Peer Effects in Education, Sport, and Screen Activities: Local Aggregate or Local Average?

Eleonora Patacchini

La Sapienza University of Rome, CEPR, EIEF and IZA

Xiaodong Liu

University of Colorado at Boulder

Yves Zenou

University of Stockholm, CEPR and IZA

Purpose

Theoretical and empirical investigation of the role of peers for understanding behavior using a network perspective

We propose alternative behavioural foundations and test their empirical salience in different contexts

A test for model selection is implemented, with appropriate estimators

Outline of the presentation

- Motivation
- Related literature
- Theoretical set-up
- Empirical model and estimation strategy
- Data and estimation results
- Robustness check: a simulation experiment

Motivation

The integration of models of social interactions within economic theory is an active and interesting area of research (Handbook of Social Economics, Benhabib, Bisin, and Jackson, 2011)

Observation that many individual outcomes vary much more *between* social groups than *within* them

Application contexts:

- *social interaction* (consumption, academic achievement, unemployment, crime,...)
- *strategic interaction* (finance, production, IO, international trade...)
- *spatial externalities* (agglomeration effects, agriculture,...)
- *neighborhood effects* (new economic geography, housing market,...)

Theoretically: models of social interactions are widely used

Empirically: convincing tests of such models are still quite limited

- appropriate data sets difficult to find
- identification and measure of such peer effects is a quite difficult exercise

Methodological perspective

1. graph theory tools and spatial statistics techniques

2. inference conditional on the choice of the "neighborhood set"

3. spectral analysis of matrices

4. nested and non-nested model comparison

Criticalities in the use of statistical techniques in the social space

1. data on a regular grid, exogenous topology $\longrightarrow$ endogenous location of units

2. robusteness checks on neighborhood definition $\longrightarrow$ difficulties in the interpretation of underlying parameters

3. indicators of structural properties of matrices $\longrightarrow$ definition of interactions

4. nested and non-nested models analysis $\longrightarrow$ difficult in presence of complex schemes of cross-sectional dependence

Crucial issues in the economic literature on peer effects

1. Definition of the peers

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 at school: classroom or school level using average school achievements

2. Identification of the effects

- Reflection problem (Manski, 1993)
- Endogenous network formation/contextual effects
- Correlated individual unobservables

Does the "social multiplier" really exist?

Distinguishing the relative importance of social interactions, as opposed to preferences and attitudes, is of *positive* interest per se, but is fundamental from a *normative* perspective

3. Mechanism

Theoretical models of individual behaviour with social interactions

Bridge theory and empirics

Implementation of appropriate statistical procedures

Identification of peer effects through social networks

Models of Social Interactions

Focus: effect of the *average* level of activity of the group

$$\boldsymbol{y} = \boldsymbol{\phi y}_g + \boldsymbol{X\beta} + \boldsymbol{u}$$

Models of Social Networks

Focus: effect of the *structure* of such a group

$$\boldsymbol{y} = \boldsymbol{\phi G y} + \boldsymbol{X\beta} + \boldsymbol{u}$$

$\mathbf{G}$: the $n-$square adjacency matrix of a network g formalizes the structure of interactions of the agents in the social space

Empirical approach: assume a particular structure to the social interactions, use high quality data on social groups and draw inference based on that assumption (Bramoullé, Djebbari, Fortin, J. Econometrics, 2009; Calvò-Armengol, Patacchini, Zenou, Rev. Econ. Stud., 2009, Liu and Lee, J. Econometrics, 2010, Liu, Lee, Patacchini and Zenou, ?, 201?)

Theoretical approach: assume a particular functional form for social effects (linear in the efforts of other friends in the network with a quadratic cost function), use a Nash equilibrium concept and conduct a full-fledged equilibrium analysis that relates topology to outcome (Calvò-Armengol, Patacchini, Zenou, Rev. Econ. Stud., 2009)

Related literature - empirics of social networks

1) Reflection problem : Peer effects can be separately identified from contextual effects using the variations in the reference groups across individuals

2) Endogeneous network formation: Use high *quality data* to control for selection on observables and *network fixed effects* to control for selection on unobservables

Assumption: link formation is correlated with observed individual characteristics, contextual effects and that any remaining (troubling) source of unobserved heterogeneity can be captured at the network level

