netivreg: Estimation of Peer Effects in Endogenous Social Networks

Pablo Estrada[1] Juan Estrada[1] Kim P. Huynh[2,1] David T. Jacho-Chávez[1] Leonardo Sánchez-Aragón[3]

The views expressed in this paper are solely those of the authors and may differ from official Bank of Canada views. No responsibility for them should be attributed to the Bank of Canada.

[1]Emory University, [2]Bank of Canada, [3]ESPOL University

Motivation

- Estimation of network effects is becoming increasingly common.
  - Interest on structural coefficients: endogenous peer effects and contextual effects
  - Estimate treatment effects and spillovers under interference.
- Exogenous network formation is a commonly used assumption in empirical work.
- Recent methods allowing for the presence of network endogeneity require explicit structural restrictions on the network formation process.

Motivation

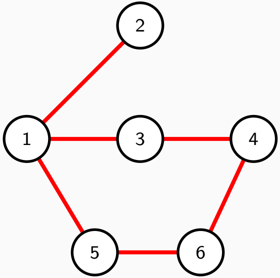
EMORY UNIVERSITY

- Estimation of network effects is becoming increasingly common.
    - Interest on structural coefficients: endogenous peer effects and contextual effects
    - Estimate treatment effects and spillovers under interference.
- Exogenous network formation is a commonly used assumption in empirical work.
- Recent methods allowing for the presence of network endogeneity require explicit structural restrictions on the network formation process.
- **Research Question:** can the *multiplex network data structure* help with the treatment of identification issues?

Preview of the Results

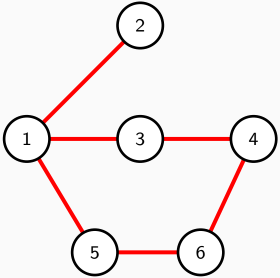
- Propose novel instruments based on the topology of multiplex networks to construct the estimator.
- Provide new identification results for peer/contextual effects that generalize existing methods by accounting for potential endogenous network formation.
- Computationally easy to implement estimator that is consistent and asymptotically normal.
- Stata implementation: `netivreg`.

Framework



- Contextual Effects (interference): i's outcome depends on the characteristics of other units.
- Endogenous Peer Effects (multiplier).

$$y_i = \alpha + \beta \sum_{i \neq j} W_{i,j} y_j + \delta \sum_{i \neq j} W_{i,j} x_j + \gamma x_i + \varepsilon_i.$$



- Contextual Effects (interference): i's outcome depends on the characteristics of other units.
- Endogenous Peer Effects (multiplier).

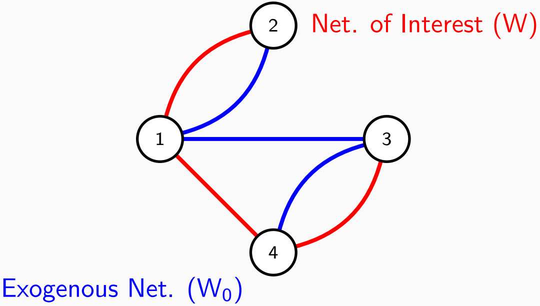
$$y_i = \alpha + \beta \sum_{i \neq j} W_{i,j} y_j + \delta \sum_{i \neq j} W_{i,j} x_j + \gamma x_i + \varepsilon_i.$$

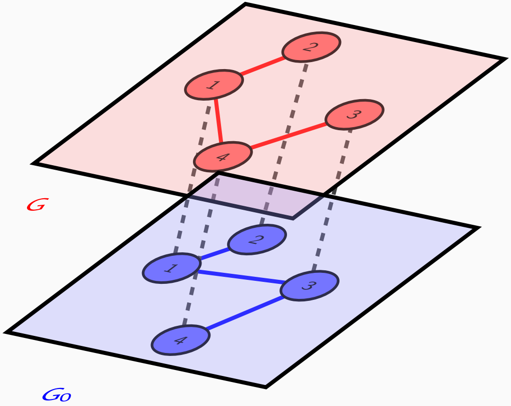
Objective: identify and consistently estimate the parameters $(\alpha, \beta, \gamma, \delta)$.

