Identification and Estimation of a Regression Model using Network Data

Eric Auerbach

Department of Economics
Northwestern University

Getting at social influence using network data

- In many settings agent behavior is shaped by social influence.
- Often the nature of this influence is not observed by the researcher.
- Instead the researcher observes a network linking pairs of agents.
- Understanding how agents form links in the network reveals underlying information about the unobserved social influence.

Three examples from the literature

1. Bramoullé, Djebbari, and Fortin (2009) study classroom peer effects in which a student's activities depends on that of his or her peers.

2. Banerjee, Chandrasekhar, Duflo, and Jackson (2013) study program participation in which information about the program spreads by word-of-mouth.

3. Ductor, Fafchamps, Goyal, and van der Leij (2014) study research productivity in which research quality depends on a researcher's professional relationships.

In each example a sample of social connections between agents characterize the relevant social influence.

This paper

- Specifies a joint model of agent behavior (regression model) and network formation.
- Establishes sufficient conditions for the parameters of the regression model to be identified using network data.
- Proposes a new procedure to estimate the parameters of the regression model: codegree differencing.

The regression model

$$y_i = \beta x_i + \lambda(w_i) + \varepsilon_i \qquad E[\varepsilon_i | x_i, w_i] = 0$$

- y_i scalar outcome (college quality)
- x_i observed explanatory variable (enrolls in college prep course)
- w_i latent social factors (ability, ambition)
- $\lambda(w_i)$ social influence (expectations about college attendence)

Examples

- $\lambda(w_i) = \sum_{k=1}^{K} \alpha_k \mathbb{1}\{w_i = k\}$ "group fixed effects"
- $\lambda(w_i) = \gamma E[y_i | w_i] + \delta E[x_i | w_i]$ "linear-in-means peer effects" (Manski 1993)

The social factors are unobserved

- The researcher observes a random sample $\{y_i, x_i\}_{i=1}^n$, but not the corresponding social factors $\{w_i\}_{i=1}^n$.

- Instead, the researcher observes a collection of network links $D := \{D_{ij}\}_{1 \leq i \neq j \leq n}$ where

$$D_{ij} = \mathbb{1}\{\text{“agents } i \text{ and } j \text{ have a social connection”}\}$$

- Identification requires a stance as to how the network links D and the social factors $\{w_i\}_{i=1}^n$ are related.

The social factors drive observed linking activity

$$D_{ij} = \mathbb{1}\{\eta_{ij} \leq f(w_i, w_j)\} \times \mathbb{1}\{i \neq j\}$$

- $f(w_i, w_j)$ is the latent intensity of the relationship between i and j.
- η_{ij} is an idiosyncratic shock.
- D_{ij} is a noisy signal of the link intensity $f(w_i, w_j)$.

Examples

- $f(w_i, w_j) = 1 - (w_i - w_j)^2$ "homophily model"
- $f(w_i, w_j) = (w_i + w_j)/2$ "degree heterogeneity model"

Two interpretations of the network formation model

- As a random utility model of link formation: Hoff, Raftery, and Handcock (2002); Goldsmith-Pinkham and Imbens (2013); Jackson (2014); Graham (2015); Dzemski (2016); Candelaria (2017); Toth (2017)

- As a network density function:
  - In a network formation game with strategic interaction: Leung (2015); Sheng and Ridder (2016); Menzel (2016); Mele and Zhu (2017)
  - For link prediction: Bickel and Chen (2009); Bickel, Chen and Levina (2011); Bickel, Choi, Chang, Zhang (2013); Chatterjee (2015); Zhang, Levina, Zhu (2016)

A review of the model with additional details

$$y_i = \beta x_i + \lambda(w_i) + \varepsilon_i$$
$$D_{ij} = \mathbb{1}\{\eta_{ij} \leq f(w_i, w_j)\} \times \mathbb{1}\{i \neq j\}$$

