Experiment Design in a Large Interfirm Network

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

Econometrics and Statistics
University of Chicago, Booth School of Business

Ongoing work with M Best, F Grosset (Columbia), A Shaikh, S Huang (U Chicago)

SNAB 2023, Anchorage, Alaska

Overview

This project aims to study how firms respond to tax audits.

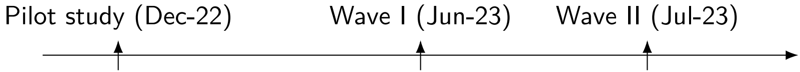
Of particular interest is to understand whether spillover effects exist.
e.g., "Do firms that are *not audited* respond to tax audits on other firms?"

We have the ability to run an experiment in collaboration with experts.

Experimental study

Our population is virtually the totality of businesses in a country of South America:

- We will run an experiment in collaboration with experts and the official Tax Authority. Randomize only **2,000 tax audit notices** (treatment).
- Data: Aggregate reported sales and purchases from both individual firms and interfirm transactions.



Intervention

[redacted]

Nota de Control Masivo [redacted]

[redacted]

De nuestra consideración:

Tenemos el agrado de dirigirnos a Usted, con el fin de informarle que la Dirección General de Fiscalización Tributaria, en virtud de sus facultades dispuestas en el artículo [redacted] de la Ley [redacted], se encuentra ejecutando procesos de controles masivos a contribuyentes que registran inconsistencias originadas en diversos cruces.

En este sentido, se han detectado diferencias entre las ventas o ajustes de débitos fiscales declaradas por Usted en su(s) Declaración(es) Jurada(s) del IVA donde registra montos menores a los informados por sus clientes en sus libros de compras [redacted].

Conforme a los resultantes de cruces y análisis tributarios resultantes del Big Data, esta inconsistencia, representa una irregularidad en la cuantía de ventas o ajustes de débito fiscal registrados por Usted. Es importante mencionar que los análisis de Big Data son realizados sobre todas las informaciones disponibles en la Administración Tributaria y el conocimiento de los perfiles de riesgo conforme a las detecciones de irregularidades producto de controles puntuales.

Por ello se solicita la revisión y rectificación de sus declaraciones de ventas y/o ajustes de débitos fiscales **exclusivamente de los rubros indicados**, en un plazo no mayor a 10 (diez) dias hábiles, a partir de la notificación de la presente, conforme a los detalles expuestos en el siguiente cuadro:

PERÍODO FISCAL		VENTAS S/ DDJJ DE IVA		VENTAS S/ LIBRO DE COMPRAS [redacted] DE CLIENTES INFORMANTES (*)		VENTAS O AJUSTES NO DECLARADOS		
AÑO	MES	[redacted]	[redacted]	[redacted]	[redacted]	GRAV 10	GRAV 5	TOTAL
[redacted]		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

(*) EL DETALLE DE LOS DATOS DE LAS FACTURAS QUE RESPALDAN LAS COMPRAS INFORMADAS POR EL CLIENTE SE ENCUENTRA EN ANEXO ADJUNTO A LA PRESENTE NOTA

En caso de constatar alguna irregularidad en las exposiciones, remitir sus observaciones **únicamente** desde la dirección de correo declarado en el [redacted]

Atentamente,

[redacted]

***ACLARACIONES IMPORTANTES*:**

"According to the results of cross-referencing and tax analysis derived from Big Data, this inconsistency represents an irregularity in the amount of sales or registered fiscal debit adjustments by you."

Setup

We have N firms indexed by $i = 1, \ldots.\ N \approx 478,000$.

- $\mathbf{C} = (C_{ij})$: C_{ij} = total purchases (\$) of firm i (buyer) from firm j (seller).
  - **Reported by i (buyer). Large $N \times N$ matrix.**
- $\mathbf{V} = (V_{ij})$. Same as **C** but from seller's perspective.
- $S = (S_j)$: S_j = total sales (\$) of firm j.
  - **Reported by j (self-report).**

From these we define the (symmetric) inter-firm network :

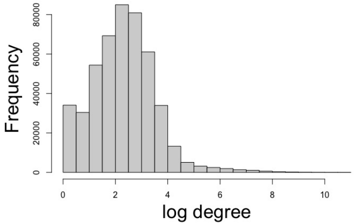
$$'\mathbf{A} = \mathbf{C} + \mathbf{V}'.$$

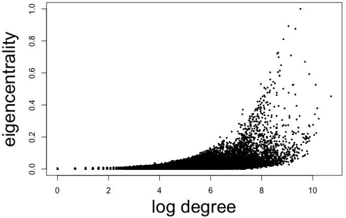
- $\mathcal{N}_i = \{j : A_{ij} = 1\}$, neighborhood of i;
- $\text{deg}_i = \sum_j A_{ij}$, degree of i; etc.

