Dynamic Aspects of Teenage Friendships and Educational Attainment[*]

Eleonora Patacchini[†] Edoardo Rainone[‡] Yves Zenou[§]

April 25, 2011

Abstract

We study peer effects in education. We first develop a network model that predicts a relationship between own education and peers' education as measured by direct links in the social network. We then test this relationship using the four waves of the AddHealth data, looking at the impact of school friends nominated in the first wave in 1994-1995 on own educational outcome reported in the fourth wave in 2007-2008. We find that there are strong and persistent peer effects in education since a standard deviation increase in peers' education attainment translates into roughly a 10 percent increase of a standard deviation in the individual's education attainment (roughly 3.5 more months of education). We also find that peer effects are in fact significant only for adolescents who were friends in grades 10-12 but not for those who were friends in grades 7-9. This might indicate that social norms are important in educational choice since the individual's choice of college seems to be influenced by that of friends in the two last years of high school.

Key words: Social networks, education, peer effects, identification strategy
JEL Classification: C21, I21, Z13.

[*] We would like to thank the participants of the Workshop on Social Networks, Oxford, December 2010.

[†] La Sapienza University of Rome, EIEF, IZA and CEPR. E-mail: eleonora.patacchini@uniroma1.it

[‡] La Sapienza University of Rome E-mail: edoardo.rainone@uniroma1.it

[§] Stockholm University, Research Institute of Industrial Economics (IFN), GAINS, IZA and CEPR. E-mail: yves.zenou@ne.su.se.

1 Introduction

The influence of peers on education outcomes has been widely recognized both in economics and sociology. The extremely difficult task is to disentangle neighborhood effects from peer effects and there is no consensus on the importance of peer effects on own achievement in this literature (see, e.g. Goux and Maurin, 2007, and the two recent literature surveys by Durlauf, 2004, Ioannides and Topa, 2010, and Ioannides, 2011). The constraints imposed by the available disaggregated data force many studies to analyze peer effects at a quite aggregate and arbitrary level, such as at the school, grade or neighborhood level.[1] This leaves little chance to separate endogenous from exogenous (contextual) effects. Besides, the detections and measure of social interactions effects is hampered by a possible endogenous group (neighborhood) membership or by omitted variables problems. If the variables that drive the sorting of individuals into neighborhoods are not fully observable, potential correlations between (unobserved) factors and the target neighborhood level variables are a major sources of bias.

A popular strategy is to use an instrumental variable approach. Indeed, several studies eliminate the problem of correlation in unobservables at the neighborhood level by using metropolitan-area level variables and exploiting cross-metropolitan variations (see e.g. Evans et al., 1992; Cutler and Glaeser, 1997; Card and Rothstein, 2007; Weinberg, 2004). It is hard to guarantee, however, that metropolitan level variables do not directly affect outcomes. Bayer et al. (2008) adopt the converse research design by using data from the US Census that characterize residential location down to the city block. They exploit block-level variation in neighbor attributes, assuming the absence of correlation in unobservables across blocks within block groups.[2]

Other studies are based on specific social experiments or quasi-experimental data (e.g. Katz et al., 2001; Sacerdote, 2001; Zimmerman 2003). However, important concerns on the external validity of these strategies in the identification of neighborhood effects remain (see Moffitt, 2001, for a detailed discussion).

Recent papers (Bramoullé et al., 2009; Calvó-Armengol et al., 2009; Lin 2010; Liu et al. 2011; Patacchini and Zenou, 2012) systematically analyze the identification of peer effects in *social networks* and show to what extent they can be separately identified from contextual

[1] Usually, peer effects in education have been tested using a rather aggregate measure of peers such as the "neighborhood", which has been measured by the high school (Evans et al., 1992), the census tract (Brooks-Gun et al., 1993), and the ZIP code (Datcher, 1982; Corcoran et al., 1992) where individuals reside.

[2] Another popular strategy is to estimate peer effects in education using comparisons across cohorts within schools. See, in particular, Bifulco et al. (2011).

effects using the variations in the reference groups across individuals, which is typical in social contact network structure.[3] In particular, Lin (2010) and Liu et al. (2011) present a network model specification and an empirical strategy that is closely related to the one presented in this paper. Using data from the first wave of the AddHealth survey, these studies provide an assessment of peer effects in student academic performance (GPA) and in crime, respectively. Following Lee et al. (2010), Lin (2010) adopts a maximum likelihood estimation approach, whereas following Liu et Lee (2010), Liu et al (2011) use the 2SLS and generalized method of moments (GMM) approaches.

Both approaches (Lin, 2010 and Liu et al., 2011), however, are based on the same identification strategy. They are valid under the assumption that 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, through the inclusion of network fixed effects. They cannot deal with the possible presence of unobservable within group individual characteristics, like unobserved individual preferences, that drive both group choice and individual outcomes.

In this paper, we exploit the longitudinal structure of the AddHealth data, which allows a more than 10-years time interval between friendship choice and educational outcomes More specifically, we assess whether and to what extent peers (i.e. friends) during the teenage period play a role for the individual's future education attainment. Possible unobserved student's characteristics driving friends' choice at school (i.e. common interests in sports or other activities, cheap talking) are unlikely to remain important determinants of individual decisions later on in life.

To the best of our knowledge, this is the first paper that exploits this comprehensive set of information to assess peer effects in education in this dynamic perspective.[4] In addition, we measure peer groups as precisely as possible by exploiting the directed nature of the

[3] A similar argument, i.e. the use of out-group effects, to achieve the identification of the endogenous group effect in the linear-in-means model has also been used by Weinberg et al. (2004), Cohen-Cole (2006), Laschever (2009), and De Giorgi et al. (2010).

[4] Using the Wisconsin Longitudinal Study of Social and Psychological Factors in Aspiration and Attainment (WLS), Zax and Rees (2002) also analyze the role of friendships in school on future earnings. Their paper is quite different than ours since they do not have a theoretical model driving the empirical analysis and do not tackle the issue of endogenous sorting of individuals into groups. Using the British National Child Development Study (NCDS), Patacchini and Zenou (2011) investigate the effects of neighborhood quality (in terms of education) when a child is thirteen on his/her educational outcomes when he/she is adult. Similarly, Gould et al. (2011), using Israeli data, estimate the effect of the early childhood environment on a large array of social and economic outcomes lasting almost 60 years. In both studies, peer effects are measured by the neighborhood where people live and not by friendship nominations.

nomination data and, furthermore, we allow peer effects to be heterogeneous by exploiting the nomination order. More specifically, we weight each individual contact according to the nomination order so that individuals nominated first have more weights than those nominated later.

Our empirical investigation is guided by a theoretical social network model[5] that extends that of Ballester et al. (2006) to the case of heterogenous agents in education.[6] We develop a *local aggregate* model where it is the *sum of the efforts of the peers* that positively affects individuals' utility. We show that, in equilibrium, the topology of the network totally characterizes peer effects so that different positions in the network imply different effort levels. We are able to derive the best-reply function of each individual as a function of peers' effects, own and peers characteristics and network specific effect. We then test this equation using the AddHealth data. We exploit four unique features of the AddHealth data: *(i)* the nomination-based friendship information, which allows us to reconstruct the precise geometry of social contacts, *(ii)* the directed nature of the nominations to measure precisely peer groups, *(iii)* the nomination order, which enables us to consider heterogenous influences within peer groups, *(iv)* the longitudinal dimension, which provides a temporal interval between friends' nomination and educational outcome.

We find that there are *strong and persistent peer effects in education*. In other words, the "quality" of friends (in terms of future educational achievement) from high school has a positive and significant impact on own future education level. In terms of magnitude, we find that a standard deviation increase in peers' aggregate years of education (roughly two more high-school graduate friends) translates into roughly a 10 percent increase of a standard deviation in the individual's education attainment (roughly 3.5 more months of education). This is a strong effect, especially given our long list of controls and the fact that friendship networks might have changed over time. It is even stronger when the peer influence is allowed to be heterogenous in terms of order of nomination. The influence of peers at school seem to be carried over time. We also analyze if the peer effect results are stronger for friends in earlier grades than in later ones. For that, we split our sample between students who were in grades 7-9 and those who were in grades 10-12. We find that peer effects are significant for the latter but not for the former. This might indicate that social norms are important in educational choice since the individual's choice of college seems to be influenced by the choice of college of friends in the two last years of high school. In other words, individuals

[5] See Goyal (2007) and Jackson (2008) for an overview on the theory of social networks.

[6] For peer effect models in education, see the seminal contributions of De Bartoleme (1990) and Benabou (1993).

are more likely to adopt and pursue an objective (here educational choice) if this choice is popular among their peers, especially in the last years at school.

2 Theoretical framework

2.1 The model

We develop a network model of peer effects, where the network reflects the collection of active bilateral influences.

The network $N_r = \{1, \dots, n_r\}$ is a finite set of agents in network r ($r = 1, ..., \overline{r}$), where $\overline{r}$ is the total number of networks. We keep track of social connections by a matrix $\mathbf{G}_r = \{g_{ij,r}\}$, where $g_{ij,r} = 1$ if i and j are direct friends, and $g_{ij,r} = 0$, otherwise. Friendship are reciprocal so that $g_{ij,r} = g_{ji,r}$. All our results hold for *non-symmetric* and *weighted* networks but, for the ease of the presentation, we focus on symmetric and unweighted networks in the theoretical model (even though we use directed weighted networks in the empirical analysis). We also set $g_{ii,r} = 0$.