3) Estimation via maximum likelihood or IVs where IVs are the characteristics of indirect peers

In addition, in this paper:

4) we test the exogeneity of G using an *over-identifying restrictions (OIR) test* suited for this spatial framework (H_0 : G is exogenous (Lee, 2007))

5) we test for weak-instruments using a *first stage F test* suited for this spatial framework

6) we overcome the weak-instrument problem by using quadratic moment conditions (*GMM estimation*)

7) we implement a *J test for model selection* and *appropriate estimators* suited for this spatial framework

Related literature cont. -theory

Several papers have formalized the local-aggregate model (e.g. Ballester et al., 2006, 2010; Bramoullé and Kranton, 2007; Galeotti et al., 2009) and the local-average model (e.g. Bernheim, 1994, Akerlof, 1980, 1997, Glaeser and Scheinkman, 2003; Patacchini and Zenou, 2012)

Choice-theoretical justification for the local-aggregate (i.e. conformist) model (Clark and Oswald, 1998)

Fewer papers have tested these models - in education or crime- (Calvó-Armengol et al., 2009, Patacchini and Zenou, 2008, Lin, 2010; Boucher et al., 2010)

In this paper:

We constrast the two approaches and test them for different activities

Network models of peer effects with ex-ante heterogenous agents

Each agent i in network r selects an effort $y_{i,r} \geq 0$ and obtains a payoff given by the utility function

1) Local aggregate model (L AGG)

$$u_{i,r}(Y_r, g_r) = \underbrace{\left(a_{i,r} + \eta_r + \varepsilon_{i,r}\right) y_{i,r}}_{\text{Benefits from own effort}} \underbrace{-\frac{1}{2} y_{i,r}^2}_{\text{Costs}} \underbrace{+ \phi_1 \sum_{j=1}^{n} g_{ij,r} y_{i,r} y_{j,r}}_{\text{Benefits from own and friends' effort}}$$

2) Local average model (L AVE)

$$u_{i,r}(Y_r, g_r) = \underbrace{\left(a_{i,r} + \eta_r + \varepsilon_{i,r}\right) y_{i,r}}_{\text{Benefits from own effort}} \underbrace{-\frac{1}{2} y_{i,r}^2}_{\text{Costs}} \underbrace{- \frac{d}{2} \left(y_{i,r} - y_{i,r}^m\right)^2}_{\text{Benefits from own and friends' effort}}$$

Two individuals i and j are directly connected (i.e. best friends) in $\mathbf{G} = \left\{g_{ij,r}\right\}$ if and only if $g_{ij} = 1$, and $g_{ij} = 0$ otherwise

Payoffs are interdependent and agents choose their levels of activity simultaneously

If $\phi_1 \mu_1(\mathbf{G})$, $\phi_2 = d/(1+d) < 1$, unique Nash equilibrium of this peer effect game $y^*_{i,r}$

$$1) \text{ L AGG} : y^*_{i,r} = \phi_1 \sum_{j=1}^{n} g_{ij,r} y^*_{j,r} + a_{i,r} + \eta_r + \varepsilon_{i,r}$$

$$2) \text{ L AVE} : y^*_{i,r} = \phi_2 \frac{1}{g_{i,r}} \sum_{j=1}^{n_r} g_{ij,r} y^*_{j,r} + \frac{a_{i,r} + \eta_r + \varepsilon_{i,r}}{(1-\phi_2)}$$

Each individual provides effort proportional to that of her/his reference group of best friends and to her/his idiosyncratic characteristics

Empirical counterpart

Theoretical models: behavioral foundation for the so-called *spatial lag model*

$$y_{i,r} = \phi_1 \sum_{j=1}^{n_r} g_{ij,r} y_{j,r} + \underbrace{x'_{i,r}\beta_1 + \frac{1}{g_{i,r}} \sum_{j=1}^{n_r} g_{ij,r} x'_{j,r}\gamma_1}_{a_{i,r}} + \eta_{1r} + \epsilon_{1i,r},$$

outcome $\mathbf{G} = \left\{g_{ij}\right\} : n \times n$ *non row-normalized* adjacency matrix

$$y_{i,r} = \phi_2 \frac{1}{g_{i,r}} \sum_{j=1}^{n_r} g_{ij,r} y_{j,r} + \underbrace{x'_{i,r}\beta_2 + \frac{1}{g_{i,r}} \sum_{j=1}^{n_r} g_{ij,r} x'_{j,r}\gamma_2}_{a_{i,r}} + \eta_{2r} + \epsilon_{2i,r},$$

outcome $\mathbf{G} = \left\{g_{ij}\right\} : n \times n$ *row-normalized* adjacency matrix

