Peer effects in bedtime decisions among adolescents: a social network model with sampled data

XIAODONG LIU[†], ELEONORA PATACCHINI[‡] AND EDOARDO RAINONE[§]

[†]*Department of Economics, University of Colorado, Boulder, CO 80309, USA.*
E-mail: xiaodong.liu@colorado.edu

[‡]*Department of Economics, Cornell University, 426 Uris Hall, Ithaca, NY 14853, USA.*
E-mail: ep454@cornell.edu

[§]*Bank of Italy, via Nazionale 91, 00184, Rome, Italy.*
E-mail: edoardo.rainone@bancaditalia.it

First version received: May 2016; final version accepted: September 2016

Summary Using unique information on a representative sample of US teenagers, we investigate peer effects in adolescent bedtime decisions. We extend the nonlinear least-squares estimator for spatial autoregressive models to estimate network models with network fixed effects and sampled observations on the dependent variable. We show the extent to which neglecting the sampling issue yields misleading inferential results. When accounting for sampling, we find that, besides the individual, family and peer characteristics, the bedtime decisions of peers help to shape one's own bedtime decision.

Keywords: *Missing data, Nonlinear least-squares, Social interactions, Spatial autoregressive model.*

1. INTRODUCTION

Sleep that knits up the ravell'd sleeve of care,
The death of each day's life, sore labour's bath,
Balm of hurt minds, great nature's second course,
Chief nourisher in life's feast.
(Shakespeare, *Macbeth*, Act 2, Scene 2)

Nearly a third of a person's life is spent in slumber. Yet, sleeping behaviour has received relatively little attention in economics.¹ In particular, there is virtually no study examining how sleeping behaviour is affected by social incentives. While sleep is primarily a function of the body's internal biological clock (circadian rhythm), individual choice also plays a big role in determining the timing and duration of sleep. In many circumstances, individual choice in a

¹ The interest in the topic of sleep has been rapidly growing in the last few years. Most notably, there is an ongoing lab-in-the-field experiment in India by Rao et al. (2015), which focuses on how sleep deprivation might have tremendous consequences for the poor.

certain activity cannot be adequately explained by personal characteristics and the intrinsic utility derived from this activity. Rather, its rationale can be found in how this activity is valued by peers. There is indeed strong evidence that peer effects play an important role in shaping individual choices in many activities.² Hence, peer effects might also be important for understanding sleeping behaviour, which is the residual activity.³

In this paper, we exploit the unique information contained in the National Longitudinal Survey of Adolescent Health (Add Health) to study bedtime patterns among adolescents in the United States. Sleeping behaviour during teenage years is of particular interest because of its effect on human capital formation. Research outside of economics suggests that lack of sleep reduces school attendance, increases tardiness, and lowers the grades of adolescent students.⁴ Furthermore, lack of sleep in youth is correlated with health and behavioural problems such as moodiness, depression, difficulty controlling behaviour, and increased frustration, all of which make learning in school difficult (National Sleep Foundation, 2006). Sleep also affects productivity on the job, which in some cases represents a public safety concern.

The Add Health data contain unique information on friendship relationships among a nationally representative sample of students in grades 7–12 in the United States, together with basic information on individual, family, neighbourhood and school characteristics (the in-school survey). The survey also includes a questionnaire (the in-home survey) administered to a random subsample of the students participating the in-school survey, which collects information on more sensitive topics (health issues, crime, drug use, sexual behaviour, etc.) including bedtime on weekdays during the school year. However, the use of this additional information from the in-home survey comes at a cost. The in-home sampling scheme results in missing observations on the behaviour of students who were not sampled, and thus induces measurement error in the endogenous peer effect regressor formulated as the average behaviour of a student's friends. As a result, the existing estimation methods for social network models – see, e.g. Bramoullé et al. (2009) and Lee et al. (2010) – are not generally valid.⁵

Recently, social network studies have attracted a great deal of attention. Network models are widely used to represent relational information among interacting units and the implications of these relations. Most inference for social network models assumes that all the possible links are observed and that the relevant information of all individuals (nodes) is available. This is clearly not true in practice, as most network data are collected through surveys. In a recent paper, Sojourner (2013) has considered a linear-in-means social interaction model with missing observations on covariates. He shows that random assignment of individuals to peer groups can help to overcome this missing data problem. However, Chandrasekhar and Lewis (2011) consider the estimation of network models with missing observations on network links. They propose a set of analytical corrections for commonly used network statistics and a two-step estimation procedure using graphical reconstruction. In this paper, we focus on the case in which all the network links and the covariates can be observed but only the dependent variable of the sampled nodes can be observed.

