Peer Effects and Social Networks in Education

Antoni Calvó-Armengol Eleonora Patacchini
Yves Zenou

February 6, 2009

1 Introduction

Behavior of individual agents is affected by that of their peers.

Education, crime, labor markets, fertility, participation to welfare programs, etc.

Detection and measure of peer effects: difficult exercise.

Peer effects: an average intra-group externality that affects identically all the members of a given group.

Group boundaries: arbitrary and at a quite aggregate level

Peer effects in crime: neighborhood level using local crime rates

Peer effects in school: classroom or school level using average school achievements

This paper: smallest unit of analysis for cross influences: the dyad (two-person group)

Collection of dyadic bilateral relationships: social network

Theory and Empirics

Theory: Explicit network analysis of peer effects

Each agent belongs to a network of peer influences, Ex ante heterogeneity

Payoffs are interdependent and agents choose their levels of activity simultaneously

Nash equilibrium of this peer effect game

Nash equilibrium proportional to Katz-Bonacich network centrality

Katz-Bonacich: counts, for each agent, the *total* number of direct and indirect paths of any length in the network

Paths are weighted by a geometrically decaying factor (with path length).

Katz-Bonacich centrality: not parameter free (depends on network topology and on decaying factor).

Empirics: Test predictions of our peer-effect model

Dataset of friendship networks in the United States from the National Longitudinal Survey of Adolescent Health (AddHealth).

Role of network location in education.

Empirical issues: endogenous network formation, unobserved individual, school and network heterogeneity.

Richness of the information provided by the AddHealth data

Use of both within and between network variations

11,491 pupils distributed over 181 networks.

Direct estimation of the model:

A one-standard deviation increase in the Katz-Bonacich index translates into roughly 7 percent of a standard deviation in education outcome.

Influence of peers on education outcomes

Standard approach: instrumental variables (e.g. Evans et al., 1992)

or a natural experiment (e.g. Angrist and Lavy, 1999; Sacerdote, 2001; Zimmerman, 2003)

Nearly no studies that have adopted a more structural approach to test a specific peer effect model in education

(Glaeser et al. (1996) peer effecs in criminal behavior).

Here:

Stress the role of the structure of social networks in explaining individual behavior.

Build a theoretical model of peer effects

Direct empirical test of our model on the network structure of peer effects

We characterize the exact conditions on the geometry of the peer network, so that the model is fully identified

2 A network model of peer effects

Population of agents $N = \{1, ..., n\}$.

The network Network g, $g_{ij} = 1$ if i and j are direct friends, and $g_{ij} = 0$, otherwise.

Individual i exerts a direct peer influence on j if and only if $g_{ij} = 1$.

$g_{ij} = g_{ji}$ and $g_{ii} = 0$.

y_i^0 effort of individual i absent of any peer influence.

z_i effort of individual i whose returns depend on others' peer efforts.

Each agent i selects both efforts $y_i^0 \geq 0$ and $z_i \geq 0$.

Utility function

$$u_i(\mathbf{y}^0, \mathbf{z}; \mathbf{g}) = \theta_i \, y_i^0 - \frac{1}{2}\left(y_i^0\right)^2 + \mu g_i z_i - \frac{1}{2} z_i^2 + \phi \sum_{j=1}^{n} g_{ij} z_i z_j .$$

$$\theta_i(\mathbf{x}) = \underbrace{\sum_{m=1}^{M} x_i^m}_{i\text{'s observable heterogeneity}} + \underbrace{\frac{1}{g_i} \sum_{m=1}^{M} \sum_{j=1}^{n} g_{ij} \, x_j^m}_{i\text{'s friends (average) observable heterogeneity}}$$

x_i^m set of M variables accounting for observable differences in individual, neighborhood and school characteristics of individual i.

Bilateral influences, for $i \neq j$:

$$\frac{\partial^2 u_i(\mathbf{y}^0, \mathbf{z}; \mathbf{g})}{\partial z_i \partial z_j} = \phi g_{ij} \geq 0. \tag{1}$$

When i and j are direct friends, the cross derivative is $\phi > 0$ and reflects a strategic complementarity in efforts.

Example: 3 agents

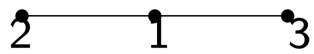


Figure 1. Three agents on a line.

Through the interaction with the central agent 1, peripheral agents end up reaping complementarities indirectly from each other.

Equilibrium decisions in each dyad cannot be analyzed independently of each other.

Each dyad exerts a strategic externality on the other one, and the equilibrium effort level of each agent reflects this externality, and the role each agent may play as a driver for the externality.

3 Analysis of the model

The Katz-Bonacich network centrality To each network $\mathbf{g}$, we associate its adjacency matrix $\mathbf{G} = [g_{ij}]$.

Symmetric zero-diagonal square matrix that keeps track of the direct connections in $\mathbf{g}$.

The kth power $\mathbf{G}^k = \mathbf{G}^{(k \overset{times}{\dots})}\mathbf{G}$ of the adjacency matrix $\mathbf{G}$ keeps track of indirect connections in $\mathbf{g}$.

The coefficient $g_{ij}^{[k]}$ in the (i, j) cell of $\mathbf{G}^k$ gives the number of paths of length k in $\mathbf{g}$ between i and j.