These are sensitive, anonymized tax data. No covariates.

Summary statistics

- $\sim$0.5m nodes (businesses)
- $\sim$8.7m edges (transactions between pairs of firms); density $=$ 0.0076%
- Degree distribution: $(Q_1, Q_2, Q_3, \text{max}) = (5, 10, 21, 45000)$





- Eigenvalue centrality has roughly a linear relationship with degree (as expected).
- However, we also observe **significant heterogeneity**.

Interfirm Network

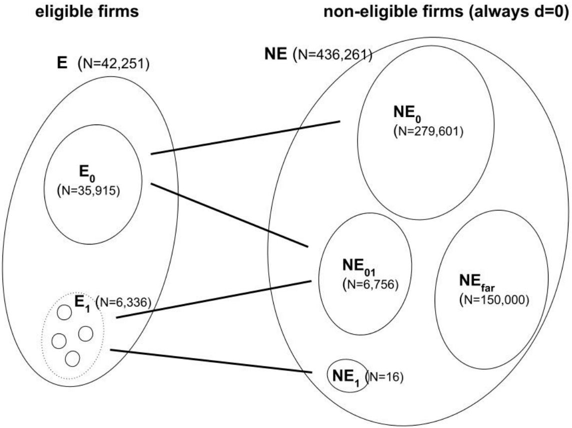


Figure: Main subgraphs of the firm transaction network

Primary outcome

In a perfect world, there is complete agreement between all account books.
No agreement indicates possible tax evasion.

Outcomes are denoted by $Y = (Y_j)$ where

$$Y_j := Y_j(\mathbf{C}, S) = \sum_i C_{ij} - S_j.$$

$$= \text{total purchases reported from buyers of } j - \text{sales reported from } j. \quad (1)$$

A firm j has an incentive to under-state its S_j (less revenue $\rightarrow$ less tax).
A firm i has an incentive to over-state its C_{i*} (more expenses $\rightarrow$ less tax).

Both lead to $Y > 0$.

A firm is *eligible* for treatment if $Y_j^{\text{pre}} > 0$ measured in a pre-2022 tax survey.

Approaches to experiment design on networks

- Typically, researchers aim to split the network into 'clusters', and randomize treatment within clusters with different intensities; e.g., two-stage designs, saturation designs.(Hudgens and Halloran, 2008), (Crepon et al, 2013), (Ugander et al, 2013), (Bakshy et al, 2015), (Eckles et al, 2017).
- Optimal design is heavily model-based: Assume a model for Y wrt treatment, covariates, then minimize standard error of estimators ("D-optimality"); e.g., see (Baird et al, 2017).

Our approach

Our network is too big for common clustering algorithms.

More importantly, we want to be **model-agnostic**.

To that end, we first build procedures (randomization tests) that are **finite-sample exact**; i.e., valid for any finite $n > 0$.

Then, use the procedures to inform the design space P_θ. Use models **only to** construct alternative hypotheses (power calculations).

Causality through Potential Outcomes

We adopt the potential outcomes framework (Neyman, 1923), (Rubin, 1974).

For any fixed treatment vector $d \in \{0, 1\}^N$, the potential outcome of unit j is

$$Y_j(d) := Y_j(\mathbf{C}(d), S(d)) = \sum_i C_{ij}(d) - S_j(d).$$

Causality through Potential Outcomes

We adopt the potential outcomes framework (Neyman, 1923), (Rubin, 1974).