² The integration of social interaction models within economic theory is an active and interesting area of research; see Benhabib et al. (Eds.) (2011), *Handbook of Social Economics*.

³ Biddle and Hamermesh (1990) study the demand for sleep in this perspective without social incentives.

⁴ Wolfson and Carskadon (2003) is an excellent summary of the medical research in this area.

⁵ This issue is typically neglected in most empirical papers using the information on friends together with the in-home survey in the Add Health dataset.

The social network model considered in this paper has the specification of a spatial autoregressive (SAR) process in which the average behaviour of the peers is formulated as spatial lags. Kelejian and Prucha (2010) consider the estimation of the SAR model with missing observations on the dependent variable and covariates. They suggest two-stage least-squares (2SLS) estimators that are based on a subset of the cross-sectional units for which the dependent variable and covariates are observed, and the spatial lags are either completely observed or partially observed with an asymptotically negligible measurement error. More recently, Wang and Lee have considered the estimation of the SAR model with missing observations on the dependent variable for cross-sectional data (Wang and Lee, 2013a) and for random effect panel data (Wang and Lee, 2013b). They propose the generalized method of moments (GMM) estimator, the nonlinear least-squares (NLS) estimator and the 2SLS estimator with imputation. They show that the three estimators are consistent and robust against unknown heteroscedasticity.

In this paper, we extend the NLS estimator in Wang and Lee (2013a) to estimate social network models with network fixed effects and we provide the first empirical application of this method. We show that the conventional 2SLS estimator is inconsistent without accounting for sampling. In our empirical study, the 2SLS estimator fails to detect the presence of (endogenous) peer effects. When sampling is taken into account, we instead find that the sleeping behaviour of friends is important in shaping one's own sleeping behaviour.⁶ The contributions of this paper are summarized as follows.

- (a) We extend the NLS estimator of Wang and Lee (2013a) to estimate network models with network fixed effects. The proposed NLS estimator is easy to implement in applied works.
- (b) We conduct Monte Carlo experiments to investigate the finite sample performance of the proposed NLS estimator and we evaluate the bias of the conventional 2SLS estimator when estimating a social network model with sampled observations on the dependent variable.
- (c) We provide the first empirical application of this method using a unique dataset of friendship networks, and we find that adolescents respond to the sleeping behaviour of their peer group, holding constant other factors.

The rest of the paper is organized as follows. We begin by describing our data in Section 2. In Section 3, we present the network model, together with the identification and estimation strategy. We discuss our estimation results in Section 4, whereas in Section 5 we provide some robustness checks. We conclude in Section 6.

2. DATA AND DESCRIPTIVE EVIDENCE

Our data source is the Add Health dataset. This dataset has been designed to study the impact of the social environment (i.e. friends, family, neighbourhood and school) on adolescents' behaviour 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 1994–1995. Every student attending the sampled schools on the interview day is asked to compile a questionnaire (the in-school survey) containing questions on respondents' demographic and behavioural characteristics, education, family background and friendship. Most notably, students are asked to identify their

⁶ The validity of the NLS method relies on the exogeneity of the network structure. Based on diagnostic tests, we argue that network structure is exogenous conditioning on individual, family and peer characteristics, together with network-component fixed effects.