Not the shortest possible route between two agents

Example Network g with three individuals (star)

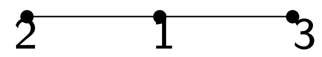


Figure 1

Adjacency matrix :

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$k \geq 1$

$$\mathbf{G}^{2k} = \begin{bmatrix} 2^k & 0 & 0 \\ 0 & 2^{k-1} & 2^{k-1} \\ 0 & 2^{k-1} & 2^{k-1} \end{bmatrix} \qquad \text{and} \qquad \mathbf{G}^{2k+1} = \begin{bmatrix} 0 & 2^k & 2^k \\ 2^k & 0 & 0 \\ 2^k & 0 & 0 \end{bmatrix}$$

$$\mathbf{G}^3 = \begin{bmatrix} 0 & 2 & 2 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \end{bmatrix}$$

$\mathbf{G}^3$: two paths of length three between 1 and 2: $12 \longrightarrow 21 \longrightarrow 12$ and $12 \longrightarrow 23 \longrightarrow 32$.

no path of length three from i to i

The vector of Katz-Bonacich centralities:

$$\mathbf{b}(\mathbf{g},\phi) = \phi\mathbf{G1} + \phi^2\mathbf{G}^2\mathbf{1} + \phi^3\mathbf{G}^3\mathbf{1} + \cdots = \sum_{k=0}^{+\infty} \phi^k \mathbf{G}^k \cdot (\phi\mathbf{G1}).$$

$\mathbf{1}$ vector of ones.

$\mathbf{G1}$ vector of node connectivities

$\mathbf{G}^k\mathbf{1}$ give the total number of paths of length k that emanate from the corresponding network node.

ϕ small enough so that this infinite sum is well-defined.

Example

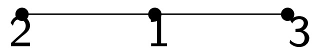


Figure 1

Adjacency matrix :

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$$\mathbf{G1} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \qquad \mathbf{G}^3\mathbf{1} = \begin{bmatrix} 4 \\ 2 \\ 2 \end{bmatrix}$$

Observe

$$\sum_{k=0}^{+\infty} \phi^k \mathbf{G}^k = (\mathbf{I} - \phi \mathbf{G})^{-1}$$

$\mathbf{I}$ identity matrix.

Vector Katz-Bonacich centralities:

$$\mathbf{b}(\mathbf{g}, \phi) = (\mathbf{I} - \phi \mathbf{G})^{-1} \cdot (\phi \mathbf{G} \mathbf{1}) \,. \qquad (2)$$

Katz-Bonacich centrality of a given node is zero when the network is empty.

Katz-Bonacich centrality is null when $\phi = 0$

Katz-Bonacich centrality increasing and convex with ϕ.

Katz-Bonacich centrality bounded from below by ϕ times the node connectivity, that is, $b_i(\mathbf{g}, \phi) \geq \phi g_i$.

Katz-Bonacich centrality well-defined for low enough values of ϕ, so that the infinite sum $\mathbf{1} + \phi\mathbf{G}\mathbf{1} + \phi^2\mathbf{G}^2\mathbf{1} + \cdots$ converges to a finite value.

The exact strict upper bound for the scalar ϕ is given by the inverse of the largest eigenvalue of $\mathbf{G}$ (Debreu and Herstein, 1953).

Example.

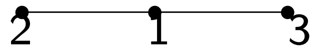


Figure 1

Largest eigenvalue of $\mathbf{G}$ is 2, Exact strict upper bound for ϕ is $1/2$.

The vector of Bonacich network centralities is:

$$\mathbf{b}\left(\mathbf{g},\phi\right)=\begin{bmatrix} b_1\left(\mathbf{g},\phi\right) \\ b_2\left(\mathbf{g},\phi\right) \\ b_3\left(g,\phi\right) \end{bmatrix}=\frac{\phi}{1-2\phi^2}\begin{bmatrix} 2+2\phi \\ 1+2\phi \\ 1+2\phi \end{bmatrix}.$$

Equilibrium behavior Nash equilibrium

$$\max_{y_i^0, z_i} u_i(\mathbf{y}^0, \mathbf{z}; \mathbf{g}) = \theta_i\, y_i^0 - \frac{1}{2}\left(y_i^0\right)^2 + \mu g_i z_i - \frac{1}{2} z_i^2 + \phi \sum_{j=1}^{n} g_{ij} z_i z_j.$$

Best reply function for each $i = 1, ..., n$:

$$y_i^{0*}(\mathbf{x}) = \theta_i(\mathbf{x}) = \sum_{m=1}^{M} x_i^m + \frac{1}{g_i} \sum_{m=1}^{M} \sum_{j=1}^{n} g_{ij}\, x_j^m$$

$$z_i^*(\mathbf{g}) = \mu g_i + \phi \sum_{j=1}^{n} g_{ij} z_j$$

Individual outcome (educational achievement) is the sum of these two different efforts:

$$y_i^* (\mathbf{x}, \mathbf{g}) = \underbrace{y_i^{0*} (\mathbf{x})}_{\text{idiosyncratic}} + \underbrace{z_i^* (\mathbf{g})}_{\text{peer effect}} .$$

We can decompose additively individual behavior into an exogenous part and an endogenous peer effect component that depends on the individual under consideration:

Denote by $\omega(\mathbf{g})$ the largest eigenvalue of the adjacency matrix $\mathbf{G} = \left[g_{ij}\right]$ of the network.

