Estimating Social Networks with Missing Links

Arthur Lewbel, Xi Qu, and Xun Tang

Northwestern University, March 2023

Introduction

- In social networks, individual outcomes depend on:
  - own characteristics (*direct* effects)
  - neighbor characteristics (*contextual* effects)
  - neighbor outcomes (*peer* effects)
- In practice, existing network links may be missing from the sample due to:
  - recall errors in survey responses
  - lapses in data input
- Our goal: estimate these effects despite missing links

Introduction

- Conventional 2SLS:
  - Structural form: $y = \lambda Gy + X\beta + \varepsilon$, where G_{ij} indicates whether i and j are linked.
  - Suppose G is perfectly reported in a sample.
  - Peer outcomes Gy are endogenous due to simultaneity.
  - Apply 2SLS using GX or G^2X as instruments for Gy - e.g. Lee (2007), Bramoulle et al (2009)
  - IV exogeneity and relevance are guaranteed if $E(\varepsilon|X, G) = 0$.

Introduction

- How do missing links affect inference?
  - Suppose the sample only reports $H \neq G$, with H randomly missing links from G
  - Feasible structural form: $y = \lambda Hy + X\beta + u$, with $u = \varepsilon + \lambda(G - H)y$
  - Endogenous peer outcomes: Hy correlated with u through measurement errors in H *and* through simultaneity
  - Also, X is now endogenous (correlated with u via y).
  - Hence HX (and H^2X) are not valid IV b/c H and X correlate with u.

Related Literature

- Lee (2007), Bramoulle, Djebbari, and Fortin (2009)
  - introduce conventional IV methods
- Boucher and Houndetoungan (2020)
  - use knowledge (or estimates) of distribution of networks
  - draw networks from the distribution to construct IVs
- Griffith (2021)
  - missing links due to censoring (caps on # of links reported)
  - characterized the omitted variable bias in feasible regression
  - for model with no peer effects, estimate the bias under an *order invariance* condition
- Lewbel, Qu, and Tang (2022): identification when no links are observed

Introduction

- We illustrate the main idea when links are randomly missing at rate $p \in (0, 1)$.
- Adjusted 2SLS:
  - scale Hy by $1/(1-p)$ restores exogeneity of X in feasible structural form
  - find alternative, valid IV for Hy: e.g., $H'X$
  - requires knowledge of p, which can be estimated if there are multiple measures of same links
  - is $\sqrt{n}$-CAN

Introduction

- Extensions:
  - add contextual effects
  - allow for heterogeneous missing rates
  - include group-level fixed effects
- Adjusted 2SLS: works with a single, large network
  - need notion of sparsity or weak dependence
  - e.g., many groups (blocks) with few links across groups, which are not reported

Introduction: Preview of Application

- We apply our method to data from Banerjee, Chandrasekhar, Duflo, and Jackson (2013)
  - surveys from 4,134 households in 43 villages
  - two measures of links imputed ("*VisitCome*" vs "*VisitGo*")
  - dependent variable: participation in microfinance program
  - evidence of missing links: symmetrized measures differ
- Findings:
  - missing rate $p \approx 0.18$
  - "endorsement effect": $\lambda \approx 0.046$. An additional participating neighbor increases own participation by 4.6%.
  - ignoring missing links using traditional 2SLS yields 9% upward bias in λ estimates

Social Network with Missing Links

- Model:
  - A large number of small, independent networks

    $y = \lambda Gy + X\beta + \varepsilon$, $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times K}$, $\varepsilon \in \mathbb{R}^n$, $E(\varepsilon|X, G) = 0$.
  - links $G_{ij} \in \{0, 1\}$ (*not* row-normalized); $G_{ii} = 0$.
  - reduced form: $y = M(X\beta + \varepsilon)$, $M \equiv (I - \lambda G)^{-1}$.
  - data reports H instead of G, with $H_{ii} = 0$.
  - feasible structural form:

$$y = \lambda Hy + X\beta + \underbrace{[\varepsilon + \lambda(G - H)y]}_{u}.$$

Model Assumptions

- (A1) $E(H_{ij}|G, X) = E(H_{ij}|G_{ij}, X)$.
- (A2) Links missing at random:
  - $E(H_{ij}|G_{ij} = 1, X) = 1 - p$ for $p \in (0, 1)$;
  - $E(H_{ij}|G_{ij} = 0, X) = 0$.
- Under (A1)-(A2), $E(H|G, X) = (1 - p)G$.
- Exogeneity: (A3) $E(\varepsilon|X, G, H) = 0$.

Restore Exogeneity of Covariates

- Step 1. Suppose p were known. Reparametrize the feasible structural form:

$$y = \lambda \tfrac{Hy}{1-p} + X\beta + \underbrace{\varepsilon + \lambda \left(Gy - \tfrac{Hy}{1-p} \right)}_{\equiv v}.$$

- (A1)-(A3) imply:
  - $E(Gy|X, G) = GMX\beta$
  - $E(Hy|X, G) = E(H|G, X)MX\beta = (1 - p)GMX\beta$
- Together they imply $E(v|X, G) = 0$.
- In this reparametrized structural form, X is *no longer endogenous*.

Bias in (Unscaled) 2SLS

- Let $R \equiv (Hy, X)$, $Z \equiv (\zeta(X), X)$, where $\zeta(\cdot)$ is nonlinear function of X.
- Suppose:

  (IV-R) $E(Z'R)$ and $E(Z'Z)$ both have full rank.

  Then:

$$y = \frac{\lambda}{1-p} Hy + X\beta + \underbrace{\varepsilon + \lambda\left(Gy - \frac{Hy}{1-p}\right)}_{\equiv v},$$

$$\implies E(Z'y) = E(Z'R)\left(\frac{\lambda}{1-p}, \beta'\right)' + \underbrace{E(Z'v)}_{=0}.$$

- Missing links in H lead to "*augmentation bias*" on peer effects in 2SLS.
- We provide sufficient conditions for the rank condition (IV-R).

Construct Instruments from H

- Recall we can *not* use HX as instruments. But $H'X$ is!
- (A4) Given (G, X), $H_{ij} \perp H_{kl}$ for all $(i, j) \neq (k, l)$.
  - rules out symmetric H (*undirected* links).
- We show $Z = (H'X, X)$ satisfies $E(Z'v) = 0$.

Construct Instruments from H

- Recall we can *not* use HX as instruments. But $H'X$ is!
- (A4) Given (G, X), $H_{ij} \perp H_{kl}$ for all $(i,j) \neq (k,l)$.
  - rules out symmetric H (*undirected* links).
- We show $Z = (H'X, X)$ satisfies $E(Z'v) = 0$.
  - $E\left[(H^2)_{ij}|G,X\right] = (1-p)^2\left(G^2\right)_{ij}$, and $E\left[HG|G,X\right] = E(H|G,X)G = (1-p)G^2$;

Construct Instruments from H

- Recall we can *not* use HX as instruments. But $H'X$ is!
- (A4) Given (G, X), $H_{ij} \perp H_{kl}$ for all $(i, j) \neq (k, l)$.
  - rules out symmetric H (*undirected* links).
- We show $Z = (H'X, X)$ satisfies $E(Z'v) = 0$.
  - $E\left[(H^2)_{ij}|G, X\right] = (1-p)^2\left(G^2\right)_{ij}$, and $E\left[HG|G, X\right] = E(H|G, X)G = (1-p)G^2$;
  - Hence $E(HGy|G, X) = E(H^2y|G, X)/(1-p)$. So, $E(X'Hv|G, X) = 0$.

