

Randomization Inference for Spillover Effects

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

Econometrics and Statistics
University of Chicago, Booth School of Business

Introduction

The standard methods of causal inference tacitly assume no interference; i.e., treatment on an individual unit cannot affect other units.

This assumes a simple, static world.

However, many interesting problems exist in settings where units interact in a complex way.

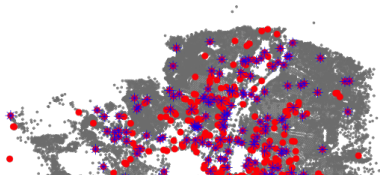
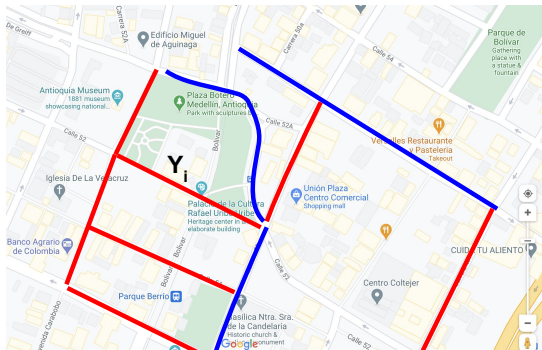
—spillovers, peer effects, contagion, equilibrium effects, etc.

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

New methods and tools are needed. Many applications:
e.g., policy making, marketplace algorithms, climate science, healthcare, etc.

Motivation: Crime spillovers in Medellin, Colombia

Crime spillovers from nearby **treated streets** on **control streets**?



Causal Inference

Suppose data $\{(Y_i, Z_i, X_i)\}, i = 1, \dots, N$.

Here, $Y = \text{outcome}$, $Z = \text{treatment}$, $X = \text{covariates (features)}$.

We want to understand the causal effect of Z on Y .

Some options:

- 1 *Model-based approach: Regress $Y \sim Z + X$.
Validate with IV, “parallel trends”, etc.
- 2 *Design-based approach: Exploit known variation in Z (e.g., from an experiment). The “potential outcomes” are fixed.
e.g., Randomized studies. Remains the gold standard of causal inference.
- 3 Causal graphs: Not today.
- 4 DSGE-style / structural models: “Model-based approach on steroids”.
Still popular in macro policy making.

Pitfalls of model-based approach

A model-based approach requires correct specification, and is open to potential biases.

A more pernicious problem is how the method **quantifies uncertainty**.

Example: Suppose a completely randomized design (50% treated/control):

Unit (i)	Treatment (Z_i)	Outcome (Y_i)
1	1	8
2	0	$3 + \epsilon$
3	0	$3 - \epsilon$
4	1	8

Regress $Y_i \sim Z_i$. The estimate of “causal effect” is +5.

What is the standard error?

Pitfalls of model-based approach

A model-based approach requires correct specification, and is open to potential biases.

A more pernicious problem is how the method **quantifies uncertainty**.

Example: Suppose a completely randomized design (50% treated/control):

Unit (i)	Treatment (Z_i)	Outcome (Y_i)
1	1	8
2	0	$3 + \epsilon$
3	0	$3 - \epsilon$
4	1	8

Regress $Y_i \sim Z_i$. The estimate of “causal effect” is +5.

What is the standard error? $O(\epsilon)$. (arbitrary level of certainty).

↪ Standard error estimation is **conflated** with model fit.
(here, the data fit a line very well).

Design-based approach

A design-based approach exploits the actual variation in the experiment.

The idea is to predict outcomes under **counterfactual treatment assignments**. Then compare with what was observed.

When possible, this is a more accurate way of quantifying uncertainty.

Design-based approach

A design-based approach exploits the actual variation in the experiment.

The idea is to predict outcomes under **counterfactual treatment assignments**. Then compare with what was observed.

When possible, this is a more accurate way of quantifying uncertainty.

To illustrate, suppose counterfactual assignment $Z' = (0, 1, 1, 0)$.

— According to our experiment design, this assignment is equally probable to the observed one.

What would be the outcomes Y' under Z' ?

Unit (i)	Treatment (Z'_i)	Outcome (Y'_i)
1	0	?
2	1	?
3	1	?
4	0	?

Design-based approach

If the treatment does not affect outcomes, then Y' would be equal to the observed Y .