Preferences Individuals in network r decide how much effort to exert in education (e.g. how many hours to study). We denote by $y_{i,r}$ the educational effort level of individual i in network r and by $\mathbf{y}_r = (y_{1,r}, ..., y_{n,r})'$ the population effort profile in network r. Each agent i selects an effort $y_{i,r} \geq 0$, and obtains a payoff $u_{i,r}(\mathbf{y}_r, g_r)$ that depends on the effort profile $\mathbf{y}_r$ and on the underlying network g_r, in the following way:

$$u_{i,r}(\mathbf{y}_r, g_r) = (a_{i,r} + \eta_r + \varepsilon_{i,r})\ y_{i,r} - \frac{1}{2}y_{i,r}^2 + \phi \sum_{j=1}^{n} g_{ij,r} y_{i,r} y_{j,r} \qquad (1)$$

where $\phi > 0$. Two key aspects characterize the utility function $u_{i,r}(\mathbf{y}_r, g_r)$ of individual i in network r. There is the idiosyncratic exogenous part $(a_{i,r} + \eta_r + \varepsilon_{i,r})\ y_{i,r} - \frac{1}{2}y_{i,r}^2$ and the endogenous peer effect aspect $\phi \sum_{j=1}^{n} g_{ij,r} y_{i,r} y_{j,r}$. In (1), η_r denotes the unobservable network characteristics and $\varepsilon_{i,r}$ is an error term, meaning that there is some uncertainty in the benefit part of the utility function. There is also an ex ante *idiosyncratic heterogeneity*, $a_{i,r}$, which is assumed to be deterministic, perfectly *observable* by all individuals in the network and corresponds to the observable characteristics of individual i (like e.g. sex, race, age, parental education, etc.) and to the observable average characteristics of individual i's best friends, i.e. average level of parental education of i's friends, etc. (contextual effects). To be more precise, $a_{i,r}$ can be written as:

$$a_{i,r} = \sum_{m=1}^{M} \beta_m x_{i,r}^m + \frac{1}{g_{i,r}} \sum_{m=1}^{M} \sum_{j=1}^{n_r} g_{ij,r}\, x_{j,r}^m \gamma_m \tag{2}$$

where x_i^m is a set of M variables accounting for observable differences in individual characteristics of individual i, β_m, γ_m are parameters and $g_{i,r} = \sum_{j=1}^{n} g_{ij,r}$ is the total number of friends individual i has in network r. The benefits from the utility are given by $(a_{i,r} + \eta_r + \varepsilon_{i,r})\ y_{i,r}$ and are increasing in own educational effort $y_{i,r}$. In this first part, there is also a cost of providing educational effort, $\frac{1}{2}y_{i,r}^2$, which is also increasing in effort $y_{i,r}$. The second part of the utility function is: $\phi \sum_{j=1}^{n_r} g_{ij,r} y_{i,r} y_{j,r}$, which reflects the influence of friends' behavior on own action. The peer effect component is also heterogeneous, and this *endogenous heterogeneity* reflects the different locations of individuals in the friendship network r and the resulting effort levels. More precisely, bilateral influences are captured by the following cross derivatives, for $i \neq j$:

$$\frac{\partial^2 u_{i,r}(\mathbf{y}_r, g_r)}{\partial y_{i,r} \partial y_{j,r}} = \phi g_{ij,r} \geq 0. \tag{3}$$

When i and j are direct friends, the cross derivative is $\phi > 0$ and reflects strategic complementarity in efforts. When i and j are not direct friends, this cross derivative is zero. In particular, $\phi > 0$ means that if two students are friends, i.e. $g_{ij,r} = 1$, and if j increases her effort, then i will experience an increase in her (marginal) utility if she also increases her effort. Interestingly, utility increases with the *number* of friends each person has, weighted by efforts $x_{j,r}$.

To summarize, when individual i exerts some effort in education, the benefits of the activity depends on individual characteristics $a_{i,r}$, some network characteristics η_r and on some random element $\varepsilon_{i,r}$, which is specific to individual i. In other words, $a_{i,r}$ is the observable part (by the econometrician) of i's characteristics while $\varepsilon_{i,r}$ captures the unobservable characteristics of individual i. Note that the utility (1) is concave in own decisions, and displays decreasing marginal returns in own effort levels. In sum,

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

2.2 Nash equilibrium

We now characterize the Nash equilibrium of the game where agents choose their effort level $y_{i,r} \geq 0$ simultaneously. At equilibrium, each agent maximizes her utility (1). The corresponding first-order conditions are:

$$\frac{u_{i,r}(\mathbf{y}_r, g_r)}{\partial y_{i,r}} = a_{i,r} + \eta_r + \varepsilon_{i,r} - y_{i,r} + \phi \sum_{j=1}^{n} g_{ij,r} y_{j,r} = 0.$$

Therefore, we obtain the following best-reply function for each $i = 1, ..., n$:

$$y_{i,r} = \phi \sum_{j=1}^{n} g_{ij,r} y_{j,r} + a_{i,r} + \eta_r + \varepsilon_{i,r} \tag{4}$$

Denote by $\mu_1(\mathbf{G}_r)$ the spectral radius of $\mathbf{G}_r$. We have:

Proposition 1 *If $\phi\mu_1(\mathbf{G}_r) < 1$, the peer effect game with payoffs (1) has a unique Nash equilibrium in pure strategies given by (4)*

Proof. Apply Theorem 1, part b, in Calvó-Armengol et al. (2009) to our problem. ▮

We would like now to test this model, especially equation (4), using data from adolescent friendships in the US. In other words, we would like to see how strong are peer effects in education by estimating the magnitude of ϕ.

3 Data description

Our analysis is made possible by the use of a unique database on friendship networks from the National Longitudinal Survey of Adolescent Health (AddHealth).[7]

The AddHealth survey has been designed to study the impact of the social environment (i.e. friends, family, neighborhood and school) on adolescents' behavior in the United States by collecting data on students in grades 7-12 from a nationally representative sample of roughly 130 private and public schools in years 1994-95. Every pupil attending the sampled schools on the interview day is asked to compile a questionnaire (in-school data) containing questions on respondents' demographic and behavioral characteristics, education, family

[7] This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

background and friendship. This sample contains information on roughly 90,000 students. A subset of adolescents selected from the rosters of the sampled schools, about 20,000 individuals, is then asked to compile a longer questionnaire containing more sensitive individual and household information (in-home and parental data). Those subjects of the subset are interviewed again in 1995–96 (wave II), in 2001–2 (wave III), and again in 2007-2008 (wave IV).[8]

From a network perspective, the most interesting aspect of the AddHealth data is the friendship information, which is based upon actual friends nominations. It is collected at wave I, i.e. when individuals were at school. Indeed, pupils were asked to identify their best friends from a school roster (up to five males and five females).[9] As a result, one can reconstruct the whole geometric structure of the friendship networks. Such a detailed information on social interaction patterns allows us to measure the peer group more precisely than in previous studies. 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 in network r, and $g_{ij,r} = 0$, otherwise.[10] In addition, we also exploit the nomination order to weight differently the influence of each peer within peer groups, i.e. we consider heterogenous peer effects. To the best of our knowledge this information has not been used before. More specifically, we weight each individual contact using a function which is linearly decreasing with the corresponding order in the nomination list and also accounts for the total number of nominations made by the individual. Each non-zero entry $w_{ij,r}$ of the adjacency matrix $\mathbf{G}_r$ would be:

$$w_{ij,r} = 1 - \frac{(\vartheta - 1)}{g_{i,r}}$$

where ϑ denotes the order of nomination given by individual i to friend j in his/her nomination list while $g_{i,r} = \sum_{j=1}^{n} g_{ij,r}$ is the total number of nominations made by individual i. By doing so, we allow for the fact that each individual can be affected differently by different peers within his/her peer group. For example, imagine that individual 1 has nominated three friends (i.e. $g_{i,r} = 3$), say first friend 2, then 4 and then 3. In that case, $\vartheta = 1$ for individual 2, $\vartheta = 2$ for individual 4 and $\vartheta = 3$ for individual 3. We will therefore have the

[8] The AddHealth website describes survey design and data in details. http://www.cpc.unc.edu/projects/addhealth

[9] The limit in the number of nominations is not binding (even by gender). Less than 0.1% of the students in our sample show a list of ten best friends.

[10] As highlighted by Wasserman and Faust (1994), centrality indices for directional relationships generally focus on choices made.

following weights: $w_{12,r} = 1$, $w_{14,r} = 2/3$ and $w_{13,r} = 1/3$ and therefore in the first row of the adjacency matrix $\mathbf{G}_r$ (for individual 1), there will be a 0 for individuals that 1 has not nominated and a 1, 2/3 and 1/3 for individuals 2, 4 and 3, respectively.

By matching the identification numbers of the friendship nominations to respondents' identification numbers, one can also obtain information on the characteristics of nominated friends. In addition, the longitudinal structure of the survey provides information on both respondents and friends during the adulthood. In particular, the questionnaire of wave IV contains detailed information on the highest education qualification achieved. We measure education attainment in completed years of full time education.[11] Social contacts (i.e. friendship nominations) are, however, only collected in Wave I. Our final sample of in-home wave I students (and friends) that are followed over time and have non missing information on our target variables both in waves I and IV consists of 1,319 individuals distributed over 138 networks. The minimum number of individuals in a network is 4 while its maximum is 100.[12] The mean and the standard deviation of network size are roughly 9 and 14 individuals, respectively.[13]

Table A.1 in Appendix 1 provides the descriptive statistics and definitions of the variables used in our study.[14] Among the individuals selected in our sample, 53% are female and 17% are blacks. The average parental education is high school graduate. Roughly 10% have parents working in a managerial occupation, another 10% in the office or sales sector, 20% in a professional/technical occupation, and roughly 30% have parents in manual occupations. More than 70% of our individuals come from household with two married parents, from an household of about four people on average. At wave IV, 45% of our adolescents are now married and roughly half of them (47%) have a son or a daughter. The mean intensity in religion practice slightly decreases during the transition from adolescence to adulthood. On

[11] More precisely the Wave IV questionnaire asks the highest education qualification achieved (distinguishing between 8th grade or less, high school, vocational/technical training, bachelor's degree, graduate school, master's degree, graduate training beyond a master's degree, doctoral degree, post baccalaureate professional education). Those with high school and above qualification are also asked to report the exact year when the highest qualification was achieved. Such an information allows us to construct a reliable measure of each individual's completed years of education.