In matrix notation ($\mathbf{G}$ *row-normalized* or *non row-normalized*):

$$\mathbf{y} = \phi\mathbf{G}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\eta} + \boldsymbol{\epsilon}$$

The reduced-form equation:

$$\mathbf{y} = [\mathbf{I} - \phi\mathbf{G}]^{-1}\mathbf{X}\boldsymbol{\beta} + [\mathbf{I} - \phi\mathbf{G}]^{-1}\mathbf{G}\mathbf{X}\boldsymbol{\gamma} + [\mathbf{I} - \phi\mathbf{G}]^{-1}(\boldsymbol{\eta} + \boldsymbol{\epsilon}).$$

If ϕ smaller than the norm of the inverse of the largest eigenvalue of $\mathbf{G}$ (Debreu and Herstein, 1953),

$$\sum_{k=0}^{+\infty} \phi^k \mathbf{G}^k = [\mathbf{I} - \phi\, \mathbf{G}]^{-1}$$

The parameter ϕ is a decay factor that scales down the relative weight of longer paths

$\sum_{k=0}^{+\infty} \phi^k g_{ij}^{[k]}$: number of paths starting at i and ending at j, where paths of length k are weighted by ϕ^k

The adjacency matrix $\mathbf{G} = [g_{ij}]$ keeps track of the direct connections. The kth power $\mathbf{G}^k = \mathbf{G}^{(k \ times \atop \ldots)}\mathbf{G}$ keeps track of indirect connections: the coefficient in the (i, j) cell of $\mathbf{G}^k$ gives the number of paths of length k between i and j

A path between i and j needs not to follow *the shortest possible route* between those agents

Generalized J test for network models with group fixed effects

Let us write the two models in matrix notation, for the entire sample:

$$\begin{aligned} H_1 &: Y = \phi_1 GY + X^*\delta_1 + \iota \cdot \eta_1 + \epsilon_1, \\ H_2 &: Y = \phi_2 G^*Y + X^*\delta_2 + \iota \cdot \eta_2 + \epsilon_2, \end{aligned}$$

The test of model H_1 against model H_2

Augmented model of H_1

$$Y = \alpha_1 Y_{H_2} + \phi_1 GY + X^*\delta_1 + \iota \cdot \eta_1 + \epsilon_1,$$

where Y_{H_2} is a predictor of Y under H_2 such that $Y_{H_2} = \phi_2 G^*Y + X^*\delta_2 + \iota \cdot \eta_2$ (see Kelejian and Prucha, 2007; Kelejian and Piras, 2011)

Hypotheses $H_0 : \alpha_1 = 0$ against $H_a : \alpha_1 \neq 0$

α_1 insignificant: evidence against model H_2

The test of model H_2 against model H_1

Augmented model of H_2

$$H_2 : Y = \alpha_2 Y_{H_1} + \phi_2 G^* Y + X^* \delta_2 + \iota \cdot \eta_2 + \epsilon_2,$$

where Y_{H_1} is a predictor of Y under H_1 such that $Y_{H_1} = \phi_1 GY + X^* \delta_1 + \iota \cdot \eta_1$.

Hypotheses $H_0 : \alpha_2 = 0$ against $H_a : \alpha_2 \neq 0$

α_2 insignificant: evidence against model H_1

J test for model selection: general validity for testing non-nested models

In our case: a hybrid model encompassing both *local-aggregate* and *local-average* effects

$$Y = \phi_1 GY + \phi_2 G^* Y + X^* \delta + \iota \cdot \eta + \epsilon$$

This model can be estimated by 2SLS and GMM methods (Liu and Lee ,2010) tailored to this case: all the available information is used in an efficient way

The asymptotic properties of the estimators are derived

Data

Dataset of friendship networks in the United States from the National Longitudinal Survey of Adolescent Health (AddHealth), roughly 90,000 students in grades 7-12 from roughly 130 private and public schools in years 1994-95

