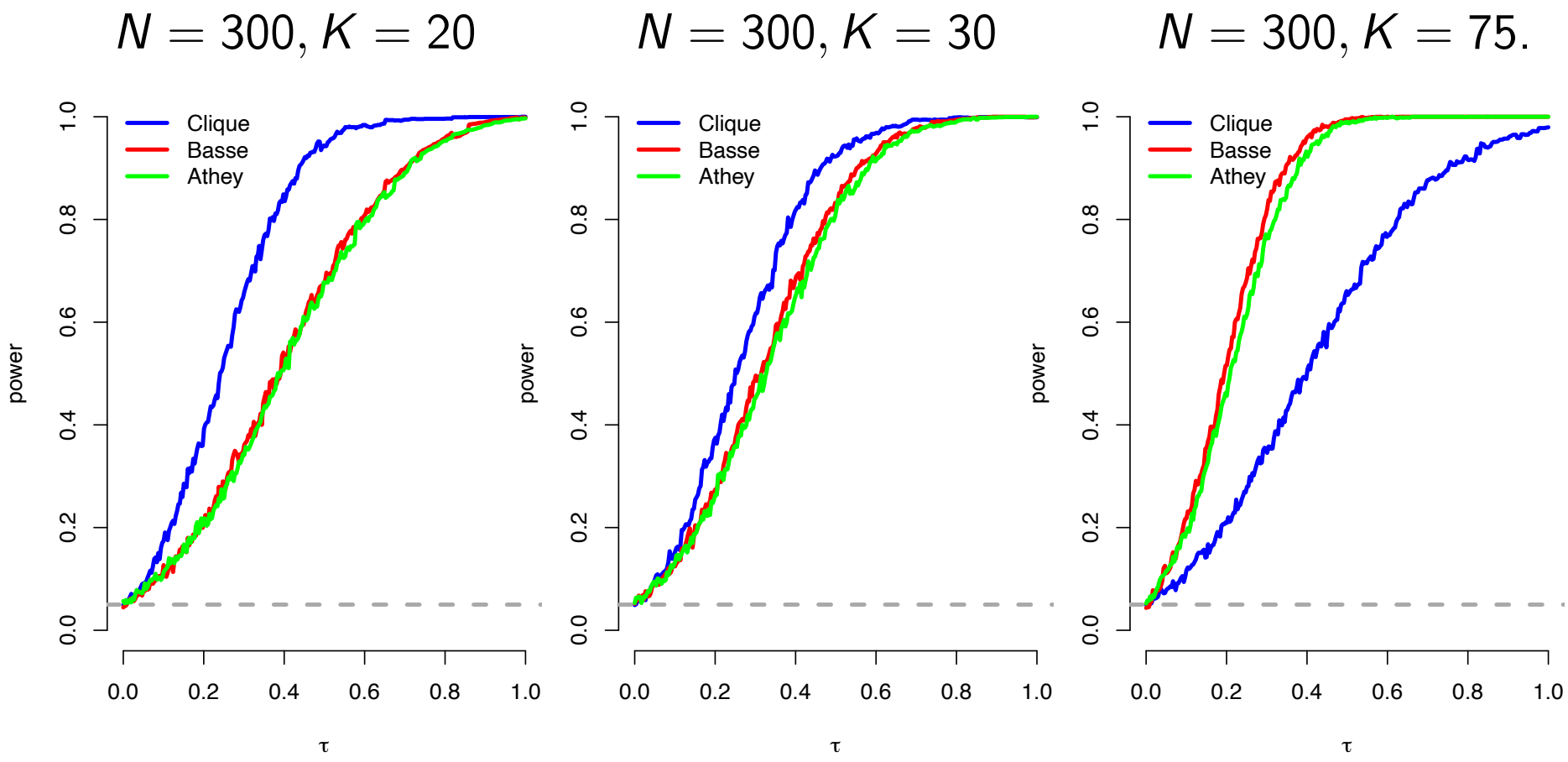
where:
$f_i(Z) =$`control`, if $Z_i = 0$ and $\sum_{j \in [i]} Z_j = 0$;
$f_i(Z) =$`exposed`, if $Z_i = 0$ and $\sum_{j \in [i]} Z_j = 1$, and $[i]$ denotes i's cluster.

1. Athey et. al. (2018): sample one focal per household. Run randomization test*.

2. Basse et. al. (2019): For treated households, sample one untreated focal unit (uniformly). For untreated households, sample one focal. Run permutation test on the focals.

