Randomization tests for spillovers under interference: A graph-theoretic approach

Panagiotis (Panos) Toulis
panos.toulis@chicagobooth.edu

Econometrics and Statistics
University of Chicago, Booth School of Business

Interference

Interference exists when the outcomes of some unit depend on the treatment of others.

—(Hong and Raudenbush, 2006); (Hudgens and Halloran, 2008); (Aronow, 2012); (Bowers, 2013); (Toulis and Kao, 2013); (Ogburn and VanderWeele, 2014); (Eckles et. al., 2016); (Aronow and Samii, 2017); (Ogburn et. al., 2017); (Savje et al, 2017); (Athey et. al, 2018), (Basse and Feller, 2018); (Basse et. al., 2019); (Jagadeesan et. al., 2020) (Forastiere et. al., 2020);

Includes spillovers, peer effects, contagion, equilibrium effects, etc.

Pervasive in most social studies. Can be either a nuisance to be addressed by design, or the quantity of interest.

Motivation for this work: Crime spillovers across streets from policing experiment in Medellin, Colombia.

Current approaches

Several model-based approaches exist. Typically include regressions of unit outcomes on group/peer treatments and outcomes.
—(Durlauf and Young, 2001); (Brock and Durlauf, 2001); (Jackson, 2010); (Graham, 2008)

Model-based approach has risks due to identification and interpretation issues.
—(Deaton, 1990); (Manski, 1993); (Boozer and Cacciola, 2001); (Moffit, 2001); (Angrist, 2014)

Design-based approaches have emerged as a robust alternative. They mostly aim to generalize the classical Fisher randomization test.
—(Aronow, 2012); (Athey et. al., 2018); (Basse et. al., 2019)

The main benefits of randomization-based approaches are finite-sample validity and robustness.
—Criticism mainly focuses on generalizability of randomization results.

Outline

1. Setup and notation
2. Classical Fisher randomization test (FRT)
3. Interference
   - Hypotheses of interest
   - Treatment exposures
   - Main challenge for FRTs
   - Conditional FRTs
   - Current approaches
4. Main method
   - The null-exposure graph
   - “Clique-based” FRT
5. Application in Medellin
6. Considerations: computation, test power

Setup and notation

There is a set $\mathbb{U} = \{1, \ldots, N\}$ of N units indexed by i.

Denote:

$Z = (Z_1, \ldots, Z_N) \in \{0,1\}^N =: \mathbb{Z}$	binary treatment
$Y(z) = (Y_1(z), \ldots, Y_N(z)) \in \mathbb{R}^N$	potential outcomes under $z \in \mathbb{Z}$
$Z^{\text{obs}} \in \mathbb{Z},\ Y^{\text{obs}} \in \mathbb{R}^N$	observed quantities
Z^*, Y^*	randomization draws
$P(Z) \in [0,1]$	design, assumed known

As usual, potential outcomes are assumed to be fixed, and randomness comes only from $P(Z)$.

Some questions of interest

There is **no treatment effect** when for all i:

$$H_0 : Y_i(z) = Y_i(z'), \text{ for all } z, z' \in \mathbb{Z}.$$

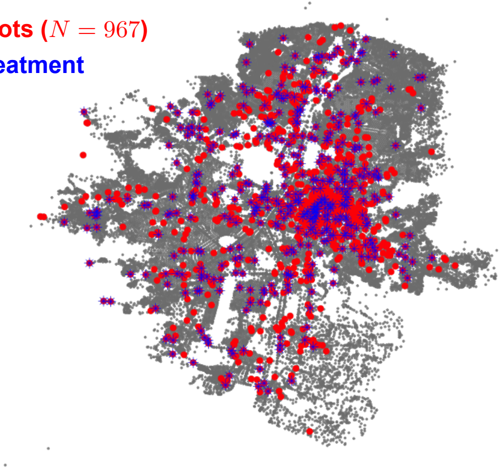
There is **no interference** when for all i:

$$H_0 : Y_i(z) = Y_i(z'), \text{ for all } z, z' \in \mathbb{Z} \text{ such that } z_i = z'_i$$ —[aka "SUTVA"].

Suppose units are in social network. There is only **neighborhood interference** when for all i:

$$H_0 : Y_i(z) = Y_i(z'), \text{ for all } z, z' \in \mathbb{Z} \text{ such that } z_i = z'_i \text{ and } z_{\text{neighbor}_i} = z'_{\text{neighbor}_i}.$$

Medellin application



$i =$ "hotspot"; $Z_i =$ policing level at unit i; $Y_i =$ crime "score".

We will test whether there are spillovers on control streets from nearby treated streets.

Treatment exposures

One compact to way to represent these null hypotheses is through "treatment exposures". For some unit i the exposure under assignment z is given by:

$$f_i(z) : \mathbb{Z} \to \mathbb{F},$$

where $\mathbb{F}$ = set of possible exposures —e.g., number of "neighbors treated", "pure control", etc.

