Social Networks with Link Misclassification

Arthur Lewbel, Xi Qu, and Xun Tang

University of Oxford, March 1, 2024

Introduction

- In social networks, individual outcomes depend on:
  - own characteristics (*direct* effects)
  - others' characteristics (*contextual* effects)
  - others' outcomes (*peer* effects)
- Links reported in samples are subject to misclassification:
  - recall errors in survey responses
  - errors in data entry
- We propose estimators for social effects that are robust to link misclassification.

Introduction

- Conventional 2SLS:
  - Structural form (SF): $y = \lambda Gy + X\beta + \varepsilon$, where $G_{ij} = 1$ if i and j are linked, and 0 otherwise.
  - Suppose G is perfectly reported in a sample.
  - Peer outcomes Gy are endogenous due to simultaneity.
  - Conventional 2SLS using GX or G^2X as instruments for Gy - e.g. Lee (2007), Bramoulle et al (2009)
  - IV exogeneity and relevance hold with $E(\varepsilon|X, G) = 0$.

Introduction

- How do misclassified links affect inference?
  - Suppose the sample only reports $H \neq G$, with H randomly misclassifying links in 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 are *both* correlated 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 (2022)
  - 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): estimation when the sample does not report link status
- Lewbel, Qu, and Tang (2023): 2SLS applies when errors in link measures are small enough

Preview: Basic Idea

- We illustrate the main idea when links are randomly misclassified with rates

$$p_0 = E(H_{ij}|G_{ij} = 0),\ p_1 = E(1 - H_{ij}|G_{ij} = 1).$$

- Adjusted 2SLS:
  - replaces H with an *adjusted* $\mathcal{H}(p)$ in structural form, using $p \equiv (p_0, p_1)$; this restores exogeneity in X
  - uses new IVs for $\mathcal{H}(p)y$: $H'X$ or $\mathcal{H}(p)'X$
  - is implemented using closed-form estimates of (p_0, p_1)
  - applies in various scenarios: (a)symmetric G, single or multiple (un)symmetrized measures H

Preview: Extensions

- Extensions:
  - add contextual effects
  - allow for heterogeneous misclassification rates
  - include group-level fixed effects
- Adjusted 2SLS: works with a single, large network
  - approximate groups (blocks) with sparse, unreported links between blocks
  - links within blocks are misclassfied with non-diminishing rates

Preview: Application

- We apply our method to data from Banerjee, Chandrasekhar, Duflo, and Jackson (2013)
  - surveys from over 4.1k households in 43 villages
  - two measures of links imputed ("*VisitCome*" vs "*VisitGo*")
  - evidence of link misclassification: symmetrized measures differ

VisitCome vs VisitGo

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

Preview: Application

- Dependent variable: whether participate in micro-finance program (sample average participation rate is 18.4%)
- Main findings:
  - low misclassification rates; mostly due to missing links (p_0 near zero; p_1 around 0.11 and 0.14).
  - "endorsement effect": $\lambda \approx 0.051$ (additional participating neighbor increases own participation by 5.1%)
  - ignoring link misclassification results in upward bias in peer effect estimates

Social Network with Link Measures

- Model:
  - many small, independent networks

    $$y = \lambda G y + X\beta + \varepsilon,\ E(\varepsilon|X, G) = 0,$$
    $$y \in \mathbb{R}^n,\ X \in \mathbb{R}^{n\times K},\ \varepsilon \in \mathbb{R}^n,$$
    $$G_{ij} \in \{0, 1\},\ G_{ii} = 0.$$

  - reduced form: $y = M(X\beta + \varepsilon)$, $M \equiv (I - \lambda G)^{-1}$.
  - data reports H instead of G, with $H_{ii} = 0$.

Model Assumptions

- (A1) $E(H_{ij}|G, X) = E(H_{ij}|G_{ij}, X)$.
  - caution: fails if G is asymmetric while H is symmetrized
- (A2) Random misclassification
  - $E(1 - H_{ij}|G_{ij} = 1, X) = p_1$, $E(H_{ij}|G_{ij} = 0, X) = p_0$.
- (A3) $E(\varepsilon|X, G, H) = 0$.

Restore Exogeneity in X

- Consider an (infeasible) *adjusted* structural form:

$$y = \lambda \mathcal{H}(p) y + X\beta + \underbrace{\varepsilon + \lambda \left(G - \mathcal{H} \right) y}_{\equiv v},$$

where

$$\mathcal{H}(p) \equiv \frac{H - p_0(\iota\iota' - I)}{1 - p_0 - p_1}.$$

- Under (A1), (A2), (A3),
  - $E(H_{ij} | G_{ij}, X) = p_0(1 - G_{ij}) + (1 - p_1) G_{ij}$ for $i \neq j$
  - $E(\mathcal{H}(p) | X, G) = G$
  - $E(\mathcal{H}(p) y | X, G) = E(\mathcal{H}(p) | G, X) MX\beta = GMX\beta = E(Gy | X, G)$
  - $E(v | X, G) = 0.$

Adjusted 2SLS

- Let $R \equiv (\mathcal{H}(p)y, X)$, $Z \equiv (\zeta(X), X)$, where $\zeta(\cdot)$ is nonlinear function of X.
- Suppose:

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

  Then
  $$E(Z'y) = E(Z'R)(\lambda, \beta')' + \underbrace{E(Z'v)}_{=0}.$$
- So, 2SLS works after this adjustment, with *proper* IVs.
- We provide sufficient conditions for (IV-R).