For any fixed treatment vector $d \in \{0,1\}^N$, the potential outcome of unit j is

$$Y_j(d) := Y_j(\mathbf{C}(d), S(d)) = \sum_i C_{ij}(d) - S_j(d).$$

The *randomized* treatment vector is denoted $D \in \{0,1\}^N, \;\; D \sim P(D)$.
(assume P known for now, will discuss design of P soon)

$D_j = 1$ if firm j receives tax audit notice.

Only one potential outcome is **observed**: $Y = Y(D)$. All other treatment/outcomes remain' counterfactual (D', Y').

Potential outcomes are useful to (i) define the causal problem, (ii) separate the problem from the model, (iii) clarify assumptions.

Interference

In classical causal inference, there are only two potential outcomes for control-treatment: "$Y_j(0), Y_j(1)$".

This is unrealistic here. In fact, a key problem is to estimate spillover effects.

Interference

In classical causal inference, there are only two potential outcomes for control-treatment: "$Y_j(0), Y_j(1)$".

This is unrealistic here. In fact, a key problem is to estimate spillover effects.

Assumption ("neighborhood interference")

For any treatment vector $d \in \{0,1\}^N$ *and firm* j

$$Y_j(d) := Y_j(\underbrace{d_j}_{\textit{own}}, \underbrace{d_{\mathcal{N}_j}}_{\textit{neighbors}}).$$

That is, the treatment of k-hops away neighbors ($k > 1$) does not matter for the potential outcome of any firm.

Known as "effective treatment" or "exposure mapping" (Manski, 2013), (T. and Kao, 2013) (Aronow and Samii, 2017).

Null hypotheses

Potential outcomes are useful to express the scientific questions:

a *"Does the treatment has any effect whatsoever?"*

$$H_0^{\text{global}} : Y_j(d) = Y_j(d'), \text{ for all } j, d, d'. \quad (2)$$

b *"Is there a direct effect?"*

$$H_0^{\text{dir}} : Y_j(1, d_{\mathcal{N}_j}) = Y_j(0, d'_{\mathcal{N}_j}), \text{ for all } j, d, d' \text{ for which } d_{\mathcal{N}_j} = d'_{\mathcal{N}_j}. \quad (3)$$

c *"Is there a spillover effect?"*

$$H_0^{\text{spill}} : Y_j(0, d_{\mathcal{N}_j}) = Y_j(0, d'_{\mathcal{N}_j}), \text{ for all } j, d, d'. \quad (4)$$

Fisherian randomization tests of causal effects

Design priority to use randomization tests. Here is a quick recap of this method (Fisher, 1935).

Consider the global null

$$H_0^{\text{global}} : Y_j(d) = Y_j(d'), \text{ for all } j, \; d, d'. \tag{5}$$

Fisherian randomization tests of causal effects

Design priority to use randomization tests. Here is a quick recap of this method (Fisher, 1935).

Consider the global null

$$H_0^{\text{global}} : Y_j(d) = Y_j(d'), \text{ for all } j,\ d, d'. \tag{5}$$

1. Choose test statistic : $T = t(Y, D, \mathbf{A})$.
2. Randomization p-value based on **resampled** $D' \sim P$ iid:

$$\text{pval} = E[t(Y, D', \mathbf{A}) > T],\ D' \sim P. \tag{6}$$

Fisherian randomization tests of causal effects

Design priority to use randomization tests. Here is a quick recap of this method (Fisher, 1935).

Consider the global null

$$H_0^{\text{global}} : Y_j(d) = Y_j(d'), \text{ for all } j, \; d, d'. \tag{5}$$

1. Choose test statistic : $T = t(Y, D, \mathbf{A})$.
2. Randomization p-value based on **resampled** $D' \sim P$ iid:

$$\text{pval} = E[t(Y, D', \mathbf{A}) > T], \; D' \sim P. \tag{6}$$

- Under the null, $Y' = Y$. So, $t(Y, D', \mathbf{A}) \overset{H_0}{=} t(Y', D', \mathbf{A}) \overset{d}{=} T$ by design.
- The key condition is that we can impute all counterfactual outcomes under the null hypothesis. The "null is sharp".

Why Fisherian Randomization Tests (FRTs) ?