Proposition 3.1 *Suppose that $\phi\omega(\mathbf{g}) < 1$. Then, the individual equilibrium outcome is uniquely defined and given by:*

$$y_i^* (\mathbf{x}, \mathbf{g}) = \theta_i (\mathbf{x}) + \frac{\mu}{\phi} b_i (\mathbf{g}, \phi) . \tag{3}$$

Rewriting (3) as

$$y_i^* (\mathbf{x}, \mathbf{g}) = \left(\mathbf{1} + \frac{\mu}{\phi} \frac{b_i (\mathbf{g}, \phi)}{\theta_i^0 (\mathbf{x})} \right) y_i^{0*} (\mathbf{x}) ,$$

Peer influence acts as a multiplier on the behavior of the isolated individual.

Condition $\phi\omega(\mathbf{g}) < 1$: Network complementarities must be small enough than own concavity.

Network complementarities: measured by the compound index $\phi\omega(\mathbf{g})$

ϕ intensity of each non-zero cross effect

$\omega(\mathbf{g})$ pattern of such cross effects.

This prevents multiple equilibria to emerge and, in the same time, rules out corner solutions.

Bonacich centrality: right network index to account for equilibrium behavior.

Example. Consider again the network g in Figure 1.

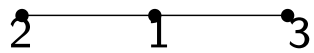


Figure 1

When $\phi < 1/2$, the unique Nash equilibrium is:

$$y_1^* = y_1^0 + \mu \left(\frac{2 + 2\phi}{1 - 2\phi^2} \right)$$

$$y_2^* = y_2^0 + \mu \left(\frac{1 + 2\phi}{1 - 2\phi^2} \right)$$

$$y_2^* = y_3^0 + \mu \left(\frac{1 + 2\phi}{1 - 2\phi^2} \right)$$

Alternative formulation of the model Our utility function: each individual i chooses two different effort levels, y_i^0 and z_i.

Assume utility function with only one type of effort z_i:

$$u_i(\mathbf{z};\mathbf{g}) = \mu g_i z_i - \frac{1}{2} z_i^2 + \phi \sum_{j=1}^{n} g_{ij} z_i z_j. \tag{4}$$

Best reply function for each individual i

$$z_i^* (\mathbf{g}) = \mu g_i + \phi \sum_{j=1}^{n} g_{ij} z_j$$

When $\phi\omega(\mathbf{g}) < 1$

$$z_i^* (\mathbf{g}) = \frac{\mu}{\phi} b_i (\mathbf{g}, \phi)$$

Educational achievement of each individual i:

$$y_i^* (\mathbf{x}, \mathbf{g}) = \theta_i(\mathbf{x}) + z_i^* (\mathbf{g})$$

$\theta_i(x)$ contextual effects

3.1 Discussion

What happens when the condition $\phi\omega(g) < 1$ does not hold?

Theory

(i) We cannot characterize the Nash equilibrium since the Katz-Bonacich centrality measure is not anymore defined;

(ii) The existence of equilibrium becomes an issue since the strategy space is unbounded. To obtain existence, we need to bound the strategy space in some arbitrary way.

(iii) Even if existence is guaranteed, uniqueness does not always follows. In fact, as it is well known in the literature on supermodular games, multiple equilibria are rather the rule than the exception.

Empirical analysis

Difficult to interpret the results.

Multiple equilibria

No centrality measure and thus unable to distinguish between the effects of individual's network location and individual's idiosyncratic characteristics on educational achievements.

Only 9 percent of the networks in our database violate this condition

473 discarded people (descriptive statistics on these discarded people do not differ significantly from those on the whole sample).

4 Data

Unique database on friendship networks from the National Longitudinal Survey of Adolescent Health (AddHealth).

Adolescents' behavior in the United States: Data on students in grades 7-12 from a nationally representative sample of roughly 130 private and public schools in years 1994-95.

Respondents' demographic and behavioral characteristics, education, family background and friendship.

Friendship networks

Friendship information is based upon actual friends nominations.

Pupils were asked to identify their best friends from a school roster (up to five males and five females).

The limit in the number of nominations is not binding.

Less than 1% of the students in our sample show a list of ten best friends.

Less than 3% a list of five males and roughly 4% name five females.

On average, they declare to have 5.48 friends with a small dispersion around this mean value (the standard deviation is equal to 1.29).

The corresponding figures for male- and female-friends are 2.78 (with standard deviation equal to 1.85) and 3.76 (with standard deviation equal to 1.04).

A link exists between two friends if at least one of the two individuals has identified the other as his/her best friend (undirected networks)

We also consider directed networks: 14% of relationships are not reciprocal.

For each school, we obtain all the networks of (best) friends.

Education achievements

Grade achieved by each student in mathematics, history and social studies and science, ranging from D or lower to A, the highest grade (recoded 1 to 4).

School performance index

Final sample: 11,964 pupils distributed over 199 networks.