One general class of hypotheses is then:

$$H_0^{\mathcal{F}} : Y_i(z) = Y_i(z'), \text{ for all } z, z' \in \mathbb{Z} \text{ such that } f_i(z), f_i(z') \in \mathcal{F} \subseteq \mathbb{F}$$

- $H_0^{\mathbb{F}}$ is the "global null".
- $H_0^{\{a,b\}}$ is a "contrast hypothesis" —Medellin example.
- $\cap_{a \in \mathbb{F}} H_0^{\{a\}}$ covers the previous examples. Equivalent to:

$$H_0 : Y_i(z) = Y_i(z'), \text{ for all } z, z' \in \mathbb{Z} \text{ such that } f_i(z) = f_i(z').$$

Fisher randomization test (FRT, 1935)

We start with the simplest "global null" hypothesis of no effect:

$$H_0 : Y_i(z) = Y_i(z'), \text{ for all } z, z' \in \mathbb{Z}.$$

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

1. $T^{\text{obs}} = t(Y^{\text{obs}}, Z^{\text{obs}})$.
2. Sample $Z^* \sim P(Z^*)$, store $T_R = t(Y^{\text{obs}}, Z^*)$.
3. p-value $= \mathbb{E}\left[\mathbb{1}\left\{T_R \geq T^{\text{obs}}\right\}\right]$.

Fisher randomization test (FRT, 1935)

We start with the simplest "global null" hypothesis of no effect:

$$H_0 : Y_i(z) = Y_i(z'), \text{ for all } z, z' \in \mathbb{Z}.$$

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

1. $T^{\text{obs}} = t(Y^{\text{obs}}, Z^{\text{obs}})$.
2. Sample $Z^* \sim P(Z^*)$, store $T_R = t(Y^{\text{obs}}, Z^*)$.
3. p-value $= \mathbb{E}\left[\mathbb{1}\left\{T_R \geq T^{\text{obs}}\right\}\right]$.

Proof of validity:

$$t(Y^{\text{obs}}, Z^*) \overset{H_0}{=} t(Y^*, Z^*) \overset{d}{=} t(Y^{\text{obs}}, Z^{\text{obs}})$$

—"$T_R \sim T^{\text{obs}}$ (under null)"

More about the FRT

Advantages of FRT include:

- Simple and exact. The test is valid in finite samples.
- Minimal assumptions. No model for Y.
- Robust. Same answer under some transformations of Ys.

Main critique of FRT:

- Can only test strong/uninteresting nulls.
- Cannot generalize out of sample.

More about the FRT

Advantages of FRT include:

- Simple and exact. The test is valid in finite samples.
- Minimal assumptions. No model for Y.
- Robust. Same answer under some transformations of Ys.

Main critique of FRT:

- Can only test strong/uninteresting nulls.
- Cannot generalize out of sample.

$\star$ How to test a more complex hypothesis, such as “SUTVA”?

Testing for interference

To test for any interference we can test the "SUTVA" assumption:

$$H_0 : Y_i(z) = Y_i(z'), \text{ for all } z, z' \in \mathbb{Z} \text{ such that } z_i = z'_i.$$

The classical FRT runs into trouble here.

Testing for interference

To test for any interference we can test the "SUTVA" assumption:

$$H_0 : Y_i(z) = Y_i(z'), \text{ for all } z, z' \in \mathbb{Z} \text{ such that } z_i = z'_i.$$

The classical FRT runs into trouble here.

To see this, note that in the test we need to have $Z_i^* = Z_i^{\text{obs}}$ in order to be able to use H_0 to impute missing outcomes.

This immediately implies that

$$Z^* = Z^{\text{obs}}. \tag{1}$$

The randomization distribution is degenerate, and cannot effectively test H_0.
—In other words, H_0 is not sharp and so introduces constraints in the randomization — Equation (1).

Testing for interference: Resolving the problem

$$H_0 : Y_i(z) = Y_i(z'), \text{ for all } z, z' \in \mathbb{Z} \text{ such that } z_i = z'_i.$$

One way to resolve the problem is to execute FRT on **subsets** of the units and assignments. —(Aronow, 2012); (Athey et. al., 2018); (Basse et. al., 2019)

Example (Suppose $P(Z)$ is Bernoulli design)

- Pick $U \subset \mathbb{U}$, $|U| = N/2$, at random.—[focal units]
- Choose test statistic $t()$ that depends only on outcomes from units in U.
- Run FRT by shuffling the treatments only of units in $\mathbb{U} \setminus U$.

This now works because the support of the randomization distribution is a large set of assignments:

$$\mathbb{Z}(U) = \{z \in \mathbb{Z} : z_i = Z_i^{\text{obs}} \text{ for all } i \in U\}.$$

The *unconditional* FRT did not work because $\mathbb{Z}(\mathbb{U}) = \{Z^{\text{obs}}\}$.

