

Heterogeneous Endogenous Effects in Networks


Sida Peng

Microsoft Research

Motivation

- Learning who is influential in a network is important.
 - Some students influence their classmates' smoking decisions.
 - Hedge funds' portfolios have larger impact on others' portfolios.
 - Online opinion leaders influence other users to tweet news.
 - Experienced workers can boost the productivity of their co-workers.
- Policy targeting the right individuals is more efficient than targeting entire population.

Motivation

- Learning who is influential in a social network via survey can be costly, may contain bias and often infeasible.
- Alternative: model-based approach
 - Taking the network as given, individual's decision depends on its neighbors' decision.
 - Spatial Autoregressive models (SARs) are the most widely used tool to model and estimate peer effects. 

Limitation of SAR

- SAR assumes a **constant rate** of influence.
 - ⇒ **Homogeneous** influence in the network.
 - ⇒ **More connections** (centrality) implies **higher influence**.
- **Panel data** is required to introduce group level heterogeneity.
 - Most network data are single **cross-sections**.

My Model

- My model generalizes existing SARs:
 - allows for heterogeneity at individual level.
 - allows for multiple types of connections (friendship, borrowing/lending).
 - nests standard SARs as special case.
- Main features:
 - Identifies **leaders** (with non-zero endogenous effect) and **followers** (with zero endogenous effect).
 - Identifies types of connections relevant to decision-making.
 - Does not require panel data.

Literature and Contributions

- My paper extends the literature on key players in network:
 - e.g. Ballester et al. (Econometrica, 2005)
- My paper expands the use of **LASSO** in network:
 - e.g. Manresa (2013), de Paula et al. (2015)
- Technical Contributions:
 - Derive statistical properties for my LASSO estimator.
 - ▷ e.g. Belloni et al. (Biometrika, 2011)
 - Derive uniformly valid inference including confidence intervals.
 - ▷ e.g. van de Geer et al. (Ann. Stat., 2014)



Empirical Application

- A non-profit (BSS) provides small loans to poor women in rural India.
- “Predefined leaders” are selected by BSS to spread information about the micro-finance program.
 - The fact that a villager is selected as a “predefined leader” does not *a priori* guarantee her *influence*.
- My results show: connectedness \neq influence.
 - Barbers, hotel workers and tailors, who have many connections and selected as “predefined leaders” are not truly influential.

Outline

Background

Network Primer

Standard SAR

Model

Heterogeneous Endogenous Effects Model

Assumptions

Estimation

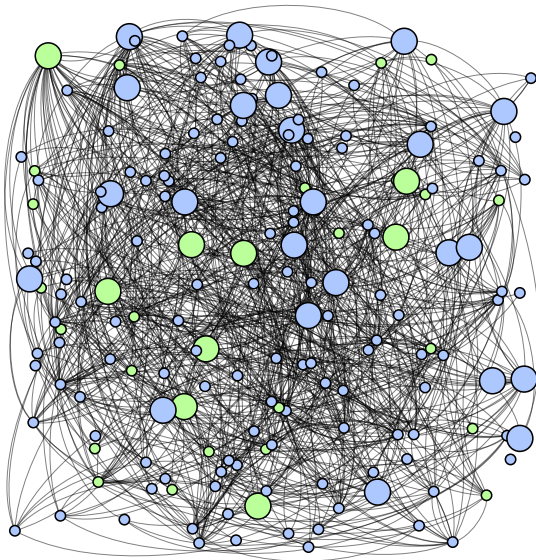
Estimator

Asymptotic Distribution

Results

Empirical Application

Network Primer



- Green vs Blue participant vs non-participant
- Big nodes “pre-defined leaders”
- Edges: know each other

Network Primer

- Network data represent how individuals are connected, typically in a matrix form.
 - Each column/row represents an individual (node).
 - Each entry $M_{ij} = 1$ if individual i and j are connected.
 - The set of neighbors for individual i is defined as:
$$N_i = \{j : M_{ij} = 1\}.$$
 - Different interactions can be viewed as multiple networks on the same set of individuals: $M^{(k)}$, $k = 1, 2, \dots, q$.
- My method picks out the most important types of interactions that predict the decision.
- My method picks out influential individuals who have non-zero endogenous effects on their neighbors' decision.

Outline

Background

Network Primer

Standard SAR

Model

Heterogeneous Endogenous Effects Model

Assumptions

Estimation

Estimator

Asymptotic Distribution

Results

Empirical Application

Standard SAR

Decision to Join the Micro-finance Program

d_i : individual i 's decision to participate

$$d_i = \alpha \sum_{j \in N_i} d_j + x_i \beta_0 + \epsilon_i$$

- Standard SAR model assumes a constant rate of influence α .
- Every individual in the network is assumed to influence her neighbors at the same rate.

Standard SAR

Decision to Join the Micro-finance Program

d_i : individual i 's decision to participate

$$d_i = \alpha \sum_{j \in N_i} d_j + x_i \beta_0 + \epsilon_i$$

- Standard SAR model assumes a constant rate of influence α .
- Every individual in the network is assumed to influence her neighbors at the same rate.

Standard SAR

- Heterogeneity can be specified exogenously:

Decision to Join the Micro-finance Program

d_i : individual i 's decision to participate

$$d_i = \alpha \sum_{j \in N_i} d_j w_{ij} + x_i \beta_0 + \epsilon_i$$

- where w_{ij} are spatial weights between individual i and j .
- My model allows heterogeneity to be identified by the data.
- Relaxes the assumption that individuals with high centrality are the key players in the network.

Outline

Background

Network Primer

Standard SAR

Model

Heterogeneous Endogenous Effects Model

Assumptions

Estimation

Estimator

Asymptotic Distribution

Results

Empirical Application

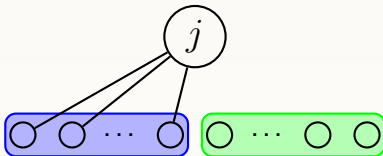
Heterogeneous Endogenous Effects Model

- Consider the decision to join the micro-finance program.

$$\text{SAR: } d_i = \alpha \sum_{j \in N_i} d_j + x_i \beta_0 + \epsilon_i$$

$$\text{My Model: } d_i = \sum_{j \in N_i} d_j \eta_j + x_i \beta_0 + \epsilon_i$$

- where each η_j represents endogenous effect of individual j .