Construct Instruments from H ('ctd)

- (A4) requires the noisy measure H be asymmetric. What if only symmetric measures are available?
- Suppose there are two symmetric measures $H^{(1)}, H^{(2)}$
  - (A4) Given (G, X), $H_{ij}^{(1)} \perp H_{kl}^{(2)}$ for all $(i,j) \neq (k,l)$.
  - e.g., two independent measures of the same network.
  - We can show that

$$E[(H^{(2)}X)'v^{(1)}] = 0.$$

Recover the missing rate

- We now show how to identify the missing rate p when
  - either (a) asymmetric noisy measure H of a symmetric G;
  - or (b) two independent measures $H^{(1)}, H^{(2)}$ of the same G (all matrices can be symmetric or asymmetric)
- Solution in (a):
  - suppose $\Pr(G_{ij} = G_{ji}) = 1$, $\Pr(H_{ij} \neq H_{ji}) > 0$
  - construct $\tilde{H}_{ij} = \max\{H_{ij}, H_{ji}\}$ with missing rate p^2
  - $E(H_{ij}) = (1-p)E\left(G_{ij}\right)$, $E(\tilde{H}_{ij}) = (1-p^2)E(G_{ij})$
  - then $p = E\left[\psi(\tilde{H})\right] / E[\psi(H)] - 1$, where $\psi(H)$ is a linear function of H (e.g. average of all entries)
- Solution in (b) follows from a similar argument.

Adjusted 2SLS Estimator

- Step 1. Use the analog principle to estimate missing rates $\widehat{p}$ in (a).
- Step 2. (Single H case) Use $(H'X, X)$ as instruments for $\left(\frac{Hy}{1-p}, X\right)$ in 2SLS:

$$\hat{\theta} \equiv \left(\mathbf{A}'\mathbf{B}^{-1}\mathbf{A}\right)^{-1}\mathbf{A}'\mathbf{B}^{-1}(\mathbf{Z}'Y),$$

where $\mathbf{A} \equiv \mathbf{Z}'\mathbf{W}(\widehat{p})$ and $\mathbf{B} \equiv \mathbf{Z}'\mathbf{Z}$, with $\mathbf{W}$, $\mathbf{Z}$ stacking

$$W_s(p) \equiv \left(\frac{H_s y_s}{1-p}, X_s\right), \; Z_s \equiv (H_s' X_s, X_s)$$

over the observed group s in the sample.

- We derived asymptotic variance, taking into account estimation error in $\widehat{p}$.

Adjusted S2SLS Estimator

- In the case with multiple measures $H^{(t)}$, $t = 1, 2$, we apply system 2SLS.
- Stack the moments: $E\left[\tilde{Z}_s'(\tilde{y}_s - \tilde{W}_s\theta)\right] = 0$, where

$$\tilde{Z}_s \equiv \begin{pmatrix} Z_s^{(1)} & 0 \\ 0 & \tilde{Z}_s^{(2)} \end{pmatrix}; \; \tilde{y}_s \equiv \begin{pmatrix} y_s \\ y_s \end{pmatrix}; \; \tilde{W}_s \equiv \begin{pmatrix} W_s^{(1)} \\ W_s^{(2)} \end{pmatrix}$$

and for each group s observed in the sample and $t = 1, 2$,

$$Z_s^{(t)} \equiv \left(H_s^{(3-t)}X_s, X_s\right), \; W_s^{(t)} \equiv \left(\frac{H_s^{(t)}y_s}{1 - p^{(t)}}, X_s\right).$$

- Provided $E\left(\tilde{Z}_s'\tilde{W}_s\right)$ has full rank, we can identify θ from the stacked moments. Thus we can do S2SLS:

$$\tilde{\theta} \equiv \left[\tilde{W}'\tilde{Z}\left(\tilde{Z}'\tilde{Z}\right)^{-1}\tilde{Z}'\tilde{W}\right]^{-1}\tilde{\mathbf{W}}'\tilde{\mathbf{Z}}\left(\tilde{Z}'\tilde{Z}\right)^{-1}\tilde{\mathbf{Z}}'\tilde{\mathbf{y}}.$$

Extension 1

- Allowing for group fixed effects,

$$y = \lambda Gy + X\beta + \alpha + \varepsilon,$$

  where G is measured as H with missing links.