Definition (Conditional FRT / Conditioning mechanism)

Let $C = (U, \mathcal{Z})$, where $U \subseteq \mathbb{U}$ and $\mathcal{Z} \subseteq \mathbb{Z}$, with distribution $P(C|Z)$. Let $t(y, z; C)$ denote a test statistic that uses outcomes only from units in U that can be imputed for all $z \in \mathcal{Z}$. Consider the test:

1. $C \sim P(C|Z^{\text{obs}})$.
2. $T^{\text{obs}} = t(Y^{\text{obs}}, Z^{\text{obs}}; C)$.
3. Sample $Z^* \sim r(Z^*)$, store $T_R = t(Y^{\text{obs}}, Z^*; C)$.
4. p-value $= \mathbb{E}\left[\mathbb{1}\left\{T_R \geq T^{\text{obs}}\right\}\right]$.

This procedure is a conditional FRT. Distribution $P(C|Z)$ is the *conditioning mechanism*, and C is the *conditioning event*.

Definition (Conditional FRT / Conditioning mechanism)

Let $C = (U, \mathcal{Z})$, where $U \subseteq \mathbb{U}$ and $\mathcal{Z} \subseteq \mathbb{Z}$, with distribution $P(C|Z)$. Let $t(y, z; C)$ denote a test statistic that uses outcomes only from units in U that can be imputed for all $z \in \mathcal{Z}$. Consider the test:

1. $C \sim P(C|Z^{\text{obs}})$.
2. $T^{\text{obs}} = t(Y^{\text{obs}}, Z^{\text{obs}}; C)$.
3. Sample $Z^* \sim r(Z^*)$, store $T_R = t(Y^{\text{obs}}, Z^*; C)$.
4. p-value $= \mathbb{E}\left[\mathbb{1}\left\{T_R \geq T^{\text{obs}}\right\}\right]$.

This procedure is a conditional FRT. Distribution $P(C|Z)$ is the *conditioning mechanism*, and C is the *conditioning event*.

The conditional FRT is valid (Basse et. al., 2019) as long as:

$$r(Z^*) = P(Z^*|C) \propto \underbrace{P(C|Z^*)}_{\text{conditioning mech.}} \cdot \underbrace{P(Z^*)}_{\text{design}}$$

Proof of validity:

$$t(Y^{\text{obs}}, Z^*; C) \overset{H_0,\, C}{=} t(Y^*, Z^*; C) \overset{d}{=} t(Y^{\text{obs}}, Z^{\text{obs}}; C)$$

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

Conditioning mechanisms

⋆ The key is therefore the conditioning mechanism, $P(C|Z)$.

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

But how to construct one?

(Athey et. al., 2018) considered mechanisms of the form:

$$P(C = (U, \mathcal{Z})|Z) = P(U) \cdot \mathbb{1}\{\mathbb{Z}(U) = \mathcal{Z}\},$$

for some choice of $P(U)$ (e.g., random sample, "ϵ-nets").
—Works in all cases but may lead to loss of power; e.g., when $\mathbb{Z}(U) = \{Z^{\text{obs}}\}$. Also, generally not a permutation test.

(Basse et. al., 2019) constructed a mechanism under clustered interference such that:

$$t(Y, Z; C) \stackrel{d}{=} t(Y, Z; \pi C)$$

where $\pi C = (\pi U, \mathcal{Z})$ denotes permutation of the focal units.
—Simple and good power; but not general.

The Medellin example

In the Medellin example, define treatment exposures:

$$f_i(z) = \begin{cases} \texttt{short}, & z_i = 0, \text{dist}_i < 125\text{m} \\ \texttt{control}, & z_i = 0, \text{dist}_i > 500\text{m} \\ \texttt{neither}, & \text{otherwise.} \end{cases}$$

where $\text{dist}_i = \min_{j \neq i : z_j = 1} d(j, i)$ = distance to closest treated street.

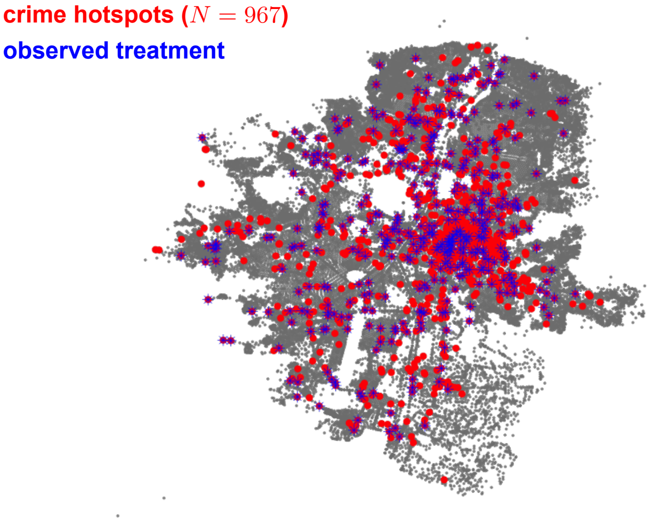
We wish to test $H_0^{\{a,b\}}$ with $a = \texttt{short}$ and $b = \texttt{control}$, i.e.,

$$H_0 : \; Y_i(z) = Y_i(z') \text{ for every } i, z, z',$$

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

⋆ What should be the conditioning mechanism?