Descriptive statistics

Table 1: Descriptive statistics

(11,964 pupils; 199 networks)

	Mean	St. Dev.	Min	Max
Female	0.41	0.35	0	1
Black or African American	0.17	0.31	0	1
Other races	0.12	0.15	0	1
Age	15.29	1.85	10	19
Religion practice	3.11	1.01	1	4
Health status	3.01	1.77	0	4
School attendance	3.28	1.86	1	6
Student grade	9.27	3.11	7	12
Organized social participation	0.62	0.22	0	1
Motivation in education	2.23	0.88	1	4
Relationship with teachers	0.12	0.34	0	1
Social exclusion	2.26	1.81	1	5
School attachment	2.59	1.76	1	5
Parental care	0.69	0.34	0	1
Household size	3.52	1.71	1	6
Two married parent family	0.41	0.57	0	1
Single parent family	0.23	0.44	0	1
Public assistance	0.12	0.16	0	1
Mother working	0.65	0.47	0	1

	Mean	St. Dev.	Min	Max
Parental education	3.69	2.06	1	5
Parent age	40.12	13.88	33	75
Parent occupation manager	0.11	0.13	0	1
Parent occupation professional or technical	0.09	0.21	0	1
Parent occupation office or sales worker	0.26	0.29	0	1
Parent occupation manual	0.21	0.32	0	1
Parent occupation military or security	0.09	0.12	0	1
Parent occupation farm or fishery	0.04	0.09	0	1
Parent occupation retired	0.06	0.09	0	1
Parent occupation other	0.11	0.16	0	1

Table 1: Descriptive statistics (continued)

	Mean	St. Dev.	Min	Max
Neighborhood quality	2.99	2.02	1	4
Residential building quality	2.95	1.85	1	4
Neighborhood safety	0.51	0.57	0	1
Residential area suburban	0.32	0.38	0	1
Residential area urban - residential only	0.18	0.21	0	1
Residential area commercial properties - retail	0.12	0.15	0	1
Residential area commercial properties - industrial	0.13	0.18	0	1
Residential area type other	0.19	0.25	0	1
Friend attachment	0.49	0.54	0	1
Friend involvement	1.88	1.56	0	3
Friend contacts	0.89	0.12	0	1
Physical development	3.14	2.55	1	5
Self esteem	3.93	1.33	1	6

4.1 Descriptive evidence

Mean and the standard deviation of network size are 60.42 and 24.48, respectively.

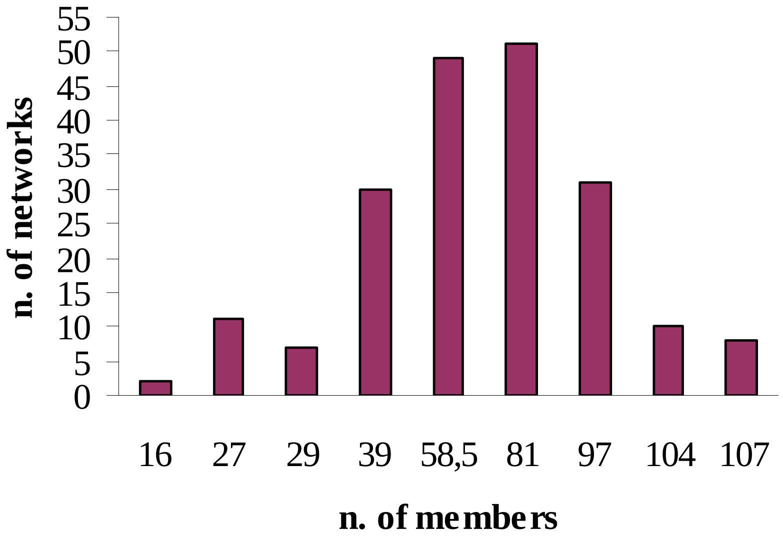


Figure 2. The empirical distribution of adolescent networks

Smallest network in our sample.

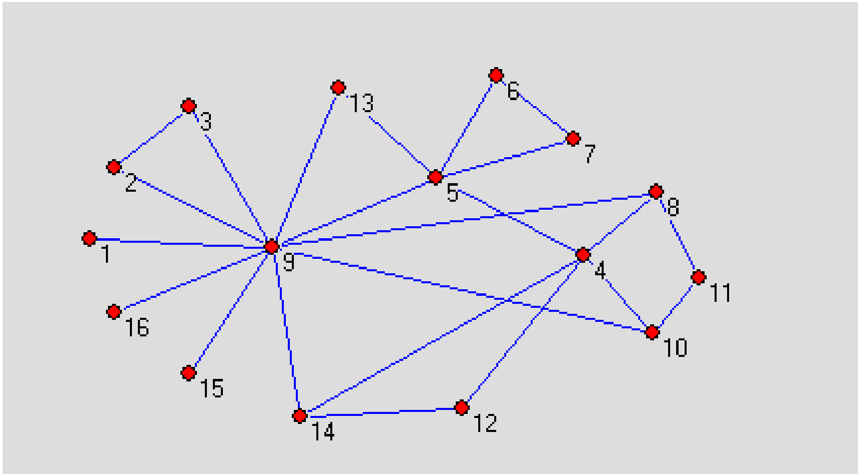


Figure 3. Smallest network of adolescents ($n = 16$)

The largest network in our sample is almost seven times bigger and has 107 members.

5 Empirical strategy

Proposition 3.1: Actual empirical relationship between $b_i(\mathbf{g}, \phi)$ and the observed effort level y_i^*.

Maximally connected components:

Two agents in a network are either directly linked or indirectly linked through a sequence of intermediate agents (connectedness).

Two agents in different networks cannot be connected through any such sequence (maximality).

$\mathbf{g}_\kappa$ network encompassing n_κ different individuals.

K such networks in the economy. $\sum_{\kappa=1}^{K} n_\kappa = n$.