Design-based approach

If the treatment does not affect outcomes, then Y' would be equal to the observed Y .

That is, the observed data would be as follows:

Unit (i)	Treatment (Z'_i)	Outcome (Y'_i)
1	0	8
2	1	$3 + \epsilon$
3	1	$3 - \epsilon$
4	0	8

In this case, we **would have** calculated an effect of -5 instead of $+5$.

We can repeat this procedure for all 6 possible randomizations.

Observing an effect of $+5$, although extreme, has a $1/6 > 16\%$ chance of happening.

No significance. (cf. linear model).

General Idea: Fisher's Randomization Test

Design $D(z) \in [0, 1]$ = probability distribution of treatment.

Let $Y_i(0), Y_i(1)$ be the “potential outcomes” of unit i under control and treatment, respectively.

This is known as a stability assumption (“SUTVA”).

Suppose the treatment has no effect on the outcomes:

$$H_0 : Y_i(0) = Y_i(1).$$

How to test?

- 1 Choose test statistic, $t(z, y)$; e.g., diff in means, or OLS using X as control.
- 2 Build the randomization distribution: $F_R = \{t(z', Y) : z' \sim D\}$.
- 3 $pval = 1 - F_R(t(Z, Y))$.

An assessment of FRT

Major benefits:

- The test is exact in finite samples. No asymptotics.
- Not necessary to have correct Y -model specification.
- The test is robust. Same answer under transformations of Y .
(cf. regression/ML on $\log Y$ may yield completely different results than on Y)

Some disadvantages:

- Can only test “strong” hypotheses.
(Currently, a lot of research activity in this area).
- Cannot generalize to population.
(Personal opinion: this is a feature, not a bug.)

Complex Systems – Interference

A crucial assumption in causal inference (model- or design-based) has been SUTVA: For every unit i , there are only two potential outcomes $Y_i(0), Y_i(1)$ under treatment or control, i.e.,

$$Y_i = \begin{cases} Y_i(0) & \text{when } Z_i = 0 \\ Y_i(1) & \text{when } Z_i = 1. \end{cases}$$

However, in many problems there is **treatment interference**. (spillovers, peer effects, contagion, dynamics etc.)

Under interference, a unit is exposed to “something more” than Z_i .

It is exposed to a sum effect from the entire population treatment, Z .

Think of a vaccine trial. A control unit (unvaccinated) is “protected” by treated units (vaccinated) in proximity.

Some more examples earlier.

Example 1 - Hypotheses for spillovers

Under interference, every unit is exposed to “something more” than Z_i .

A popular convention is to call this **treatment exposure**, $f_i(Z) \in \mathbb{F}$.

Although not necessary, it is useful to think that the outcomes are the same between any two z, z' as long as $f(z) = f(z')$.

— effective treatment (Manski, 2009), exclusion restriction, etc.

Examples of treatment exposure:

- $f_i(z) = z_i$. Standard setting. No interference.
- $f_i(z) = z_i + \gamma \sum_{j \in \text{household}_i} z_j$. Clustered design.
- $f_i(z) = z_i + \gamma \sum_{j \in \text{city}_i} z_j / |\text{city}_i|$. Saturation design.
- $f_i(z) = (z_i, z_{\text{household}_i})$. Multivalued exposure.

Wait, could I just fit a regression?

Indeed, a popular approach is to fit:

$$Y_i = \alpha + \beta Z_i + \gamma \underbrace{f_i(Z)}_{\text{exposure}} + \delta' X_i + \epsilon_i.$$

- As before, model specification is crucial.
- $f_i(Z)$ may have a complex correlation structure with other covariates, and possibly an underlying network.
- Cannot accurately quantify uncertainty, in general.
(cf. simple linear example in the introduction)
- Asymptotics on $\hat{\gamma}$ may well be intractable.

Finally, it is not uncommon to use a model with Y s on the “left and right” of the regression. This is almost **never a good idea**. (Angrist, 2019)

Example 1 - Hypotheses for spillovers

In many settings, we want to test whether the exposures in a set \mathbb{F}_0 are equivalent.

This may be expressed as:

$$H_0 : Y_i(z) = Y_i(z') \text{ for all } i, z, z' \text{ st } f_i(z), f_i(z') \in \mathbb{F}_0.$$

(Manski, 2009), (Aronow, 2012), (T. and Kao, 2013), (Bowers et al., 2013), (Athey et al., 2019), (Basse et al, 2019), (Puelz et al, 2021).