Main Challenges: Observational Data

- Simultaneity of the peer effects regressors (reflection problem)
- The decision of forming a peer connection can be correlated with unobserved characteristics or there could exists common shocks (correlated effects)
- The network structure could induce correlation between X and ε (unobserved homophily)

$$y = \alpha^0 \boldsymbol{\iota} + \beta^0 Wy + \delta^0 WX\boldsymbol{\delta}^0 + X\boldsymbol{\gamma}^0 + \varepsilon, \text{ with } \mathbb{E}[\varepsilon \mid W, X] \neq 0 \text{ and } \mathbb{E}[\varepsilon \mid W_0, X] = 0.$$





Ideal Experiment

- Individuals are (quasi-) randomized into groups (for example classrooms) defining W_0.

Ideal Experiment

- Individuals are (quasi-) randomized into groups (for example classrooms) defining W_0.
- Only the fact that two individuals share a classrooms does not necessarily generate social effects.

Ideal Experiment

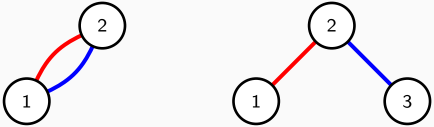
- Individuals are (quasi-) randomized into groups (for example classrooms) defining W_0.
- Only the fact that two individuals share a classrooms does not necessarily generate social effects.
- It is possible to observe a relevant network (for example friendship) defining W.

Ideal Experiment

- Individuals are (quasi-) randomized into groups (for example classrooms) defining W_0.
- Only the fact that two individuals share a classrooms does not necessarily generate social effects.
- It is possible to observe a relevant network (for example friendship) defining W.
- This method can be used to causally estimate network friendship effects.

Main Identifying Assumptions

1. Monolayer Linear model and Bi-layer multiplex network data $\mathcal{M} = 2$ (W and W_0).
2. Conditional distribution $\mathcal{F}(\varepsilon \mid X, \mathcal{M})$ is such that $\mathbb{E}[\varepsilon|W, X] \neq 0$ and $\mathbb{E}[\varepsilon|W_0, X] = 0$.
3. The networks generating the adjacency matrices W and W_0 are correlated in the sense that it is possible to find connections in common ($E_0 \cap E_1 \neq 0$) and distance two paths that change edge type ($(i, j) \in E_0$ and $(j, k) \in E_1$).



Identification

Let $\mathbf{\Pi}$ be the projection coefficients from a regression of WS on $\mathsf{W}_0\mathsf{S}$, where $\mathsf{S} = [\mathsf{y} \quad \mathsf{X}]$.

<u>Theorem:</u>

Let Assumptions 1, 2, 3, and $\gamma^0(\pi_{11}\beta^0 + \pi_{12}\delta^0) + \pi_{21}\beta^0 + \pi_{22}\delta^0 \neq 0$ hold. If the matrices I, W_0, $\mathsf{W}_0{}^2$ are linearly independent, then the parameters $\alpha^0, \beta^0, \gamma^0$ and δ^0 are identified.

Remark

Note that this is a generalization of the identification result in Proposition 1 of Bramoullé et al. (2009, JoE), i.e., if $\mathsf{W}_0 = \mathsf{W}$, one has $\mathbf{\Pi} = \mathsf{I}$, and the condition reduces to $\gamma^0\beta^0 + \delta^0 \neq 0$ and the matrices I, W and W^2 being linearly independent.