[12] We do not consider networks at the extremes of the network size distribution (i.e. composed by 2-3 individuals or by more than 100) because peer effects can show extreme values (too hig or too low) in these edge networks.

[13] On average, these adolescents declare having 1.46 friends with a standard deviation of 1.4.

[14] Information at the school level, such as school quality and teacher/pupil ratio is also available but we don't use it since our sample of networks are within schools and we use fixed network effects in our estimation strategy.

average, during their teenage years, our individuals felt that adults care about them and had a good a good relationship teachers. Roughly, 30% of our adolescents were highly performing individuals at school, i.e. had the highest mark in mathematics.

Before we start our empirical analysis, we would like to look at simple correlations between the education attainment of an individual and the friends that he/she has nominated when she/he was adolescent at school. Figure A.1 documents this correlations by differentiating between *direct* best friends ($k = 1$), friends of friends ($k = 2$), etc. One clearly sees that the correlation curve is decreasing and is steeper when different weights are put on friends according to their nomination order. This indicates that direct friends's education outcomes have much more impact on own education outcome than indirect friends and that this relation is stronger when the order of nomination is taken into account. For example, it can be seen from Figure 1 that the correlation in education between an individual and his/her direct friend is twice as high as between an individual and his/her indirect friend of length 8 (i.e. $k = 8$).

4 Empirical analysis

4.1 Empirical model

Let $\bar{r}$ be the total number of networks in the sample ($\bar{r} = 138$ in our dataset), n_r be the number of individuals in the rth network, and $n = \sum_{r=1}^{\bar{r}} n_r$ be the total number of individuals ($n = 1,319$ in our dataset). For $i = 1, \cdots, n_r$ and $r = 1, \cdots, \bar{r}$, the empirical model corresponding to (4) can be written as:

$$y_{i,r,t+1} = \phi \sum_{j=1}^{n_r} g_{ij,r,t} y_{j,r,t+1} + x'_{i,r,t,t+1}\delta + \frac{1}{g_{i,r,t}} \sum_{j=1}^{n_r} g_{ij,r,t} x'_{j,r,t}\gamma + \eta_{r,t} + \epsilon_{i,r,t+1}, \tag{5}$$

where $y_{i,r,t+1}$ is the highest education level reached by individual i at time $t+1$ who belonged to network r at time t, where time $t+1$ refers to wave IV in 2007-2008 while time t refers to wave I in 1994-95. Similarly, $y_{j,r,t}$ is the highest education level reached by individual j at time $t+1$ who has been nominated as his/her friend by individual i at time t in network r. All the other variables have the same meaning as in equation (4) with the added new time subscript t or $t+1$ or both. For example, $x'_{i,r,t,t+1} = (x^1_{i,r,t,t+1}, \cdots, x^m_{i,r,t,t+1})'$ indicates the different individual characteristics both at times t (e.g. self esteem, mathematics score, quality of the neighborhood, etc.) and $t+1$ (marital status, age, children, etc.) of individual i. Some characteristics are clearly the same at times t and $t+1$, such as race, parents' education, gender, etc. Finally, $\epsilon_{i,r}$'s are i.i.d. innovations with zero mean and variance σ^2 for all i and r.

Observe that even if our theoretical model is static, as can be seen by (4), we use here a dynamic formulation of (4) for both econometric (this prevents reverse causality problems) and economic issues since we want to know how persistent are peer effects over time. We also adopt this strategy because we do not have information about the friendship network at time $t+1$; we have it only at time t. Observe also that, in the data, $y_{i,r,t+1}$ is the highest education level reached by individual i while, in the model, $y_{i,r}$ is the educational effort level of individual i. It seems reasonable to approximate education effort by education attained since the two are strongly correlated.

In the next two sections, to avoid too cumbersome notations, we omit the time index.

4.2 Identification strategy

The identification of peer effects (ϕ in model (5)) raises different challenges.

In *linear-in-means* models, simultaneity in behavior of interacting agents introduces a perfect collinearity between the expected mean outcome of the group and its mean characteristics. Therefore, it is difficult to differentiate between the effect of peers' choice of effort and peers' characteristics that do impact on their effort choice (the so-called *reflection problem*; see Manski, 1993). Basically, the reflection problem arises because, in the standard approach, individuals interact in groups, that is individuals are affected by all individuals belonging to the same group and by nobody outside the group. In other words, groups completely overlap. In the case of social networks, instead, this is nearly never true since the reference group has individual-level variation. Formally, as shown by Bramoullé et al. (2009), social effects are identified (i.e. there is no reflection problem) if $\mathbf{I}$, $\mathbf{G}_r$ and $\mathbf{G}_r^2$ are linearly independent where $\mathbf{I}$ is the identity matrix and $\mathbf{G}_r^2$ keeps track of indirect connections of length 2 in network r. In other words, if i and j are friends and j and k are friends, it does not necessarily imply that i and k are also friends. Denote $\mathbf{X}_r = (x_{1,r}, \cdots, x_{n_r,r})'$ and $\mathbf{Y}_r = (y_{1,r}, \cdots, y_{n_r,r})'$. Then, because of these intransitivities, $\mathbf{G}_r^2\mathbf{X}_r$, $\mathbf{G}_r^3\mathbf{X}_r$, etc. are not collinear with $\mathbf{G}_r\mathbf{X}_r$ and they can therefore act as valid instruments. Take, for example, individuals i, j and k in network r such that $g_{ij,r} = 1$ and $g_{jk,r} = 1$ but $g_{ik,r} = 0$. In that case, for individual i, the characteristics of peers of peers $\mathbf{G}_r^2\mathbf{X}_r$ (i.e. $x_{k,r}$) is a valid instrument for peers' behavior $\mathbf{G}_r^2\mathbf{Y}_r$ (i.e. $y_{j,r}$) since $x_{k,r}$ affects $y_{i,r}$ only indirectly through its effect on $y_{j,r}$ (distance 2). The architecture of social networks implies that these attributes will affect each individual outcome only through their effect on his/her friends' outcomes. Even in linear-in-means models, the Manski's (1993) reflection problem is thus eluded.[15] Peer effects in social networks are thus identified and can be estimated using 2SLS or maximum likelihood (Lee 2007; Calvó-Armengol et al., 2009; Lin, 2010).[16]

Although this setting allows us to solve the reflection problem, the estimation results might still be flawed because of the presence of *unobservable factors* affecting both individual and peer behaviors. It is indeed difficult to disentangle the endogenous peer effects from the correlated effects, i.e. effects arising from the fact that individuals in the same network tend to behave similarly because they face a common environment. If individuals are not randomly assigned into networks, this problem might originate from the possible sorting of agents. If the variables that drive this process of selection are not fully observable, potential correlations between (unobserved) network-specific factors and the target regressors are major sources of bias. A number of papers using network data have dealt with the estimation of peer effects with correlated effects (e.g., Clark and Loheac 2007; Lee 2007; Calvó-Armengol et al., 2009; Lin, 2010; Lee et al., 2011). This approach is based on the use of *network fixed effects* and extends Lee (2003) 2SLS methodology. Network fixed effects can be interpreted as originating from a two-step model of link formation where agents self-select into different networks in a first step and, then, in a second step, link formation takes place within networks based on observable individual characteristics only. An estimation procedure alike to a panel within group estimator is thus able to control for these correlated effects. One can get rid of the network fixed effects by subtracting the network average from the individual-level variables.[17] As detailed in the next section, this paper follows this approach.

Finally, one might question the presence of problematic unobservable factors that are not network-specific, but rather individual-specific. In this respect, the richness of the information provided by the AddHealth questionnaire on adolescents' behavior allow us to

[15] These results are formally derived in Bramoullé et al. (2009) (see, in particular, their Proposition 3) and used in Calvó-Armengol et al. (2009) and Lin (2010). Cohen-Cole (2006) presents a similar argument, i.e. the use of out-group effects, to achieve the identification of the endogenous group effect in the linear-in-means model (see also Weinberg et al., 2004; Laschever, 2009).

[16] More technical results can be found in Liu and Lee (2010). Liu et al. (2011) explicitly study the case of a non row-normalized adjacency matrix and provides the conditions on the parameters that guarantee the identification of peer effects (similarly to the conditions derived by Bramoullé et al., 2009, who derive them for the case of a row-normalized adjacency matrix).

[17] Bramoullé et al. (2009) also deal with this problem in the case of a row-normalized $\mathbf{G}_r$ matrix. In their Proposition 5, they show that if the matrices $\mathbf{I}$, $\mathbf{G}_r$, $\mathbf{G}_r^2$ and $\mathbf{G}_r^3$ are linearly independent, then by subtracting from the variables the network average (or the average over neighbors, i.e. direct friends), social effects are again identified and one can disentangle endogenous effects from correlated effects. In our dataset this condition of linear independence is always satisfied.

find proxies for typically unobserved individual characteristics that may be correlated with our variable of interest. Specifically, to control for differences in leadership propensity across adolescents, we include an indicator of *self-esteem* and an indicator of the *level of physical development* compared to peers, and we use *mathematics score* as an indicator of ability. Also, we attempt to capture differences in attitude towards education, parenting and more general social influences by including indicators of the student's school attachment, relationship with teachers, parental care and social inclusion.