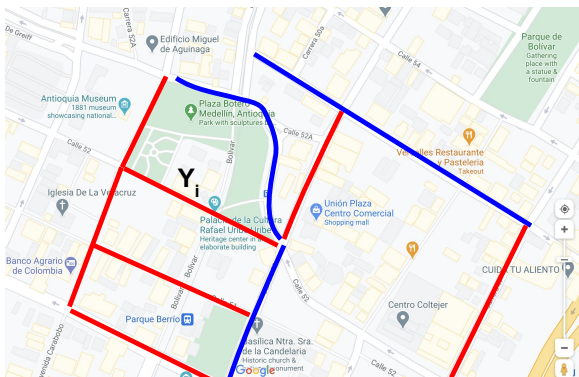
When $\mathbb{F}_0 = \mathbb{F}$ then the problem reduces to the classical FRT.

If $\mathbb{F}_0 \subset \mathbb{F}$ we run into problems. (the null is “weak”)

I will illustrate with the Medellin example. (Collazos et al, 2019).

Illustration from Medellin

Crime spillovers from nearby **treated streets** on **control streets**?

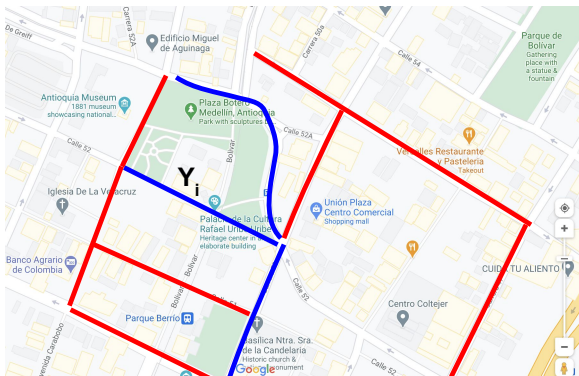


Here, $\mathbb{F}_0 = \{ \text{"control-spillovers"}, \text{"pure-control"} \}$ where

- "control-spillovers" : $z_i = 0$ and $\sum_{j:d(i,j)<125m} z_j > 0$;
- "pure-control" : $z_i = 0$ and $\sum_{j:d(i,j)<125m} z_j = 0$.

FRT problems under interference

Suppose we resample z' in the FRT as shown below:



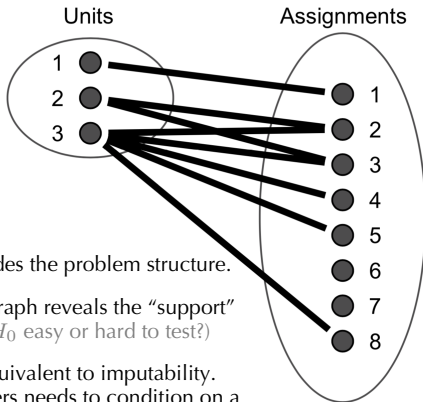
The exposure of i is not in \mathbb{F}_0 . Thus, $Y_i(z')$ cannot be imputed under H_0 .

Main insight of recent literature: We have to condition on a subset of units/assignments where imputation is possible —“focal units” in (Athey et al, 2019).

Conditioning the FRT for spillovers

Puelz et al. (2021) developed a general method to construct such valid conditioning for FRTs under spillovers.

Connect every pair (i, z) iff $f_i(z) \in \mathbb{F}_0 \Rightarrow$ **null exposure graph**.



- The NE graph encodes the problem structure.
- The density of the graph reveals the “support” for testing H_0 . (is H_0 easy or hard to test?)
- An edge in NE is equivalent to imputability. Any FRT for spillovers needs to condition on a **biclique** of the graph.

FRT for spillovers

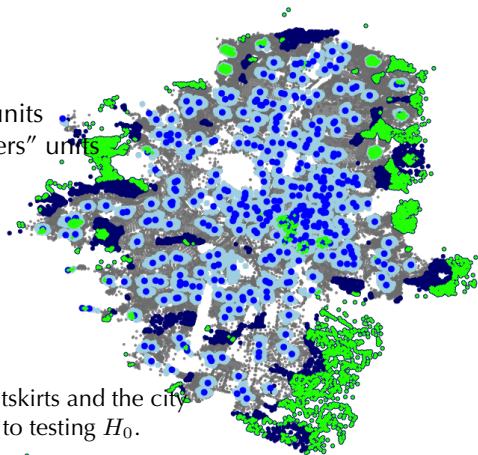
This leads to the following **modifications** of the classical FRT.