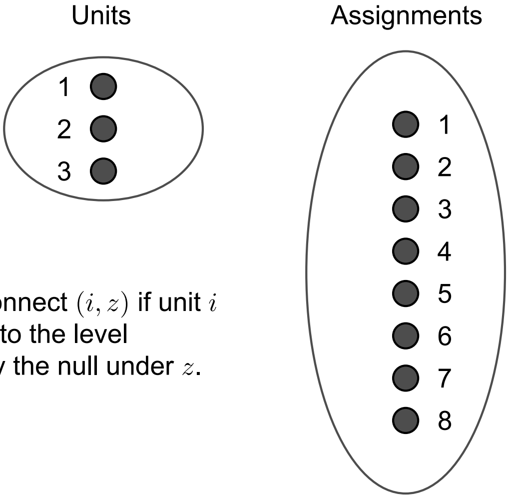
The Medellin example: What's a good cond. mechanism?



⋆ The correct kind of conditioning is unclear.

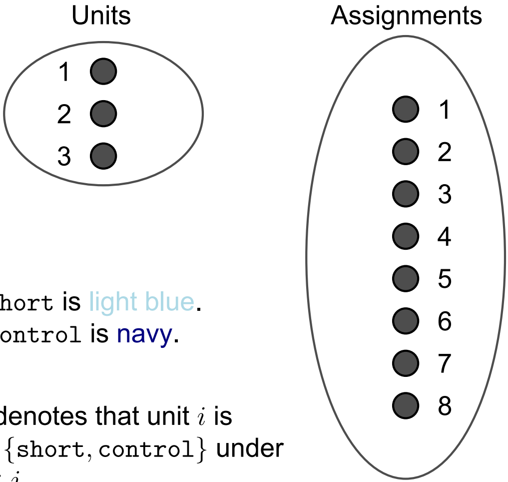
The null exposure graph

We set all the units one side and all the assignments on the other.



Then we connect (i, z) if unit i is exposed to the level specified by the null under z.

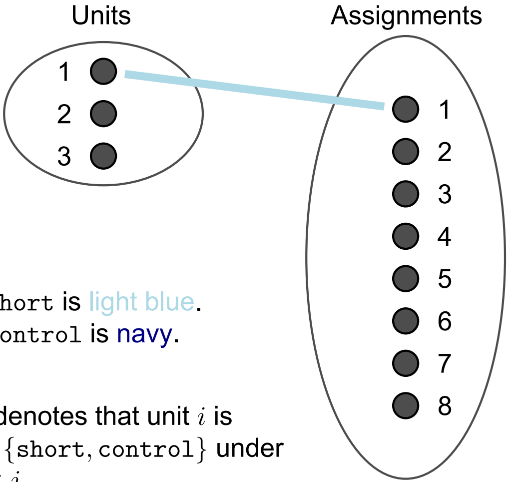
The null exposure graph



Exposure $\texttt{short}$ is light blue.
Exposure $\texttt{control}$ is navy.

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

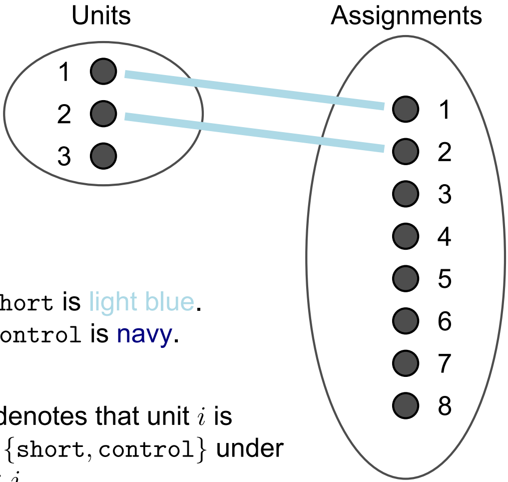
The null exposure graph



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.