best friends from a school roster – up to five males and five females. However, the limit in the number of nominations is not binding (not even by gender): 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% a list of five females. On average, they declare that they have 4.35 friends with a small dispersion around this mean value (standard deviation equal to 1.41), and in the large majority of cases (more than 90%) the nominated best friends are in the same school. Hence, it is possible to reconstruct the entire geometry of the friendship networks within each school. In addition, by matching the identification numbers of the friendship nominations to respondents' identification numbers, it is possible to obtain information on the characteristics of nominated friends. This sample contains information on roughly 90,000 students. These features make these data almost unique. It is extremely rare to have information on the universe of network contacts (here school friends), together with their detailed characteristics.⁷ The survey design also includes a longer questionnaire (the in-home survey) containing questions related to more sensitive individual and household information, which is administered to a subset of students. We use the core sample of the in-home survey, which provides information on a random subset of adolescents, about 12,000 individuals.⁸ The in-home questionnaire contains detailed information about the timing and duration of sleep. The questions have been slightly reformulated over time to measure sleeping patterns more precisely. Indeed, the students participating the in-home survey are interviewed again one year later, in 1995–1996 (wave II).⁹ We derive the information on sleeping patterns by using the wave II question: 'During the school year, what time do you usually go to bed on week nights?'.^{10,11}

Our final sample consists of about 8,000 individuals in 33 networks. The large reduction in sample size with respect to the original sample is mainly due to the network construction procedure – roughly 20% of the students do not nominate any friends and another 20% cannot be correctly linked.¹² In addition, we focus on networks with sizes between 10 and 400 individuals, as peer effects may operate very differently in very small and very large networks; see, e.g. Calvó-Armengol et al. (2009).

Figure 1 plots the empirical distribution of bedtimes, where the kernel is the Epanechnikov kernel and the bandwidth is 40.429. We report the distribution of students by the time they go to sleep. The graph shows a notable dispersion around the mean bedtime value (mean equal to 10:30 p.m. and standard deviation equal to 59 minutes). About 50% of the students go to bed between 10 p.m. and 11.30 p.m.

⁷ The information on social network contacts collected in other existing surveys is about 'ego-networks', i.e. the respondent is asked to name a few personal contacts and provides (self-reported) information about an extremely limited number of their characteristics.

⁸ The core sample contains roughly the 60% of the individuals interviewed in the in-home survey (which consists of about 20,000 individuals). The difference is because of the fact that, in the in-home sampling design, some types of individuals are oversampled.

⁹ Those subjects are also interviewed again in 2001–2002 (wave III) and again in 2007–2008 (wave IV). For the purpose of this paper, we do not use this longitudinal information. The friendship nominations were only collected when the students were at school (i.e. in waves I and II).

¹⁰ The questions on sleep behaviour formulated in wave I do not differentiate between the school period and summer time. The same issue also applies to the other question on sleeping behaviour in wave II: 'How many hours of sleep do you usually get?'. Finally, a third question is available: 'Do you usually get enough sleep?', which measures a subjective perception, and thus increases measurement errors. In addition, the answers to both questions are not continuous variables, as required by the NLS estimation.

¹¹ We rescaled each hour in 100 units, so for instance half an hour is transformed to 50 units. We dropped individuals declaring that they go to sleep before 5 p.m. and after 6 a.m.

¹² This is common when working with Add Health data. However, the representativeness of the sample is preserved.

Figure 1. Kernel density estimate of bedtime.

A detailed description of the explanatory variables (covariates) and sample summary statistics can be found in Appendix A. Among the individuals selected in our sample, 53% are female, 12% are black and 9% are Asian. The average student is in grade 9, has a good school performance, has parents with more than a high-school education and lives in a family with four people. The variables on extracurricular activities show that more than 20% of the students play baseball or softball, almost 25% play basketball, roughly 10% play soccer and another 10% play volleyball. They are also quite involved in activities other than sports.

3. ECONOMETRIC ANALYSIS

Our aim is to assess the empirical relationship between individual bedtime decisions and the bedtime decisions of peers using the unique information provided by the Add Health data. This exercise requires facing the traditional challenges in identifying social interaction effects, while also overcoming a further (and so far neglected) issue stemming from the sampling design of the Add Health survey. We present the network model in Section 3.1, and we discuss the estimation of network models with sampling on the dependent variable in Section 3.2.