Unique advantages of FRTs:

- Nonasymptotic. P-value is **exact** for any finite sample $n > 0$.
- Model-agnostic. Valid for any $t(\cdot)$ and can use any complex model (e.g., regression, ML, network model, etc.)
- Robust. Same answer under reasonable transformations of the data.
- Inference. We can invert the tests for exact inference; c.f. "conformal prediction" (same idea but for prediction intervals).

Main critique of FRT:

- Usually test only "strong nulls" (important advances recently).
- Cannot generalize out of sample.

Randomization tests under interference

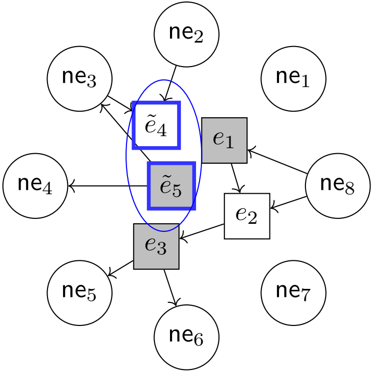
Under interference, it is not easy to test nulls such as H_0^{main} or H_0^{spill}.

This is because these don't imply $Y' = Y$ for all units as before (i.e., non-sharp)

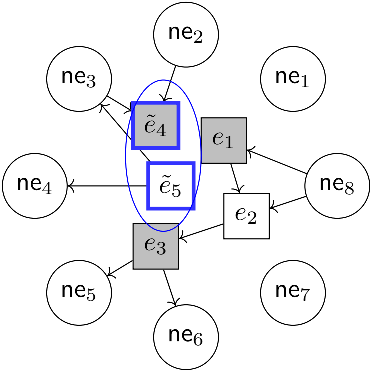
Recent literature suggests to condition on a subset of units / assignments such that $Y'_S = Y_S$ holds for the subset S. (Aronow, 2012), (Athey et al, 2019), (Basse et al, 2019), (Puelz et al, 2022)

Just consider only keeping those "focal units" (in S) and discarding the rest. Then, run standard FRT on the remaining data.

Test for the direct effect: Illustration



Observed treatment



Resampled treatment

The focal units are included in the blue ellipse $(\tilde{e}_4, \tilde{e}_5)$. These form an anticlique. Our test for H_0^{dir} simply permutes the treatment of these units.

Experimental designs

These randomization tests assume the experiment design P is known.
We now discuss the design of P.

We have considered three types of designs:

1. Bernoulli design.
2. Cluster randomization (by firm degree).
3. "SplitGraph". Our current design taking into account the particular problem structure (e.g., anticliques).

We optimize each design under a fixed *counterfactual model.*

Counterfactual models

For purchases:

$$\mathbf{C}(d) = \mathbf{Diag}(1 + \beta * 1\{w(d) = 1\} + \epsilon)\mathbf{C}(0)$$

- An *untreated* firm that is connected to a treated firm increases its purchase reports by β_j% compared to control (spillover)

For sales:

$$S(d) = S^{\text{pre}} + d * \gamma * 1\{Y^{\text{pre}} > 0\} * Y^{\text{pre}} + \epsilon.$$

- A treated firm that misreported at baseline reduces its sales discrepancy by γ_j% compared to control (direct).

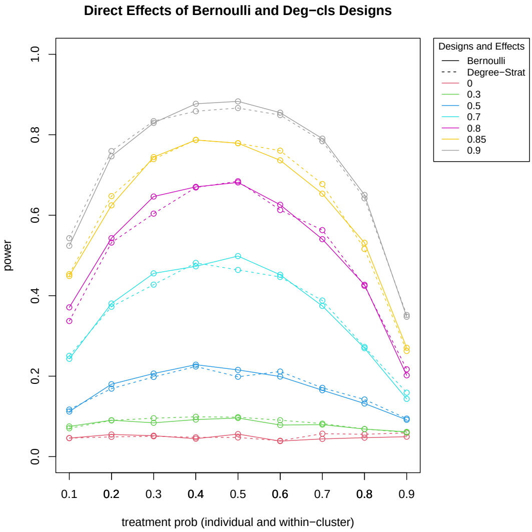
Bernoulli design

- Bernoulli(p): Treatment $P(D_i = 1) = p$ independently for each firm i.
- Clustered version: We first cluster the firms by different methods (e.g., Leiden algorithm). Within each cluster, we run a Bernoulli(p) experiment.