In addition, we present an IV approach that uses as instruments only variables lagged in time to ensure that the instruments are not correlated with the contemporaneous error term. Observe that any unobserved source of heterogeneity that can be captured at the network level is already taken into account by the inclusion of network fixed effects.

4.3 Econometric methodology

Our econometric methodology follows closely Liu and Lee (2010). Let us expose this approach and highlight the modification that is implemented in this paper.

Let $\mathbf{Y}_r = (y_{1,r}, \cdots, y_{n_r,r})'$, $\mathbf{X}_r = (x_{1,r}, \cdots, x_{n_r,r})'$, and $\boldsymbol{\epsilon}_r = (\epsilon_{1,r}, \cdots, \epsilon_{n_r,r})'$. Denote the $n_r \times n_r$ sociomatrix by $\mathbf{G}_r = [g_{ij,r}]$, the row-normalized of $\mathbf{G}_r$ by $\mathbf{G}_r^*$, and the n_r-dimensional vector of ones by $\mathbf{l}_{n_r}$. Then model (5) can be written in matrix form as:

$$\mathbf{Y}_r = \phi \mathbf{G}_r \mathbf{Y}_r + \mathbf{X}_r^* \beta + \eta_r \mathbf{l}_{n_r} + \boldsymbol{\epsilon}_r, \tag{6}$$

where $\mathbf{X}_r^* = (\mathbf{X}_r, \mathbf{G}_r^* \mathbf{X}_r)$ and $\beta = (\delta', \gamma')'$.

For a sample with $\bar{r}$ networks, stack up the data by defining $\mathbf{Y} = (\mathbf{Y}_1', \cdots, \mathbf{Y}_{\bar{r}}')'$, $\mathbf{X}^* = (\mathbf{X}_1^{*\prime}, \cdots, \mathbf{X}_{\bar{r}}^{*\prime})'$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1', \cdots, \boldsymbol{\epsilon}_{\bar{r}}')'$, $\mathbf{G} = \mathrm{D}(\mathbf{G}_1, \cdots, \mathbf{G}_{\bar{r}})$, $\boldsymbol{\iota} = \mathrm{D}(\mathbf{l}_{n_1}, \cdots, \mathbf{l}_{n_{\bar{r}}})$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_{\bar{r}})'$, where $\mathrm{D}(\mathbf{A}_1, \cdots, \mathbf{A}_K)$ is a block diagonal matrix in which the diagonal blocks are $m_k \times n_k$ matrices $\mathbf{A}_k$'s. For the entire sample, the model is

$$\mathbf{Y} = \mathbf{Z}\theta + \boldsymbol{\iota} \cdot \boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{7}$$

where $\mathbf{Z} = (\mathbf{GY}, \mathbf{X}^*)$ and $\theta = (\phi, \beta')'$.

We treat η as a vector of unknown parameters. When the number of networks $\bar{r}$ is large, we have the incidental parameter problem. Let $\mathbf{J} = \mathrm{D}(\mathbf{J}_1, \cdots, \mathbf{J}_{\bar{r}})$, where $\mathbf{J}_r = \mathbf{I}_{n_r} - \frac{1}{n_r}\mathbf{l}'_{n_r}\mathbf{l}_{n_r}$. The network fixed effect can be eliminated by a transformation with $\mathbf{J}$ such that:

$$\mathbf{JY} = \mathbf{JZ}\theta + \mathbf{J}\boldsymbol{\epsilon}. \tag{8}$$

Let $\mathbf{M} = (\mathbf{I} - \phi\mathbf{G})^{-1}$. The equilibrium outcome vector $\mathbf{Y}$ in (7) is then given by the reduced form equation:

$$\mathbf{Y} = \mathbf{M}(\mathbf{X}^*\delta + \boldsymbol{\iota} \cdot \boldsymbol{\eta}) + \mathbf{M}\boldsymbol{\epsilon}. \tag{9}$$

It follows that $\mathbf{GY} = \mathbf{GMX}^*\beta + \mathbf{GM}\boldsymbol{\iota}\boldsymbol{\eta} + \mathbf{GM}\boldsymbol{\epsilon}$. $\mathbf{GY}$ is correlated with $\boldsymbol{\epsilon}$ because $\mathrm{E}[(\mathbf{GM}\boldsymbol{\epsilon})'\boldsymbol{\epsilon}] = \sigma^2\mathrm{tr}(\mathbf{GM}) \neq 0$. Hence, in general, (8) cannot be consistently estimated by OLS.[18] If $\mathbf{G}$ is row-normalized such that $\mathbf{G}\cdot\mathbf{l}_n = \mathbf{l}_n$, where $\mathbf{l}_n$ is a n-dimensional vector of ones, the endogenous social interaction effect can be interpreted as an average effect. With a row-normalized $\mathbf{G}$, Lee et al. (2010) have proposed a partial-likelihood estimation approach for the estimation based on the transformed model (8). However, for this empirical study, we are interested in the *aggregate endogenous effect* instead of the *average effect*. Hence, row-normalization is not appropriate. Furthermore, we are also interested in the centrality of networks that are captured by the variation in row sums in the adjacency matrix $\mathbf{G}$. Row-normalization could eliminate such information. If $\mathbf{G}$ is not row-normalized as it is in this empirical study, the (partial) likelihood function for (8) could not be derived, and alternative estimation approaches need to be considered. Liu and Lee (2010) use an instrumental variable approach and propose different estimators based on different instrumental matrices, denoted here by $\mathbf{Q}_1$, $\mathbf{Q}_2$ and $\mathbf{Q}_3$. They first consider the 2SLS estimator based on the conventional instrumental matrix for the estimation of (8): $\mathbf{Q}_1 = \mathbf{J}(\mathbf{GX}^*, \mathbf{X}^*)$ (*finite-IVs 2SLS*). For the case that the adjacency matrix $\mathbf{G}$ is not row-normalized, Liu and Lee (2010) then propose to use additional instruments (IVs) $\mathbf{JG}\boldsymbol{\iota}$ and enlarge the instrumental matrix: $\mathbf{Q}_2 = (\mathbf{Q}_1, \mathbf{JG}\boldsymbol{\iota})$ (*many-IVs 2SLS*). The additional IVs of $\mathbf{JG}\boldsymbol{\iota}$ are based on the row sums of $\mathbf{G}$ (i.e. the outdegrees of a network) and thus use the information on centrality of a network. They show that those additional IVs could help model identification when the conventional IVs are weak and improve upon the estimation efficiency of the conventional 2SLS estimator based on $\mathbf{Q}_1$. However, the number of such instruments depends on the number of networks. If the number of networks grows with the sample size, so does the number of IVs. The 2SLS could be asymptotic biased when the number of IVs increases too fast relative to the sample size (see, e.g., Bekker, 1994; Bekker and van der Ploeg, 2005; Hansen et al., 2008). Liu and Lee (2010) have shown that the proposed many-IV 2SLS estimator has a properly-centered asymptotic normal distribution when the average group size needs to be large relative to the number of networks in the sample. As detailed in Section 3, in this empirical study, we have a number of small networks. Liu and Lee (2010) also propose a bias-correction procedure based on the estimated leading-order many-IV bias: $\mathbf{Q}_3$ (*bias-corrected 2SLS*). The bias-corrected many-IV 2SLS estimator is properly centered, asymptotically normally distributed, and efficient when the average group size is sufficiently large. It is thus the more

[18] Lee (2002) has shown that the OLS estimator can be consistent in the spatial scenario where each spatial unit is influenced by many neighbors whose influences are uniformly small. However, in the current data, the number of neighbors are limited, and hence that result does not apply.

appropriate estimator in our case study (see Liu and Lee (2010) for a detailed derivation and an analysis of the asymptotic properties of the different estimators).[19]

In this paper, we use these estimators and also implement a modification of this approach, which takes advantage of the longitudinal structure of our data. The exact equivalent of (5) can be written in matrix form as:

$$\mathbf{Y}_{r,t+1} = \phi \mathbf{G}_{r,t} \mathbf{Y}_{r,t+1} + \mathbf{X}^*_{r,t}\beta_1 + \mathbf{X}^*_{r,t+1}\beta_2 + \boldsymbol{\eta}_{r,t}\mathbf{l}_{n_r} + \boldsymbol{\epsilon}_{r,t+1}.$$

Our modification of the IV approach proposed by Liu and Lee (2010) consists in including in the different instrumental matrices only values lagged in time (i.e. observed in wave I). So, for instance the first instrumental matrix for the finite-IVs 2SLS will thus be: $\mathbf{Q}_1' = \mathbf{J}(\mathbf{G}_t\mathbf{X}^*_t, \mathbf{X}^*_t)$. Such a strategy should ensure that the instruments are not correlated with the contemporaneous (wave IV) error term ϵ_{t+1}, thus strengthening our identification strategy.

5 Estimation results

5.1 General results

Table 1 collects the estimation results of model (5) when using the different estimators discussed in the previous section, without using the information of the nomination order, i.e. all nominated friends receive the same weight equals to 1.

As explained above, for the estimation of ϕ, we pool all the networks together by constructing a block-diagonal network matrix with the adjacency matrices from each network on the diagonal block. Hence we implicitly assume that the ϕ in the empirical model is the same for all networks. The difference between networks is controlled for by network fixed effects. Indeed, the estimation of ϕ for each network might be difficult (in terms of precision) for the small networks. Furthermore, as stated above, it is a crucial empirical concern to control for unobserved network heterogeneity by using network fixed effects.