Generalization

- Heterogeneous Endogenous Effects model

$$d_i = \sum_{j \in N_i} d_j \eta_j + x_i \beta_0 + \epsilon_i$$

- η_j can be identified as fixed effect of individual j .

- Pair-wise Heterogeneous Endogenous Effects model

$$d_i = \sum_{j \in N_i} d_j \eta_{ij} + x_i \beta_0 + \epsilon_i$$

- Panel data is required to identify η_{ij} .
- Different types of LASSO estimator will be proposed.

Heterogeneous Endogenous Effects

Heterogeneous Endogenous Effects Model

$$D_n = \left(M_n \circ D_n \right) \eta_0 + X_n \beta_0 + \epsilon_n,$$

- ▶ D_n is n by 1 dependent variables of interest.
- ▶ M_n is n by n adjacency matrix.
- ▶ η_0 is n by 1 parameter.
- ▶ X_n is n by k individual characteristic matrix.
- ▶ β_0 is k by 1 parameter.
- ▶ Operator “ \circ ” is defined as:

$$A \circ B = A \cdot \text{diag}(B)$$

Two Main Problems

- There are $n + k$ unknowns but only n observations.
 - Key players are a small fraction of the total population.
 - LASSO can be used to estimate the structure.
- $\sum_{j \in N_i} d_j \eta_j$ is correlated with ϵ_i .
 - I propose a set of instruments for $(M_n \circ D_n)$.

Two Main Problems

- There are $n + k$ unknowns but only n observations.
 - Key players are a small fraction of the total population.
 - LASSO can be used to estimate the structure.
- $\sum_{j \in N_i} d_j \eta_j$ is correlated with ϵ_i .
 - I propose a set of instruments for $(M_n \circ D_n)$.

Outline

Background

Network Primer

Standard SAR

Model

Heterogeneous Endogenous Effects Model

Assumptions

Estimation

Estimator

Asymptotic Distribution

Results

Empirical Application

Assumption 1: Sparsity


Assumption 1

Let $S_n \subset \{1, 2, \dots, n\}$ denote the set of influential individuals (i.e. $\eta_j \neq 0$). Let $s_n = |S_n|$ be the number of elements in S_n . Then,

$$s_n = o\left(\frac{\sqrt{n}}{\log n}\right), \quad \text{as } n \rightarrow \infty$$

- Only a small number of the individuals are leaders and the rest are followers.
- Sparsity assumption can be relaxed in an extension of the model.

Instruments

- Solve D_n as a function of exogenous variables. 
- For simplicity, let X_n be n by 1 vector and β_0 be a scale.

$$E(D_n) = X_n\beta_0 + (M_n \circ X_n)(\beta_0\eta_0) + \sum_{i=2}^{\infty} (M_n \circ \eta_0)^i \beta_0 X_n$$

- Consider first the ideal case: set of influential individuals (S_n) is known.
 - Define $(\cdot)_{S_n}$ as matrix restricted to those columns indexed in S_n .

Instruments

- The non-zero columns in $(M_n \circ X_n)(\beta_0 \eta_0)$ are $(M_n \circ X_n)_{S_n}$.
- The exogenous characteristics of influential individuals can be used as instruments for their neighbors.
- $(M_n \circ X_n)_{S_n}$ is correlated with D_n .
- $(M_n \circ X_n)_{S_n}$ is not correlated with ϵ_n .

$\Rightarrow (M_n \circ X_n)_{S_n}$ and X_n are valid instruments.

Assumption 3: Independence

$[X_n, (M_n \circ X_n)_S]$ is full rank.

Instruments

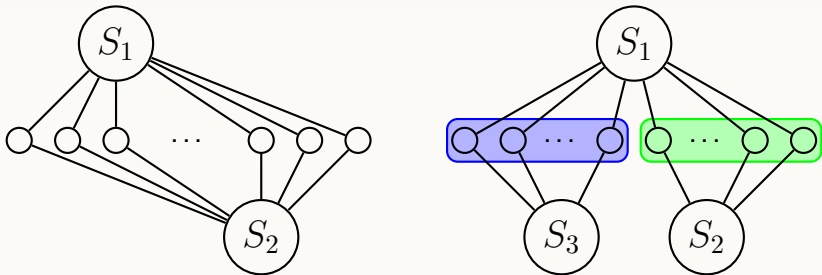
- The non-zero columns in $(M_n \circ X_n)(\beta_0 \eta_0)$ are $(M_n \circ X_n)_{S_n}$.
- The exogenous characteristics of influential individuals can be used as instruments for their neighbors.
- $(M_n \circ X_n)_{S_n}$ is correlated with D_n .
- $(M_n \circ X_n)_{S_n}$ is not correlated with ϵ_n .

$\Rightarrow (M_n \circ X_n)_{S_n}$ and X_n are valid instruments.

Assumption 3: Independence

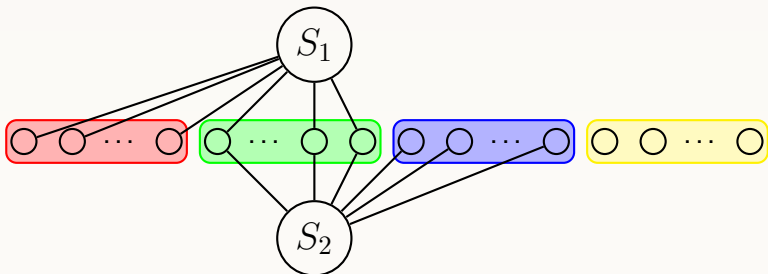
$[X_n, (M_n \circ X_n)_S]$ is full rank.

Assumption 3: Independence



- If influential individuals are connected to every node (left) or to a complete partition of nodes (right), perfect collinearity occurs and Assumption 3 is violated.

Assumption 3: Independence



- Influence of S_1 can be estimated by comparing red and yellow groups, while influence of S_2 can be estimated by comparing blue and yellow groups.
- One can interpret the η_0 coefficients as fixed effect for influential individuals.

S_n unknown

- Typically, S_n is not known to us.
- But $(M_n \circ X_n)$ contain the valid instruments $(M_n \circ X_n)_{S_n}$.
- Use LASSO to select instruments in the first stage.