3.1. *The network model*

Consider a population of n individuals partitioned into $\bar{r}$ networks.¹³ For the n_r individuals in the r th network, their connections with each other are represented by an $n_r \times n_r$ adjacency matrix

¹³ In this paper, a network (or a network component) contains a set of individuals (nodes) such that any two individuals in the same network are connected to each other (directly or indirectly) and no individuals from different networks are connected.

$G_r^* = [g_{ij,r}^*]$ where $g_{ij,r}^* = 1$ if individuals i and j are friends and $g_{ij,r}^* = 0$ otherwise.¹⁴ Let $G_r = [g_{ij,r}]$ be the row-normalized G_r^* such that $g_{ij,r} = g_{ij,r}^* / \sum_{k=1}^{n_r} g_{ik,r}^*$.

Given the network adjacency matrix G_r , we assume $y_{i,r}$, the bedtime of individual i in network r , is given by the following network model:

$$y_{i,r} = \phi \sum_{j=1}^{n_r} g_{ij,r} y_{j,r} + \sum_{k=1}^p x_{ik,r} \beta_k + \sum_{k=1}^p \left(\sum_{j=1}^{n_r} g_{ij,r} x_{jk,r} \gamma_k \right) + \eta_r + \epsilon_{i,r}. \quad (3.1)$$

In this model, $\sum_{j=1}^{n_r} g_{ij,r} y_{j,r}$ is the average bedtime of i 's direct friends with its coefficient ϕ representing the (endogenous) peer effect, and $x_{ik,r}$, for $k = 1, \dots, p$, are exogenous explanatory variables. For $k = 1, \dots, p$, $\sum_{j=1}^{n_r} g_{ij,r} x_{jk,r}$ is the average value of the k th explanatory variable taking over i 's direct friends with its coefficient γ_k representing the (exogenous) contextual effect. Finally, individuals in the same network tend to behave similarly because they face a common environment. This is usually referred to as the correlated effect – see Manski (1993) – and is captured by the network-specific parameter η_r . The error term $\epsilon_{i,r}$ is i.i.d. with zero mean and finite variance σ^2 .¹⁵ In Appendix B, we provide a microfoundation for this empirical model.

Let $x_{i,r} = (x_{i1,r}, \dots, x_{ip,r})'$, $\beta = (\beta_1, \dots, \beta_p)'$ and $\gamma = (\gamma_1, \dots, \gamma_p)'$. In matrix form, (3.1) can be rewritten as

$$Y_r = \phi G_r Y_r + X_r \beta + G_r X_r \gamma + \eta_r l_{n_r} + \epsilon_r, \quad (3.2)$$

where $Y_r = (y_{1,r}, \dots, y_{n_r,r})'$, $X_r = (x_{1,r}, \dots, x_{n_r,r})'$, $\epsilon_r = (\epsilon_{1,r}, \dots, \epsilon_{n_r,r})'$ and l_{n_r} is an $n_r \times 1$ vector of ones.

Let $\text{diag}\{A_j\}_{j=1}^m$ denote a generalized diagonal block matrix with the diagonal blocks being A_j s, where A_j may or may not be a square matrix. Then, for all $\bar{r}$ networks, we can stack the data such that (3.2) becomes

$$Y = \phi GY + X\beta + GX\gamma + L\eta + \epsilon, \quad (3.3)$$

where $Y = (Y_1', \dots, Y_{\bar{r}}')'$, $G = \text{diag}\{G_r\}_{r=1}^{\bar{r}}$, $X = (X_1', \dots, X_{\bar{r}}')'$, $L = \text{diag}\{l_{n_r}\}_{r=1}^{\bar{r}}$, $\eta = (\eta_1, \dots, \eta_{\bar{r}})'$ and $\epsilon = (\epsilon_1', \dots, \epsilon_{\bar{r}}')'$.