Bias in Unadjusted 2SLS

- What if apply 2SLS to a structural form (SF) using unadjusted H?

Bias in Unadjusted 2SLS

- What if apply 2SLS to a structural form (SF) using unadjusted H?
  - Then $y = \check{R}\theta + u$, where $\check{R} \equiv (Hy, X)$, $\theta \equiv (\lambda, \beta')'$ and

$$u = v + \left(\tfrac{p_0+p_1}{1-p_0-p_1}\right)\lambda Hy - \left(\tfrac{p_0}{1-p_0-p_1}\right)\lambda(\iota\iota' - I)y,$$

with v being errors in SF using $\mathcal{H}(p)$, and $E(v|X, G) = 0$.

Bias in Unadjusted 2SLS

- What if apply 2SLS to a structural form (SF) using unadjusted H?
  - Then $y = \check{R}\theta + u$, where $\check{R} \equiv (Hy, X)$, $\theta \equiv (\lambda, \beta')'$ and

    $$u = v + \left(\frac{p_0+p_1}{1-p_0-p_1}\right)\lambda Hy - \left(\frac{p_0}{1-p_0-p_1}\right)\lambda(\iota\iota' - I)y,$$

    with v being errors in SF using $\mathcal{H}(p)$, and $E(v|X,G) = 0$.
  - Will show how to construct valid IV. But such IVs won't resolve misclassification bias (b/c 2nd and 3rd terms in u).

Bias in Unadjusted 2SLS

- What if apply 2SLS to a structural form (SF) using unadjusted H?
  - Then $y = \check{R}\theta + u$, where $\check{R} \equiv (Hy, X)$, $\theta \equiv (\lambda, \beta')'$ and

    $$u = v + \left(\frac{p_0+p_1}{1-p_0-p_1}\right)\lambda Hy - \left(\frac{p_0}{1-p_0-p_1}\right)\lambda(\iota\iota' - I)y,$$

    with v being errors in SF using $\mathcal{H}(p)$, and $E(v|X,G) = 0$.
  - Will show how to construct valid IV. But such IVs won't resolve misclassification bias (b/c 2nd and 3rd terms in u).
- A special case:

Bias in Unadjusted 2SLS

- What if apply 2SLS to a structural form (SF) using unadjusted H?
  - Then $y = \check{R}\theta + u$, where $\check{R} \equiv (Hy, X)$, $\theta \equiv (\lambda, \beta')'$ and
  $$u = v + \left(\frac{p_0+p_1}{1-p_0-p_1}\right)\lambda Hy - \left(\frac{p_0}{1-p_0-p_1}\right)\lambda(\iota\iota' - I)y,$$
  with v being errors in SF using $\mathcal{H}(p)$, and $E(v|X, G) = 0$.
  - Will show how to construct valid IV. But such IVs won't resolve misclassification bias (b/c 2nd and 3rd terms in u).
- A special case:
  - One-sided randomly missing: $p_0 = 0$, $p_1 > 0$.

Bias in Unadjusted 2SLS

- What if apply 2SLS to a structural form (SF) using unadjusted H?
  - Then $y = \check{R}\theta + u$, where $\check{R} \equiv (Hy, X)$, $\theta \equiv (\lambda, \beta')'$ and

$$u = v + \left(\frac{p_0+p_1}{1-p_0-p_1}\right)\lambda Hy - \left(\frac{p_0}{1-p_0-p_1}\right)\lambda(\iota\iota' - I)y,$$

    with v being errors in SF using $\mathcal{H}(p)$, and $E(v|X,G) = 0$.
  - Will show how to construct valid IV. But such IVs won't resolve misclassification bias (b/c 2nd and 3rd terms in u).
- A special case:
  - One-sided randomly missing: $p_0 = 0$, $p_1 > 0$.
  - Thus $E(Z'u) = E(Z'\check{R})(\frac{p_1}{1-p_1}\lambda, 0')'$.

Bias in Unadjusted 2SLS

- What if apply 2SLS to a structural form (SF) using unadjusted H?
  - Then $y = \check{R}\theta + u$, where $\check{R} \equiv (Hy, X)$, $\theta \equiv (\lambda, \beta')'$ and
  $$u = v + \left(\frac{p_0+p_1}{1-p_0-p_1}\right)\lambda Hy - \left(\frac{p_0}{1-p_0-p_1}\right)\lambda(\iota\iota' - I)y,$$
  with v being errors in SF using $\mathcal{H}(p)$, and $E(v|X, G) = 0$.
  - Will show how to construct valid IV. But such IVs won't resolve misclassification bias (b/c 2nd and 3rd terms in u).