LASSO Estimator


$$(\tilde{\beta}, \tilde{\eta}) = \min_{\beta, \eta} \left\| D_n - X_n \beta - (M_n \circ X_n) \eta \right\|_2 + \lambda |\eta|_1$$

- The l_1 norm introduced as a penalty in the minimization problem enforces sparsity in the estimator.

Assumption 4: LASSO assumptions

- I assume **Irrepresentable Condition** and **Beta Min Condition** to guarantee consistent selection for LASSO, Zhao and Yu (JMLR, 2006).

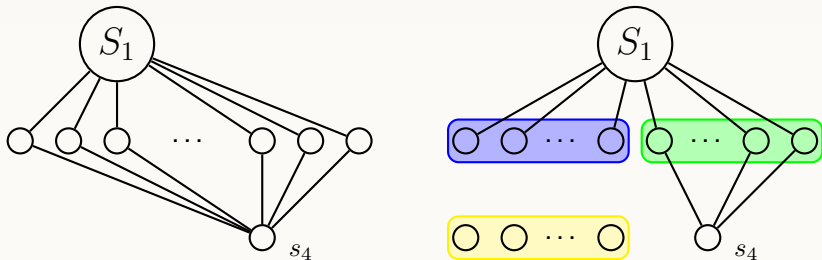
$$P(\{i : \hat{\eta}_i \neq 0\} = \{i : \eta_i \neq 0\}) = 1 \quad \text{as } n \rightarrow \infty$$

- Irrepresentable Condition imposes restriction on neighbors of influential and non-influential individuals. 

 - It implies bounds on the lowest eigenvalue of design matrix.



- Beta Min Condition assumes all endogenous effects are larger than a threshold.

Irrepresentable Condition



- If influential individuals have the same neighbors as an influential individual (left), Irrepresentable Condition is violated.
- Influence of S_1 can be estimated by comparing blue and yellow groups.
- s_4 is identified as non-influential once we compare blue and green.

Summary and Extensions

- Under Assumptions 1-5, my model is identified as a linear system with a unique solution.
- Further extensions can allow for more flexible network structures.
 - Network with multiple cliques. 
 - Existence of multiple networks (i.e. multiple types of connections among individuals). 

Outline

Background

- Network Primer
- Standard SAR

Model

- Heterogeneous Endogenous Effects Model
- Assumptions

Estimation

- Estimator
- Asymptotic Distribution

Results

- Empirical Application

Two Stage LASSO

Two Stage LASSO

- Estimate the first stage:

$$(\tilde{\beta}, \tilde{\eta}) = \arg \min_{\beta, \eta} \left\| D_n - X_n \beta - (M_n \circ X_n) \eta \right\|_2 + \lambda |\eta|_1$$

and obtain a LASSO fitting \hat{D}_n :

$$\hat{D}_n = X_n \tilde{\beta} + (M_n \circ X_n) \tilde{\eta}$$

- Estimate the second stage:

$$(\hat{\beta}, \hat{\eta}) = \arg \min_{\beta, \eta} \left\| D_n - (M_n \circ \hat{D}_n) \eta - X_n \beta \right\|_2 + \lambda |\eta|_1,$$

De-sparse 2SLSS Estimator

- $(\hat{\beta}, \hat{\eta})$ as defined in the previous slides are asymptotically biased.
 - Post-model-selection inference is not uniformly valid.
- To eliminate bias, I construct the following estimators for my Two-stage LASSO estimator:

$$\hat{e} = \hat{\eta} + \hat{\Theta}(M_n \circ \hat{D}_n)'(D_n - X\hat{\beta} - (M_n \circ X_n)\hat{\eta})/n$$

$$\hat{b} = \hat{\beta} - (X_n'X_n)^{-1}X_n'(M_n \circ \hat{D}_n)\hat{\Theta}(M_n \circ \hat{D}_n)'(D_n - (M_n \circ \hat{D}_n)\hat{\eta} - X_n\hat{\beta})/n$$

- $\hat{\Theta}$ is constructed by nodewise regression as in van de Geer (2014).
- Define the set of estimated influential individuals as

$$\hat{S}_n = \{i | \hat{\eta} \neq 0\}$$

Outline

Background

- Network Primer
- Standard SAR

Model

- Heterogeneous Endogenous Effects Model
- Assumptions

Estimation

- Estimator
- Asymptotic Distribution**

Results

- Empirical Application

Asymptotic Distribution

Theorem 1

Under Assumptions 1-5, when $\lambda \propto \sqrt{\log n/n}$

- $\lim_{n \rightarrow \infty} \mathbb{P}(\hat{S}_n = S) = 1$
- $\hat{e} \rightarrow \eta_0$
- $\hat{b} \rightarrow \beta_0$

Theorem 2

Under Assumptions 1-5, when $\lambda \propto \sqrt{\log n/n}$, for any $\iota : \|\iota\|_0 < \infty$

$$\sqrt{n}\iota'(\hat{e} - \eta_0) \rightarrow N(0, \sigma^2 \iota' \Theta_1 \text{diag}(\Gamma) \Omega \text{diag}(\Gamma) \Theta_1' \iota),$$

$$\sqrt{n}(\hat{b} - \beta_0) \rightarrow N(0, \sigma^2 \Theta_2 \text{diag}(\Gamma) \Omega \text{diag}(\Gamma) \Theta_2'),$$

Asymptotic Distribution

Theorem 1

Under Assumptions 1-5, when $\lambda \propto \sqrt{\log n/n}$

- $\lim_{n \rightarrow \infty} \mathbb{P}(\hat{S}_n = S) = 1$
- $\hat{e} \rightarrow \eta_0$
- $\hat{b} \rightarrow \beta_0$


Theorem 2

Under Assumptions 1-5, when $\lambda \propto \sqrt{\log n/n}$, for any $\iota : \|\iota\|_0 < \infty$

$$\sqrt{n}\iota'(\hat{e} - \eta_0) \rightarrow N(0, \sigma^2 \iota' \Theta_1 \text{diag}(\Gamma) \Omega \text{diag}(\Gamma) \Theta_1' \iota),$$

$$\sqrt{n}(\hat{b} - \beta_0) \rightarrow N(0, \sigma^2 \Theta_2 \text{diag}(\Gamma) \Omega \text{diag}(\Gamma) \Theta_2'),$$

Asymptotic Distribution

- Similar asymptotics can be constructed for extensions with cliques and multiple networks.
- My proof extends the existing literature on inference for LASSO estimator in two aspects:
 - I derive asymptotic distributions for my LASSO estimators.
 - ▷ Extend the "de-sparse" LASSO in van de Geer (2014) to square-root LASSO and under two-stage setting.
 - I derive statistical properties for square-root sparse group LASSO used in extension with multiple networks. 