Rank Condition

EMORY UNIVERSITY

$$y = \alpha^0 \boldsymbol{\iota} + \mathrm{W}\mathrm{S}\boldsymbol{\theta}^0 + \mathrm{X}\boldsymbol{\gamma}^0 + \boldsymbol{\varepsilon} \quad \text{for} \quad \mathrm{S} = [\mathrm{y} \quad \mathrm{X}] \quad \text{and} \quad \boldsymbol{\theta}^0 = [\beta^0 \quad \boldsymbol{\delta}^0]$$
$$y = \alpha^0 \boldsymbol{\iota} + \mathrm{W}_0\mathrm{S}\boldsymbol{\theta}^* + \mathrm{X}\boldsymbol{\gamma}^0 + \mathrm{e}, \quad \text{for} \quad \boldsymbol{\theta}^* = \Pi\boldsymbol{\theta}^0.$$

Estimation Procedure

1. Estimate Π by OLS (WS on $\mathrm{W}_0\mathrm{S}$).
2. 2SLS of $[\boldsymbol{\iota}, \mathrm{X}, \mathrm{W}_0\mathrm{y}, \mathrm{W}_0\mathrm{X}]$ with instrument $\mathrm{Z} = \left[\boldsymbol{\iota}, \mathrm{X}, \mathrm{W}_0{}^2\mathrm{X}, \mathrm{W}_0\mathrm{X}\right]$. Calculate $\widehat{\boldsymbol{\theta}} = \widehat{\Pi}^{-1}\widehat{\boldsymbol{\theta}}^*$.
3. IV of $\left[\boldsymbol{\iota}, \mathrm{X}, \widehat{\mathrm{W}\mathrm{y}}, \widehat{\mathrm{W}\mathrm{X}}\right]$ with instruments $\widehat{\mathrm{Z}}^* = \left[\boldsymbol{\iota}, \mathrm{X}, \left[E\left(\mathrm{W}_0\mathrm{y}|\mathrm{X}, \mathrm{W}_0\right), \mathrm{W}_0\mathrm{X}\right]\widehat{\Pi}\right]$.

Estimator and Properties

$$\widehat{\psi}_{G3SLS} = \left(\widehat{\mathrm{Z}}^{*\top}\widehat{\mathrm{D}}\right)^{-1}\widehat{\mathrm{Z}}^{*\top}\mathrm{y},$$

$$\sqrt{n}(\widehat{\psi}_{G3SLS} - \psi) \xrightarrow{d} N(0, \mathrm{V}_{\psi})$$

Stata Implementation

Empirical Application: Specification

W: Coauthors - W_0: Alumni

Empirical Application: Specification

W: Coauthors - W_0: Alumni

$$y_{i,r,t} = \alpha + \beta \sum_{j \neq i} w_{\ell;i,j,t} y_{j,r,t} + \sum_{j \neq i} w_{\ell;i,j,t} \widetilde{x}_{j,r,t}^{\top} \delta + x_{\ell;i,r,t}^{\top} \gamma + \lambda_r + \lambda_t + \lambda_0 + \varepsilon_{i,r,t}$$

Peer Effects (β)

log(# Citations)

Contextual Effects (δ)

Editor

Different Gender

Direct Effects (γ)

Editor

Different Gender

Authors

Pages

References

Fixed Effects (λs)

Journal

Year

Institutions Component

Empirical Application: Specification

W: Coauthors - W_0: Alumni

$$y_{i,r,t} = \alpha + \beta \sum_{j \neq i} w_{\ell;i,j,t} y_{j,r,t} + \sum_{j \neq i} w_{\ell;i,j,t} \widetilde{x}_{j,r,t}^{\top} \delta + x_{\ell;i,r,t}^{\top} \gamma + \lambda_r + \lambda_t + \lambda_0 + \varepsilon_{i,r,t}$$

Peer Effects (β)

log(# Citations)

Contextual Effects (δ)

Editor

Different Gender

Direct Effects (γ)

Editor

Different Gender

Authors

Pages

References

Fixed Effects (λs)

Journal

Year

Institutions Component

```
netivreg lcitations editor diff_gender n_pages n_authors n_references isolated
(edges = edges0), wx(diff_gender editor) cluster(c_coauthor) first second
```

Data Structure

Stata 16 Capabilities: (1) Python Integration for Sparse Matrices and (2) Multiframes