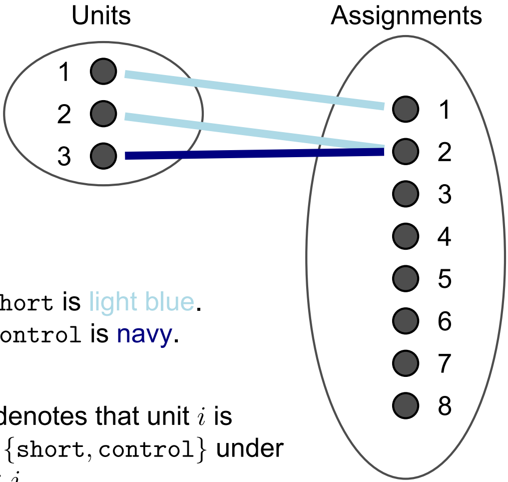
The null exposure graph



Exposure $\texttt{short}$ is light blue.
Exposure $\texttt{control}$ is navy.

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

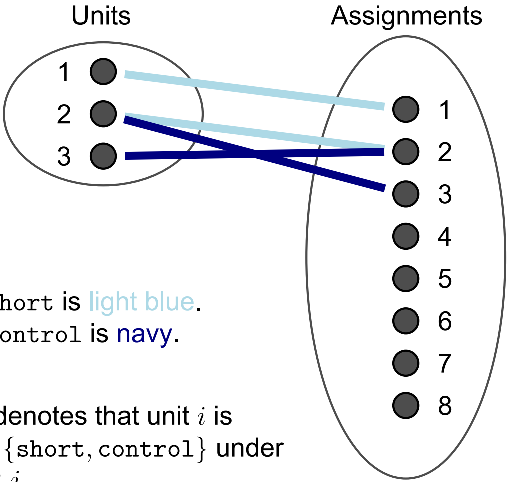
The null exposure graph



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.

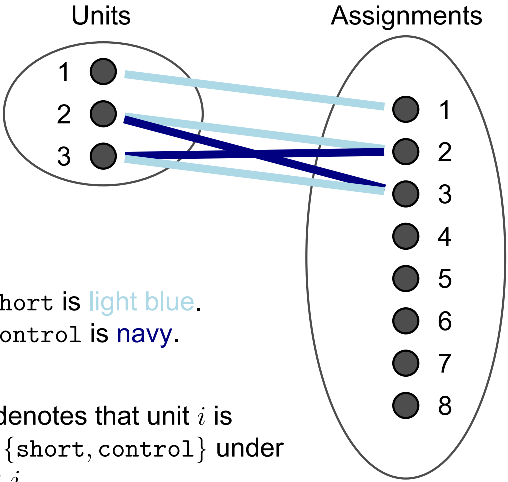
The null exposure graph



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.

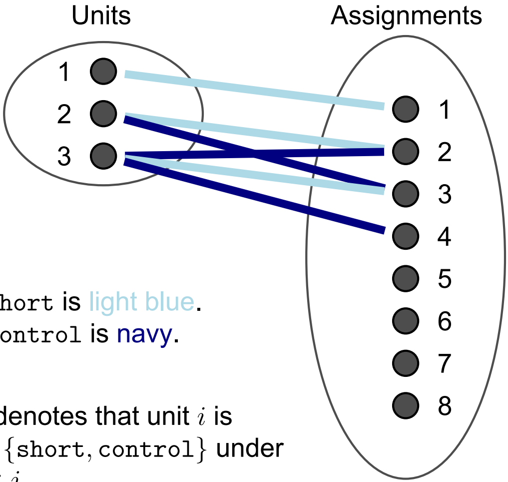
The null exposure graph



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.

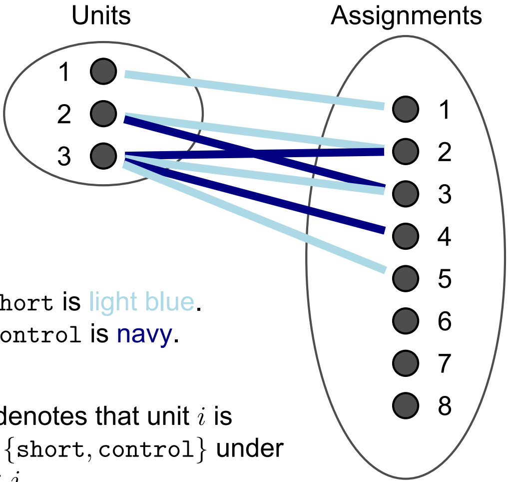
The null exposure graph



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.