Outline

Background

- Network Primer
- Standard SAR

Model

- Heterogeneous Endogenous Effects Model
- Assumptions


Estimation

- Estimator
- Asymptotic Distribution

Results

- Empirical Application

Background

- Bharatha Swamukti Samsthe (BSS) is a non-profit organization providing small loan products to poor women.
- The loan is around 10,000 rupees (approximately \$200) with an annualized rate about 28% and is repaid in 50 weeks.
- In 2006, 75 villages were surveyed 6 month before the BSS's entry.
- By the time of 2011, BSS had entered 43 of those villages.
- BSS provided data on who joined the program (D_n) and individual characteristic (X_n). 

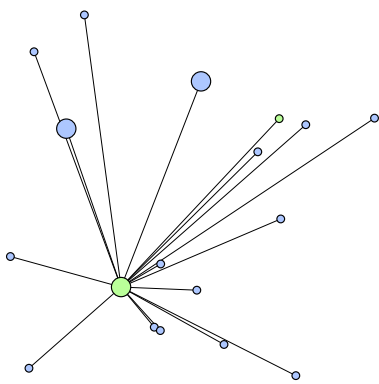
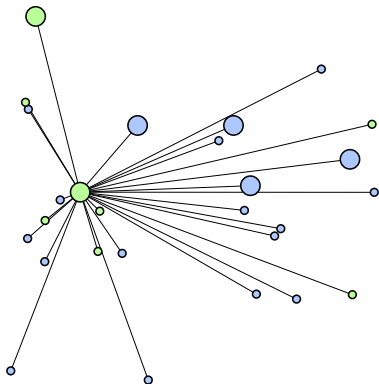
Multiple Networks

- I consider 8 different networks among families:
 - visits each other's house in his or her free time
 - borrow rice (kerosene) from each other
 - non-relative friends
 - ask for help under medical emergency
 - borrow money from each other
 - ask for advice or help with decision
 - visit temple/mosque/church together
 - relatives
- I estimate a linear probability model as an illustration of my method.

Influential vs Non-Influential Individual

Influential ($\eta_i \neq 0$)

Non-influential ($\eta_i = 0$)

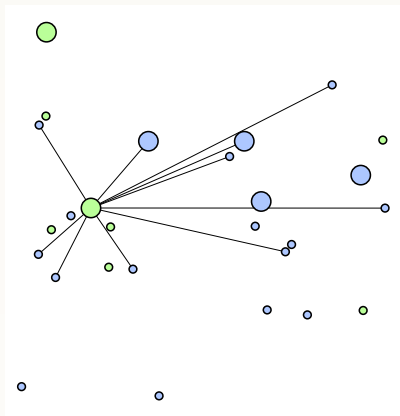
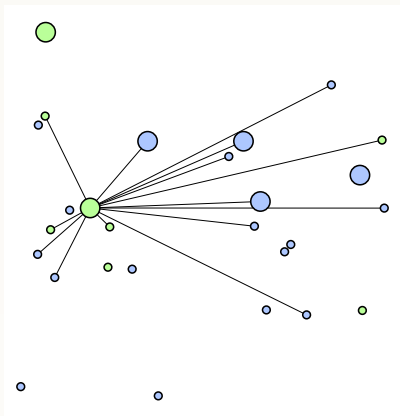


Green: participant; Blue: non-participant

Influential vs Non-Influential Network

Visit-go and come ($\eta_i^l \neq 0$)

Help decision ($\eta_i^l = 0$)



Green: participant; Blue: non-participant

Empirical Evidence: Influential Networks

- Among 37 villages, which networks are influential.

Second Stage: influential networks

	visit go-come	borrow-lend keroric	borrow-lend money	friendship	medical help	help decision	relatives	temple company
Cross-validation:								
probability	51%	43%	41%	41%	30%	32%	30%	14%
identified	18	12	13	14	9	14	9	6
De-sparse:								
probability	51%	46%	51%	43%	32%	41%	43%	19%
identified	3	3	3	2	3	3	3	2

▶ Centrality

▶ Magnitude

Empirical Evidence: Influential Networks

- Detection based on LASSO using Cross-validation.

Second Stage: influential networks

	visit go-come	borrow-lend keroric	borrow-lend money	friendship	medical help	help decision	relatives	temple company
Cross-validation:								
probability	51%	43%	41%	41%	30%	32%	30%	14%
identified	18	12	13	14	9	14	9	6
De-sparse:								
probability	51%	46%	51%	43%	32%	41%	43%	19%
identified	3	3	3	2	3	3	3	2

► Centrality

► Magnitude

Empirical Evidence: Influential Networks

- probability: Empirical probability of at least one leader being detected in a given network.

Second Stage: influential networks

	visit go-come	borrow-lend keroric	borrow-lend money	friendship	medical help	help decision	relatives	temple company
Cross-validation:								
probability identified	51% 18	43% 12	41% 13	41% 14	30% 9	32% 14	30% 9	14% 6
De-sparse:								
probability identified	51% 3	46% 3	51% 3	43% 2	32% 3	41% 3	43% 3	19% 2

► Centrality

► Magnitude

Empirical Evidence: Influential Networks

- identified: Average number of leaders detected conditioning on at least one leader being detected.

Second Stage: influential networks

	visit go-come	borrow-lend keroric	borrow-lend money	friendship	medical help	help decision	relatives	temple company
Cross-validation:								
probability	51%	43%	41%	41%	30%	32%	30%	14%
identified	18	12	13	14	9	14	9	6
De-sparse:								
probability	51%	46%	51%	43%	32%	41%	43%	19%
identified	3	3	3	2	3	3	3	2

▸ Centrality

▸ Magnitude

Empirical Evidence: Influential Networks

- Detection based on De-sparse LASSO estimator controlling FDR at 5%.