- $\{x_i, w_i, \varepsilon_i\}_{i=1}^n$ iid with $E[\varepsilon_i | x_i, w_i] = 0$
- w_i and η_{ij} have $\mathcal{U}[0,1]$ marginals. $E[D_{ij} | w_i, w_j] = f(w_i, w_j)$
- $\{\eta_{ij}\}_{i,j=1}^n$ and $\{x_i, w_i, \varepsilon_i\}_{i=1}^n$ have mutually independent entries.
- Goldsmith-Pinkham and Imbens (2013); Chan (2014); Jackson (2014); Hsieh and Lee (2014, 2016); Arduini, Patacchini, and Rainone (2016); Johnsson and Moon (2016); c.f. Badev (2017), Griffith (2017)

A typical differencing argument when w_i is observed

- Focus first on β. Recall $y_i = \beta x_i + \lambda(w_i) + \varepsilon_i$.

- Suppose w_i is observed with finite support. Then

$$(y_i - y_j)\mathbb{1}_{w_i=w_j} = \beta(x_i - x_j)\mathbb{1}_{w_i=w_j} + (\varepsilon_i - \varepsilon_j)\mathbb{1}_{w_i=w_j}$$

- The parameter β is identified if $E\left[(x_i - x_j)^2\mathbb{1}_{w_i=w_j}\right] > 0$ with

$$\beta = E\left[(y_i - y_j)(x_i - x_j)\mathbb{1}_{w_i=w_j}\right] / E\left[(x_i - x_j)^2\mathbb{1}_{w_i=w_j}\right]$$

- When w_i has continuous support and λ is continuous can replace $\mathbb{1}_{w_i=w_j}$ with $\mathbb{1}_{w_i\approx w_j}$ where $w_i \approx w_j$ means $|w_i - w_j|$ is close to 0.

D cannot determine if $w_i \approx w_j$ when f is unrestricted

- The distribution of $D_{ij} = \mathbb{1}\{\eta_{ij} \leq f(w_i, w_j)\}$ does not generally contain information about whether $w_i \approx w_j$.

- The problem is that there always exists f' and (w_i', w_j') such that
  - $D_{ij} = \mathbb{1}\{\eta_{ij} \leq f'(w_i', w_j')\}$
  - $w_i' \not\approx w_j'$ and $w_i \approx w_j$

- The implication is that when f is unrestricted D cannot determine whether $|w_i - w_j|$ is close to 0.

When can D determine if $\lambda(w_i) \approx \lambda(w_j)$?

- The distribution of D cannot tell us anything about $|w_i - w_j|$ when f is unrestricted.

- Suppose $\lambda(w_i)$ has finite support. Then

$$(y_i - y_j)\mathbb{1}_{\lambda(w_i)=\lambda(w_j)} = \beta(x_i - x_j)\mathbb{1}_{\lambda(w_i)=\lambda(w_j)} + (\varepsilon_i - \varepsilon_j)\mathbb{1}_{\lambda(w_i)=\lambda(w_j)}$$

- Under what assumptions can D tell us something about $|\lambda(w_i) - \lambda(w_j)|$?

The main identification condition

- Agents with different social influences make different linking decisions.
- That is $\lambda(w_i) \neq \lambda(w_j)$ implies $f(w_i, w_t) \neq f(w_j, w_t)$ for some agent t.
- Equivalently $f(w_i, w) = f(w_j, w)$ for (almost) every w implies that $\lambda(w_i) = \lambda(w_j)$.

The main identification condition using network distance

$$d(w_i, w_j) = \left(\int (f(w_i, \tau) - f(w_j, \tau))^2 d\tau \right)^{1/2} = \|f(w_i, \cdot) - f(w_j, \cdot)\|_2$$

- $f(w_i, \cdot)$ is agent i's linking function.
- $d(w_i, w_j)$ compares the linking functions of agents i and j
- The main identification condition is then

$$d(w_i, w_j) = 0 \implies |\lambda(w_i) - \lambda(w_j)| = 0$$

How to interpret the main identification condition

- The social influence function λ is continuous with respect to the network distance d.

- This condition is automatically satisfied when

1. The social factors are identified by the agent's distribution of network links. That is $d(w_i, w_j) = 0 \implies w_i = w_j$.