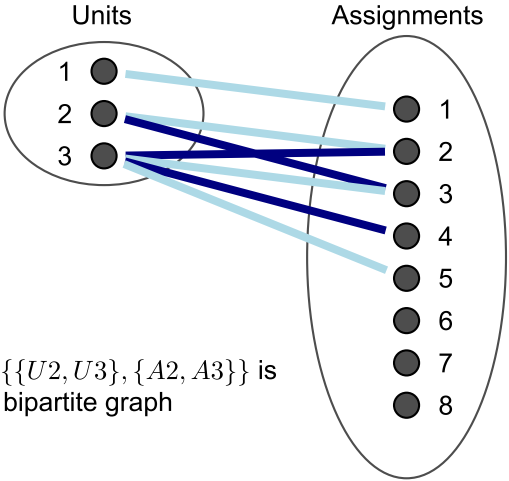
The null exposure graph



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

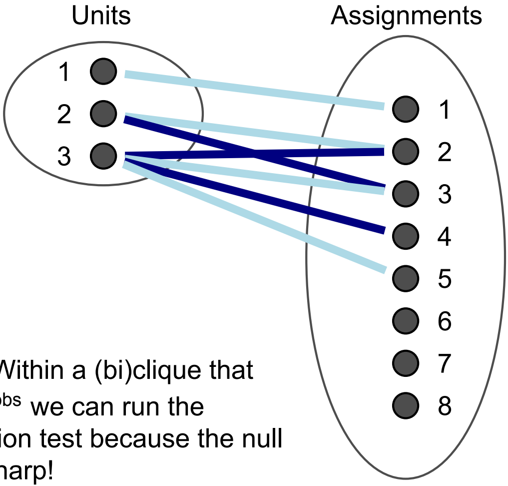
edge (i, j) denotes that unit i is exposed to $\{\texttt{short}, \texttt{control}\}$ under assignment j.

The null exposure graph



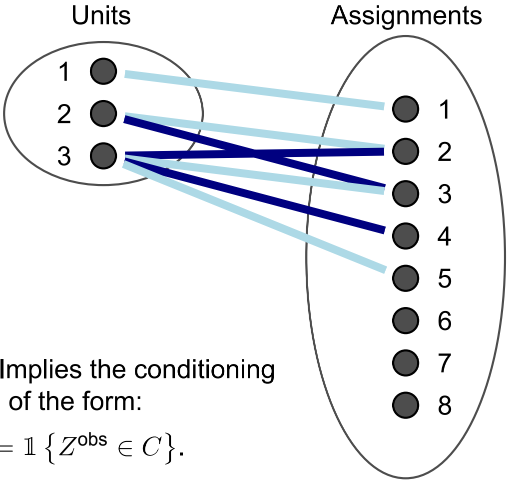
Notice that $\{\{U2, U3\}, \{A2, A3\}\}$ is a complete bipartite graph (biclique).

The null exposure graph



Key idea: Within a (bi)clique that contains Z^{obs} we can run the randomization test because the null becomes sharp!

The null exposure graph



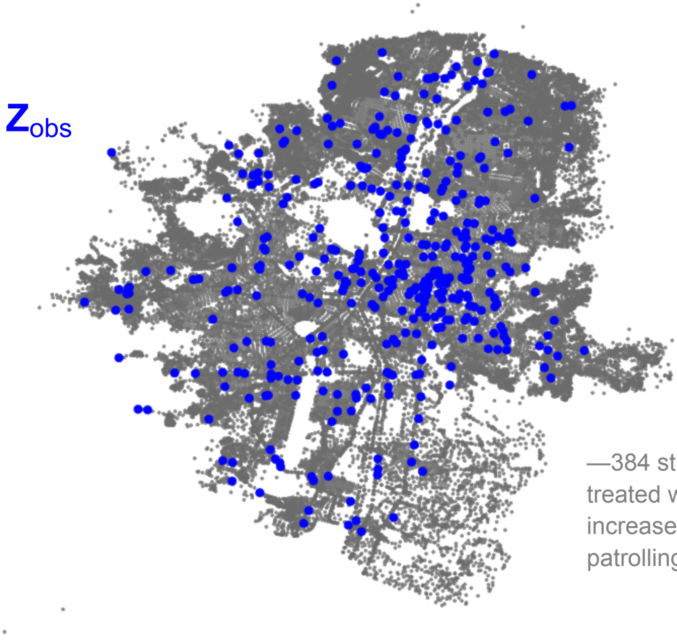
Key idea: Implies the conditioning mechanism of the form:

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

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

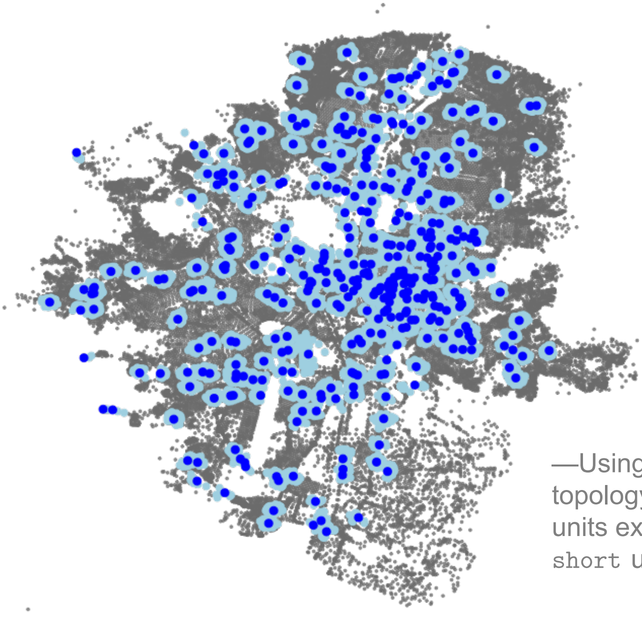
Returning to the map

The observed assignment



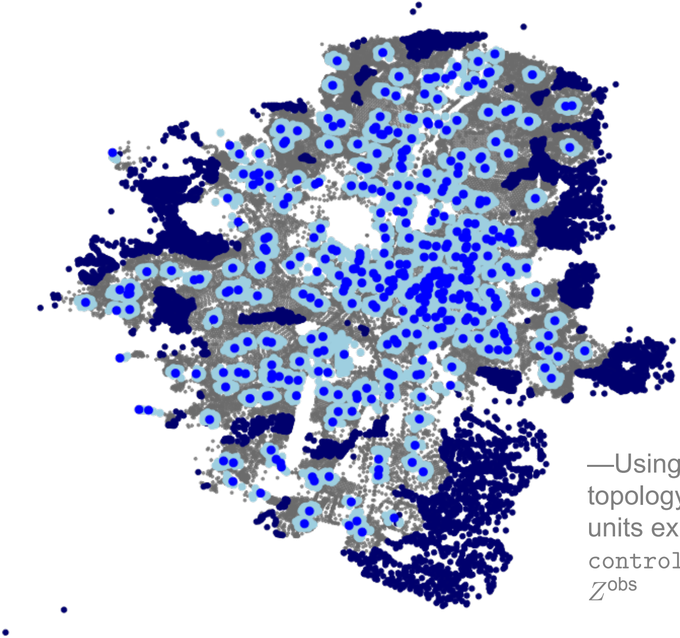
—384 streets are treated with increased police patrolling

Short-range spillover units (`short`)



—Using network topology, color units exposed to `short` under Z^{obs}

Pure control units (`control`)



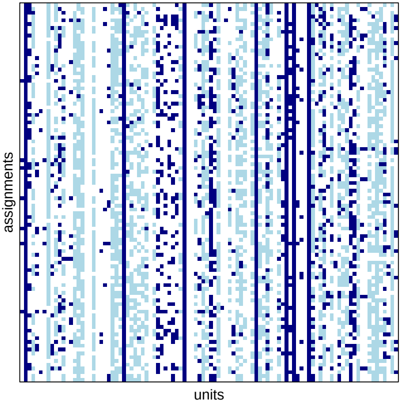
—Using network topology, color units exposed to `control` under Z^{obs}

We can remake these pictures for every assignment Z drawn from design $P(Z)$ ...

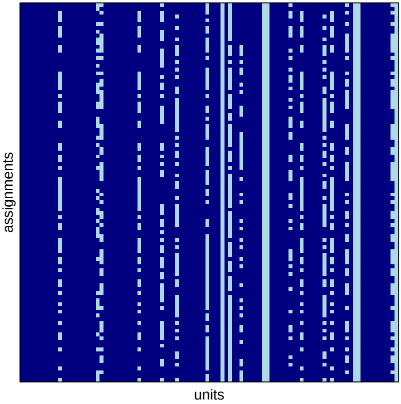
We can remake these pictures for every assignment Z drawn from design $P(Z)$ ...

—The output is our null exposure graph!

Null exposure graph and clique



null exposure graph



clique (zoomed-in)

Null-exposure graphs: summary

- 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 can run a conditional randomization test within a clique. Equivalently, the conditioning mechanism is:

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

Null-exposure graphs: summary

- 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 can run a conditional randomization test within a clique. Equivalently, the conditioning mechanism is:

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

$\star$ But which clique to condition on?

A naive test (which doesn't work)

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

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

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

2. Condition the randomization test on C^*, i.e., resample assignments according to

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

A naive test (which doesn't work)

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

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^*, i.e., resample assignments according to

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

Proof of invalidity:

The correct conditional distribution is:

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

Main method: Clique-based randomization test

1. **Decompose:** Compute *biclique decomposition* $\mathcal{C}$ of G_f.
2. **Condition:** Pick out clique containing Z^{obs}, call it C.
3. **Summarize:** Compute $T^{\text{obs}} = t(Y^{\text{obs}}, Z^{\text{obs}}; C)$, then

$$\text{p-value} = \mathbb{E}\left[\mathbb{1}\left\{t(Y^{\text{obs}}, Z^*; C) \geq T^{\text{obs}}\right\} \mid C\right]$$