Second Stage: influential networks

	visit go-come	borrow-lend keroric	borrow-lend money	friendship	medical help	help decision	relatives	temple company
Cross-validation:								
probability	51%	43%	41%	41%	30%	32%	30%	14%
identified	18	12	13	14	9	14	9	6
De-sparse:								
probability	51%	46%	51%	43%	32%	41%	43%	19%
identified	3	3	3	2	3	3	3	2

► Centrality

► Magnitude

Empirical Evidence: Leaders

- Overlap between LASSO detected leaders and “predefined leaders”.

Second Stage: coverage of predefined leaders

	Coverage	Total Number of Discovery
Cross-validation	19%	22
De-sparse	13%	6

Empirical Evidence: Leaders

- Percentage of leaders that are also “predefined leaders”.

Second Stage: coverage of predefined leaders

	Coverage	Average Number of Discoveries
Cross-validation	19%	22
De-sparse	13%	6

Empirical Evidence: Leaders

- Total number of leaders detected by lasso.

Second Stage: coverage of predefined leaders

	Coverage	Average Number of Discoveries
Cross-validation	19%	22
De-sparse	13%	6

Empirical Evidence: Leaders

- There are on average 27 predefined leaders in each village.

Second Stage: coverage of predefined leaders

	Coverage	Average Number of Discoveries
Cross-validation	19%	22
De-sparse	13%	6

Empirical Evidence: Leaders

Second Stage: who they are

	Predefined leaders	Selected by LASSO	Participate
Agriculture labour	-0.01 (0.01)	0.07*** (0.01)	0.05* (0.03)
Anganwadi Teacher	0.04 (0.06)	0.12** (0.06)	0.07 (0.13)
Blacksmith	-0.08 (0.09)	0.16* (0.09)	-0.23 (0.20)
Construction/mud work	0.01 (0.03)	0.06** (0.03)	0.22*** (0.05)
Police officer	-0.15 (0.19)	0.33* (0.19)	-0.04 (0.40)
Mechanic	0.01 (0.06)	0.13** (0.06)	-0.12 (0.13)
Skilled labour/work for company	0.05 (0.05)	0.08* (0.05)	0.03 (0.10)
Small business	0.20*** (0.02)	0.06*** (0.02)	0.13*** (0.05)
Tailor garment worker	0.09*** (0.03)	0.03 (0.03)	0.12* (0.06)
Hotel worker	0.33*** (0.08)	0.08 (0.07)	0.43*** (0.16)
Poojari	0.37*** (0.14)	-0.15 (0.29)	0.15 (0.13)
Veterinary clinic	0.86*** (0.33)	0.04 (0.33)	1.91*** (0.70)
Barber/salon	0.49*** (0.10)	0.04 (0.10)	-0.00 (0.21)
Doctor/Health assistant	0.27** (0.11)	0.09 (0.10)	0.27 (0.22)

Standard errors in parentheses * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Village fixed effects are controlled
45 different careers controlled

Empirical Evidence: Leaders

Second Stage: who they are

	Predefined leaders	Selected by LASSO	Participate
Agriculture labour	-0.01 (0.01)	0.07*** (0.01)	0.05* (0.03)
Anganwadi Teacher	0.04 (0.06)	0.12** (0.06)	0.07 (0.13)
Blacksmith	-0.08 (0.09)	0.16* (0.09)	-0.23 (0.20)
Construction/mud work	0.01 (0.03)	0.06** (0.03)	0.22*** (0.05)
Police officer	-0.15 (0.19)	0.33* (0.19)	-0.04 (0.40)
Mechanic	0.01 (0.06)	0.13** (0.06)	-0.12 (0.13)
Skilled labour/work for company	0.05 (0.05)	0.08* (0.05)	0.03 (0.10)
Small business	0.20*** (0.02)	0.06*** (0.02)	0.13*** (0.05)
Tailor garment worker	0.09*** (0.03)	0.03 (0.03)	0.12* (0.06)
Hotel worker	0.33*** (0.08)	0.08 (0.07)	0.43*** (0.16)
Poojari	0.37*** (0.14)	-0.15 (0.29)	0.15 (0.13)
Veterinary clinic	0.86*** (0.33)	0.04 (0.33)	1.91*** (0.70)
Barber/salon	0.49*** (0.10)	0.04 (0.10)	-0.00 (0.21)
Doctor/Health assistant	0.27** (0.11)	0.09 (0.10)	0.27 (0.22)

Standard errors in parentheses * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Village fixed effects are controlled
45 different careers controlled

Empirical Evidence: Counter Factual

- Target leaders and make them participate.

Counter Factual			
	In data	Predefined Leaders	LASSO Leaders
Participation Rate(non-leaders)	16%	20%	33%

Conclusion

- Propose heterogeneous endogenous effects model.
- Propose instruments and a de-sparse two-stage LASSO estimator.
- Construct uniformly valid inference for my estimator.
- Empirical application: Micro-finance in rural India
 - “visit go-come”, “borrow-lend keroric”, “borrow-lend mony”, “friendship” networks are influential.
 - LASSO-detected leaders are different from predefined leaders.

Notation

- Let

$$\Gamma = \lim_{n \rightarrow \infty} (I - M_n \circ \eta_0)^{-1} X_n \beta$$

$$\Theta_1 = \lim_{n \rightarrow \infty} \hat{\Theta}, \quad Z = (M_n \circ \hat{D}_n), \quad \tilde{Z} = X_n (X_n' X_n)^{-1} X_n' Z,$$

$$\Theta_2 = \lim_{n \rightarrow \infty} \frac{1}{n} \left(I - Z \hat{\Theta} \tilde{Z}' / n \right)' X_n (X_n' X_n)^{-1} X_n' \left(I - Z \hat{\Theta} \tilde{Z}' / n \right)$$