- A special case:
  - One-sided randomly missing: $p_0 = 0$, $p_1 > 0$.
  - Thus $E(Z'u) = E(Z'\check{R})(\frac{p_1}{1-p_1}\lambda, 0')'$.
  - Plim of $\hat{\lambda}$ in naive 2SLS: $\left(1 + \frac{p_1}{1-p_1}\right)\lambda = \frac{\lambda}{1-p_1}$.

Bias in Unadjusted 2SLS

- What if apply 2SLS to a structural form (SF) using unadjusted H?
  - Then $y = \check{R}\theta + u$, where $\check{R} \equiv (Hy, X)$, $\theta \equiv (\lambda, \beta')'$ and

$$u = v + \left(\frac{p_0+p_1}{1-p_0-p_1}\right)\lambda Hy - \left(\frac{p_0}{1-p_0-p_1}\right)\lambda(\iota\iota' - I)y,$$

    with v being errors in SF using $\mathcal{H}(p)$, and $E(v|X, G) = 0$.
  - Will show how to construct valid IV. But such IVs won't resolve misclassification bias (b/c 2nd and 3rd terms in u).
- A special case:
  - One-sided randomly missing: $p_0 = 0$, $p_1 > 0$.
  - Thus $E(Z'u) = E(Z'\check{R})(\frac{p_1}{1-p_1}\lambda, 0')'$.
  - Plim of $\hat{\lambda}$ in naive 2SLS: $\left(1 + \frac{p_1}{1-p_1}\right)\lambda = \frac{\lambda}{1-p_1}$.
  - We have an "*augmentation*" bias!

Construct IVs from H

- Recall HX is not valid IV; but we'll show $\mathcal{H}(p)'X$ 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 $Z = (\mathcal{H}(p)'X, X)$ satisfies $E(Z'v) = 0$.

Construct IVs from H

- Recall HX is not valid IV; but we'll show $\mathcal{H}(p)'X$ 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 $Z = (\mathcal{H}(p)'X, X)$ satisfies $E(Z'v) = 0$.
  - $E\left[(\mathcal{H}(p)^2)_{ij} | G, X\right] = \left(G^2\right)_{ij}$ under (A4).

Construct IVs from H

- Recall HX is not valid IV; but we'll show $\mathcal{H}(p)'X$ 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 $Z = (\mathcal{H}(p)'X, X)$ satisfies $E(Z'v) = 0$.
  - $E\left[(\mathcal{H}(p)^2)_{ij} | G, X\right] = \left(G^2\right)_{ij}$ under (A4).
  - $E\left[\mathcal{H}(p)G | G, X\right] = E(\mathcal{H}(p)|G, X)G = G^2$.

Construct IVs from H

- Recall HX is not valid IV; but we'll show $\mathcal{H}(p)'X$ 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 $Z = (\mathcal{H}(p)'X, X)$ satisfies $E(Z'v) = 0$.
  - $E\left[(\mathcal{H}(p)^2)_{ij} | G, X\right] = \left(G^2\right)_{ij}$ under (A4).
  - $E\left[\mathcal{H}(p) G | G, X\right] = E(\mathcal{H}(p) | G, X) G = G^2$.
  - $E(\mathcal{H}(p) Gy | G, X) = E(\mathcal{H}(p)^2 y | G, X)$ under (A3) $\Rightarrow E[(\mathcal{H}(p)'X)'v | G, X] = 0$.

Construct IVs from H

- Recall HX is not valid IV; but we'll show $\mathcal{H}(p)'X$ 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 $Z = (\mathcal{H}(p)'X, X)$ satisfies $E(Z'v) = 0$.
  - $E\left[(\mathcal{H}(p)^2)_{ij}|G,X\right] = \left(G^2\right)_{ij}$ under (A4).
  - $E\left[\mathcal{H}(p)G|G,X\right] = E(\mathcal{H}(p)|G,X)G = G^2$.
  - $E(\mathcal{H}(p)Gy|G,X) = E(\mathcal{H}(p)^2y|G,X)$ under (A3) $\Rightarrow E[(\mathcal{H}(p)'X)'v|G,X] = 0$.
  - $H'X$ also satisfies IV exogeneity (b/c $E(v|G,X) = 0$).