- To test H_0 : “Are exposures in \mathbb{F}_0 equivalent?”
 - 1 Calculate NE graph. This is uniquely determined by the H_0 being tested.
 - 2 Calculate a “biclique decomposition” of NE.
Let C be the one that contains the realized assignment, Z , and
 U = units in C ; (focal units)
 $\tilde{D} = D(z|C)$ = design conditional on assignments of biclique.
 - 3 Choose test statistic, $t(z, y)$ using only units in U .
 - 4 Build randomization distribution: $F_R = \{t(z', Y) : z' \sim \tilde{D}\}$.
 - 5 $\text{pval} = 1 - F_R(t(Z, Y))$.

This inherits all the nice properties of classical FRTs in testing for spillovers.

Medellin application

- treated units
- “pure-control” units
- “control-spillovers” units
- focal units



Only units in the outskirts and the city center are pertinent to testing H_0 .

The picture reveals a complex conditioning structure for this particular H_0 .

A regression approach uses all data, even from units not pertinent to H_0 . Its validity **crucially relies** on correct specification.

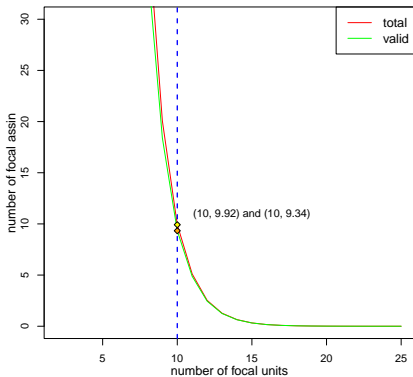
Spin-off 1: Diagnostic

This gives us an idea to “warn” the user when H_0 is hard to test.

Spin-off 1: Diagnostic

This gives us an idea to “warn” the user when H_0 is hard to test.

Example: “Effects of a large-scale social media advertising campaign on holiday travel and COVID-19 infections: a cluster randomized controlled trial” (Breza et al, 2021)



Spin-off 2: Improving the experimental design

We could use the NE graph to optimize the experimental design for a given null hypothesis, H_0 .

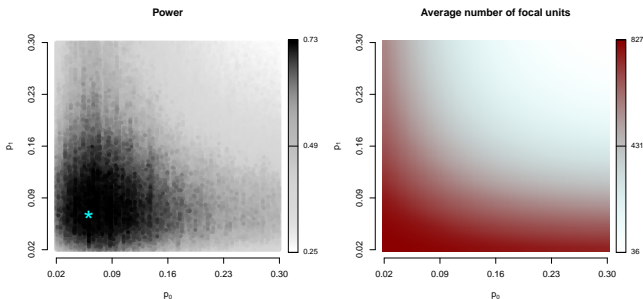
Spin-off 2: Improving the experimental design

We could use the NE graph to optimize the experimental design for a given null hypothesis, H_0 .

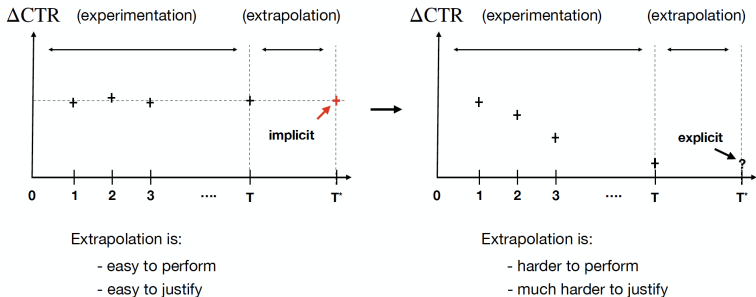
Example: Suppose a design space= $(p_0, p_1) \in [0, 1]^2$ where p_0 =treatment prob. in city-center, and p_1 = treatment prob. in outskirts.

Left: Power calculated under a simulated model for Y over the design space. (darker=higher power).

Right: Average clique sizes in NE graph over the design space.