- Do with-in transformation, and then applies our method.
- This works because of model linearity, and that $E(H|G, X)$ is linear in G.

Extension 2

- Structural model with contextual effects is

$$y = \lambda Gy + X\beta + GX\gamma + \varepsilon.$$

- Adjusted feasible structural form is

$$y = \lambda \tfrac{Hy}{1-p} + X\beta + \tfrac{HX}{1-p}\gamma + \eta,$$

where $\eta \equiv \varepsilon - \lambda(\frac{H}{1-p} - G)y - (\frac{H}{1-p} - G)X\gamma$.

- Under (A1)-(A3), $E(\eta | X, G) = 0$.
- Under (A4), use $(H'X, H'\zeta(X))$ as instruments for (Hy, HX).
- Or, one can do efficient method of moments, by plugging in estimates for p.

Heterogeneous Missing Rates

- Now let the missing rates vary with X.
- Relax (A2) with:

$$E(H_{ij}|G_{ij} = 1, X) = 1 - p_{ij}(X) \text{ and } E(H_{ij}|G_{ij} = 0, X) = 0.$$

- Then

$$E(H|G, X) = Q(X) \circ G \text{ with } Q_{ij}(X) \equiv 1 - p_{ij}(X),$$

where denote "$\circ$" Hadamard product.

- Step 1: estimate $p_{ij}(X)$ using sample analogs as before.

Heterogeneous Missing Rates

- Step 2: apply 2SLS to

$$y = \lambda\left(\tilde{Q}\circ H\right)y + X\beta + \underbrace{\varepsilon + \lambda[G - \tilde{Q}\circ H]y}_{v^*}$$

where $\tilde{Q}_{ij} \equiv 1/(1-p_{ij})$, and

$$\begin{aligned} E(v^*|G,X) &= \lambda[GMX\beta - \tilde{Q}\circ E(H|G,X)MX\beta] \\ &= \lambda[GMX\beta - \tilde{Q}\circ(Q\circ G)MX\beta] = 0. \end{aligned}$$

Now we need nonlinear function $\zeta(X)$ as instruments for $(\tilde{Q}\circ H)y$.

- One can do efficient *method of moment* instead, using $E(v^*|X) = 0$.

Single, Large Network

- Our method applies to single, large network if there is "weak dependence" between individuals "sufficiently far" from each other.
- Nearly block-diagonal (NBD)
  - sample partitioned into *approximate* groups, or "blocks"
  - links within each block are *dense;* links across blocks are *sparse*
- Measurement errors in NBD networks
  - within-block links are reported, but randomly missing at rate p
  - no links reported across blocks

Single, Large Network

- A key condition:

$$\sum_{i=1}^{N} \sum_{j \notin s(i)} E(|H_{i,j} - G_{i,j}|) = O(S^{\rho}) \text{ for } \rho < 1,$$

where $j \notin s(i)$ means j is not in the same block as i, with S being # of blocks and $N = \sum_{s=1}^{S} n_s$ the sample size.

- We show that 2SLS applied to unscaled peer outcomes, denoted $\hat{\theta}_a$ is such that

$$\hat{\theta}_a - \theta_a = O_p(S^{-1/2} \vee S^{\rho-1}),$$

where $\theta_a \equiv (\lambda/(1-p), \beta')'$. And with $\rho < 1/2$,

$$\sqrt{S}\,(\hat{\theta}_a - \theta_a) \xrightarrow{d} \mathcal{N}(0, \Omega).$$

- In our empirical application we assume this near block diagonal structure.

Application: Microfinance in Indian Villages

- Data source: Banerjee et al (2013). 4,134 households from 43 villages in the State of Karnataka, India.
- Dependent variable y: participation in a microfinance program. Average participation rate is 18.9%
- Covariates X are demographhcs at the household and individual level.
- From survey responses, Banerjee et al (2013) provide various symmetrized social network measures.