Construct IVs from H

- What if H is a symmetrized measure (e.g. $H_{ij} = H_{ji}$ by construction)?
- Need *two* symmetrized measures $H^{(1)}, H^{(2)}$
  - (A4) Given (G, X), $H^{(1)}_{ij} \perp H^{(2)}_{kl}$ for all $(i,j) \neq (k,l)$.
  - e.g., two independent surveys of the same, latent G
  - Analogous argument shows

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

where $v^{(t)}$ is error in structural form using adjusted measure

$$\frac{H^{(t)} - p_0^{(t)}(\iota\iota' - I)}{1 - p_0^{(t)} - p_1^{(t)}}.$$

Identify and Estimate Misclassification Rates (MR): (p_0, p_1)

- For adjusted 2SLS, we need estimates for p_0, p_1.
- We obtain these estimates using:
  - either (a) two independent $H^{(1)}, H^{(2)}$ (symmetrized or not) for the same G (symmetric or not);
  - or (b) a single unsymmetrized H for symmetric G

Identify and Estimate MR: (p_0, p_1)

- $\phi_{ij}(X)$: demographic info related to link formation (*not* modeling link formation per se).
- E.g., $\phi_{ij}(X) \equiv 1\{X_{i,1} = X_{j,1}\}$; let ω_1 denote "$\phi_{ij}(X) = 1$."
- Scenario (a): two measures $H^{(1)}, H^{(2)}$
  - parameters of interests: $p_1^{(t)}, p_0^{(t)}$ for $t = 1, 2$
  - nuisance: $\pi_1 \equiv \frac{1}{n(n-1)} \sum_{i \neq j} \Pr\{G_{ij} = 1 | \omega_1\}$ and π_0
  - we do *not* seek to learn about link formation from π_1, π_0.

Identify and Estimate MR: (p_0, p_1)

- Summarize joint distribution $H_{ij}^{(1)}, H_{ij}^{(2)}$:

$$\frac{1}{n(n-1)}\sum_{i\neq j} E\left(H_{ij}^{(1)}H_{ij}^{(2)}\middle|\omega_1\right) = \left(1-p_1^{(1)}\right)\left(1-p_1^{(2)}\right)\pi_1 + p_0^{(1)}p_0^{(2)}\left(1-\pi_1\right),$$

$$\frac{1}{n(n-1)}\sum_{i\neq j} E\left(H_{ij}^{(t)}\middle|\omega_1\right) = \left(1-p_1^{(t)}\right)\pi_1 + p_0^{(t)}\left(1-\pi_1\right) \text{ for } t = 1, 2;$$

and likewise conditioning on ω_0.

- We get closed-form expressions for $p_1^{(t)}, p_0^{(t)}$ as functions of identifiable moments on the left-hand side.

Identify and Estimate MR: (p_0, p_1)

- This idea also extends to Scenario (b), with a single, unsymmetrized measure H for a symmetric G.
  - For *unordered* $\{i, j\}$, let $H^{(1)}_{\{i,j\}} \equiv H_{ij}$, $H^{(2)}_{\{i,j\}} \equiv H_{ji}$.
  - Method in (a) applies with $\frac{1}{n(n-1)}$, $\sum_{i \neq j}$, $H^{(t)}_{ij}$ replaced by $\frac{2}{n(n-1)}$, $\sum_{i>j}$, $H^{(t)}_{\{i,j\}}$ respectively.

Identify and Estimate MR: (p_0, p_1)

- We can recover MR using *any* generic definition of $\phi_{ij}(X)$ and partition of its support
  - necessary condition for identification: $\pi_1 \neq \pi_0$.
- Another extension: use aggregate moments in the argument.
  - e.g., $E\left[\delta(H^{(t)})|\sigma(X)\right]$ with $\delta(H)$: $\#$ of links in H; $\sigma(X)$: gender ratio.
  - estimators easy to computation with a closed form.

Identification Summary

	Reported Network Measures					
	Single, unsym'zed		Multiple, sym'zed		Multiple, unsym'zed	
	(IV)	(MR)	(IV)	(MR)	(IV)	(MR)
Sym. G	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Asym. G	$\surd$	?	violates (A1)		$\surd$	$\surd$

Adjusted 2SLS: Single Measure

- Step 1. Use analog principle to estimate misclassification rates $\hat{p} \equiv (\hat{p}_1, \hat{p}_0)$.
- Step 2. (Single H) Use $(H'X, X)$ as IV for $(\mathcal{H}(p)y, X)$:

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

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

$$R_s(\widehat{p}) \equiv (\mathcal{H}_s(\widehat{p})y_s, X_s)\,,\; Z_s \equiv (H_s'X_s, X_s)$$

over all group s in the sample.

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

Adjusted 2SLS: Multiple Measures

- With two measures $H^{(t)}$, stack the moments: $E\left[\tilde{Z}_s'(\tilde{y}_s - \tilde{R}_s\theta)\right] = 0$, where

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

and for each group s in the sample,

$$Z_s^{(t)} \equiv \left(H_s^{(3-t)} X_s, X_s\right), \; R_s^{(t)} \equiv \left(\mathcal{H}_s^{(t)}(\widehat{p}) y_s, X_s\right).$$

- Provided $E\left(\tilde{Z}_s'\tilde{R}_s\right)$ has full rank, we can identify θ from the stacked moments. Apply 2SLS:

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

Extension: Group Fixed Effects

- Let α denote group-level fixed effects,

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

  where G is measured by H.

- Apply *with-in* transformation to y, X and network measure(s).
  - Constructing IVs requires two measures $H^{(1)}, H^{(2)}$.
- This works because $E(\mathcal{H}(p)|G, X) = G$ and the with-in transformations are linear.