W (Coauthors)

	source	target
4	5	478
5	5	665
6	5	705
7	8	113
8	8	133
9	8	177
10	8	189
11	8	639
12	8	658
13	10	356
14	10	527
15	11	26
16	11	639
17	13	213
18	13	428

W_0 (Alumni)

	source	target
4	4	136
5	4	407
6	5	10
7	5	95
8	5	97
9	5	130
10	5	144
11	5	152
12	5	161
13	5	194
14	5	301
15	5	324
16	5	357
17	5	383
18	5	416

(y, X)

	id	lcitations	editor	diff_gender	isolated	n_pages	n_authors	n_references	journal	year
4	21	2.302585	0	0	1	15	2	27	aer	2000
5	31	3.806663	0	0	1	21	1	39	aer	2000
6	38	3.555348	0	0	1	17	2	31	aer	2000
7	51	3.583519	0	0	1	20	1	48	aer	2000
8	59	3.988984	0	0	1	17	2	50	aer	2000
9	68	2.197225	0	0	1	15	2	31	aer	2000
10	76	2.197225	0	0	0	11	2	18	aer	2000
11	86	3.218876	0	0	0	24	1	32	aer	2000
12	96	4.836282	0	0	0	24	2	57	aer	2000
13	105	4.691348	0	0	1	25	2	40	aer	2000
14	122	4.770685	0	0	1	30	1	30	aer	2000
15	139	3.850147	0	0	1	16	1	49	aer	2000
16	144	2.564949	0	0	0	21	2	16	aer	2000
17	151	4.26268	1	0	0	21	3	36	aer	2000
18	162	2.890372	0	0	1	26	1	26	aer	2000

Step1: Projection Coefficients

$$WS = W_0 S\Pi + U,$$

Projection of W on W0

	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
W_lcitations						
W0_lcitations	.4956186	.0321772	15.40	0.000	.4324379	.5587993
W0_diff_gender	.0127121	.5132519	0.02	0.980	-.9950719	1.020496
W0_editor	.0085967	.7897166	0.01	0.991	-1.542033	1.559227
W_diff_gender						
W0_lcitations	.137265	.0033955	40.43	0.000	.1305979	.1439321
W0_diff_gender	.1422822	.0541602	2.63	0.009	.0359371	.2486273
W0_editor	.0325262	.0833338	0.39	0.696	-.131102	.1961544
W_editor						
W0_lcitations	.4249148	.0025367	167.51	0.000	.419934	.4298957
W0_diff_gender	.1027705	.0404624	2.54	0.011	.0233214	.1822195
W0_editor	.1367464	.0622576	2.20	0.028	.0145019	.2589909

Step2: 2SLS estimation

2SLS of $[\iota, X, W_0 y, W_0 X]$ with instrument $Z = [\iota, X, W_0{}^2 X, W_0 X]$

2SLS Regression

Number of obs =	729
Wald chi2(62) =	-1.1e+17
Prob > chi2 =	1.0000
R-squared =	0.1317
Root MSE =	1.846

lcitations	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
W_y						
lcitations	.9496092	.5481734	1.73	0.084	-.126744	2.025962
X						
diff_gender	.2224841	.1317096	1.69	0.092	-.0361313	.4810994
editor	.1691513	.1181452	1.43	0.153	-.06283	.4011327
n_pages	.0282953	.0048171	5.87	0.000	.0188369	.0377538
n_authors	.0747385	.0603238	1.24	0.216	-.043709	.1931859
n_references	.0119404	.0025597	4.66	0.000	.0069143	.0169665
isolated	-.2131575	.0942419	-2.26	0.024	-.3982041	-.0281109