Empirical Application: Network Measures

- We use two of symmetrized measures of links reported in the data: $H^{(1)}$ is who visits you (*VisitCome*) and $H^{(2)}$ is who you visit (*VisitGo*).
- $H^{(1)}$ and $H^{(2)}$ are both measures of the same underlying G, because if household A visits household B, as recorded in $H^{(1)}$ then household B must have been visited by household A, as recorded in $H^{(2)}$.
- These two matrices empirically differ substantially, showing both are noisy measures of G.
- We assume the observed differences between $H^{(1)}$ and $H^{(2)}$ are missing links, and any of the reported zeros in both could also be missing links.

Table 2(a): Summary of Dependent and Explanatory Variables

Variable	definition	obs.	mean	s.d.	min	max
y	dummy for participation	4149	0.1894	0.3919	0	1
room	number of rooms	4149	2.4389	1.3686	0	19
bed	number of beds	4149	0.9229	1.3840	0	24
age	age of household head	4149	46.057	11.734	20	95
edu	education of household head	4149	4.8383	4.5255	0	15
lang	whether to speak other language	4149	0.6799	0.4666	0	1
male	whether the hh head is male	4149	0.9161	0.2772	0	1
leader	whether it has a leader	4149	0.1393	0.3463	0	1
shg	whether in any saving group	4149	0.0513	0.2207	0	1
sav	whether to have a bank account	4148	0.3840	0.4864	0	1
election	whether to have an election card	4149	0.9525	0.2127	0	1
ration	whether to have a ration card	4149	0.9012	0.2985	0	1

Table 2(b): Summary of Category Variables

Variable	definition	obs.	per.	Variable	definition	obs.	per.
religion				*latrine*			
-	Hinduism	3943	95.04	-	Owned	1195	28.80
-	Islam	198	4.77	-	Common	20	0.48
-	Christianity	7	0.19	-	None	2934	70.72
roof				*own*	property ownership		
-	Thatch	82	1.98	-	Owned	3727	89.83
-	Tile	1388	33.45	-	Owned & shared	32	0.77
-	Stone	1172	28.25	-	Rented	390	9.40
-	Sheet	868	20.92				
-	RCC	475	11.45				
-	Other	164	3.95				
electricity	electricity provision			*caste*			
				-	Scheduled caste	1139	27.54
-	Private	2662	64.18	-	Scheduled tribe	221	5.34
-	Government	1243	29.97	-	OBC	2253	54.47
-	No power	243	5.86	-	General	523	12.65

Table 3 Degree Distribution in Two Network Measures

Degree	0	1	2	3	4	5	6	7	8	9	10
$H^{(1)}$	2	21	110	227	357	505	526	546	506	379	269
$H^{(2)}$	4	24	112	245	384	522	534	577	491	386	255
Degree	11	12	13	14	15	16	17	18	19	20	≥ 21
$H^{(1)}$	224	145	90	74	54	33	27	15	9	6	24
$H^{(2)}$	179	137	102	59	46	28	22	13	9	3	17

- Scaled feasible structural linear probability model:

$$y = \lambda \frac{H^{(t)} y}{1 - p^{(t)}} + X\beta + villageFE + v^{(t)}.$$

- Estimates of missing rates

$$\hat{p}^{(1)} = 0.1681 \text{ and } \hat{p}^{(2)} = 0.1909.$$

- Next 2SLS estimates and inference are based on single growing network.

Two-stage Least Squares Estimates:

We report five different estimates, as follows:

(a) Standard network 2SLS treating $H^{(1)}$ as true G.

(b) Our adjusted 2SLS using $H^{(2)}X$ as instruments for the scaled feasible structural model:

$$y = \lambda \frac{H^{(1)}y}{1-p^{(1)}} + X\beta + villageFE + v^{(1)}.$$

(c) Standard network 2SLS treating $H^{(2)}$ as true G.