Example 2: Experiments for long-term effects under learning/habituation



An online service aiming to improve CTR needs to carefully design an experiment to extrapolate for long-term effects.

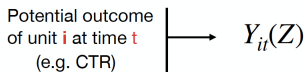
For example, (Honhold et al, 2015) proposed designs to estimate “ad blindness” at Google.

Using potential outcomes

Potential outcomes can serve as a foundation again.

They have a temporal component here.

Let $Y_{it}(Z_i)$ denote the outcome of unit i at time t under assignment $Z_i = (Z_{i1}, \dots, Z_{iT})$, a sequence of treatment from $t = 1$ to $t = T$.



Assumption 1 (no-interference)

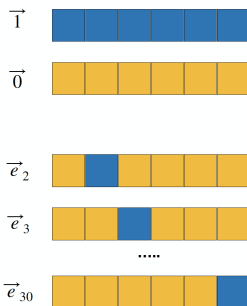
$$Y_{it}(Z) = Y_{it}(Z_i) = Y_{it}(\text{[1: Blue, 2: Yellow, 3: Yellow, ..., T: Blue]})$$

Assumption 2 (non-anticipating outcomes)

$$Y_{it}(Z_i) = Y_{it}(Z_{i:t}) \quad \text{e.g.} \quad Y_{i3}(Z_i) = Y_{i3}(\text{[1: Blue, 2: Yellow, 3: Yellow]})$$

Design space

In this setting, a unit is to be exposed to a sequence of treatments.
This will help us define and estimate habituation effects.



Here, **1** = active treatment at all time points; **0** = control at every t .
 e_t = “pulse treatment” at t . It is $e_t = (0, \dots, 1, \dots, 0)$, i.e., “1” only at t .

Using potential outcomes

The following decomposition is the target of inference:

$$\lambda_t = \frac{1}{N} \sum_i [Y_{it}(\mathbf{1}) - Y_{it}(e_t)], \quad \delta_t = \frac{1}{N} \sum_i [Y_{it}(e_t) - Y_{it}(\mathbf{0})]$$

That is, λ_t = habituation effect, and δ_t = instantaneous treatment effect.

We would like to design an experiment to estimate $\{(\lambda_t, \delta_t)\}_{t=1}^T$.

The “loss function” is simply $L(\theta) = \sum_t (\hat{\lambda}_{t,\theta} - \lambda_t)^2 + (\hat{\delta}_{t,\theta} - \delta_t)^2$.

Here, θ are the experimental parameters, and the “hats” are sample estimators of λ_t, δ_t .

Minimax Design

Theorem (Basse et. al., 2022)





If \mathbb{Y} is permutation invariant, then the minimax design is a completely randomized design assigning units to various treatment arms as follows:


$$\begin{aligned}N_1 &= O(N/\sqrt{T}) \\N_0 &= O(N/\sqrt{T}) \\N_{e_t} &= O(N/T), \quad t = 2, \dots, T.\end{aligned}\tag{1}$$

This result shows that the minimax design needs to be imbalanced in the presence of temporal effects.

For instance, $Z_i = 0$ still gives information about $Y_{it'}(e_t)$ for any $t' < t$ because of the no anticipation assumption.

Example ($T = 30, N = 10000$)

			Balanced CRD		Minimax optimal (also CRD!)
$\vec{1}$		→	322	<<	1040
$\vec{0}$		→	322	<<	1040
\vec{e}_2		→	322	>	273
\vec{e}_3		→	322	>	273

\vec{e}_{30}		→	322	>	273

Optimality gap: May range from $O(1)$ to $O(T)$ depending on the actual outcome model.

Concluding remarks

Causal inference in complex systems is under-developed.

Standard practice does not account for interference, or treatment dynamics, habituation, etc.

But it should!

The methods in this talk aim to address the complexities of some real-world problems.

But these methods are but a tiny sample of what is possible, and have important limitations.

More challenges ahead: Marketplace dynamics, game theory etc.

Thank you!

Basse, Ding, Toulis, “Minimax designs for causal effects in temporal experiments with treatment habituation” (Biometrika, 2022)

Puelz, Basse, Feller, Toulis “A graph-theoretic approach to randomization tests of causal effects under interference” , (JRSS-B, 2021)

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

Toulis and Parkes, “Long-term causal effects via behavioral game theory” (NIPS, 2016)