Causal inference on endogenous social network formation

Maximilian Kasy

Department of Economics, University of Oxford

June 2025

(with Elizabeth Linos and Sanaz Mobasseri)

Introduction

- Social networks are everywhere, and they are consequential.
- How do network ties form?
  - Based on exogenous factors (e.g. shared characteristics, place).
  - Based on existing ties (e.g. triadic closure).
- Causal identification for network formation is hard.
  - Unobservables, reverse causality, equilibrium.
- Statistical inference for networks is conceptually subtle.
  - Many small networks? Sampling from a large network?

Empirical example

- Network of employees at a global professional services firm.
- Edges $\approx$ employees working together.
- Employees choose their collaborators.
- *Random initial assignment*:
  Within offices, new hires are randomly assigned to teams.
- *Network dynamics:*
  We observe the evolution of the network over time.

Setup and notation

- Time periods $t = 1, 2$, individuals $i, j \in \{1 \dots n\}$.
- Adjacency matrices A^t with $A^t_{ij} \in \{0, 1\}$.
- Structural (causal) relationship $A^2 = f(A^1)$.
- Randomization of initial network: A^1 uniform from $\mathcal{A}$.
- Design-based identification and inference:
  - We condition on sample $\{1 \dots n\}$, and on potential outcomes f.
  - Only source of randomness: Sampling of A^1 from $\mathcal{A}$.

Assumption 1

- *Structural relationship*: $A^2 = f(A^1)$.
- *Panel data*: Both A^1 and A^2 are observed.
- *Randomization*: $P(A^1 = A|f) = \frac{1}{|\mathcal{A}|}$ for all $A \in \mathcal{A}$.
- *Exclusion restriction*: $d(A^1) = d \Rightarrow y(f(A^1)) = Y^d$.
- *Support*: $P(d(A^1) = d) > 0$.

Identification: Inverse probability weighting

- Denote $Y = y(A^2)$ and $D = d(A^1)$.
- Define the IPW estimator
$$\widehat{Y}^d = Y \cdot \frac{\mathbf{1}(D = d)}{P(D = d)}.$$
- Under Assumption 1,
$$E\left[\widehat{Y}^d | f\right] = Y^d.$$

Examples

- Outcome: Presence of a tie between i ad j.

$$Y = A^2_{ij}$$

- Treatments: $D = d_{ij}(A^1)$.
    - *Triadic closure*: Presence of an indirect tie between i and j.

$$D = \mathbf{1}\left(\sum_k A^1_{ik} A^1_{kj} > 0\right)$$

    - *Matthew principle*: Degree of node i.

$$D = \sum_k A^1_{ik}$$

- Randomization:
  $\mathcal{A}$ is the set of matrices obtained by swapping new hires within offices.

Assumption 2

- *Permutations*:
  - For permutation π of $\{1, \ldots, n\}$, A_π id the matrix with entries $A_{\pi(i),\pi(j)}$.
  - Let Π be an algebraic group of permutations. The set $\mathcal{A}$ is given by

$$\mathcal{A} = \{A_\pi : \pi \in \Pi\}.$$

- *Equivariance*:
  - For all $\pi \in \Pi$, $\mathcal{E}$ is invariant under π,

$$(i, j) \in \mathcal{E} \Rightarrow (\pi(i), \pi(j)) \in \mathcal{E}.$$

  - For all $\pi \in \Pi$ and $(i, j) \in \mathcal{E}$, d_{ij} is equivariant under π,

$$d_{ij}(A_\pi) = d_{\pi(i),\pi(j)}(A).$$

Weighted linear regression

- Denote $p_{ij}(d) = P(D_{ij} = d)$, and $P_{ij} = p_{ij}(D_{ij})$.
- Define

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{(i,j)\in\mathcal{E}} \frac{1}{P_{ij}} \left(Y_{ij} - D_{ij} \cdot \beta\right)^2,$$

$$\beta = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \left[\left(\sum_{d\in\mathcal{D}} d \cdot d' \right)^{-1} \cdot \left(\sum_{d\in\mathcal{D}} Y_{ij}^d \cdot d \right) \right].$$

- Under Assumptions 1 and 2,

$$E[\widehat{\beta} | f] = \beta.$$

Sample average treatment effect

- Special case: Binary treatment.
- $D_{ij} = (1, X_{ij})$, with $X_{ij} \in \{0, 1\}$.
- Then
$$\beta_2 = \sum_{(i,j)\in\mathcal{E}} (Y_{ij}^1 - Y_{ij}^0)$$
is the sample average treatment effect.
- Support condition: For all $(i, j) \in \mathcal{E}$,
$$0 < P(X_{ij} = 1) < 1.$$
- Example: Triadic closure.