Here, we resample with respect to

$$r(Z^*) \propto \underbrace{\mathbb{1}\{Z^* \in C\}}_{\text{cond. mechanism}} \cdot \underbrace{P(Z^*)}_{\text{design}}$$

Main method: Clique-based randomization test

1. **Decompose:** Compute *biclique decomposition* $\mathcal{C}$ of G_f.
2. **Condition:** Pick out clique containing Z^{obs}, call it C.
3. **Summarize:** Compute $T^{\text{obs}} = t(Y^{\text{obs}}, Z^{\text{obs}}; C)$, then

$$\text{p-value} = \mathbb{E}\left[\mathbb{1}\left\{t(Y^{\text{obs}}, Z^*; C) \geq T^{\text{obs}}\right\} \mid C\right]$$

Here, we resample with respect to

$$r(Z^*) \propto \underbrace{\mathbb{1}\left\{Z^* \in C\right\}}_{\text{cond. mechanism}} \cdot \underbrace{P(Z^*)}_{\text{design}}$$

Proof of validity:

The correct conditional distribution is:

$$P(Z^*|C) = \frac{P(C|Z^*)P(Z^*)}{P(C)} = \frac{\mathbb{1}\left\{C \in \mathcal{C}\right\}\mathbb{1}\left\{Z^* \in C\right\}P(Z^*)}{P(C)} = r(Z^*).$$

—first eq. from Bayes; second from definition of conditioning mechanism.

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 fine for hundred nodes/thousands edges.
- Our method is **constructive**, still can be optimized.
  —i.e., different biclique decompositions will have different power properties, but all are **valid**.

Power

The size of the clique is crucial for the test power.

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,

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

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.

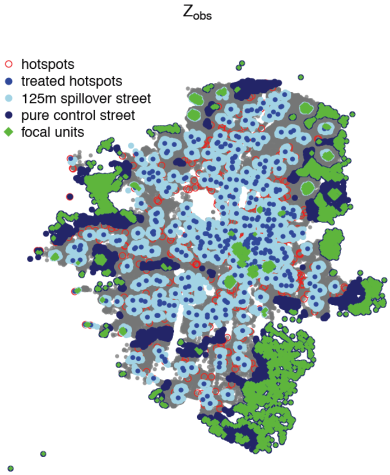
Spatial interference: Medellin data

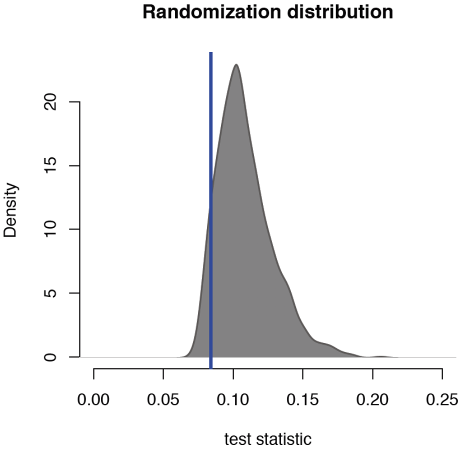
Statistics of the null-exposure graph:

- #units = 37,055.
- #assignments = 10,000 (design is uniform over this fixed set).
- #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

- Structure is placed on null hypothesis through exposure functions.
- New method is presented for testing causal effects under general interference by representing the problem through the null exposure graph and conditioning on bicliques of this graph.
- Translates the testing problem into graphical operations on the null exposure graph.
- Future work: optimal design.

Concluding thoughts

- Structure is placed on null hypothesis through exposure functions.
- New method is presented for testing causal effects under general interference by representing the problem through the null exposure graph and conditioning on bicliques of this graph.
- Translates the testing problem into graphical operations on the null exposure graph.
- Future work: optimal design.

Thank you!

Varying radius of short-range effect

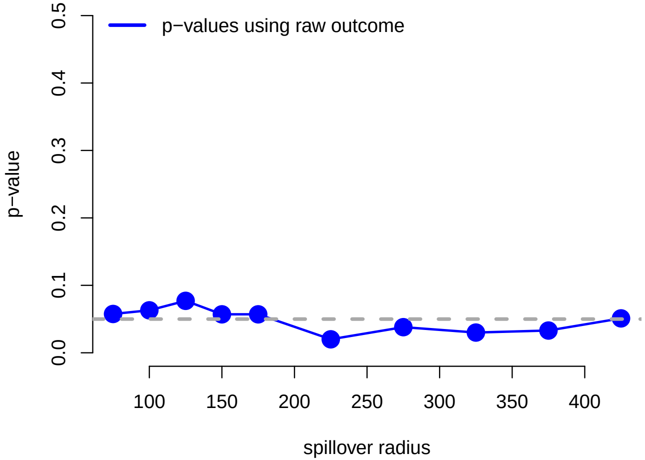


Figure: P-values for clique tests with varying spillover radius.