Richness of the information provided by the AddHealth data

Pupils were asked to identify their best friends from a school roster

Information on the characteristics of nominated friends

Selected network size range: between 50 and 150 students for all outcomes

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

Friendship relationships are not always reciprocal

Peer group definition

Knowing exactly who nominates whom in a network, we exploit the directed nature of the nominations data

We focus on choices made and we denote a link from i to j as $g_{ij,r} = 1$ if i has nominated j as his/her friend, and $g_{ij,r} = 0$, otherwise (out-degree)$\longrightarrow$ *directed networks*

Robustness checks: *undirected networks* and *a simulation experiment*

On average, students in our sample declare to have 1.46 friends with a standard deviation of 1.4

Definition of target variables and controls

Three different outcomes: (i) school performance; (ii) sport activities, such as playing baseball, softball, basketball, soccer, or football; (iii) screen activities, such as playing video or computer games.

Index of performance (or involvement) in each category using the answers to different related questions (factor analysis)

- *Screen activity index:* is derived using questions on how many hours a week students watch television, videos or play video or computer games.

The answers take values between 0 and 99 hours for each activity.

The final composite ranges between 0 and 13.01, with mean equals to 1.37 and standard deviation equals to 1.23

Sample size: 3196 students over 33 networks

- *Sport activity index:* is derived using questions on how often students go wheeling, play a team sport or do more general physical exercise during the past week.

Each response is coded using an ordinal scale ranging from 0 (i.e. never participate) to 1 (i.e. participate 1 or 2 times), 2 (participate 3 or 4 times) up to 3 (i.e. participate 5 or more times).

The final index ranges between 0 and 2.95, with mean equals to 1.53 and standard deviation equals to 1.05.

Sample size: 2934 students over 32 networks

- *GPA index:* measured using the respondent's scores in the more recent grading period in in several subjects (English or language arts, history or social science, mathematics, and science).

The scores are coded as 1=D or lower, 2=C, 3=B, 4=A.

The final index ranges between 0 and 4.40, with mean equals to 3.03 and standard deviation equals to 1.10.

Sample size: 1443 students over 13 networks

Description of Control Variables

Individual socio-demographic variables	
Female	Dummy variable taking value one if the respondent is female.
Black	Ethnic group dummies, white is the reference category
Other races	"
Grade	Grade of the student in the current year.
Self esteem (Screen, Sport)	Response to the question: "Compared with other people your age, how intelligent are you", coded as 1= moderately below average, 2= slightly below average, 3= about average, 4= slightly above average, 5= moderately above average, 6= extremely above average.
Self esteem (Gpa)	Response to the question: "I have a lot of good qualities", coded as 1= strongly agree, 2= agree, 3= neither agree nor disagree, 4= disagree, 5= strongly disagree.
Math_sc_A (Screen, Sport)	Mathematics score dummies, including a category capturing missing values. D is the reference category
Math_sc_B (Screen, Sport)	"
Math_sc_C (Screen, Sport)	"
Math_sc_mis (Screen, Sport)	"
Teacher troubles	Response to the question: "How often have you had trouble getting along with your teachers?" 0= never, 1= just a few times, 2= about once a week, 3= almost everyday, 4=everyday
School attachment	Response to the question: "I feel like you are part of your school", all coded as 1= strongly agree, 2= agree, 3=neither agree nor disagree, 4= disagree, 5= strongly disagree.
Family background variables	
Family size	Number of people living in the household
Parental education (Screen, Sport)	Schooling level of the (biological or non-biological) parent who is living with the child, distinguishing between "eighth grade or less", "more than eighth grade, but did not graduate from high school", "high school graduate", "completed a GED", "went to a business, trade, or vocational school after high school", "went to college but did not graduate", "graduated from college or a university", " professional training beyond a four-year college", coded as 1 to 8. If both parents are in the household the education of the father is considered. It is coded as zero if no parent lives with child or the reported level is "unknown".
Parental education (Gpa)	Schooling level of the (biological or non-biological) parent who is living with the child, distinguishing between "never went to school", "not graduate from high school", "high school graduate", "graduated from college or a university", "professional training beyond a four-year college", coded as 1 to 5. If both parents are in the household the education of the father is considered. It is coded as zero if no parent lives with child or the reported level is "unknown".
Parent occupation manager	Parent occupation dummies. Closest description of the job of (biological or non-biological) parent that is living with the child is manager. If both parents are in the household, the occupation of the father is considered. "none" is the reference group