Randomization inference

- Consider the null hypothesis that Y_{ij}^d does not depend on d, for all $(i, j) \in \mathcal{E}$.
- For $\pi \in \Pi$, define the permuted estimator

$$\widehat{\beta}_\pi = \underset{\beta}{\operatorname{argmin}} \sum_{(i,j)\in\mathcal{E}} \frac{1}{P_{\pi(i)\pi(j)}} \left(Y_{ij} - D_{\pi(i)\pi(j)} \cdot \beta\right)^2,$$

  and the p-value

$$p = \frac{1}{|\Pi|} \sum_{\pi\in\Pi} \mathbf{1}\left(\widehat{\beta} \leq \widehat{\beta}_\pi\right).$$

- Under the null, given Assumptions 1 and 2,

$$P(p \leq \alpha) \leq \alpha.$$

Computational implementation (1)

- Our application: New hires $i \in \mathcal{I}$ are randomly permuted within an office.
- Potential ties: Defined based on support requirement.

$$\begin{aligned} \mathcal{J} &= \{j \notin \mathcal{I} :\ \forall d \in \mathcal{D}\ \exists i \in \mathcal{I} : d_{ij}(A^1) = d\}, \\ \mathcal{E}^{max} &= \mathcal{I} \times \mathcal{J}. \end{aligned}$$

- Data can be stored in matrices of dimension $|I| \times |J|$:

$$Y = (A^2_{ij})_{i \in \mathcal{I}, j \in \mathcal{J}}, \quad D = (D_{ij})_{i \in \mathcal{I}, j \in \mathcal{J}},$$

where $D_{ij} = d_{ij}(A^1)$.

Computational implementation (2)

- Assignment probabilities:

$$p_{ij}(d) = \tfrac{1}{|\mathcal{I}|} \sum_{i'} \mathbf{1}(D_{i'j} = d), \qquad P_{ij} = p_{ij}(D_{ij}). \tag{1}$$

- "Instrument:" $Z_{ij} = \frac{1}{P_{ij}} D_{ij}$.
- Weighted regression:

$$C = \left(\sum_{i \in \mathcal{I}, j \in \mathcal{J}} Z_{ij} \cdot D'_{ij} \right)^{-1}, \quad B_{i,i'} = C \cdot \left(\sum_{j \in \mathcal{J}} Z_{ij} \cdot Y_{i'j} \right), \quad \widehat{\beta} = \sum_{i \in \mathcal{I}} B_{i,i}. \tag{2}$$

Computational implementation (3)

- Permutations π:
  Column j remains constant, any permutation of the rows i is allowed.
- Randomization inference:
  - Do not need to re-calculate the terms C and $B_{i,i'}$.
  - The permuted estimator $\widehat{\beta}_\pi$ is simply given by

$$\widehat{\beta}_\pi = \sum_{i \in \mathcal{I}} B_{\pi(i),i}.$$

Setup

Identification and inference

Empirical application

Discussion

Empirical setting

- Global firm in the professional services industry.
- Entry-level employees hired straight from degree programs.
- Work in project teams. Tie $\approx$ working together.
- Initial team assignment determined by an HR manager.
  Random within offices.
- Later team assignment based on an internal labor market.
  Junior employees aim to be recruited by senior colleagues.

Sample characteristics

Variable	Mean	Std dev
Tie formed	0.0043	0.0653
Indirect tie	0.3853	0.4867
Discretized degree	0.5071	0.4999
Female	0.4613	0.4985
Black	0.0482	0.2143

Preliminary findings

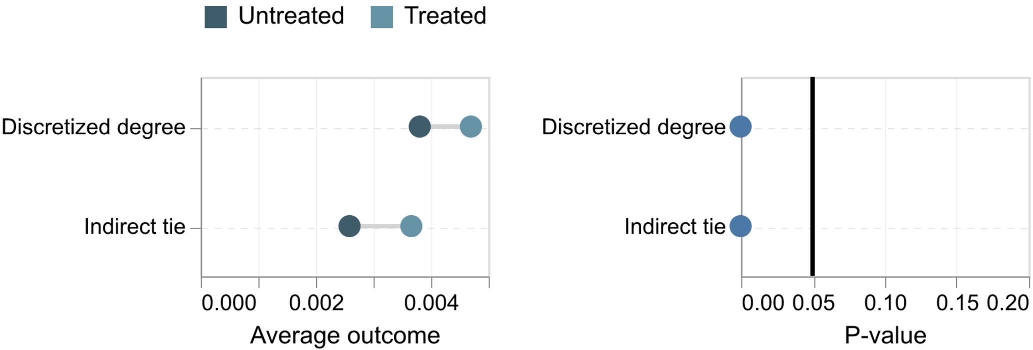
Effect estimates

Treatment	Outcome	Intercept	Effect	P-value	N new hires	N edges
Indirect tie	Tie formed	0.0026	0.0011	0.000	6,042	130,686,467
Discretized degree	Tie formed	0.0038	0.0009	0.000	4,414	105,968,417