Theory:

$$y_i(\mathbf{x},\mathbf{g}) = \underbrace{\sum_{m=1}^{M} x_i^m + \frac{1}{g_i}\sum_{m=1}^{M}\sum_{j=1}^{n} g_{ij}\, x_j^m}_{\text{idiosyncratic}} + \underbrace{z_i(\mathbf{x},\mathbf{g})}_{\text{peer effect}}.$$

$$z_i = \mu g_i + \phi \sum_{j=1}^{n} g_{ij} z_j,\ i = 1, ..., n$$

Empirical counterpart

$$\begin{aligned} y_{i,\kappa} &= \sum_{m=1}^{M} \beta_m x_{i,\kappa}^m + \frac{1}{g_{i,\kappa}} \sum_{m=1}^{M}\sum_{j=1}^{n_\kappa} \gamma_m g_{ij,\kappa} x_{j,\kappa}^m + \eta_\kappa + \varepsilon_{i,\kappa}, \qquad (5) \\ \varepsilon_{i,\kappa} &= \mu g_{i,\kappa} + \phi \sum_{j=1}^{n_\kappa} g_{ij,\kappa}\varepsilon_{j,\kappa} + \upsilon_{i,\kappa},\ \ i = 1, ..., n;\ \ \kappa = 1, ..., K, \end{aligned}$$

η_κ is an (unobserved) network-specific component (constant over individuals in the same network)

$\varepsilon_{i,\kappa}$ residual of individual i's level of activity in the network g that is *not* accounted for neither by individual heterogeneity and contextual effects nor by (unobserved) network-specific components.

$\sum_{j=1}^{n_\kappa} g_{ij,\kappa}\varepsilon_{j,\kappa}$ is the spatial lag term and ϕ is the spatial autoregressive parameter.

Matrix notation:

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{DGX}\boldsymbol{\gamma} + \boldsymbol{\eta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &= \mu\mathbf{G1} + \phi\mathbf{G}\boldsymbol{\varepsilon} + \boldsymbol{\nu},\end{aligned}$$

Our model is a variation of the Anselin (1988) spatial error model.

Using the Maximum Likelihood approach (see, e.g. Anselin, 1988), we estimate jointly $\widehat{\boldsymbol{\beta}}$, $\widehat{\gamma}$, $\widehat{\phi}$, $\widehat{\mu}$.

These values: measure the relative importance of

individual characteristics, $\widehat{\beta}_1, ..., \widehat{\beta}_m$ (e.g. parental education, school and neighborhood quality),

contextual effects, $\widehat{\gamma}_1, ..., \widehat{\gamma}_m$ (e.g. average parental education of each individual's best friends, etc.),

the individual Katz-Bonacich centrality index, $\widehat{\phi}$ and $\widehat{\mu}$,

in shaping individuals' behavior (equation (3) in Proposition 3.1).

5.1 Identification of peer effects

Assessment of the effects of peer pressure on individual behavior, i.e. the identification of endogenous social effects

Typically characterized by econometric issues

First problem: The endogenous sorting of individuals into groups

Second problem: The reflection problem (Manski, 1993).

First problem: The endogenous sorting of individuals into groups

The role of network fixed effects

Individuals sort into groups non-randomly.

If the variables that drive this process of selection are not fully observable, potential correlations between (unobserved) group-specific factors and the target regressors are major sources of bias.

Use of network fixed effects (also referred to as correlated effects or network unobserved heterogeneity)

Assume that agents self-select into different groups in a first step, and that link formation takes place within groups in a second step.

If link formation is uncorrelated with the observable variables, this two-step model of link formation generates network fixed effects.

Assuming additively separable group heterogeneity, a within group specification is able to control for these correlated effects.

We use the model specification (5), which has a network-specific component η_κ of the error term, and adopt a traditional (pseudo) panel data fixed effects estimator, namely, we subtract from the individual-level variables the network average.

Network fixed effects estimation allows us to distinguish endogenous effects from correlated effects.

Table 2 reports the estimated correlations between individual and network averages of variables that are commonly believed to induce self-selection into teenagers' friendship group, once the influence of a variety of other factors and network-fixed effects are washed out.

The estimated correlation coefficients reported in Table 2 are not statistically significant for all variables. This indicates that teenagers are not clustered by any of the attributes considered.

Table 2. Correlation between individual and network-level characteristics

Variable	
Parental education	-0.0719
	(0.0649)
Parental care	-0.0506
	(0.0602)
Mathematics score	0.1481
	(0.1901)
Motivation in education	0.1329
	(0.1505)
School attachment	-0.0507
	(0.0499)
Social exclusion	-0.1032
	(0.1344)
Individual socio-demographic variables	yes
Family background variables	yes
Protective factors	yes
Residential neighborhood variables	yes
Contextual effects	yes
School fixed effects	yes

Second problem: The reflection problem (Manski, 1993).

The role of peer groups with individual level variation

Network fixed effects: does not necessary estimate the causal effect of peers' influence on individual behavior.

In a peer group everyone's behavior affects the others, so that we cannot distinguish if a group member's action is the cause or the effect of peers' influence: reflection problem (Manski, 1993).

Here: the reference group is the number of friends each individual has and groups do overlap.

Because peer groups are individual specific, this issue is eluded.

The reduced-form equation corresponding to the spatial error term in (5) is, in matrix notation:

$$\boldsymbol{\varepsilon} = \mu \left[\mathbf{I} - \phi\mathbf{G}\right]^{-1} \mathbf{G}\mathbf{1} + \left[\mathbf{I} - \phi\mathbf{G}\right]^{-1} \boldsymbol{\nu}. \tag{6}$$

Peer effects are identified if the structural parameters (μ, ϕ) uniquely determine the reduced-form coefficients in (6).