Proposition (1) requires that ϕ is in absolute value smaller than the inverse of the largest eigenvalue of the block-diagonal network matrix $\mathbf{G}_r$, i.e. $\phi < 1/\mu(\mathbf{G}_r)$. In our case, the largest eigenvalue of $\mathbf{G}_r$ is 3.70 . Furthermore our theoretical model postulates that $\phi \geq 0$. As a result, we can accept values within the range $[0, 0.280)$. Table 1 shows, in the first column, the results obtained when using our most extensive set of instruments and, in column 2, those produced when using as instruments only variables lagged in time. All our estimates of ϕ

[19] Liu and Lee (2010) also generalize this 2SLS approach to the GMM using additional quadratic moment conditions.

are within the acceptable parameter space $[0, 0.280)$ and are all significant. Looking across columns, it appears that the results are similar and only slightly higher in magnitude in the second column. This finding (incidentally) validates the empirical identification strategies used by Lin (2010) and Liu et al. (2011). Indeed, given the extensive set of controls available in the AddHealth, the inclusion of network fixed effects, and, most importantly, because friendship networks are quite small (see Section 3), the presence of uncaptured (troubling) individual unobserved within network characteristics is very unlikely. If these factors were at work, we should have found a substantial difference between the results in the first and second column in Table 1, since the latter controls for such influences.

$[Insert\ Table\ 1\ here]$

Looking now within each column, as explained above, in our case study with small networks in the sample, the preferred estimator is the bias-corrected 2SLS one. Let us thus focus on the bias-corrected estimator. First, the effect of friends' education on own education is always significant and positive, i.e., there are *strong and persistent peer effects in education*. This shows that the "quality" of friends (in terms of future educational achievement) from high school has a positive and significant impact on own future education level, even thought it might be that individuals who were close friends in 1994-1995 (wave I) might not be friends anymore in 2007-2008 (wave IV). In terms of magnitude, we find that a standard deviation increase in peers' aggregate years of education (roughly two more high-school graduate friends) translates into roughly a 10 percent increase of a standard deviation in the individual's education attainment (roughly 3.5 more months of education). This is a strong effect, especially given our long list of controls and the fact that friendship networks might have changed over time. The influence of peers at school seem to be carried over time.

When the information on the nomination order is exploited (Table2), thus allowing to weight differently best friends, the magnitude of the effects is higher. A standard deviation increase in peers' education attainment translates into roughly a 15 percent increase of a standard deviation in the individual's education attainment (roughly 6 more months of education).[20]

$[Insert\ Table\ 2\ here]$

[20] When $\mathbf{G}_r$ is weighted the largest eigenvalue is 2.59. We can thus accept values within the range $[0, 0.385)$. All our estimates of ϕ in Table 2 are within this parameter space.

5.2 Additional results

We have seen so far that friends at school have an impact on own future educational outcome. We would like to understand better what is behind this result. Is it the case that the choice to go to college is affected by the choice of peers? Or is it the case that hanging out with friends who study hard increase one's motivation to study and therefore the choice to go to college? With our dataset, it is difficult to pin down the exact mechanism behind our peer effect results. We can, however, improve our understanding of the results.

The students who were interviewed in wave I (1994-1995) in the AddHealth survey were in fact in different grades. We would like now to see if the peer effect results are stronger for friends in earlier grades than in later ones. For that, we split our sample between students who were in grades 7-9 and those who were in grades 10-12 in wave I and estimate model (5) on these two sub-samples separately. The results are contained in Tables 3 and 4, for grades 7-9 and Tables 5 and 6 for grades 10-12. We consider again unweighted networks (Tables 3 and 5) and weighted networks (Tables 4 and 6). We find that in the early grades peer effects do not seem to be at work (Table 3), not even when the influence of peers is weighted by the order of nomination (Table 4). On the contrary, when we consider friendships occurring only in the two last years of high school (grades 10-12), then effects of peers' education on own education become significant (Tables 5 and 6) and very large in magnitude.

[*Insert Tables* 3, 4, 5, 6 *here*]

This suggests that friendships made earlier in life do not last or do not affect educational choices made after high school while this is the not case for friendships made later in life. This also indicates that social norms are important in educational choice since the individual's choice of college seems to be influenced by the choice of college of friends in the two last years of high school. In other words, individuals are more likely to adopt and pursue an objective (here educational choice) if this choice is popular among their peers, especially in the last years at school. This could represent the effect of contagion and collective socialization. This result is in line with that of Zax and Rees (2002) who, using the Wisconsin Longitudinal Study of Social and Psychological Factors in Aspiration and Attainment (WLS), find that the college aspirations of friends are positively and significantly related to respondents' later earnings.

6 Robustness checks

Our identification and estimation strategies depend on the correct specification of network links. In particular, our identification strategy hinges upon non linearities in group membership, i.e. on the presence of intransitive triads. In this section we test the robustness of our results with respect to misspecification of network topology. So far we have measured peer groups as precisely as possibly by exploiting the direction of the nomination data. However, there can be, for example, some "unobserved" network link that, if considered, would change the network topology and break some intransitivities in network links. In this section we use simulated data to answer questions such as: Do our results change if some links are misspecified? To what extent? How many links need to be misspecified before explaining away our results?

A Numerical experiment

We use a simulation approach to randomly change a certain percentage of links in each network r, p_r, one hundred times for each value of p_r going from 0 to 1 with a pace of .005. We thus draw one hundred network structures (samples) of size equal to the real one (n=$1,319$) for each value of p_r, twenty thousand network structures in total. The desired replacement rate is assumed to be the same for all networks, i.e. $p_r = p$.

The first empirical issue that we face in our procedure relates to the relationship between the strength of peer effects and network density. Because peer effects varies with network density (see, e.g. Calvò et al. 2009), our numerical exercise needs to generate a constant number of links after replacement.

Let L_r, with cardinality l_r,be the set of existent links in network r and O_r, with cardinality o_r, the set of non existent links in the same network. The number of "possible changes" coherent with our constraint is $c_r = \min(n_r, o_r)$.

In words, for each network r we can exchange only a fraction of existent links with non existent links (and viceversa) if we want to maintain constant the total number of links in our network r of a given size (network density). The percentage of randomly replaced links p_r is thus calculated over the possibly interchangeable links (excluding overlapping), rather than over the total number of network links. The actual percentage will be $q_r = p_r \cdot c_r$.

The second empirical issue here is that this theoretical portion of links that we want to change may not correspond to a discrete number of links. For example, a replacement rate of 20% in a network with 7 possibly interchangeable links would imply that 1.4 links need to be changed. Do we swap one link (i.e. one existing into non existing and one non existing into existing at random) or two links (i.e. two couples)?

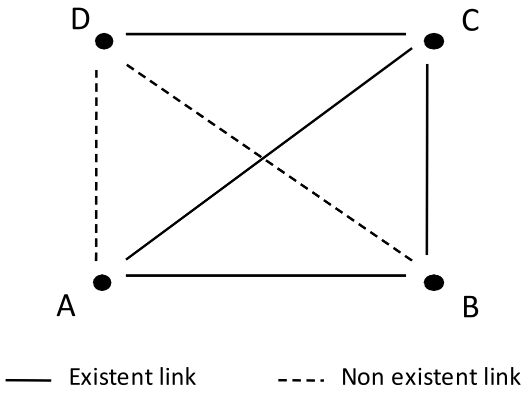


Figure 1: A simple example

We rigorously implement this decision rule as follows.

Let $p_r \in (0,1)$ be our desired replacement rate in network r. In order to obtain a number of changes as close as possible to the desired one, the actual number of changes s_r is:

$$s_r = \begin{cases} [q_r] & if \quad u > a \\ [q_r] + 1 & if \quad u < a \end{cases}$$

where $a = q_r \ - [q_r]$, and u is a random extraction from a variable uniformly distributed on $(0, 1)$.

Let us consider a simple example (Figure 1).

Suppose we have an undirected network composed of 4 nodes, {A, B, C, D}, and links {AB, AC, BC, CD}. In this situation, $l_r = 4$, $o_r = 2$ ({AD, BD}) and therefore $c_r = 2$. We can make at the maximum two changes within the set of "possible changes" {(AB AD), (AB BD), (AC AD), (AC BD), (BC AD), (BC BD),(CD AD), (CD BD)}. This means that we can extract randomly just two couples out of eight. Now suppose that our desired replacement rate p_r is 0.3 (30%), yielding to an actual replacement rate q_r of 0.6 (30% of 2). At this point our algorithm draws u. If $u < .6$ than $s_r = 1$ and we will replace one link (i.e. we extract at random one couple), otherwise nothing will happen. Clearly, given that 0.6 is closer to 1 than to 0, the probability to extract $u < .6$ is higher than the probability to extract $u < .6$, as it is desired.[21]

[21] This algorithm was been written in Matlab. The code is available upon request.

Simulated evidence

Our link replacement procedure enables us to simulate different network structures (**G** matrices in model (5)) that differ from the real one by a given (increasing) number of misspecified links. As mentioned before, for each percentage of randomly replaced links, we draw 100 network structures (samples) of size and network density equal to the real one. We then estimate model (5) replacing the real **G** matrix with the simulated ones in turn, so that in total we estimate model (5) twenty thousand times for each type of estimator (see Section 4.3).

[*Insert Figure* 2 *here*]

Figure 2 shows the results of our simulation experiment. The upper panel depicts the estimates of peer effects, whereas the lower panel shows the t-statistics with 90% confidence bands. It appears that the higher the percentage of misspecified links, the wider is the range of the peer effects estimates and the t-statistics fail more often to reject the hypothesis of no effects.

The crucial question for our purposes is what is the percentage of network structure misspecification over which peer effects are explained away.