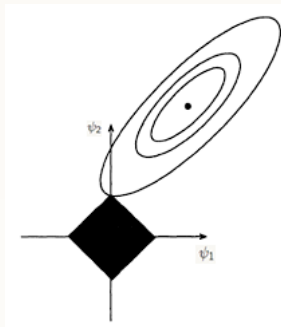
▶ Go back

LASSO Estimator

- LASSO is a model fitting and selection tool:

$$\hat{\psi} = \arg \min_{\psi} \|Y - X\psi\|_2^2 + \lambda \|\psi\|_1$$

- Sparsity due to the geometry of l_1 norm



Empirical Literature

- **Education:**

(Sacerdote 2001 QJE), (Neidell and Waldfogel 2011 RES), (Lavy and Sand 2015)

- **Finance:**

(Bonaldi et al. 2014), (Denbee et al. 2014)

- **Health:**

(Krauth, 2006 CJE), (Nakajima, 2007 RES), (Clark and Loheac 2007 JHE), (Christakis and Fowler 2007 NJM)

- **Labor and productivity:**

(Topa 2001, RES), (Mas and Moretti 2009 AER), (Guryan et al. 2009 AEJ)

- **Agricultural:**

(Holloway et al. 2002 AE), (Conley and Udry 2010 AER)

Network with Cliques

Heterogeneous Endogenous Effects Model with Cliques

$$D_n = \left(M_n \circ D_n \right) \eta_0 + M_n D_n \gamma_0 + X_n \beta_0 + \epsilon_n,$$

- γ_0 will capture all influence from local leaders.
- Sparsity assumption is only applied to global leaders.
- Global leaders will be identified via η .
- If no global leader exists, it is equivalent to standard SAR model.
- A similar set of assumptions can guarantee identification.

▶ Go back

Multiple Networks

Multiple Networks

$$D_n = \sum_{j=1}^q (M_n^j \circ D_n) \eta_0^j + X_n \beta_0 + \epsilon_n$$

- Let η_k^j represents the endogenous effect of individual k via network j .
- I allow the number of networks q to increase as n increases.
- Some networks could be completely irrelevant (i.e. $\eta_0^j = 0$).
- This model can be estimated using the square-root sparse group LASSO.

▶ Go back

Square-root Sparse Group LASSO

- I propose the use of the square-root sparse group LASSO when there exists multiple networks.

$$(\tilde{\beta}, \tilde{\eta}) = \arg \min_{\beta, \eta} \left\{ \left\| D_n - X_n \beta - \sum_{j=1}^q (M_n^j \circ X_n) \eta^j \right\|_2 + \left(\sum_{j=1}^q (\lambda_1 \|\eta^j\|_2 + \lambda_2 \|\eta^j\|_1) \right) \right\}$$

- I derive statistical properties of this estimator in order to prove consistency and asymptotics for the 2SLSS for multiple networks.

▶ Go back

Simulation

Heterogeneous Endogenous Effects Model with 10 influential individuals

p	0.1			0.2			
	n	50	200	500 ¹	50	200	500 ¹
Avgcov S_0		0.9730	0.9870	0.8805	0.6905	0.9870	0.8330
Avglength S_0		11.8263	1.5870	4.5802	0.8400	3.6207	2.1104
Avgcov S_0^c		0.9942	0.9905	0.9638	0.9827	0.9972	0.9733
Avglength S_0^c		23.2425	2.5128	4.0562	9.3871	2.9423	5.5000
Avgcov β		0.9800	0.9700	0.9300	0.9500	0.9950	0.9950
Avglength β		2.6520	0.5203	0.9008	1.2524	0.5261	0.7915

1. For 500 cases, all lasso tuning parameter is chosen using rule of thumb instead of cross-validation

Simulation

Heterogeneous Endogenous Effects Model in Watts-Strogatz networks

p^1	0.04			0.08		
	n	50	200	500	50	200
Avgcov S_0	0.9180	0.8490	0.9920	0.9410	0.8310	0.9860
Avglength S_0	5.7298	1.6333	1.6646	5.2069	3.8309	0.9132
Avgcov S_0^c	0.9543	0.9646	0.9809	0.9577	0.9581	0.9949
Avglength S_0^c	7.8860	5.2748	4.3686	3.4985	2.9435	3.4044
Avgcov β	0.9900	0.9350	0.9933	0.9350	0.9650	0.9950
Avglength β	0.8524	0.4044	0.9067	0.7532	0.5382	1.4130

1. Given the number of node $N = 50, 200, 500$, the mean degree for each node is $0.04N$ and $0.08N$. The rewriting probability is fixed at 0.4.

Simulation

Heterogeneous Endogenous Effects Model with Multiple Networks

p	0.1			0.2		
	n	50	200	500 ⁴	50	200
Avgcov S_0	0.9860	0.9940	0.9930	0.8950	0.9910	1.0000
Avglength S_0	10.2325	0.6168	1.2395	6.1020	0.7531	4.4228
Avgcov S_0^c	0.9923	0.9884	0.9809	0.9893	0.9909	0.9910
Avglength S_0^c	9.2108	2.2820	4.4911	5.9378	1.7050	6.1668
Avgcov β	0.9900	0.9650	0.9950	0.9750	0.9650	0.9900
Avglength β	8.2338	0.5556	0.8615	6.4205	0.6026	4.3825
Network 1 ¹ :						
probability ²	0.8050	0.8950	0.5250	0.7400	0.8500	0.2111
# identified ³	2.3540	3.8547	6.6095	4.0743	3.3314	4.7619
Network 2 ¹ :						
probability ²	0.0450	0.0550	0.3700	0.1300	0.0350	0.0251
# identified ³	4.3333	1.0000	3.9730	3.4615	1.0000	1.0000

1. Network 1 and Network 2 are generated separately using Erdős-Rényi algorithm.
2. Probability: empirical probability that at least one regressor in the group is significant
3. # identified: the averaged number of significant regressors in the group conditioning on at least one regressor in the group is significant.
4. All LASSO tuning parameter is chosen using rule of thumb instead of cross-validation