Parent occupation professional/technical	”
Parent occupation office or sales worker	”
Parent occupation manual	”
Parent occupation military or security	”
Parent occupation farm or fishery	”
Parent occupation other	”
Married Parents	Dummy variable taking value one if the child leaves in a family with both parents who are married
Parental care	Dummy taking value one if both parents care very much about her/him
Residential neighborhood variables	
Neighborhood quality (Screen, Sport)	Interviewer response to the question "How well kept is the building in which the respondent lives", coded as 4= very poorly kept (needs major repairs), 3= poorly kept (needs minor repairs), 2= fairly well kept (needs cosmetic work), 1= very well kept.
Neighborhood quality (Gpa)	Response to the question: “I feel safe in my neighborhood" all coded as 1= strongly agree, 2= agree, 3=neither agree nor disagree, 4= disagree, 5= strongly disagree.

Endogenous selection into activity

Concern: possible mixing of extensive and intensive margins?

Less than 1% and 5% of the students never participated in screen and sport activities, respectively

A final assessment of the comprehension within the different subjects studied at school is mandatory

Estimation Results

Increasing set of controls and different estimators:

(a) "2SLS-1": a 2SLS estimator with traditional IVs .

(b) "2SLS-2": a 2SLS estimator with enlarged IVs

(c) "C2SLS": a bias-corrected 2SLS estimator with enlarged IVs

(d) "GMM-1": an optimal GMM estimator with traditional linear moment conditions and quadratic moment conditions

(e) "GMM-2": an optimal GMM estimator with enlarged linear moment conditions and the same quadratic moment conditions

(f) "CGMM": a bias-corrected optimal GMM estimator with the same (enlarged) moment functions

Results for *directed* and *undirected* networks roughly unchanged

Estimation Results: bias corrected optimal GMM

	(1)	(2)	(3)	(4)
Basic controls	yes	yes	yes	yes
Basic+		yes	yes	yes
Basic++				yes
Basic+++				yes
Network fixed effects	yes	yes	yes	yes
		SPORT		
Aggregate (φ_1)	**0.0216*** **	**0.0217*** **	**0.0215*** **	**0.0213*** **
	(0.0065)	**(0.0066)**	**(0.0065)**	**(0.0065)**
Average (φ_2)	**0.0148**	**0.0143**	**0.0136**	**0.0129**
	(0.0266)	**(0.0266)**	**(0.0266)**	**(0.0266)**
J test	Null Hp. (α_1=0)	No local aver.	*t statistic*	**0.483**
	Null Hp. (α_2=0)	No local aggr.	*t statistic*	**3.252**
		SCREEN ACTIVITIES		
Aggregate (φ_1)	**-0.0022**	**-0.0022**	**-0.0022**	**-0.0026**
	(0.0089)	**(0.0089)**	**(0.0089)**	**(0.0089)**
Average (φ_2)	**-0.0118**	**-0.0126**	**-0.0123**	**-0.0111**
	(0.0278)	**(0.0279)**	**(0.0279)**	**(0.0279)**
J test	Null Hp. (α_1=0)	No local aver.	*t statistic*	**0.368**
	Null Hp. (α_2=0)	No local aggr.	*t statistic*	**0.312**
		EDUCATION (GPA)		
Aggregate (φ_1)	**0.0153*** **	**0.0150*** **	**0.0149*** **	**0.0146*** **
	(0.0046)	**(0.0045)**	**(0.0045)**	**(0.0045)**
Average (φ_2)	**0.2279*** **	**0.2258*** **	**0.2216*** **	**0.2177*** **
	(0.0340)	**(0.0340)**	**(0.0340)**	**(0.0339)**
J test	Null Hp. (α_1=0)	No local aver.	*t statistic*	**6.411**
	Null Hp. (α_2=0)	No local aggr.	*t statistic*	**3.235**