Extension: Contextual Effects

- SF with contextual effects:

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

- Adjusted feasible structural form is

$$y = \lambda \mathcal{H}(p)y + X\beta + \mathcal{H}(p)X\gamma + \eta,$$

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

- Under (A1)-(A3), $E(\eta|X, G) = 0$.
- Under (A4), use $(H'X, H'\zeta(X))$ as IVs for $(\mathcal{H}(p)y, \mathcal{H}(p)X)$ in adjusted 2SLS.

Extension: Heterogeneous MR

- Relax (A2) to (A2') as follows:

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

- Let

$$\mathcal{H}_{ij}(X;p)\equiv\frac{H_{ij}-p_{ij,0}(X)}{1-p_{ij,0}(X)-p_{ij,1}(X)}\ \forall i\neq j,\ \mathcal{H}_{ii}(X)=0.$$

  Then $E[\mathcal{H}(X;p)|G,X]=G$ under (A2') and (A1), (A3).

- Step 1: estimate $p_{ij}(X)$ using sample analogs, possibly with parametrization.

Extension: Heterogeneous MR

- Step 2: apply 2SLS to

$$y = \lambda \mathcal{H}(X;p)y + X\beta + \underbrace{\varepsilon + \lambda[G - \mathcal{H}(X;p)]y}_{v^*},$$

where

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

Use non-linear $\zeta(X)$, e.g. $X \circ X$ as IVs for $\mathcal{H}(X;p)y$.

- Or do *method of moment*, using efficient IVs.

A Single, Large Network

- Consider a "nearly block-diagonal" (NBD) setting
  - sample partitioned into S *approximate* groups, a.k.a. *blocks*
  - links between all n_s individuals in a block are *dense*; links across blocks are *sparse*
  - e.g., much less likely to have linked households across villages
- Measurement errors:
  - links within blocks are reported, but randomly misclassified
  - the sample does not report *any* link across blocks

Single, Large Network

- Let $\tilde{G}$ differ from G by missing all links *between* blocks. Assume:

$$(*) \quad \sum_{i=1}^{N} \sum_{j \notin s(i)} E(|\tilde{G}_{ij} - G_{ij}|) = O(S^{\rho}) \text{ for } \rho < 1,$$

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.

- Condition (*) posits the order of measurement errors outside blocks are small. Example:
    - n_s is uniformly bounded by $n_B < \infty$ for all s;
    - dyadic links across blocks formed at rate $q_S = O(S^{-\gamma})$;
    - (*) holds with $\rho = 2 - \gamma < 1$.
- Adjusted 2SLS, denoted $\hat{\theta}$, is such that

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

where $\theta \equiv (\lambda, \beta')'$. If $\rho < 1/2$, then $\hat{\theta}$ is root-n CAN.

MC Simulation

- Data-generating process:
  - $y_s = \lambda G_s y_s + X_s\beta + \alpha_s + \varepsilon_s$
  - $X_{s,i,1}\tilde{}Bernoulli(0.5)$, $X_{s,i,2}\tilde{}N(0,1)$, $\lambda = 0.05$, $\beta = (1,2)$
  - correlated fixed effect: $\alpha_s = 5\bar{X}_s\beta - 3/2 + e_s$, $e_s\tilde{}N(0,1)$
  - $\pi_1 = E(G_{ij}|X_{i1} = X_{j1}) = 0.2$, $\pi_0 = E(G_{ij}|X_{i1} \neq X_{j1}) = 0.1$
  - small MR: $\left(p_0^{(1)}, p_1^{(1)}\right) = (0.10, 0.20)$, $\left(p_0^{(2)}, p_1^{(2)}\right) = (0.08, 0.16)$
  - large MR $= 2\times$small MR
- Group size: $n \in \{25, 50, 100\}$.
- No. of groups : $S \in \{50, 100\}$.
- Report mean and std. dev of our closed-form estimates from $Q = 100$ replicated samples.

Table 1(a): MR Estimates (Small)

Small	π_1 =0.2	π_0 =0.1	$p_0^{(1)}$ =0.1	$p_1^{(1)}$ =0.2	$p_0^{(2)}$ =0.08	$p_1^{(2)}$ =0.16
$S = 50$	$\widehat{\pi}_1$	$\widehat{\pi}_0$	$\widehat{p}_0^{(1)}$	$\widehat{p}_1^{(1)}$	$\widehat{p}_0^{(2)}$	$\widehat{p}_1^{(2)}$
$n = 25$	0.2009	0.1015	0.0990	0.2020	0.0792	0.1638
	(0.0123)	(0.0081)	(0.0061)	(0.0301)	(0.0059)	(0.0349)
$n = 50$	0.1996	0.0998	0.1002	0.2000	0.0800	0.1573
	(0.0063)	(0.0042)	(0.0031)	(0.0150)	(0.0031)	(0.0186)
$n = 100$	0.2000	0.1002	0.1000	0.2007	0.0798	0.1573
	(0.0030)	(0.0021)	(0.0014)	(0.0075)	(0.0015)	(0.0086)
$S = 100$						
$n = 25$	0.1994	0.0997	0.0996	0.1968	0.0804	0.1588
	(0.0099)	(0.0060)	(0.0042)	(0.0241)	(0.0047)	(0.0245)
$n = 50$	0.2006	0.1006	0.0997	0.2011	0.0798	0.1608
	(0.0043)	(0.0029)	(0.0020)	(0.0099)	(0.0019)	(0.0112)
$n = 100$	0.2002	0.1002	0.0999	0.2001	0.0800	0.1609
	(0.0025)	(0.0017)	(0.0011)	(0.0054)	(0.0011)	(0.0067)