Simulation

Heterogeneous Endogenous Effects Model with Cliques

p	0.1			0.2		
	50	200	500 ¹	50	200	500 ¹
Avgcov S_0	0.9670	0.9580	0.9850	0.9610	0.9954	0.9980
Avglength S_0	20.3014	1.3383	1.9988	8.6764	2.0044	4.5572
Avgcov S_0^c	0.9665	0.9883	0.9975	0.9680	0.9926	0.9980
Avglength S_0^c	14.0695	3.4002	4.7511	40.5927	1.6113	4.7505
Avgcov β	0.9800	0.9950	0.9900	0.9750	0.9943	0.9950
Avglength β	2.9138	0.8404	0.5866	1.5054	0.6253	0.6881
Avgcov γ	0.9600	0.9950	0.9950	0.9950	1.0000	1.0000
Avglength γ	0.5683	0.1568	0.0257	0.4235	0.0544	0.0294
test- $\gamma = 0$ ²	0.4300	0.3750	1.0000	0.4950	1.0000	1.0000

1. All LASSO tuning parameter is chosen using rule of thumb instead of cross-validation
2. Report the empirical probability of rejecting a z-test on parameter $\gamma = 0$

▶ Go back

Watts-Strogatz

- Define the pN (even number) as the mean degree for each node and a special parameter $\omega = 0.4$.
 - construct a graph with N nodes each connected to pN neighbors, which $\frac{pN}{2}$ on each side.
 - For each node n_i , take every edge (n_i, n_j) with $i < j$ and rewrite it with probability ω . Rewrite means replace (n_i, n_j) with (n_i, n_k) where k is choosing uniformly among all nodes that is not currently connected with n_i

▶ Go back

Parameterization

- I use Erdős-Rényi algorithm and Watts-Strogatz mechanism to simulate networks.
- Individual responses are generated as $Y_n = (I - M_n \circ \eta_0)^{-1}(X_n\beta_0 + \epsilon)$ where ϵ is drawn independently from standard normal distribution.
- I vary number of influential individuals to be either 5 or 10.

Parameterization

- Use (M_n, X_n, Y_n) as observations and estimate (η_0, β_0, S_n) .
- Tuning parameters for the first stage is chosen $\propto \Phi^{-1}(1 - \alpha/(2 * n))/\sqrt{n}$.
- Tuning parameters for the second stage is chosen by cross-validation.
- Monte Carlo simulation is repeated 200 times in each case as in van de Geer (2014).

Simulation

Heterogeneous Endogenous Effects Model in Erdős-Rényi Networks ¹

p	0.1			0.2		
	50	200	500	50	200	500
Avgcov S_0	0.9780	0.9560	0.9380	0.9770	0.9480	0.9580
Avglength S_0	2.9420	3.6734	2.6136	1.0179	3.3098	2.0386
Avgcov S_0^c	0.9222	0.9861	0.9846	0.9920	0.9861	0.9884
Avglength S_0^c	18.9664	8.1006	2.5444	21.4923	3.1052	1.9782
Avgcov β	0.8700	0.9700	0.9650	0.9500	0.9650	0.9800
Avglength β	4.0056	0.4890	0.2959	0.9773	0.7905	0.5209

1. Confidence intervals are constructed at 95% nominal level.

▶ Go back



Simulation

- I report the average coverage probability and average length of confidence intervals for $S_0 = \{\eta_1, \dots, \eta_5\}$, β_0 and the rest of zeros η s: S_0^c . For example:

$$\text{Avgcov } S_0 = s_0^{-1} \sum_{j \in S_0} \mathbb{P}[\eta_j^0 \in CI_j]$$

$$\text{Avglength } S_0 = s_0^{-1} \sum_{j \in S_0} \text{length}(CI_j)$$

▶ Go back

Simulation

- I consider multiple two-sided tests of hypotheses $H_{0,j} : \eta_j = 0$ for $j = 1, 2, \dots, n$.
- I control the False Discovery Rate (FDR) using Benjamin-Hochberg method.
- For measuring power, I report the empirical version of

$$\text{Power} = s_0^{-1} \sum_{j \in S_0} \mathbb{P}[H_{0,j} \text{ is rejected}]$$

$$\text{FDR} = \sum_{j \in S_0^c} \mathbb{P}[H_{0,j} \text{ is rejected}] / \sum_{j=1}^n \mathbb{P}[H_{0,j} \text{ is rejected}]$$

▶ Go back

LASSO assumptions

Assumption 4

(Irrepresentable Condition) There exists $N \in \mathbb{N}$: $\forall n \geq N$, there is a $\vartheta \in (0, 1)$ such that

$$P \left(\left\| \text{diag}((\hat{D}_n)_{S^c}) \Sigma_n \text{diag}((\hat{D}_n)_S)^{-1} \text{sign}(\eta_0) \right\|_{\infty} \leq \vartheta \right) = 1$$

(Beta Min Condition) There exists $N \in \mathbb{N}$: $\forall n \geq N$, there is a $m > 0$ such that

$$\min(|\eta_0|)_S \geq m/\sqrt{n}$$

- where $\Sigma_n = (M_n)'_{S^c} (M_n)_S ((M_n)'_S (M_n)_S)^{-1}$
- \hat{D}_n is the LASSO fit from the first stage. [▶ Go back](#)

Shrinkage Bias and Variance

Assumption 5

(**Maximum Neighbors Condition**) There exists $N \in \mathbb{N}$: $\forall n \geq N$,

$$\|M'_n \mathbf{1}_n\|_\infty = O([\log n]^\epsilon), \quad \epsilon \in (0, 1]$$

(**Variance Condition**)

$$\frac{1}{n} M'_n W_n (I - M_n \circ \eta_0)^{-1} (I - M_n \circ \eta_0)^{-1'} W_n M_n \rightarrow \Omega$$

◦ where $W_n = \left(I - X_n (X'_n X_n)^{-1} X'_n \right)$

▶ Go back

Literature Review

- My LASSO estimator builds on:
Belloni et al. (Biometrika, 2011), Bunea et al. (IEEE, 2013), Zhu (2016)
- My post-LASSO inference builds on:
Potscher and Leeb (Econ Theory, 2008), van de Geer et al. (Ann. Stat., 2014), Belloni et al. (Biometrika, 2015)
- My paper expands the use of LASSO in network:
Manresa (2013), de Paula, Rasul and Souza (2015)
- My paper extends SAR model allowing for heterogeneity:
Kelejian and Prucha (1998), Lee (Econ Theory, 2010), Bramouille et al. (JOE, 2009), Bonhomme and Manresa, (Econometrica, 2015), Rose (2016)
- My paper extends the literature on key players in network:
Ballester et al. (Econometrica, 2005), Calvo Armengol et al. (RES, 2009), Banerjee et al. (2016)