2. Social influence only depends on the social factors through the agent's distribution of network links. That is $\lambda(w_i) = \phi(f(w_i, \cdot))$.

An additional condition gives the identification of β

- Suppose $d(w_i, w_j) = 0 \implies |\lambda(w_i) - \lambda(w_j)| = 0$

- Then

$$(y_i - y_j)\mathbb{1}_{d(w_i,w_j)=0} = \beta(x_i - x_j)\mathbb{1}_{d(w_i,w_j)=0} + (\varepsilon_i - \varepsilon_j)\mathbb{1}_{d(w_i,w_j)=0}$$

- The parameter β is identified if $E\left[(x_i - x_j)^2\mathbb{1}_{d(w_i,w_j)=0}\right] > 0$ with

$$\beta = E\left[(y_i - y_j)(x_i - x_j)\mathbb{1}_{d(w_i,w_j)=0}\right] / E\left[(x_i - x_j)^2\mathbb{1}_{d(w_i,w_j)=0}\right]$$

A review of the identification conditions for β

Assumptions (Identification)

1. *If* $d(w_i, w_j) = 0$ *then* $|\lambda(w_i) - \lambda(w_j)| = 0$
2. $E\left[(x_i - x_j)^2 \mathbb{1}_{d(w_i,w_j)=0}\right] > 0$

- The identification conditions imply

$$\beta = E\left[(y_i - y_j)(x_i - x_j)\mathbb{1}_{d(w_i,w_j)=0}\right] / E\left[(x_i - x_j)^2 \mathbb{1}_{d(w_i,w_j)=0}\right]$$

- Estimate β by regressing $(y_i - y_j)\mathbb{1}_{\widehat{d(w_i,w_j)}\approx 0}$ on $(x_i - x_j)\mathbb{1}_{\widehat{d(w_i,w_j)}\approx 0}$

Model

Identification

Estimation

Conclusion

Direct estimation of the network distance is difficult

- Identification conditions suggest estimating β by regressing $(y_i - y_j)\mathbb{1}_{\widehat{d(w_i,w_j)}\approx 0}$ on $(x_i - x_j)\mathbb{1}_{\widehat{d(w_i,w_j)}\approx 0}$.

- A complication is that $d(w_i, w_j)$ is difficult to estimate because it requires an approximation of the unknown function f.

- The usual trick is to estimate $f(w_i, w_j)$ by local averaging. That is, average D_{kl} for k, l such that $w_i \approx w_k$ and $w_j \approx w_l$.

- But we introduced $d(w_i, w_j)$ because $|w_i - w_j|$ was not identified...

I propose an alternative based on agent-pair codegrees

- Recall $f(w_i, w_t) = E[D_{it}|w_i, w_t]$ is the (conditional) probability that agents i and t are linked.

- $f(w_i, w_t)f(w_j, w_t) = E[D_{it}D_{jt}|w_i, w_j, w_t]$ is the probability that agents i and j are both linked to agent t.

- $p(w_i, w_j) = E[D_{is}D_{js}|w_i, w_j]$ is the probability that agents i and j are both linked to a randomly drawn agent.

- Equivalently, $p(w_i, w_j)$ is the inner product of agent i and j's linking functions: $\int f(w_i, \tau)f(w_j, \tau)d\tau$

Codegree distance as an alternative to network distance

$$\delta(w_i, w_j) = \left(\int (p(w_i, \tau) - p(w_j, \tau))^2 \, d\tau \right)^{1/2} = \|p(w_i, \cdot) - p(w_j, \cdot)\|_2$$

- $p(w_i, \cdot)$ is agent i's codegree function.
- $\delta(w_i, w_j)$ compares the codegree functions of agents i and j.

Two key results about codegree distance

I propose using codegree distance δ as an alternative to network distance d for two reasons:

- Result 1: $\delta(w_i, w_j) \approx 0$ (almost always) implies $d(w_i, w_j) \approx 0$.
- Result 2: $\delta(w_i, w_j)$ is straightforward to estimate using D.

.