Table 1(b): MR Estimates (Large)

Large	$\pi_1 = 0.2$	$\pi_0 = 0.1$	$p_0^{(1)} = 0.2$	$p_1^{(1)} = 0.4$	$p_0^{(2)} = 0.16$	$p_1^{(2)} = 0.32$
$S = 50$	$\widehat{\pi}_1$	$\widehat{\pi}_0$	$\widehat{p}_0^{(1)}$	$\widehat{p}_1^{(1)}$	$\widehat{p}_0^{(2)}$	$\widehat{p}_1^{(2)}$
$n = 25$	0.2032	0.1039	0.1994	0.4012	0.1586	0.3191
	(0.0370)	(0.0260)	(0.0092)	(0.0442)	(0.0112)	(0.0654)
$n = 50$	0.1987	0.0994	0.2005	0.3990	0.1602	0.3137
	(0.0174)	(0.0122)	(0.0045)	(0.0224)	(0.0052)	(0.0330)
$n = 100$	0.2004	0.1006	0.1998	0.4004	0.1598	0.3206
	(0.0084)	(0.0059)	(0.0023)	(0.0100)	(0.0025)	(0.0155)
$S = 100$						
$n = 25$	0.1987	0.0993	0.1995	0.3943	0.1604	0.3142
	(0.0257)	(0.0173)	(0.0062)	(0.0322)	(0.0075)	(0.0452)
$n = 50$	0.2011	0.1012	0.1998	0.4013	0.1594	0.3189
	(0.0123)	(0.0090)	(0.0032)	(0.0159)	(0.0039)	(0.0216)
$n = 100$	0.2004	0.1003	0.1999	0.4003	0.1599	0.3201
	(0.0059)	(0.0042)	(0.0017)	(0.0073)	(0.0017)	(0.0112)

Table 1(c): Estimation of Peer Effects: small MR

	$S=50$					$S=100$				
	Naive		Adjusted		Oracle	Naive		Adjusted		Oracle
Reg.	$H^{(1)}y$	$H^{(2)}y$	$\mathcal{H}^{(1)}y$	$\mathcal{H}^{(2)}y$	Gy	$H^{(1)}y$	$H^{(2)}y$	$\mathcal{H}^{(1)}y$	$\mathcal{H}^{(2)}y$	Gy
IV	$H^{(1)}X$	$H^{(2)}X$	$H^{(2)}X$	$H^{(1)}X$	GX	$H^{(1)}X$	$H^{(2)}X$	$H^{(2)}X$	$H^{(1)}X$	GX
$n=25$	Expected # of peers 3.75									
$\lambda=0.05$	0.0259	0.0307	0.0490	0.0467	0.0508	0.0283	0.0324	0.0517	0.0511	0.0489
s.t.d	(0.007)	(0.006)	(0.012)	(0.014)	(0.005)	(0.005)	(0.005)	(0.008)	(0.009)	(0.007)
$\beta_1=1$	1.0613	1.0523	1.0113	1.0131	1.0108	1.0614	1.0540	1.0102	1.0117	1.0112
s.t.d	(0.078)	(0.081)	(0.079)	(0.086)	(0.062)	(0.064)	(0.066)	(0.062)	(0.064)	(0.078)
$\beta_2=2$	1.9978	1.9983	1.9950	1.9951	2.0018	2.0064	2.0058	2.0041	2.0027	1.9946
s.t.d	(0.046)	(0.046)	(0.047)	(0.047)	(0.031)	(0.032)	(0.032)	(0.034)	(0.032)	(0.046)
$n=50$	Expected # of peers 7.5									
$\lambda=0.05$	0.0274	0.0312	0.0492	0.0497	0.0499	0.0274	0.0310	0.0495	0.0493	0.0499
s.t.d	(0.003)	(0.004)	(0.006)	(0.006)	(0.003)	(0.002)	(0.003)	(0.005)	(0.004)	(0.003)
$\beta_1=1$	1.1001	1.0836	1.0029	0.9971	1.0019	1.1021	1.0897	1.0010	1.0059	0.9988
s.t.d	(0.068)	(0.064)	(0.067)	(0.060)	(0.043)	(0.047)	(0.047)	(0.047)	(0.046)	(0.060)
$\beta_2=2$	2.0036	2.0032	2.0021	2.0008	1.9991	2.0017	2.0013	1.9990	1.9983	2.0010
s.t.d	(0.032)	(0.031)	(0.035)	(0.032)	(0.020)	(0.021)	(0.020)	(0.022)	(0.021)	(0.030)
$n=100$	Expected # of peers 15									
$\lambda=0.05$	0.0277	0.0313	0.0504	0.0504	0.0500	0.0278	0.0313	0.0503	0.0500	0.0501
s.t.d	(0.001)	(0.001)	(0.003)	(0.003)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)
$\beta_1=1$	1.2544	1.2210	0.9984	1.0039	1.0060	1.2589	1.2197	1.0051	0.9999	1.0008
s.t.d	(0.072)	(0.065)	(0.070)	(0.064)	(0.026)	(0.048)	(0.041)	(0.047)	(0.045)	(0.041)
$\beta_2=2$	2.0002	2.0004	1.9983	1.9988	1.9979	2.0017	2.0010	1.9983	1.9973	1.9993
s.t.d	(0.026)	(0.022)	(0.035)	(0.028)	(0.013)	(0.019)	(0.017)	(0.023)	(0.019)	(0.020)