Diagnostics		
	2SLS-1	**GMM-1**
All controls	yes	yes
Network fixed effects	yes	yes
	SPORT	
Aggregate (φ_1)	**0.0208*** **	**0.0223*** **
	(0.0077)	**(0.0066)**
Average (φ_2)	**0.0182**	**0.0209**
	(0.3125)	**(0.0267)**
1st Stage F statistic	0.883	
OIR test *p-value*		*0.887*
	SCREEN ACTIVITIES	
Aggregate (φ_1)	**-0.0021**	**-0.0008**
	(0.0103)	**(0.0091)**
Average (φ_2)	**0.3749**	**-0.0116**
	(0.3345)	**(0.0280)**
1st Stage F statistic	0.870	
OIR test *p-value*		*0.698*
	EDUCATION (GPA)	
Aggregate (φ_1)	**0.0059**	**0.0147*** **
	(0.0068)	**(0.0045)**
Average (φ_2)	**0.6952** **	**0.2447*** **
	(0.3201)	**(0.0340)**
1st Stage F statistic	1.127	
OIR test *p-value*		*0.540*

Robustness check

Our identification and estimation strategies depend on the correct specification of network links

There can be some "unobserved" network links that, if considered, would change the network topology and break some intransitivities in network links

Do our results change if some links are mispecified? To what extent?

How many links need to be mispecified before explaining away our results?

A simulation experiment

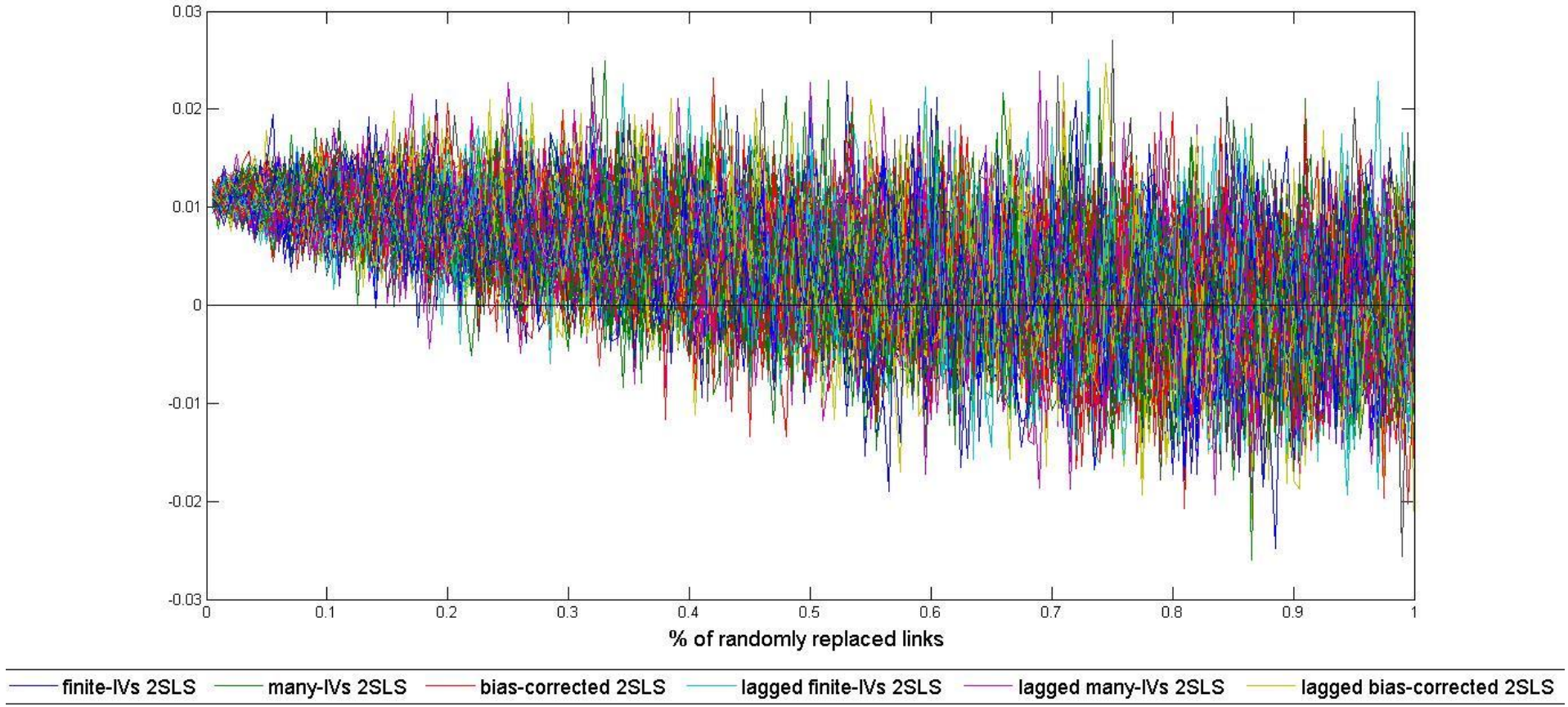
We simulate different network structures that differ from the real one by a given (increasing) number of misspecified links