Step3: Optimal Instrument

IV of $\left[\iota, X, \widehat{Wy}, \widehat{WX}\right]$ with instruments $\widehat{Z}^* = \left[\iota, X, [E(W_0 y | X, W_0), W_0 X]\, \widehat{\Pi}\right]$

```
Network IV Regression                                  Number of obs  =        729
Number of clusters (c_coauthor) =     575             Wald chi2(62)  =  6.5e+16
                                                      Prob > chi2    =   0.0000
                                                      R-squared      =   0.1723
                                                      Root MSE       =    1.339
                             (Std. Err. adjusted for 575 clusters in c_coauthor)
```

lcitations	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
W_y						
lcitations	.5200772	.3616317	1.44	0.151	-.1899963	1.230151
X						
diff_gender	.218709	.1305651	1.68	0.094	-.0376592	.4750771
editor	.1733642	.1157379	1.50	0.135	-.0538902	.4006187
n_pages	.0288947	.0044187	6.54	0.000	.0202184	.0375709
n_authors	.0719403	.0597035	1.20	0.229	-.0452891	.1891696
n_references	.0119892	.0025599	4.68	0.000	.0069628	.0170156
isolated	-.2230689	.0897056	-2.49	0.013	-.3992083	-.0469295

Conclusion

- Identification of a linear-in-means model with endogenous network.
- Computationally simple estimation that uses two-layered multiplex network structure with Stata implementation.
- Robust to different types of network endogeneity. It does not require to model unobserved heterogeneity and network formation.

Appendix

Distortions Induced by Social Effects: Consumption Examples

- If individuals care about **status** (conspicuous consumption models), the proportion of conspicuous consumption may increase with respect to other goods.
- If conspicuous consumption is considered wasteful, peer effects might have noticeable welfare consequences.
- Savings may differ from the optimal in an attempt to keeping up with the peers.

▸ Empirical Work

Aggregate Effects: Consumption Example

- Unanticipated tax changes to the rich might have aggregate consequences.
- If individuals who are not affected by the shock change their consumption after observing changes in consumption of the rich, the shock can spread through the network.
- Social multipliers depend on the size of the endogenous peer effects and the connectedness of the affected groups.

▸ Empirical Work

Angrist's (2014) Critique: Group Regressions

- **Reflection Problem:** a regression of individual outcomes on <u>group</u> mean outcomes is tautological.

- **Correlated Effects:** even the leave-one-out estimator does not provide information of human behavior. "Like students in the same school, households from the same village are similar in many ways".

- **Mechanical Relationship:** the coefficient on group averages in a multivariate model of endogenous peer effects does not reveal the action of social forces. He interprets the vale $1/(1-\beta)$ as approximately the ratio of the 2SLS to OLS estimands for the effect of individual covariates on outcomes (using dummy groups as instruments).

▸ Motivation

Angrist's (2014) Critique: Network Regressions

- Start by a saturated model $E[y_i \mid x_i] = \gamma_0 + \gamma_1 x_i$ satisfying $E[u_i \mid x_i] = 0$, for $u_i \equiv y_i - \gamma_0 - \gamma x_i$.

- Individuals are ordered from left to right. Each person i is connected only with the individual to her left $i-1$. Friends are only similar on unobservables: $u_i = \beta u_{i-1} + \varepsilon_i$.

- The outcome can be written in a linear-in-means (lmm) model form:

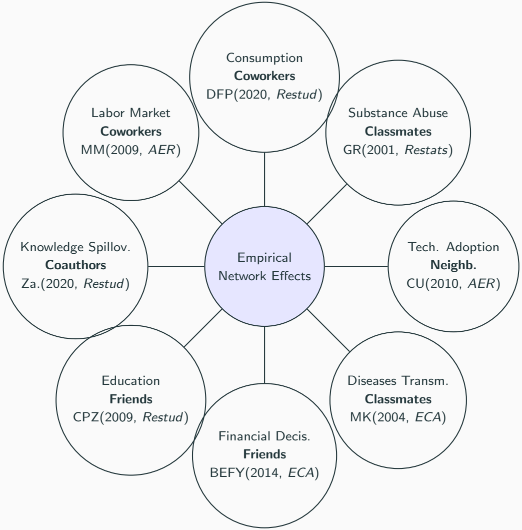
$$y_i = \gamma_0(1-\beta) + \beta y_{i-1} + \gamma x - \beta\gamma x_{i-1} + \varepsilon_i$$