Empirical Evidence: Influential Networks

Centrality Measure

	visit go-come	borrow-lend kerorice	borrow-lend money	friendship	medical help	help decision	relatives	temple company
degree x100	0.25** (0.09)	0.32** (0.11)	0.20** (0.10)	0.22** (0.10)	0.32** (0.14)	0.13 (0.11)	0.35 (0.19)	0.61 (0.32)
closeness	32.91*** (9.56)	40.27*** (10.76)	29.80** (9.39)	31.09*** (9.04)	32.56** (11.14)	18.19 (9.82)	7.95 (16.56)	231.08 (134.31)
betweenness	1.36 (1.01)	0.18 (0.82)	0.29 (0.95)	1.67 (1.02)	1.17 (0.86)	-0.57 (0.83)	0.31 (0.77)	-0.21 (0.22)
eigenvector	3.62*** (0.89)	1.52** (0.63)	0.12 (0.82)	1.39 (0.83)	-0.73 (0.77)	-0.24 (0.76)	0.78 (0.57)	3.30 (3.56)

Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Village fixed effects are controlled

▶ Go back

Empirical Evidence: X_n

Descriptive Statistics

	mean	std.	min	max
# of households (HH)	226	54	139	329
# of villagers	1062	256	663	1612
# of Rooms	2.3	0.3	1.6	2.9
Average Age	32	2	26	37
Average HH size	4.7	0.4	4.1	5.6
HH having Electric	93%	4%	81%	98%
HH having Latrine	26%	11%	4%	47%
Participation Rate	16%	8%	5%	35%

▶ Go back

Empirical Evidence: Influential Networks

- Centrality measures how individuals are well-connected.
- High centrality \Rightarrow More likely to be influenced.
- How well does centrality predict participation?

$$d_{i,village} = C_i^j \beta^j + \gamma_{village}^j + \epsilon_i^j$$

Centrality Measure

	visit go-come	borrow-lend kerorice	borrow-lend money	friendship	medical help	help decision	relatives	temple company
Degree	0.25**	0.32**	0.20**	0.22**	0.32**	0.13	0.35	0.61
Centrality X100	(0.09)	(0.11)	(0.10)	(0.10)	(0.14)	(0.11)	(0.19)	(0.32)

Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Village fixed effects are controlled

- Consistent with LASSO detection.

Invertibility

- I use exogenous characteristics of influential individuals as instruments for their neighbors.
- Following Kelejian and Prucha (1998), I want to solve for D_n as a function of exogenous variables.
- By Assumption 2:

$$D_n = (M_n \circ D_n)\eta_0 + X_n\beta_0 + \epsilon_n$$
$$\Leftrightarrow D_n = (M_n \circ \eta_0)D_n + X_n\beta_0 + \epsilon_n$$

▶ Go back

Invertibility

- I use exogenous characteristics of influential individuals as instruments for their neighbors.
- Following Kelejian and Prucha (1998), I want to solve for D_n as a function of exogenous variables.
- By Assumption 2:

$$D_n = \left(M_n \circ D_n \right) \eta_0 + X_n \beta_0 + \epsilon_n$$
$$\Leftrightarrow D_n = \left(M_n \circ \eta_0 \right) D_n + X_n \beta_0 + \epsilon_n$$

▶ Go back

Invertibility

- I use exogenous characteristics of influential individuals as instruments for their neighbors.
- Following Kelejian and Prucha (1998), I want to solve for D_n as a function of exogenous variables.
- By Assumption 2:

$$D_n = \left(M_n \circ D_n \right) \eta_0 + X_n \beta_0 + \epsilon_n$$
$$\Leftrightarrow \left(I_n - \left(M_n \circ \eta_0 \right) \right) D_n = X_n \beta_0 + \epsilon_n$$

Invertibility

- I use exogenous characteristics of influential individuals as instruments for their neighbors.
- Following Kelejian and Prucha (1998), I want to solve for D_n as a function of exogenous variables.
- By Assumption 2:

$$D_n = (M_n \circ D_n)\eta_0 + X_n\beta_0 + \epsilon_n$$
$$\Leftrightarrow D_n = \left(I_n - (M_n \circ \eta_0) \right)^{-1} (X_n\beta_0 + \epsilon_n)$$

▶ Go back

Invertibility

- I use exogenous characteristics of influential individuals as instruments for their neighbors.
- Following Kelejian and Prucha (1998), I want to solve for D_n as a function of exogenous variables.
- By Assumption 2:

$$D_n = \left(M_n \circ D_n \right) \eta_0 + X_n \beta_0 + \epsilon_n$$
$$\Leftrightarrow D_n = \sum_{i=0}^{\infty} \left(M_n \circ \eta_0 \right)^i (X_n \beta_0 + \epsilon_n)$$

Empirical Evidence: Influential Networks

- Which networks are influential based on size of influence: \hat{e}_i^j .

Second Stage: estimated \hat{e} for each network

	visit go-come	borrow-lend keroric	borrow-lend money	friendship	medical help	help decision	relatives	temple company
absolute magnitudes	0.15	0.14	0.12	0.12	0.12	0.12	0.14	0.06
percentage of positive effect	77%	67%	69%	70%	68%	77%	67%	55%

▶ Go back

Empirical Evidence: Influential Networks

- Mean of $|\hat{e}_i^j|$ where i is an influential individual network j .

Second Stage: estimated \hat{e} for each network

	visit go-come	borrow-lend keroric	borrow-lend money	friendship	medical help	help decision	relatives	temple company
absolute magnitudes	0.15	0.14	0.12	0.12	0.12	0.12	0.14	0.06
percentage of positive effect	77%	67%	69%	70%	68%	77%	67%	55%

▶ Go back

Empirical Evidence: Influential Networks

- Percentage of $\hat{e}_i^j > 0$ among all leaders in network j .

Second Stage: estimated \hat{e} for each network

	visit go-come	borrow-lend keroric	borrow-lend money	friendship	medical help	help decision	relatives	temple company
absolute magnitudes	0.15	0.14	0.12	0.12	0.12	0.12	0.14	0.06
percentage of positive effect	77%	67%	69%	70%	68%	77%	67%	55%

▶ Go back