Bramoullé *et al.* (2006) provide general results on the identification of peer effects through social networks via variations of the linear-in-means model

Here we use a similar approach.

Proposition 5.1 *Suppose that* $\phi\omega(\mathrm{g}) < 1$ *and* $\mu \neq 0$. *Peer effects are identified if and only if* $g_i^{[2]}/g_i \neq g_j^{[2]}/g_j$ *for at least two agents* i *and* j.

Peer effects are identified if we can find two agents in the economy that differ in the average connectivity of their direct friends.

This a simple property of the network, that amounts to checking that the $2 \times n$ matrix with column vectors $\mathbf{G1}$ and $\mathbf{G}^2\mathbf{1}$ is of rank two.

$\mathbf{G1} = [g_i]$ is the vector of node connectivities

$\mathbf{G}^2\mathbf{1} = \left[g_i^{[2]}\right]$ gives the total number of two-link away contacts in the network.

$g_i^{[2]}/g_i$ is the average connectivity of agent i's direct contacts.

Regular networks: Identification fails.

In our data, no network is regular and the identification requirement is always satisfied.

Peer-groups are individual specific and individuals belong to more than one group.

The role of specific controls

Proxies for typically unobserved individual characteristics that may be correlated with our variable of interest.

More self-confident and (very likely) more successful students at school are contacted by a larger number of friends, thus showing a higher value of the Katz-Bonacich measure.

Controls for differences in leadership propensity across adolescents: Indicator of self-esteem, indicator of the level of physical development compared to the peers.

Capture differences in attitude towards education and parenting: Indicators of the student's motivation in education and parental care.

Existence of possible correlations between our centrality measure and unobservable school characteristics

School fixed effects (i.e. *school dummies*).

6 Empirical results

6.1 The Katz-Bonacich network centrality index

Model (5) is estimated using the Maximum Likelihood approach.

Table 3a: Model (5) Maximum Likelihood estimation results on key variables

Dependent variable: school performance index

	ML (with network fixed effects)
Number of best friends (μ)	0.0314**
	(0.0149)
Peer effects (ϕ)	0.5667***
	(0.1433)
Individual socio-demographic variables	yes
Family background variables	yes
Protective factors	yes
Residential neighborhood variables	yes
Contextual effects	yes
School fixed effects	yes

$R^2 = 0.8987$

Notes:

- Number of observations: 2,079,871 (11,491pupils, 181networks)
- Regressions are weighted to population proportions
- Standard errors in parentheses.

Coefficients marked with one (two) [three] asterisks are significant at 10 (5) [1] percent level

Estimated μ and ϕ are both positive and highly statistically significant.

Calculate the Katz-Bonacich measure by fixing the value of ϕ at the point estimate $\widehat{\phi}$.

Derived Katz-Bonacich measures range from 0.32 to 3.48, with an average of 1.65 and a standard deviation of 2.79.

The estimated impact of this variable on education outcomes that is predicted by the theory, i.e. $\widehat{\mu}/\widehat{\phi}$ is statistically significant.

A one-standard deviation increase in the Katz-Bonacich index translates into roughly 7 percent of a standard deviation in education outcome

This effect is about 17 percent for parental education.

7 Directed networks

So far *undirected* networks, i.e. we have assumed that friendship relationships are reciprocal, $g_{ij,\kappa} = g_{ji,\kappa}$.

14 percent of relationships in our dataset are not reciprocal.

Wasserman and Faust (1994, pages 205-210) define the Katz-Bonacich centrality measure for directed networks.

In a directed graph, a node (here an individual) has two distinct ends: a head (the end with an arrow) and a tail. Each end is counted separately.

The sum of head endpoints count toward the *indegree* and the sum of tail endpoints count toward the *outdegree*.

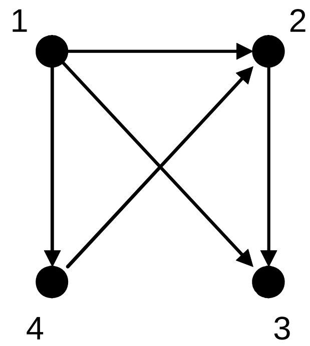
Formally, we denote a link from i to j as $g_{ij} = 1$ if j has nominated i as his/her friend, and $g_{ij} = 0$, otherwise.

The **indegree** of student i, denoted by g_i^+, is the number of nominations student i receives from other students, that is $g_i^+ = \sum_j g_{ij}$.

The **outdegree** of student i, denoted by g_i^-, is the number of friends student i nominates, that is $g_i^- = \sum_j g_{ji}$.

We consider only the indegree to define the Katz-Bonacich centrality measure.

Consider the following directed network:



Adjacency matrix, which takes into indegrees only, is equal to:

$$\mathbf{G} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Adjacency matrix $\mathbf{G} = [g_{ij}]$ is now *asymmetric*.

We can now define the Katz-Bonacich centrality measure $b(g, \phi)$ exactly as before.

Theoretical analysis unchanged since, in the proof of Theorem **??** (and thus of Proposition 3.1), the symmetry of $\mathbf{G}$ does not play any explicit role. In fact, the Bonacich-Nash linkage holds for any asymmetric matrix $\mathbf{G}$, under the condition $\phi\omega(\mathbf{g}) < 1$.