Result 1: Codegree and network distances are related

Lemma (Codegree Identification)

If f is measurable then for any $\epsilon > 0$ there exists an $\epsilon' > 0$ such that

$$\delta(w_i, w_j) \leq \epsilon' \implies d(w_i, w_j) < \epsilon$$

with probability at least $1 - \epsilon^2/4$. If f is Lipschitz continuous then

$$d(w_i, w_j) \leq C \times \delta(w_i, w_j)^{1/3}$$

with probability one.

- $\beta = E\left[(y_i - y_j)(x_i - x_j)\mathbb{1}_{\delta(w_i,w_j)=0}\right] / E\left[(x_i - x_j)^2\mathbb{1}_{\delta(w_i,w_j)=0}\right]$

A sketch of the intuition behind Result 1

$$\delta(w_i, w_j)^2 = 0$$

$$\implies \int \left(p(w_i, \tau) - p(w_j, \tau)\right)^2 d\tau = 0$$

$$(*) \implies p(w_i, \tau) = p(w_j, \tau) \text{ and } p(w_i, s) = p(w_j, s) \text{ for any } (\tau, s)$$

$$\implies p(w_i, w_i) = p(w_i, w_j) = p(w_j, w_j)$$

$$\implies \int f(w_i, \tau)^2 d\tau = \int f(w_i, \tau) f(w_j, \tau) d\tau = \int f(w_j, \tau)^2 d\tau$$

$$\implies \int \left(f(w_i, \tau) - f(w_j, \tau)\right)^2 d\tau = d(w_i, w_j)^2 = 0$$

$*$assuming f (and p) is continuous

Result 2: Codegree distance straightforward to estimate using D

Lemma (Codegree Estimation)

$\max_{i \neq j} |\hat{\delta}_{ij} - \delta(w_i, w_j)| = o_p\left(\log(n)/\sqrt{n}\right)$

- $\hat{p}_{it} = \frac{1}{n} \sum_{s=1}^{n} D_{is} D_{ts}$
- $\hat{\delta}_{ij} = \left(\frac{1}{n} \sum_{t=1}^{n} \left(\hat{p}_{it} - \hat{p}_{jt}\right)^2\right)^{1/2}$
- $\hat{\delta}_{ij}$ is the root-average-squared difference in the ith and jth columns of the squared adjacency matrix ($D \times D$).

The proposed estimator for β based on codegree distance

- Results 1 and 2 suggest estimating β by regressing $(y_i - y_j)\mathbb{1}_{\hat{\delta}_{ij}\approx 0}$ on $(x_i - x_j)\mathbb{1}_{\hat{\delta}_{ij}\approx 0}$.

- This logic is formalized by the pairwise difference estimator

$$\hat{\beta} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(y_i - y_j)(x_i - x_j)K\left(\frac{\hat{\delta}_{ij}}{h_n}\right)}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i - x_j)^2 K\left(\frac{\hat{\delta}_{ij}}{h_n}\right)}$$

where h_n is a bandwidth sequence and K is a kernel density function.

The proposed estimator for $\lambda(w_i)$ based on codegree distance

- Recall $\lambda(w_i) = E[(y_i - \beta x_i) \,|w_i]$
- Estimate $\lambda(w_i)$ by averaging $\left(y_j - \hat{\beta} x_j\right)$ for agents such that $\hat{\delta}_{ij} \approx 0$.
- This logic is formalized by the nonparametric regression

$$\widehat{\lambda(w_i)} = \frac{\sum_{j=1}^{n} \left(y_j - \hat{\beta} x_j\right) K\left(\frac{\hat{\delta}_{ij}}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{\hat{\delta}_{ij}}{h_n}\right)}$$

where h_n is a bandwidth sequence and K is a kernel density function.

Model

Identification

Estimation

Conclusion

This paper

- Specifies a joint model of agent behavior and network formation where determinants of social influence also drive link activity.
- Provides sufficient conditions for the parameters of the model of agent behavior to be identified using network data.
- Proposes a new procedure for estimating the parameters of the model of agent behavior: codegree differencing.