Trade-off between testing for direct effects and testing for spillovers.

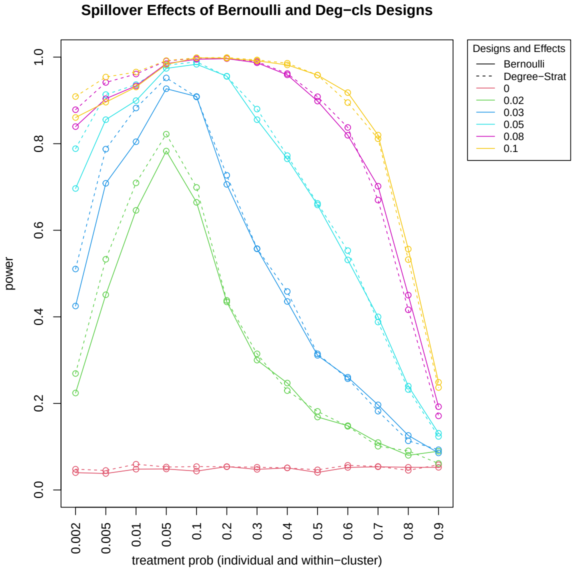
The "sweet spot" (through simulation) for the simple Bernoulli design was roughly at $p^* = 20\%$.

Power for direct effect



Power quadratic in p. Clustering does not help.

Power for spillover effect



Threshold after which network is “saturated” with spillovers. Trade-off.

"SplitGraph" — $(\mathbf{N}_t, m, p)$

Our current design. It takes into account the test structure (e.g., using anticliques for main effects).

- Firms have a 2-dimensional type:

  type = (degree, % connections to other eligibles).

- For each type t we choose N_t such that $\sum_t N_t = 2,000$ (constraint from Tax Authority). See Table below.
- For each type t, we calculate **an anticlique** via a greedy method. The max size of the anticlique is controlled by a parameter (m).
- We treat $N_t p$ within the anticlique, and $N_t(1-p)$ of the rest, completely at random.

We optimize (N_t, m, p) via a random space filling design.

Treatment schedule per type

type = (degree, % connections to other eligibles).

$(\mathbf{N}_t) =$

	out	neutral	in
small	900 (8.7%)	80 (5.0%)	99 (9.5%)
medium	750 (3.6%)	20 (3.2%)	15 (28.3%)
high	32 (1.9%)	0	0
very-high	4 (5.1%)	0	0

Row-type: degree quartile

Column-type: % connection to other eligibles.

Model-assisted optimization

Consider perturbations $\mathbf{N}_0^{(a,b,c)}$ of original schedule $\mathbf{N}_0$ according to

$\mathbf{N}_0^{(a,b,c)} =$

	out	neutral	in
small	900 + 100a - 10c	80 + 10c	99
medium	750 -100a - 20b -10c	20 + 10c	15
high	32 + 20b	0	0
very-high	4	0	0

Idea is to follow a random space filling design:

1. Sample $a_k \in [-4, 4]$, $b_k \in [0, 10]$ and $c_k \in [0, 5]$, i.i.d. uniformly.
2. Simulate experiment and randomization test. Obtain test decision F_k.
3. Fit classification model $F_k \sim (a_k, b_k, c_k)$.

Model-assisted optimization

Logistic regression and classification trees largely agree on the following.

For direct effect:

- (i) Preferable to have large anticlique sets (larger c is better) treated with small/medium intensity.

For the spillover effect:

- (i) Best to have small anticlique sets treated with high intensity. Interactions between these two parameters (m, p) are important.[1]

- (ii) Larger c is clearly worse. This means that we should **not treat** any more "high-degree" firms. Baseline values (a=0, b=0) are fine.

[1] The best pairs are (m=1, p=0.7) or (m=5, p=0.8). These lead, respectively, to increases of 22.8% and 20.5% in power.

Concluding remarks

- Randomization tests under interference need to target specific network structures (e.g., anticliques).
- Our experiment design takes that into account ("SplitGraph").
- Model-assisted optimization can be very useful. Flexible ML models can be used in conjunction with space filling designs.

In general, there are many opportunities at the intersection of modern ML/optimization and experiment design, especially in complex settings such as causal inference in networks.