Table 3b: Model (5) Maximum Likelihood estimation results on key variables

Dependent variable: school performance index

	Undirected networks	Directed networks
Number of best friends (μ)	0.0314**	0.0323**
	(0.0149)	(0.0152)
Peer effects (ϕ)	0.5667***	0.5505***
	(0.1433)	(0.1247)
Individual socio-demographic variables	yes	yes
Family background variables	yes	yes
Protective factors	yes	yes
Residential neighborhood variables	yes	yes
Contextual effects	yes	yes
School fixed effects	yes	yes
R^2	0.8987	0.8905

Notes:
- Number of observations: 2,079,871 (11,491 pupils, 181 networks)
- Control variables are those listed in Appendix 3.
- Regressions are weighted to population proportions
- Standard errors in parentheses.
Coefficients marked with one (two) [three] asterisks
are significant at 10 (5) [1] percent level

Katz-Bonacich measure still statistically significant and only sligthly lower in magnitude (5.6 percent versus 7 percent).

7.1 An alternative measure of network unit centralities

Two dimensions of centrality, connectivity and betweenness.

Bonacich centrality is an index of connectivity since it counts the number of *any* path stemming from a given node, not just optimal paths.

Degree centrality The individual-level *degree centrality* is simply each individual's number of direct friends:

$$\delta_i(\mathbf{g}_K) = g_i = \sum_{j=1}^{n} g_{ij}$$

To compare networks of different sizes, this measure is normalized to be in an interval from 0 to 1, where values 0 and 1 indicate the smallest and the highest possible centrality.

$$\delta_i^*(\mathbf{g}_K) = \frac{g_i}{n_K - 1} = \frac{\sum_{j=1}^{n} g_{ij}}{n_K - 1}$$

In our data, the normalized degree centrality index $\delta_i^*(\mathbf{g}_K)$ has a mean equal to 0,35 and a standard deviation equal to 0,18.

Closeness centrality The standard measure of centrality according to *closeness* of individual i is given by:

$$c_i(\mathbf{g}_\kappa) = \frac{1}{\sum_j d_{ij,\kappa}}$$

where d_{ij} is the geodesic distance (length of the shortest path) between individuals i and j.

As a result, the closeness centrality of individual i is the inverse of the sum of geodesic distances from i to the $n-1$ other individuals (i.e. the reciprocal of its "farness").

Compared to degree centrality, the closeness measure takes into account not only direct connections among individuals but also indirect connections.

Compared to the Katz-Bonacich centrality, the closeness measure assumes a weight of one to each indirect connection.

Relative closeness centrality measure as:

$$c_i^*(\mathbf{g}_\kappa) = \frac{n_\kappa - 1}{\sum_j d_{ij,\kappa}}$$

where $n_\kappa - 1$ is the maximum possible distance between two individuals in network κ. This measure takes value between 0 and 1.

The mean and the standard deviation in our data of this normalized index are 0.49 and 0.27.

Betweenness centrality measure of agent i in a network component $\mathbf{g}_\kappa$

$$f_i(\mathbf{g}_\kappa) = \sum_{j<l} \frac{\#\text{ of shortest paths between } j \text{ and } l \text{ through } i \text{ in } \mathbf{g}_\kappa}{\#\text{ of shortest paths between } j \text{ and } l \text{ in } \mathbf{g}_\kappa}$$

where j and l denote two given agents in $\mathbf{g}_\kappa$.

For undirected networks, a normalized version of this measure is:

$$f_i^*(\mathbf{g}_\kappa) = \frac{f_i(\mathbf{g}_\kappa)}{(n_\kappa - 1)(n_\kappa - 2)/2},$$

where n_κ is the size of the network $\mathbf{g}_\kappa$.

Betweenness is a parameter-free network measure.

In our data, the normalized betweenness measure f_i^* has a mean equal to 0.45 and a standard deviation equal to 0.51.

Table 4. Explanatory power of unit centrality measures

Dependent variable: school performance index

	OLS	OLS	OLS
Degree centrality	0.2508*	-	-
	(0.1475)		
Closeness centrality	-	0.2892	-
		(0.2599)	
Beetweenness centrality	-	-	0.0621
			(0.0698)
Individual socio-demographic variables	yes	yes	yes
Family background variables	yes	yes	yes
Protective factors	yes	yes	yes
Residential neighborhood variables	yes	yes	yes
Contextual effects	yes	yes	yes
School fixed effects	yes	yes	yes
R^2	0.7958	0.8202	0.8001

Notes:

- Number of observations: 2,079,871 (11,491 pupils, 181 networks)
- Control variables are those listed in Appendix 3.
- Network fixed-effects OLS estimators are reported. They are within-group estimates where individuals are grouped by networks
- Regressions are weighted to population proportions

Out of the three measures, only degree centrality shows a statistically slighly significant impact (i.e., at the 10% significance level).

When such an effect is translated in terms of standard deviations, its impact on educational outcomes is not even one third of the effect exerted by the Katz-Bonacich centrality index (roughly 2.1 percent versus 7 percent).

Two main explanations

1) Bonacich centrality is not an arbitrary network measure but results from a positive analysis that maps network topology to equilibrium behavior.

2) Betweenness centrality is a parameter-free network index. It only depends on the network geometry.

8 Peer effects and network structure

Relationship between peer effects and the network topology.

Estimate model (5) for each network g separately, thus using within network variation only.

Obtain K different estimates of ϕ: $\widehat{\phi}_1, ..., \widehat{\phi}_K$.

The estimated value $\widehat{\phi}_\kappa$ measures the strength of each existing bilateral influence in the network g.