$p_r = \{0.005, 0.010, 0.015, 0.020...0.95, 1\}$

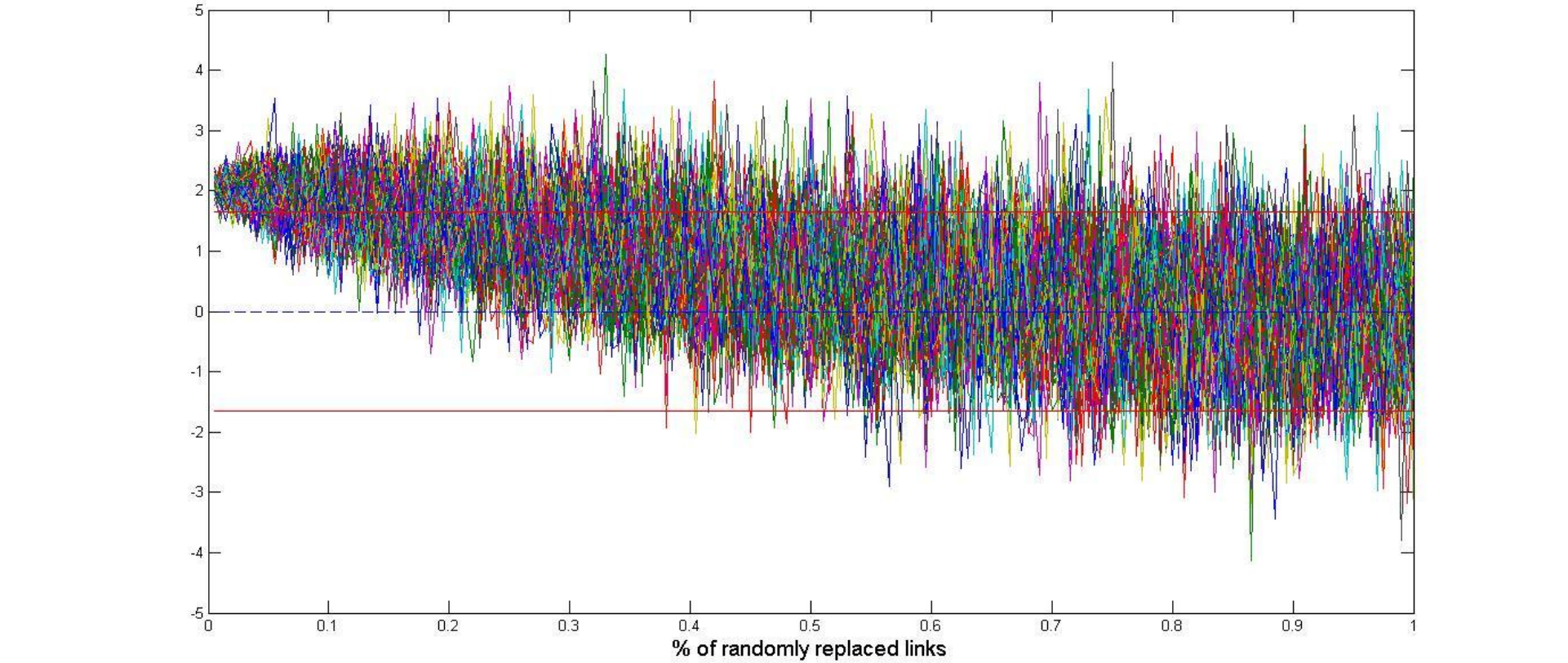
For each percentage of randomly replaced links p_r, we draw 100 network structures (samples) of size and network density equal to the real one