- **Flaw in Angrist's example:** let $\delta = -\beta\gamma$ to write this model exactly as a lmm. Note that $\delta + \gamma\beta = 0$ so that the outcome equation can be written as (for $\alpha = \gamma_0(1-\beta)$)

$$y_i = \frac{\alpha}{1-\beta} + \gamma x_i + v_i$$

Motivation

Different Network Effects



▸ Motivation

Critique to Randomly Assigned Groups

- In principle, randomization of peers would guarantee identification in a monolayer linear in means model where endogenous network formation is ruled out.
- It can completely eliminate the problem of unobserved common variables.
- However, if individuals endogenously form groups (homophily), there can be a subsequent resorting. If resorting happens faster than the effects of social interactions, identification is not possible.
- Even with random peers, researchers face a classical problem of **omitted variables** when trying to estimate contextual effects ($\mathbb{E}[x_i \varepsilon_j \mid w_{i,j} = 1] \neq 0$).

▸ Literature

Multilayers Networks in Economics

Labor Supply

- Sisters, Cousins and Neighbors networks (NST (2018, *AEJ*))

Consumption

- Coworker and Spouses networks (DFP (2020, *Restud*))

Education

- Friendship network in t and $t-1$ (GI (2013, *JBES*))
- Roommates, classmates, Study-mate, Friendship networks (CL (2015))
- Siblings and Classmates networks (NR (2017, *JAE*))

Publication Outcomes

- Coauthors, Alumni and Same Advisor networks (EHJS (2020))

▸ Multilayer

Microfoundations

- The monolayer linear model of interest corresponds with the best response of a Bayesian Game of Social Interactions as proposed by Blume, Brock, Durlauf and Jayaraman (2015, JPE).

- Quadratic utility with social pressure or strategic complementarities

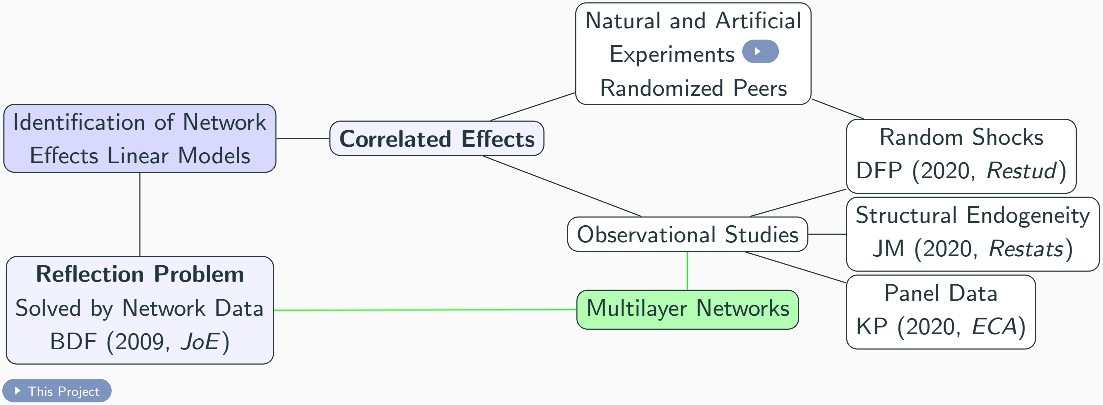
$$U_i\left(\omega_i, \omega_{-i}\right) = \left(\gamma x_i + z_i + \delta \sum_j c_{ij} x_j\right)\omega_i - \frac{1}{2}\omega_i^2 - \frac{\phi}{2}\left(\omega_i - \sum_j a_{ij}\omega_j\right)^2$$

- In their model endogeneity arises because an individual i, observing that he is connected to j, make an inference about the value of z_j that is dependent on x_j. Then, x_j will be correlated with ε_i in my equation of interest.

- Their critique of instrumental variable is that if individual i observe the instruments v_j, he can use it to predict z_j which will induce correlation between ε_i and the instrument.