Table 1(d): Estimation of Peer Effects: large MR

	$S=50$					$S=100$				
	Naive		Adjusted		Oracle	Naive		Adjusted		Oracle
Reg.	$H^{(1)}y$	$H^{(2)}y$	$\mathcal{H}^{(1)}y$	$\mathcal{H}^{(2)}y$	Gy	$H^{(1)}y$	$H^{(2)}y$	$\mathcal{H}^{(1)}y$	$\mathcal{H}^{(2)}y$	Gy
IV	$H^{(1)}X$	$H^{(2)}X$	$H^{(2)}X$	$H^{(1)}X$	GX	$H^{(1)}X$	$H^{(2)}X$	$H^{(2)}X$	$H^{(1)}X$	GX
$n=25$	Expected # of peers 3.75									
$\lambda=0.05$	0.0118	0.0180	0.0460	0.0437	0.0489	0.0136	0.0195	0.0532	0.0500	0.0508
s.t.d	(0.007)	(0.007)	(0.020)	(0.027)	(0.007)	(0.005)	(0.004)	(0.019)	(0.020)	(0.005)
$\beta_1=1$	1.0813	1.0733	1.0117	1.0173	1.0112	1.0822	1.0722	1.0005	1.0189	1.0108
s.t.d	(0.081)	(0.081)	(0.101)	(0.095)	(0.078)	(0.068)	(0.068)	(0.085)	(0.078)	(0.062)
$\beta_2=2$	1.9967	1.9980	1.9951	1.9937	1.9946	2.0045	2.0059	2.0023	2.0027	2.0018
s.t.d	(0.047)	(0.046)	(0.054)	(0.054)	(0.046)	(0.033)	(0.032)	(0.042)	(0.035)	(0.031)
$n=50$	Expected # of peers 7.5									
$\lambda=0.05$	0.0132	0.0188	0.0510	0.0510	0.0499	0.0133	0.0184	0.0491	0.0486	0.0499
s.t.d	(0.003)	(0.003)	(0.014)	(0.020)	(0.003)	(0.002)	(0.002)	(0.009)	(0.011)	(0.003)
$\beta_1=1$	1.1431	1.1273	0.9942	0.9865	0.9988	1.1458	1.1348	0.9956	1.0111	1.0019
s.t.d	(0.072)	(0.068)	(0.097)	(0.088)	(0.060)	(0.050)	(0.051)	(0.067)	(0.071)	(0.043)
$\beta_2=2$	2.0011	2.0027	1.9987	1.9995	2.0010	2.0000	2.0010	1.9967	1.9976	1.9991
s.t.d	(0.030)	(0.031)	(0.046)	(0.036)	(0.030)	(0.022)	(0.021)	(0.030)	(0.022)	(0.017)
$n=100$	Expected # of peers 15									
$\lambda=0.05$	0.0133	0.0185	0.0504	0.0500	0.0501	0.0135	0.0185	0.0500	0.0506	0.0500
s.t.d	(0.001)	(0.001)	(0.008)	(0.008)	(0.001)	(0.001)	(0.001)	(0.005)	(0.006)	(0.001)
$\beta_1=1$	1.3679	1.3357	0.9936	1.0079	1.0008	1.3726	1.3358	1.0079	0.9860	1.0060
s.t.d	(0.092)	(0.086)	(0.136)	(0.115)	(0.041)	(0.060)	(0.055)	(0.096)	(0.087)	(0.026)
$\beta_2=2$	1.9983	1.9996	1.9982	1.9986	1.9993	2.0007	2.0015	1.9995	1.9988	1.9979
s.t.d	(0.027)	(0.026)	(0.061)	(0.045)	(0.020)	(0.210)	(0.019)	(0.046)	(0.035)	(0.014)

Application: Microfinance in Indian Villages

- Data source: Banerjee et al (2013). Over 4.1k households from 43 villages in Karnataka, India.
- Dependent variable y: participation in a micro-finance program. Average participation rate is 18.9%
- Covariates X are demographics 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 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 differ substantially in data, showing both are noisy measures of G.
- We assume the 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 Variables (No. obs: 4149)