We then estimate our model replacing the real $\mathbf{G}$ matrix with the simulated ones in turn, so that in total we estimate our model twenty thousand times for each type of estimator

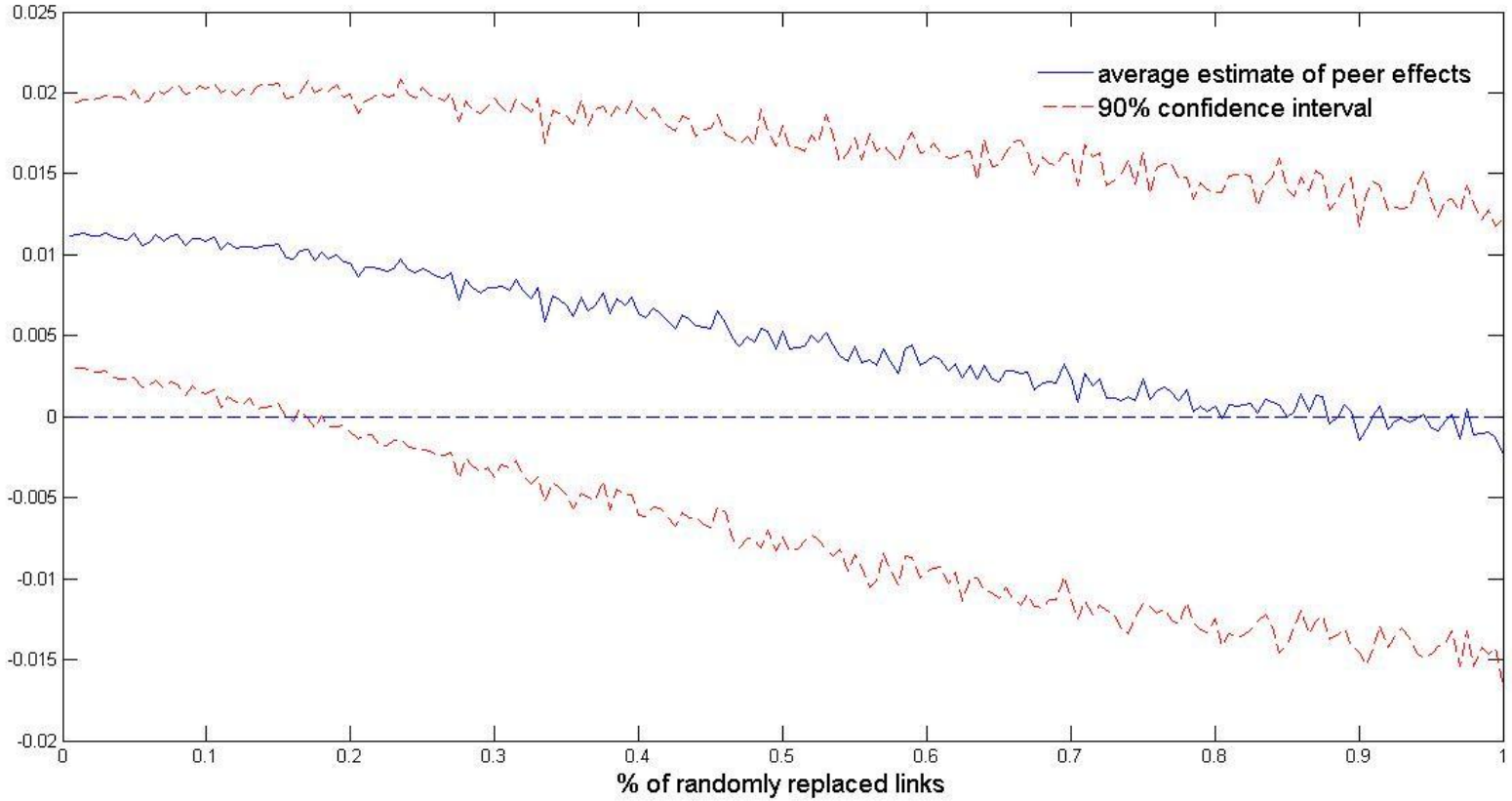
Estimates of peer effects



T statistics



Lagged bias-corrected 2SLS estimates of peer effects



THANK YOU FOR YOUR

ATTENTION