[*Insert Figure* 3 *here*]

Figure 3 plots the averages of the estimates of peer effects for each replacement rate with 90% confidence bands. Standard errors has been calculated assuming drawing independence and taking into account the variation between estimates for each replacement rate. [22] We find that peer effects remain statistically significant up to a percentage of randomly replaced (interchangeable) links about 20%. This implies that even if we do not observe or we imprecisely observe a portion of each individual's social network, our results on the existence of peer effects hold. The portion of network topology that can be misspecified is not extremely small.

7 Policy implications

The presence of peer effects provides opportunities for policies aiming at improving social welfare (Hoxby, 2000). If one wants to implement an effective education policy, it needs to

[22] Specifically, the standard error at each replacement rate, say i, is computed as follows: $\sigma_i = \sqrt{W_i + B_i}$ where $W_i = \frac{1}{n}\sum_{j=1}^{n}\sigma_{ij}^2$, $B_i = \frac{1}{n}\sum_{j=1}^{n}(\phi_{ij} - \bar{\phi}_i)^2$, σ_{ij}^2 is the estimated variance of the jth estimator at the ith replacement rate, ϕ_{ij} is the jth estimate at the ith replacement rate and $\bar{\phi}_i$ is the mean across the n estimates. In this experiment $n = 100$.

internalize peer effects. For instance, education vouchers could lead to a more efficient human capital investment profile (see e.g., Epple and Romano 1998; Nechyba 2000). Policies such as school desegregation, busing, magnet schools, Moving to Opportunity programs[23] could also be effective if the government understands the magnitude and nature of peer effects in student outcomes. For example, if low ability students benefit from the presence of superior peers, while high ability students are not harmed by the presence of disadvantaged peers, then mixing students of different ability levels can generate social gains.

As noted by Manski (1993, 2000) and Moffitt (2001), it is, however, important to *separately* identify peer or endogenous effects from contextual or exogenous effects. This is because endogenous effects generate a *social multiplier* while contextual effect don't. In the context of education, this means that a special program targeting some individuals will have multiplier effects: the individual affected by the program will improve his/her grades and will influence the grades of his/her peers, which, in turn, will affect the grades of his/her peers, and so on. On the other hand, if only contextual effects are present, then there will be no social multiplier effects from any policy affecting only the "context" (for example, improving the quality of the teachers at school). Therefore, the identification of these two effects is of paramount importance for policy purposes. Another important policy issue in the estimation of social interactions is the separation between peer effects and confounding effects. Indeed, the formation of peer group is not random and individuals do select into groups of friends. It is therefore important to separate the endogenous peer effects from the correlated effects (Manski, 1993), i.e. the same educational outcomes may be due to common unobservable variables (such as, for example, the fact that individuals from the same network like bowling together) faced by individuals belonging to the same network rather than peer effects. This is also very important for education policies since, for example, if high grades are due to the fact that teenagers like to bowling together, then obviously the implications are very different than if it is due to peer effects.

One of the main aims of this paper was to clearly identify the peer effects from the contextual affects and from the correlated effects. For that, we first developed a theoretical model where all these effects were clearly separated. We then estimated the results of the model by using an econometric techniques, which utilizes the structure of the network as well as network fixed effects to identify each of these effects. We analyzed the impact of the friends' educational attainment on an individual's educational attainment where friendship was determined when this individual was a school adolescent while educational attainment was measured when the individual and his/her friends were adults. We find that there are

[23] See Lang (2007) for an overview of these policies in the U.S.

strong and persistent peer effects in education and that the relevant peers are the friends in grade 10-12. This suggests that individuals are more likely to adopt and pursue college studies if this choice is popular among their peers, especially in the last years at school. This could represent the effect of contagion and collective socialization and mean that any education policy targeting some specific individuals will have multiplier effects.

References

[1] Ballester, C., Calvó-Armengol, A. and Y. Zenou (2006), "Who's who in networks. Wanted: the key player," *Econometrica* 74, 1403-1417.

[2] Bayer, P., Ross, S.L. and G. Topa (2008), "Place of work and place of residence: Informal hiring networks and labor market outcomes," *Journal of Political Economy* 116, 1150-1196.

[3] Bekker, P. (1994), "Alternative approximations to the distributions of instrumental variable estimators," *Econometrica* 62, 657-681.

[4] Bekker, P. and J. van der Ploeg (2005), "Instrumental variable estimation based on grouped data," *Statistica Neerlandica* 59, 239-267.

[5] Benabou, R. (1993), "Workings of a city: location, education, and production", *Quarterly Journal of Economics* 108, 619-52.

[6] Bifulco, R., Fletcher, J.M. and S.L. Ross (2011), "The effect of classmate characteristics on post-secondary outcomes: Evidence from the Add Health," *American Economic Journal: Economic Policy*, forthcoming.

[7] Bramoullé, Y., Djebbari, H. and B. Fortin (2009), "Identification of peer effects through social networks," *Journal of Econometrics* 150, 41-55.

[8] Brooks-Gunn, J., Duncan, G.J., Kato Klebanov, P. and N. Sealand (1993), "Do neighborhoods influence child and adolescent development," *American Journal of Sociology* 99, 353-395.

[9] Calvó-Armengol, A., Patacchini, E. and Y. Zenou (2009), "Peer effects and social networks in education," *Review of Economic Studies* 76, 1239-1267.

[10] Card, D. and J. Rothstein (2007), "Racial segregation and the black-white test score gap," *Journal of Public Economics* 91, 2158-2184.

[11] Clark, A.E. and Y. Loheac (2007), "It wasn't me, it was them! Social influence in risky behavior by adolescents," *Journal of Health Economics* 26, 763-784.

[12] Cohen-Cole, E. (2006), "Multiple groups identification in the linear-in-means model," *Economics Letters* 92, 157-162.

[13] Corcoran, M., Gordon, R., Laren, D. and G. Solon (1992), "The association between men's economic status and their family and community origins" *Journalf of Human Resources* 27, 575-601.

[14] Cutler, D. M. and E. L. Glaeser (1997), "Are ghettos good or bad?" *Quarterly Journal of Economics* 112, 827-872.

[15] Datcher, L. (1982), "Effects of communuty and family background on achievement," *Review of Economics and Statistics* 64, 32-41.

[16] De Giorgi, G., Pellizzari, M. and S. Redaelli (2010), "Identification of social interactions through partially overlapping peer groups," *American Economic Journal: Applied Economics* 2, 241-275.

[17] De Bartolome, C.A.M. (1990), "Equilibrium and inefficiency in a community model with peer group effects," *Journal of Political Economy* 98, 110-133.

[18] Durlauf, S.E. (2004), "Neighborhood effects," In: J.V. Henderson and J-F. Thisse (Eds.), *Handbook of Regional and Urban Economics Vol. 4*, Amsterdam: Elsevier Science, pp. 2173-2242.

[19] Epple, D. and R.E. Romano (1998), "Competition between private and public schools: Vouchers and peer group effects," *American Economic Review* 88, 33-62.

[20] Evans, W.N, Oates, W.E and R.M. Schwab (1992), "Measuring peer group effects: A study of teenage behavior," *Journal of Political Economy* 100, 966-991.

[21] Gould, E.D., Lavy, V. and D. Paserman (2011), "Sixty years after the magic carpet ride: The long-run effect of the early childhood environment on social and economic outcomes," *Review of Ecoconomic Studies*, forthcoming.

[22] Goux, D. and E. Maurin (2007), "Close neighbours matter: Neighbourhood effects on early performance at school," *Economic Journal* 117, 1193-1215.

[23] Goyal, S. (2007), *Connections: An Introduction to the Economics of Networks*, Princeton: Princeton University Press.

[24] Hansen, C., Hausman, J. and W. Newey (2008), "Estimation with many instrumental variables," *Journal of Business and Economic Statistics* 26, 398-422.

[25] Hoxby, C. (2000), "Peer effects in the classroom, learning from gender and race variation," NBER Working Paper no. 7867.

[26] Ioannides, Y.M. (2011), "Neighborhood effects and housing," In: J. Benhabib, A. Bisin, and M.O. Jackson (Eds.), *Handbook of Social Economics*, Amsterdam: Elsevier Science, forthcoming.

[27] Ioannides, Y.M. and G. Topa (2010), "Neighborhood effects: Accomplishments and looking beyond them," *Journal of Regional Science* 50, 343-362.

[28] Jackson, M.O. (2008), *Social and Economic Networks*, Princeton: Princeton University Press.

[29] Katz, L.F., Kling, J.R. and J.B. Liebman (2001), "Moving to opportunity in Boston: Early results of a randomized mobility experiment," *Quarterly Journal of Economics* 116, 607-654.

[30] Lang, K. (2007), *Poverty and Discrimination*, Princeton: Princeton University Press.

[31] Laschever, R. (2009), "The doughboys networks: Social interactions and labor market outcomes of World War I veterans," Unpusblished manuscript, University of Illinois at Urbana-Champaign.

[32] Lee, L-F. (2002), "Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models," *Econometric Theory* 18, 252-277.

[33] Lee, L-F. (2003), "Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances," *Econometric Reviews* 22, 307-335.

[34] Lee, L-F. (2007), "Identification and estimation of econometric models with group interactions, contextual factors and fixed effects," *Journal of Econometrics* 140, 333-374.

[35] Lee, L-F., Liu, X. and X. Lin (2010), "Specification and estimation of social interaction models with network structures," *Econometrics Journal* 13, 145-176.

[36] Lin, X. (2010), "Identifying peer effects in student academic achievement by a spatial autoregressive model with group unobservables," *Journal of Labor Economics* 28, 825-860.

[37] Liu, X. and L-F. Lee (2010), "GMM estimation of social interaction models with centrality," *Journal of Econometrics* 159, 99-115.

[38] Liu, X., Patacchini, E., Zenou, Y. and L-F. Lee (2011), "Criminal networks: Who is the key player?" CEPR Discussion Paper No. 8185.