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

Table 2(b): Summary of Category Variables

Variable	value	obs.	per.	Variable	value	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				*property*	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				*caste*			
-	No power	243	5.86	-	Scheduled caste	1139	27.54
-	Private	2662	64.18	-	Scheduled tribe	221	5.34
-	Government	1243	29.97	-	OBC	2253	54.47
				-	General	523	12.65

Table 3 Degree Distribution in 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

- Adjusted SF of a linear prob model:

$$y = \lambda \mathcal{H}^{(t)} y + X\beta + villageFE + v^{(t)}.$$

- MR Estimates

$$\begin{aligned} \hat{p}_0^{(1)} &= 0.002,\ \hat{p}_1^{(1)} = 0.143; \\ \hat{p}_0^{(2)} &< 0.001,\ \hat{p}_1^{(2)} = 0.108. \end{aligned}$$

- Adjusted 2SLS estimates are calculated from a single, large network.

We report five versions of 2SLS estimates:

(a) & (c): "Naive" 2SLS treating $H^{(1)}$ & $H^{(2)}$ as true G.

(b) & (d): adjusted 2SLS using $H^{(3-t)}X$ as IVs for $H^{(t)}y$, $t = 1, 2$.

(e): adjusted 2SLS exploiting stacks moments implied in (b) & (d).

Table 4: Two-stage Least Square Estimates

	OLS	(a)	(b)	(c)	(d)	(e)
R.h.s. Endogeneity Instruments		$H^{(1)}y$ $H^{(1)}X$	$\mathcal{H}^{(1)}y$ $H^{(2)}X$	$H^{(2)}y$ $H^{(2)}X$	$\mathcal{H}^{(2)}y$ $H^{(1)}X$	$\mathcal{H}^{(t)}y$ Combined
$\widehat{\lambda}$		0.0523***	0.0499***	0.0550***	0.0542***	0.0515***
		(0.0079)	(0.0086)	(0.0097)	(0.0082)	(0.0083)
leader	0.0515***	0.0371**	0.0355**	0.0414**	0.0403**	0.0379**
	(0.0175)	(0.0187)	(0.0188)	(0.0184)	(0.0184)	(0.0185)
age	-0.0012***	-0.0017***	-0.0017***	-0.0016***	-0.0017***	-0.0017***
	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
ration	0.0502**	0.0438**	0.0430**	0.0420**	0.0412**	0.0422**
	(0.0212)	(0.0201)	(0.0202)	(0.0195)	(0.0194)	(0.0198)
$electricity - gov$	0.0441**	0.0338**	0.0326**	0.0349**	0.0339**	0.0333**
	(0.0152)	(0.0157)	(0.0158)	(0.0156)	(0.0155)	(0.0156)
$electricity - no$	0.0162	0.0226	0.0233	0.0240	0.0248	0.0240
	(0.0275)	(0.0296)	(0.0296)	(0.0300)	(0.0298)	(0.0297)
$caste - tribe$	-0.0411	-0.0278	-0.0263	-0.0270	-0.0255	-0.0260
	(0.0294)	(0.0309)	(0.0305)	(0.0301)	(0.0298)	(0.0301)
$caste - obc$	-0.0822***	-0.0505**	-0.0468**	-0.0472**	-0.0435***	-0.0456***
	(0.0163)	(0.0217)	(0.0214)	(0.0218)	(0.0210)	(0.0212)
$caste - gen$	-0.1142***	-0.0718***	-0.0669***	-0.0669***	-0.0620**	-0.0650***
	(0.0239)	(0.0238)	(0.0244)	(0.0244)	(0.0235)	(0.0241)
$religion - Islam$	0.1225***	0.0967***	0.0938***	0.0880***	0.0843***	0.0895***
	(0.0332)	(0.0325)	(0.0325)	(0.0346)	(0.0349)	(0.0335)
$religion - Chri$	0.1569	0.1427	0.1410	0.1462	0.1450	0.1431
	(0.1440)	(0.1295)	(0.1279)	(0.1310)	(0.1299)	(0.1287)
Controls	√	√	√	√	√	√
VillageFE	√	√	√	√	√	√
R^2	0.0862	0.1339	0.1353	0.1356	0.1366	0.1358
Obs	4134	4134	4134	4134	4134	4134

Note: s.e. clustered at village level are 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

- Empirical findings:
  - misclassification rates are low on average; mostly due to missing links (p_0 near zero; p_1 around 0.11 and 0.14).
  - $\lambda \approx 0.051$: additional participating "neighbor" increases own participation prob by 5.1%
  - ignoring link misclassification by using traditional 2SLS yields peer effect λ estimates biased upward.

Conclusion

- We propose a simple method for applying 2SLS when some links are randomly misclassified.
- We estimate peer effects on participation in a microfinance program in India.
  - we find low rates of link misclassification.
  - errors in link measures are empirically important.

THANK YOU!