3. Clique test – proposed method.

Power comparison: $Y_i(\texttt{exposed}) = Y_i(\texttt{control}) + \tau$

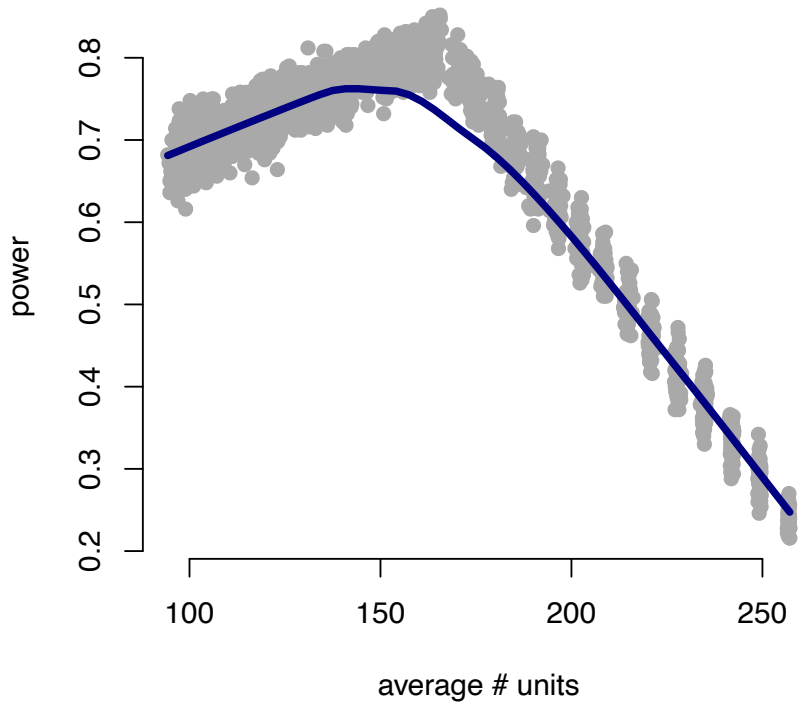


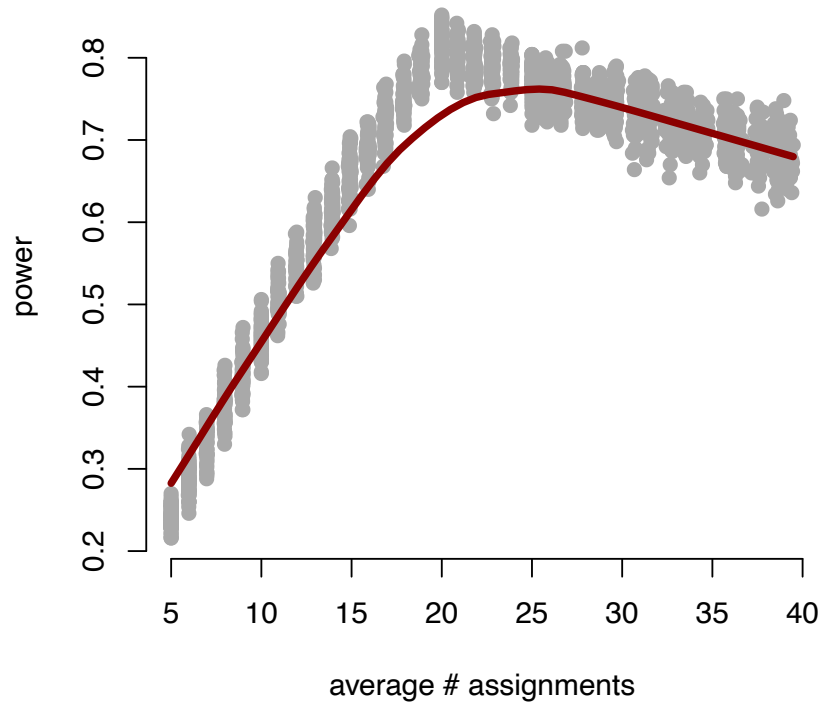
The clique test improves upon existing methods as the cluster size increases (smaller K)!

↪*It achieves more flexible conditioning (i.e., many units/cluster).*

Power characteristics

Trade-off between #units and #assignments in the cliques.





Power theory

Theorem (high level)

For $C = (U, \mathcal{Z})$ let $|C| = (n, m)$ imply that $|U| = n$ and $|\mathcal{Z}| = m$. Suppose:

(A1) *n is scale parameter $(1/\sqrt{n})$ for null distribution of test statistic;*

(A2) *spillover effect τ is additive;*

(A3) *the m test statistic values are i.i.d. from the null;*

(A4) *the null distribution cdf can be ϵ-approximated by a sigmoid.*

Then,

$$\mathrm{E}\big(\text{reject} \mid H_1, |C| = (n, m)\big) \geq \frac{1}{1 + Ae^{-a\tau\sqrt{n}}} - O(m^{-r}) - \epsilon,$$

where $a, A > 0$, $r \in (1/2, 1]$.

Interpretation:

- Number of focal units controls “sensitivity" of the test.
- Number of focal assignments controls maximum power.

Treatment exposures

For each Z, unit i is exposed to "something more" than Z_i.
Let unit i's exposure be defined by a function:

$$f_i : \{0, 1\}^N \to \mathcal{E}.$$

$\mathcal{E}$ is the set of possible exposures (short-range spillover, medium-range spillover, pure control, etc.).

$\hookrightarrow$ *Definition of $\mathcal{E}$, $\{f_i\}$ depends on the substantive scientific question.*

We can now formulate (interference-based) hypotheses in terms of **exposures**!