[39] Manski, C.F. (1993), "Identification of endogenous effects: The reflection problem," *Review of Economic Studies* 60, 531-542.

[40] Manski, C.F. (2000), "Economic analysis of social interactions," *Journal of Economic Perspectives* 14, 115-136.

[41] Moffitt, R. (2001), "Policy interventions low-level equilibria, and social interactions," In: S. Durlauf and P. Young (Eds.), *Social Dynamics*, Cambridge, MA: MIT Press, pp. 45-82.

[42] Nechyba, T.J. (2000), "Mobility, targeting, and private-school vouchers," *American Economic Review* 90, 130-46.

[43] Patacchini, E. and Y. Zenou (2011), "Intergenerational education transmission: Neighborhood quality and/or parents' involvement?", *Journal of Regional Science*, forthcoming.

[44] Patacchini, E. and Y. Zenou (2012), "Juvenile delinquency and conformism," *Journal of Law, Economic, and Organization* forthcoming.

[45] Sacerdote, B. (2001), "Peer effects with random assignment: Results from Dartmouth roomates," *Quarterly Journal of Economics* 116, 681-704.

[46] Wasserman, S., and K. Faust (1994), *Social Network Analysis. Methods and Applications*, Cambridge: Cambridge University Press.

[47] Weinberg, B.A. (2004), "Testing the spatial mismatch hypothesis using inter-city variations in industrial composition," *Regional Science and Urban Economics* 34, 505-532.

[48] Weinberg, B.A., P.B. Reagan, and J.J. Yankow (2004), “Do neighborhoods affect work behavior? Evidence from the NLSY 79,” *Journal of Labor Economics* 22, 891-924.

[49] Zax, J.S. and D.I. Rees (2002), “IQ, academic performance, environment, and earnings,” *Review of Economics and Statistics* 84, 600-616.

[50] Zimmerman, D. (2003), “Peer effects in academic outcomes: Evidence from a natural experiment,” *Review of Economics and Statistics* 9-23.

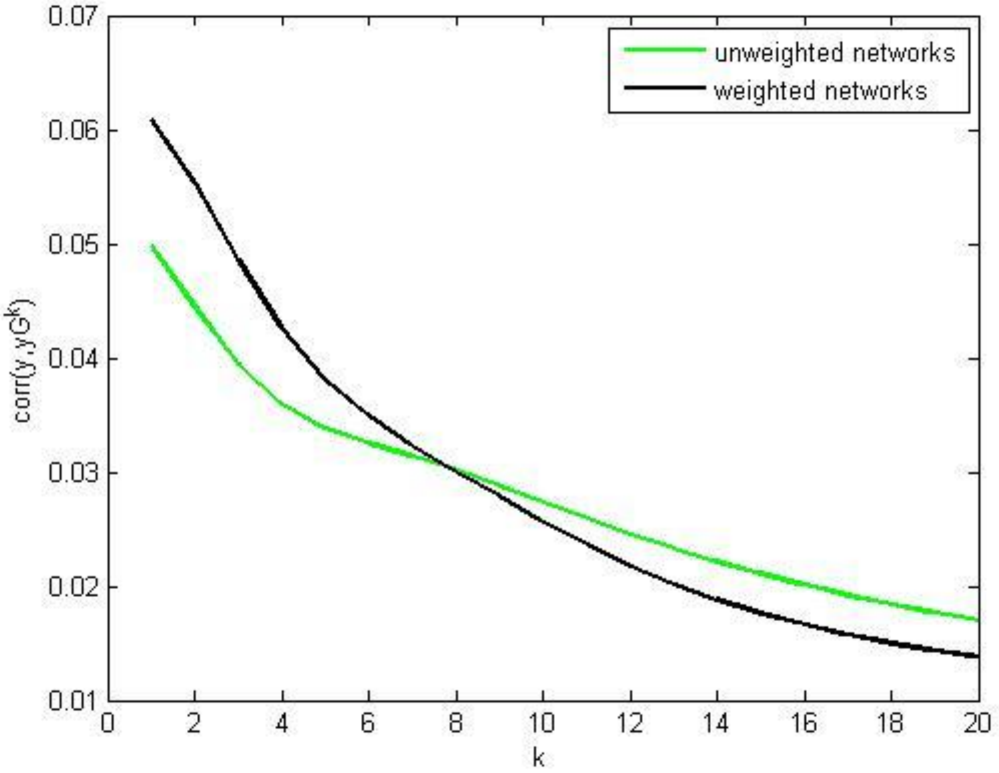
Appendix 1: Data appendix

Table A.1: Description of Data (1,319 individuals, 138 networks)

	Variable definition	Mean	St.dev	Min	Max
Wave IV (aged 24 - 32)					
Years in Education		16.31	3.19	9	26
Years in Education of peers	Aggregate value of years in education over nominated direct friends.	40.24	27.79	9	187
Married	Dummy variable taking value one if the respondent is married.	0.45	0.50	0	1
Age	Respondent's age	28.5	1.72	25	33
Son or Daughter	Dummy variable taking value one if the respondent has a son or daughter.	0.47	0.50	0	1
Religion practice	Response to the question: "How often have you attended religious services in the past 12 months?", coded as 0= never, 1= a few times, 2= once a month, 3= 2 or 3 times a month, 4=once a week, 5=more than once a week.	1.72	1.63	0	5
Wave I (grade 7 - 12)					
Individual socio-demographic variables					
Female	Dummy variable taking value one if the respondent is female.	0.53	0.50	0	1
Black or African American	Race dummies. "White" is the reference group.	0.17	0.37	0	1
Other races	Race dummies. "White" is the reference group.	0.05	0.23	0	1
Student grade	Grade of student in the current year.	9.14	1.68	7	12
Religion practice	Response to the question: "In the past 12 months, how often did you attend religious services", coded as 4= never, 3= less than once a month, 2= once a month or more, but less than once a week, 1= once a week or more. Coded as 5 if the previous is skipped because of response "none" to the question: "What is your religion?"	2.16	1.38	1	5
Mathematics score A	Dummies for scores in mathematics at the most recent grading period, coded (A, B, C, D or lower, missing).	0.29	0.45	0	1
Mathematics score B	Dummies for scores in mathematics at the most recent grading period, coded (A, B, C, D or lower, missing).	0.34	0.47	0	1
Mathematics score C	Dummies for scores in mathematics at the most recent grading period, coded (A, B, C, D or lower, missing).	0.20	0.40	0	1
Mathematics score D or lower	Dummies for scores in mathematics at the most recent grading period, coded (A, B, C, D or lower, missing).	0.11	0.32	0	1
Mathematics score missing	Dummies for scores in mathematics at the most recent grading period, coded (A, B, C, D or lower, missing).	0.05	0.21	0	1
Self esteem	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.	4.01	1.08	1	6
Physical development	Response to the question: "How advanced is your physical development compared to other boys/girls your age", coded as 1= I look younger than most, 2= I look younger than some, 3= I look about average, 4= I look older than some, 5= I look older than most	3.34	1.11	1	5
Family background variables					
Household size	Number of people living in the household.	4.39	1.35	2	11
Two married parent family	Dummy taking value one if the respondent lives in a household with two parents (both biological and non biological) that are married.	0.73	0.44	0	1
Parent education	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. We consider only the education of the father if both parents are in the household.	3.18	1.08	0	5

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	0.11	0.31	0	1
Parent occupation professional/technical	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	0.20	0.40	0	1
Parent occupation office or sales worker	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	0.11	0.31	0	1
Parent occupation manual	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	0.31	0.46	0	1
Parent occupation military or security	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	0.02	0.15	0	1
Parent occupation farm or fishery	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	0.03	0.16	0	1
Parent occupation other	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	0.13	0.34	0	1
Protective factors					
School attachment	Response to the question: "You feel like you are part of your school coded as 1= strongly agree, 2= agree, 3=neither agree nor disagree, 4= disagree, 5= strongly disagree.	1.90	0.92	1	5
Relationship with teachers	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	0.89	0.91	0	4
Social inclusion	Response to the question: "How much do you feel that adults care about you, coded as 5= very much, 4= quite a bit, 3= somewhat, 2= very little, 1= not at all	4.47	0.74	1	5
Parental care	Dummy taking value one if the respondent reports that the (biological or non-biological) parent that is living with her/him or at least one of the parents if both are in the household cares very much about her/him	0.92	0.28	0	1
Residential neighborhood variables					
Residential building quality	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.	1.52	0.79	1	4
Residential area suburban	Residential area type dummies: interviewer's description of the immediate area or street (one block, both sides) where the respondent lives. “Rural area” is the reference group.	0.30	0.46	0	1
Residential area urban - residential only	Residential area type dummies: interviewer's description of the immediate area or street (one block, both sides) where the respondent lives. “Rural area” is the reference group.	0.23	0.42	0	1
Residential area other type	Residential area type dummies: interviewer's description of the immediate area or street (one block, both sides) where the respondent lives. “Rural area” is the reference group.	0.02	0.12	0	1

Figure A.1: Correlations between own education and peers' education



Notes: Network links are defined using the choices made (out-degree). The plotted correlations are statistically significant at the 1% level.

Table 1: Estimation Results –peer effects-
-unweighted networks-

	Total IV	Lagged IV
2SLS finite IVs	0.011** (0.005)	0.015** (0.006)
2SLS many IVs	0.008* (0.005)	0.010** (0.005)
Bias-corrected 2SLS	0.009** (0.005)	0.011** (0.005)
Individual socio-demographic variables	yes	yes
Family background variables	yes	yes
Protective factors	yes	yes
Residential neighborhood variables	yes	yes
Contextual effects	yes	yes
Network fixed effects	yes	yes
1,319 individuals over 138 networks.		