(d) Our adjusted 2SLS using $H^{(1)}X$ as instruments for:

$$y = \lambda \frac{H^{(2)}y}{1-p^{(2)}} + X\beta + villageFE + v^{(2)}.$$

(e) Stacked 2SLS estimator that exploits the moments generated by both (b) and (d) above into a single combined estimator.

Table 4: Two-stage Least Squares Estimates

	(a)	(b)	(c)	(d)	(e)
r.h.s. endogeneity	$H^{(1)}$ y	$\frac{H^{(1)}}{1-\widehat{p}_1}$ y	$H^{(2)}$ y	$\frac{H^{(2)}}{1-\widehat{p}_2}$ y	$\frac{H}{1-\widehat{p}}$ y
IV used	$H^{(1)}$ X	$H^{(2)}$ X	$H^{(2)}$ X	$H^{(1)}$ X	Combined
$\widehat{\lambda}$	0.0498***	0.0456***	0.0529***	0.0484***	0.0461***
	(0.0076)	(0.0096)	(0.0092)	(0.0087)	(0.0075)
leader	0.0378**	0.0364**	0.0418**	0.0405**	0.0387**
	(0.0185)	(0.0186)	(0.0182)	(0.0182)	(0.0183)
age	-0.0016***	-0.0017***	-0.0016***	-0.0017***	-0.0017***
	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
ration	0.0441**	0.0435**	0.0423**	0.0413**	0.0426**
	(0.0201)	(0.0201)	(0.0195)	(0.0194)	(0.0197)
electricity − gov	0.0343**	0.0333**	0.0352**	0.0341**	0.0339**
	(0.0157)	(0.0157)	(0.0156)	(0.0155)	(0.0156)
electricity − no	0.0223	0.0229	0.0237	0.0247	0.0236
	(0.0297)	(0.0297)	(0.0300)	(0.0298)	(0.0298)
caste − tribe	-0.0285	- 0.0272	-0.0275	- 0.0257	- 0.0268
	(0.0312)	(0.0309)	(0.0305)	(0.0300)	(0.0305)
caste − obc	- 0.0520**	- 0.0490**	- 0.0486**	- 0.0441***	- 0.0473***
	(0.0217)	(0.0212)	(0.0215)	(0.0206)	(0.0210)
caste − gen	-0.0734***	-0.0698***	-0.0688***	-0.0628**	-0.0673***
	(0.0239)	(0.0242)	(0.0241)	(0.0234)	(0.0239)
religion − Islam	0.0980***	0.0955***	0.0893***	0.0849***	0.0910***
	(0.0323)	(0.0323)	(0.0343)	(0.0344)	(0.0332)
religion − Chri	0.1434	0.1420	0.1466	0.1452	0.1438
	(0.130)	(0.1287)	(0.1314)	(0.1300)	(0.1293)
Controls	√	√	√	√	√
VillageFE	√	√	√	√	√
R^2	0.1332	0.1345	0.1350	0.1365	0.1353
Obs	4134	4134	4134	4134	4134

Note: s.e. in parentheses. ***, **, and * indicate 1%, 5% and 10% significant.
Controls include *male* , *roof* , *room* , *bed* , *latrine* , *edu* , *lang* , *shg* , *sav* , *election* , *own* .

Empirical results: summary

- Our main empirical findings regarding peer effects on participation in a microfinance program in India:
  - missing rate $p \approx 0.18$ on average.
  - peer effect $\lambda \approx 0.046$. One more participating link (visitor) increases own participation probability by 4.6%
  - ignoring missing links by using traditional 2SLS yields peer effect λ estimates biased upward by about 9% (augmentation bias).

Conclusion

- We propose a simple method for applying 2SLS when some links are missing at random from the sample.
- We derive limiting distribution theory for our estimators.
- We provide an empirical application estimating peer effects on participation in a microfinance program in India.
  - we find strong empirical evidence of missing links.
  - we show that accounting for missing links on estimation is empirically important.

THANKS!