The estimated values, $\widehat{\phi}_1, ..., \widehat{\phi}_K$, vary widely across the $K = 181$ school-peer networks.

Network density is the fraction of ties present in a network over all possible ones. It ranges from 0 to 1 as networks get denser.

Network asymmetry is measured using the variance of connectivities. We normalize it, so that it reaches 1 for the most asymmetric network in the sample.

Network redundancy is the fraction of all transitive triads over the total number of triads. It measures the probability with which two of i's friends know each other.

Redundancy, or clustering, is much higher in social networks than in randomly generated graphs. Again, we normalize it.

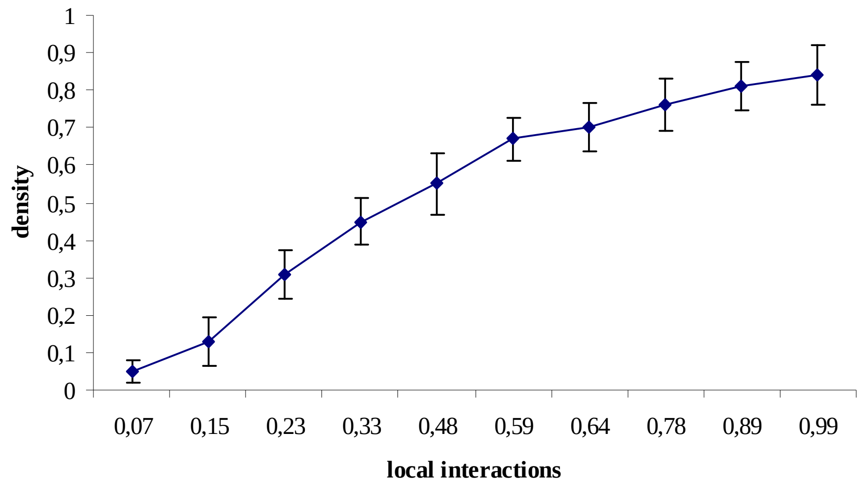


Figure 4a: Density in education networks

The strength of bilateral influences increases steadily with network density for low values, and remains roughly unchanged for higher values.

Therefore, richer networks are a sign of stronger dyadic cross effects, at least until roughly 60% of all possible networks links are created.

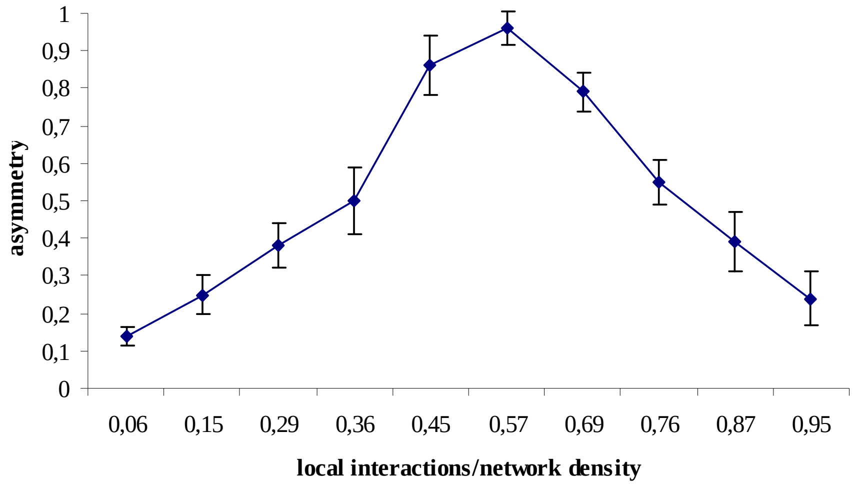


Figure 4b: Asymmetry in education networks

Network asymmetry has a non-trivial impact on the intensity of peer effects. Highly distributed and symmetric networks are compatible with both very low and very high values of the peer-to-density ratio, while highly centralized and asymmetric networks are always synonymous of an average value of peer effects.

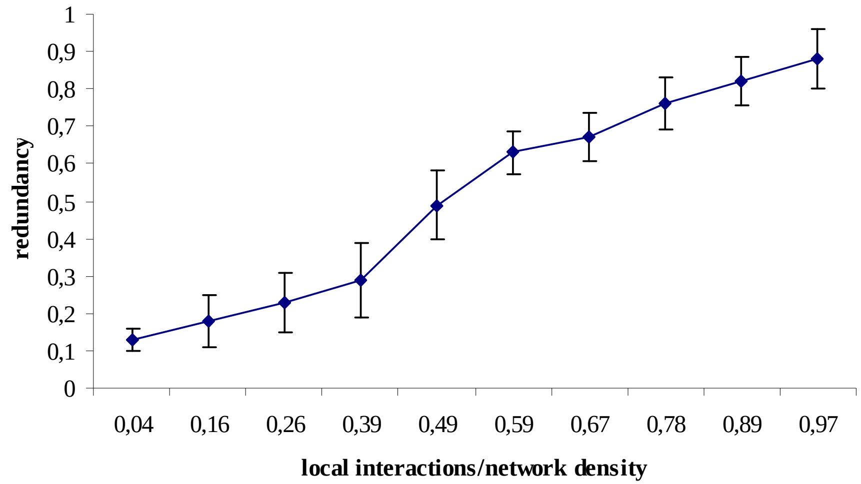


Figure 4c: Redundancy in education networks

Figure 4c: link redundancy, or clustering, has a strong positive impact on the strength of bilateral influences above a minimum threshold value.