Notes: Estimation has been performed using Matlab. Standard errors are reported in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 2: Estimation Results –peer effects-
-weighted networks-

	Total IV	Lagged IV
2SLS finite IVs	0.020** (0.009)	0.027** (0.011)
2SLS many IVs	0.013* (0.008)	0.016** (0.008)
bias-corrected 2SLS	0.014* (0.008)	0.018** (0.008)
Individual socio-demographic variables	yes	yes
Family background variables	yes	yes
Protective factors	yes	yes
Residential neighborhood variables	yes	yes
Contextual effects	yes	yes
Network fixed effects	yes	yes
1,319 individuals over 138 networks.		

Notes: Estimation has been performed using Matlab. Standard errors are reported in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 3: Estimation Results –peer effects-
Unweighted networks
Grade 7-9

	Total IV	Lagged IV
2SLS finite IVs	0.007 (0.006)	0.010 (0.007)
2SLS many IVs	0.008 (0.006)	0.009 (0.006)
bias-corrected 2SLS	0.009 (0.006)	0.011* (0.006)
Individual socio-demographic variables	yes	yes
Family background variables	yes	yes
Protective factors	yes	yes
Residential neighborhood variables	yes	yes
Contextual effects	yes	yes
Network fixed effects	yes	yes
713 individuals over 80 networks		

Notes: Estimation has been performed using Matlab. Standard errors are reported in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 4: Estimation Results –peer effects-
weighted networks

Grade 7-9

	Total IV	Lagged IV
2SLS finite IVs	0.010 (0.010)	0.011 (0.012)
2SLS many IVs	0.008 (0.009)	0.009 (0.009)
bias-corrected 2SLS	0.010 (0.009)	0.011 (0.009)
Individual socio-demographic variables	yes	yes
Family background variables	yes	yes
Protective factors	yes	yes
Residential neighborhood variables	yes	yes
Contextual effects	yes	yes
Network fixed effects	yes	yes
713 individuals over 80 networks		

Notes: Estimation has been performed using Matlab. Standard errors are reported in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 5: Estimation Results –peer effects-
Unweighted networks
Grade 10-12

	Total IV	Lagged IV
2SLS finite IVs	0.021* (0.011)	0.024** (0.011)
2SLS many IVs	0.016* (0.009)	0.016* (0.010)
bias-corrected 2SLS	0.018* (0.009)	0.018* (0.010)
Individual socio-demographic variables	yes	yes
Family background variables	yes	yes
Protective factors	yes	yes
Residential neighborhood variables	yes	yes
Contextual effects	yes	yes
Network fixed effects	yes	yes
492 individuals over 55 networks		

Notes: Estimation has been performed using Matlab. Standard errors are reported in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

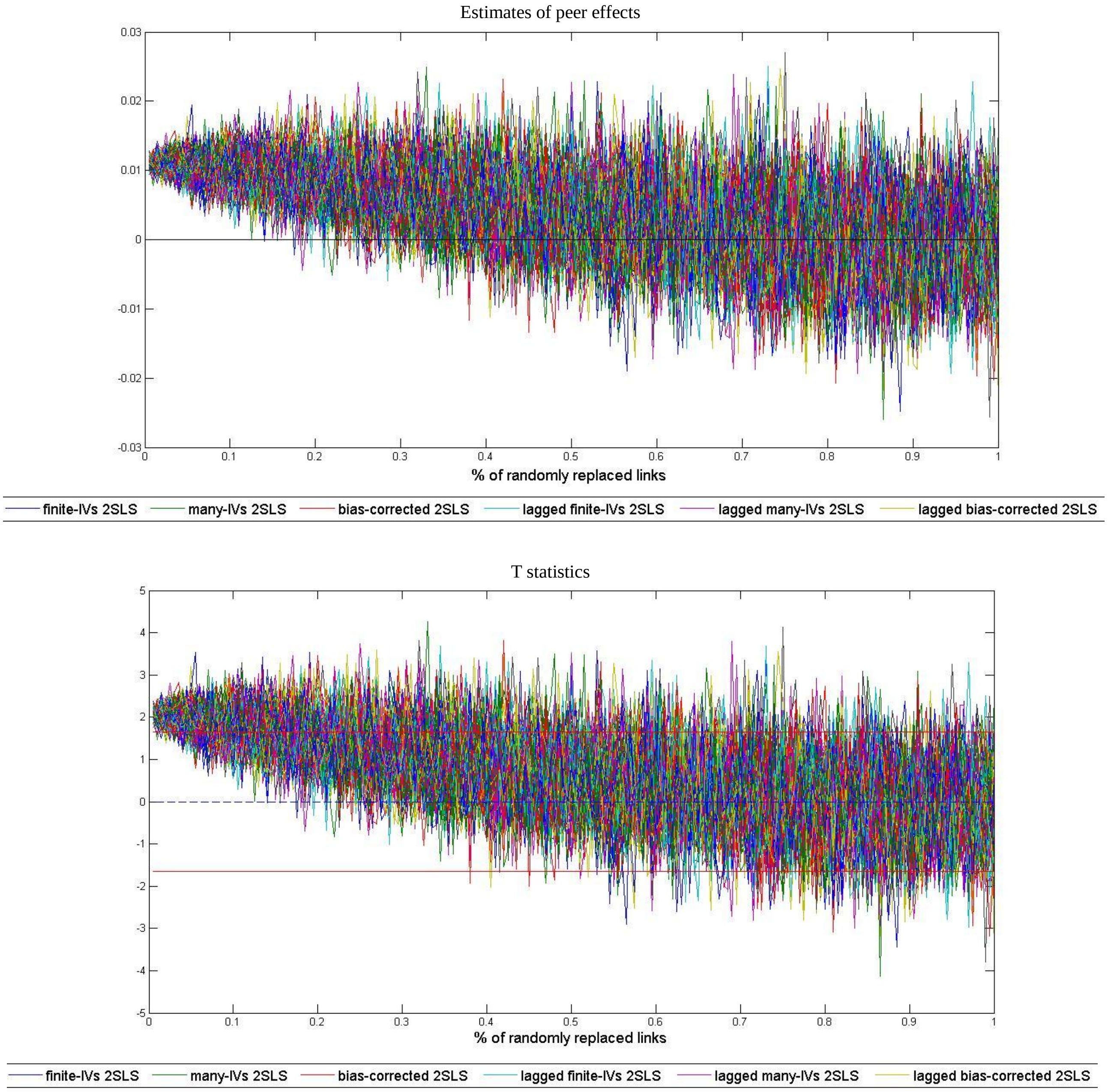
Table 6: Estimation Results –peer effects-
weighted networks

Grade 10-12

	Total IV	Lagged IV
2SLS finite IVs	0.039** (0.018)	0.044** (0.019)
2SLS many IVs	0.029** (0.014)	0.031** (0.015)
bias-corrected 2SLS	0.033** (0.009)	0.035** (0.015)
Individual socio-demographic variables	yes	yes
Family background variables	yes	yes
Protective factors	yes	yes
Residential neighborhood variables	yes	yes
Contextual effects	yes	yes
Network fixed effects	yes	yes
492 individuals over 55 networks		

Notes: Estimation has been performed using Matlab. Standard errors are reported in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Figure 2: Misspecification of network topology
Numerical simulation



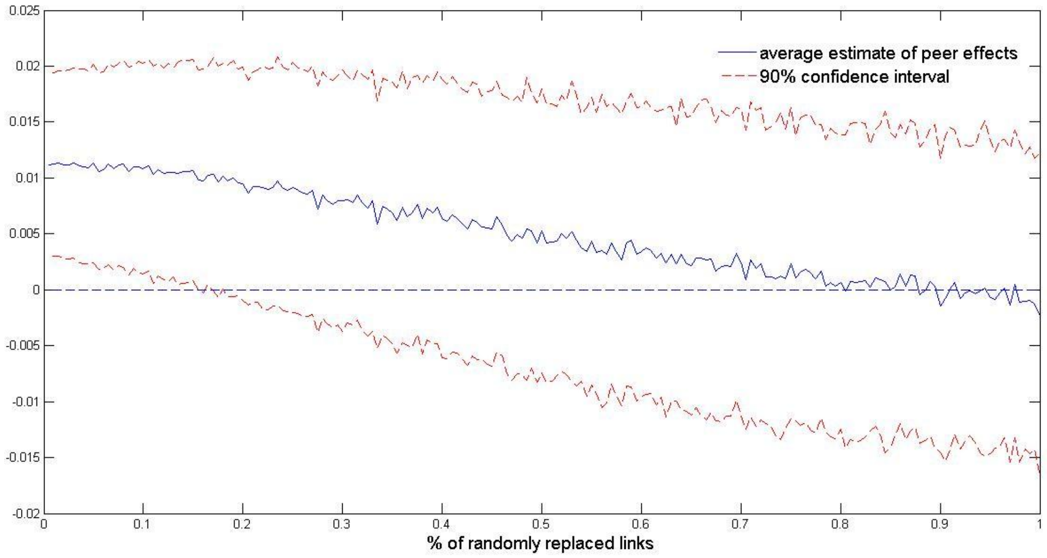
Estimates of peer effects

T statistics

Notes: For each percentage of randomly replaced links, we draw 100 samples of size and network density equal to the real one and show the estimated peer effects and t-statistics (model specification (5))

Figure 3: Simulation experiment
Summarising the evidence

Lagged bias-corrected 2SLS estimates of peer effects



Notes: For each percentage of randomly replaced links, we average the estimates of peer effects across the drawn samples. Theconfidence bands are based on the derived standard errors, accounting for within and between sample variation and assuming drawing independence.