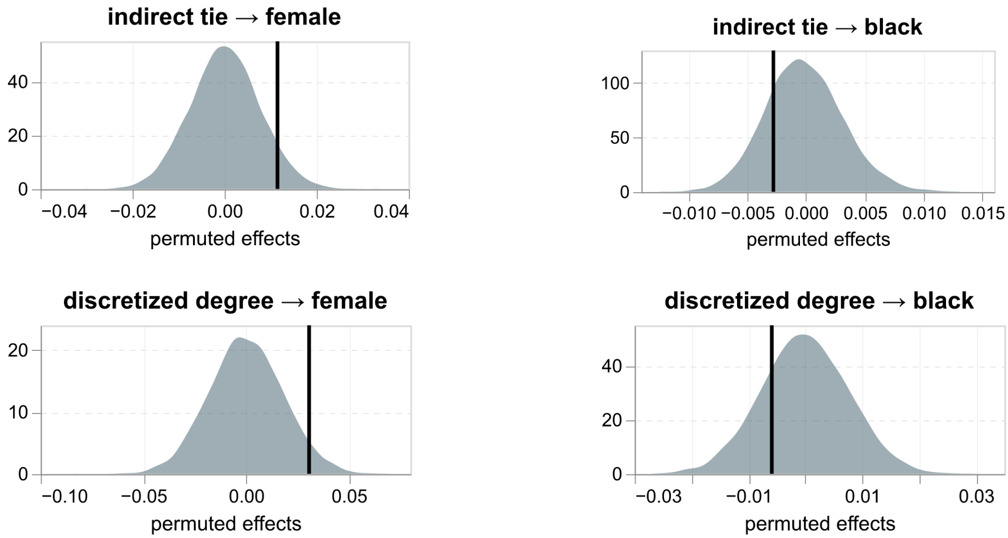
Placebo tests

Treatment	Outcome	Intercept	Effect	P-value	N new hires	N edges
Indirect tie	Female	0.4517	0.0116	0.066	6,042	130,686,467
Indirect tie	Black	0.0505	-0.0028	0.788	6,042	130,686,467
Discretized degree	Female	0.4460	0.0308	0.045	4,414	105,968,417
Discretized degree	Black	0.0554	-0.0058	0.777	4,414	105,968,417

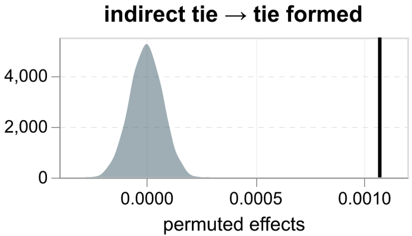
Effect estimates

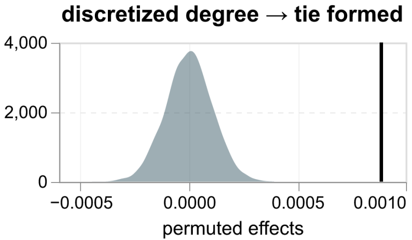


Placebo tests



Effect estimates





Setup

Identification and inference

Empirical application

Discussion

Discussion (1)

- *Design-based inference*:
  - Avoids awkwardness of sampling models for networks:
    Many small networks? Infinite super-network?
  - Avoids arbitrary asymptotics.
  - Based solely on partial randomness of initial ties.
- *Heterogeneity, super-populations, and estimands*:
  - No need to assume treatment effects are the same across ties.
  - No need to assume hypothetical super-population.
  - Inference for variants of a sample average treatment effect.

Discussion (2)

- *Dynamics versus equilibrium*:
    - Equilibrium notions for networks are ambiguous.
    - Who needs to consent to tie formation? Tie disolution?
    - Additionally: Many equilibria - equilibrium selection?
    - Focusing on network dynamics allows us to avoid taking a stance.
- *Confounders and reverse causality*:
    - Confounders: Unobserved heterogeneity can easily lead to patterns like triadic closure, Matthew effect.
    - Reverse causality: Did the tie between $1, 2$ cause the tie between $2, 3$, or the other way around? $\approx$ "reflection problem."
    - Random initial assignment solves both.

Thank you!