- Our instrument is based on x_r of individuals r connected to i in a network that is independent of the individuals' utilities. Therefore x_r is not useful to predict z_j. Model

Positioning the Research Agenda in the Literature



Assumptions

Assumption 1

There exists a $n \times n$ adjacency matrix W_0 such that: $\mathbb{E}[v|x, W_0] = 0$

Assumption 2

Let Π be the be the full-rank matrix of coefficients from the system regression

$$WS = W_0 S \Pi + U,$$
$$E\left[U | W_0 y, W_0, X\right] = O.$$

where $\mathbb{E}[S^\top w_{0;i} w_{0;i}{}^\top S] > 0$. Furthermore, the first row of $\mathbf{\Pi}$ is such that $\pi_{11}\beta + \pi_{12}\delta < 1/\lambda_{max}$, where λ_{max} is the largest eigenvalue of W_0.

Identification

Rank Condition

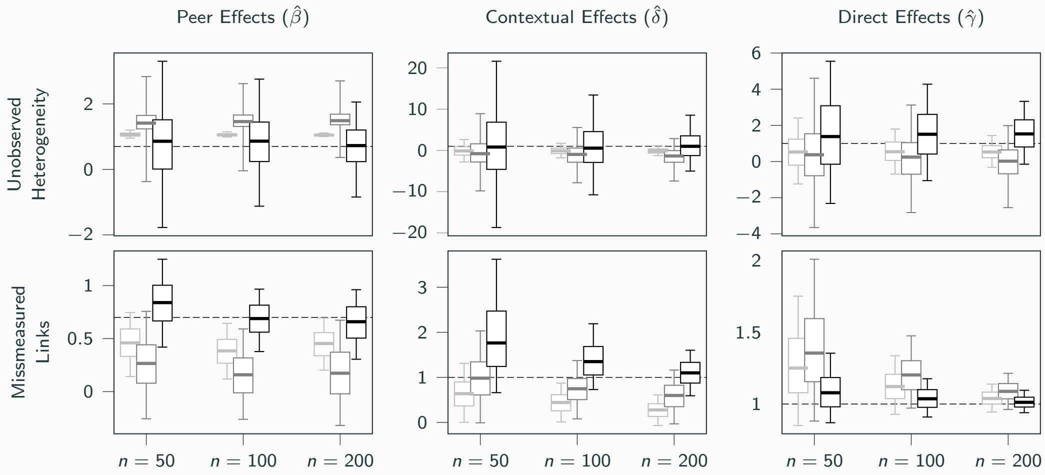
- Given that $\text{rank}(\Pi) \leq \min\{\text{rank}(E[S^\top w_{0;i} w_{0;i}^\top S]^{-1}), \text{rank}(E[S^\top w_{0;i} w_i^\top S])\}$, a necessary condition for $\text{rank}(\Pi) = k + 1$ is that $\text{rank}(E[S^\top w_{0;i} w_i^\top S]) = k + 1$ which would be equivalent to the **relevance** condition in the classical Instrumental Variable literature.

- For large enough sample, this condition imposes some restriction on the matrix $W_0 W$. This matrix contains the connections in common across the two networks in the main diagonal, and length two paths that change color in the off- diagonal.

- It cannot be zero so there have to be enough connections in common and indirect triads that change colors. This is a way to think about the **correlation** between the two matrices.

▸ Identification

Mote Carlo Experiments



Empirial Application: Data

- 1,628 articles published in the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, and the *Quaterly Journal of Economics* between 2000 and 2006. Source: RePEc, Scopus, and Journal Websites.

- Employment, education, and research interest information for 1,985 unique authors and 42 unique editors (37 of which also published papers in thee journals in this time period). Source: Web scrapping/text mining and Colussi (2018, ReStat).

- *Co-authorship* ($\ell = 1$) and *Alumni* ($\ell = 0$) networks are constructed for all 2,027 scholars.

Articles i and j are connected in network W_ℓ if at least one of the authors of article i shares a professional connection of type ℓ with at least one of article j's authors.

▸ Empirical Application