The identification and estimation of endogenous peer effects, contextual effects and correlated effects have been the main interests of social network models. The conventional identification and estimation strategy in the literature – see, e.g. Bramoullé et al. (2009), Lee et al. (2010) and Lin (2010) – relies on the assumption that $E[\epsilon_r | G_r, X_r, \eta_r] = 0$.¹⁶ Based on this assumption, Bramoullé et al. (2009) show that if intransitivities exist in networks so that I_n, G, G^2, G^3 are linearly independent, then model (3.2) is identified. For estimation, we first eliminate the incidental parameters η using a within-transformation projector $J = \text{diag}\{J_r\}_{r=1}^{\bar{r}}$, where $J_r = I_{n_r} - (1/n_r)l_{n_r}l_{n_r}'$. As $JL = 0$, premultiplying (3.3) by J , we have

$$JY = \phi JGY + JX\beta + JGX\gamma + J\epsilon.$$

¹⁴ For ease of presentation, we focus on the case in which the connections are undirected and no agent is isolated, so that G_r^* is symmetric and $\sum_{j=1}^{n_r} g_{ij,r}^* \neq 0$ for all i . The result of the paper holds for a directed network with an asymmetric G_r^* .

¹⁵ For expositional purposes, we assume the error term is homoscedastic. The proposed NLS estimator remains consistent when the error term is heteroscedastic.

¹⁶ We investigate the validity of this assumption for this empirical study in Section 5.

Let $Z = (GY, X, GX)$ and $\theta = (\phi, \beta', \gamma')'$. For the instrumental variable (IV) matrix $Q = (X, GX, G^2X)$, the 2SLS estimator is given by

$$\hat{\theta}_{2sls} = (\hat{Z}'JZ)^{-1}\hat{Z}'JY, \quad (3.4)$$

where $J\hat{Z} = JQ(Q'JQ)^{-1}Q'JZ$ is the predicted JZ from the first-stage regression.

In the following section, we focus on the sampling issue in the network model, which has been largely ignored in the literature.

3.2. Estimation of peer effects with sampling

In our study and many others, the analysis of the network model (3.1) has been made possible by the use of a unique database on friendship networks from the Add Health data; see, e.g. Patacchini and Zenou (2008) and Lin (2010), and references therein. As we explained in Section 2, every student attending the sampled schools on the interview day is asked to complete the in-school survey containing questions on respondents' demographic and behavioural characteristics, education, family background and friendship. Thus, we can observe the covariates of almost all students together with their friendship links in the network. However, as some more sensitive individual information (including information on sleeping behaviour) is in the in-home survey, we only observe the dependent variable for a subsample of the students.^{17,18}

Without loss of generality, suppose the first m_r ($m_r > 1$) individuals in network r are sampled. Suppose we can observe network connections $G_r = [g_{ij,r}]$ and covariates $x_{i,r}$ for all individuals in network r , but we can only observe $y_{i,r}$ of sampled individuals. For the sampled individuals, $i = 1, \dots, m_r$, (3.1) becomes

$$y_{i,r} = \phi \sum_{j=1}^{m_r} g_{ij,r} y_{j,r} + x'_{i,r} \beta + \sum_{j=1}^{n_r} g_{ij,r} x'_{j,r} \gamma + \eta_r + \epsilon_{i,r}^*. \quad (3.5)$$

By comparing (3.1) and (3.5), we have $\epsilon_{i,r}^* = \phi \sum_{j=m_r+1}^{n_r} g_{ij,r} y_{j,r} + \epsilon_{i,r}$. Therefore, the error term of model (3.5) contains two types of errors: the error due to unobserved individual heterogeneity $\epsilon_{i,r}$ and the measurement error due to the sampling design $\phi \sum_{j=m_r+1}^{n_r} g_{ij,r} y_{j,r}$. The measurement error could be correlated with the covariates and, as a result, the 2SLS estimator given by (3.4) may not be consistent.

To further illustrate this point, we rewrite (3.5) in matrix form. Let

$$G_r = \begin{bmatrix} G_r^S \\ G_r^N \end{bmatrix} = \begin{bmatrix} G_r^{SS} & G_r^{SN} \\ G_r^{NS} & G_r^{NN} \end{bmatrix},$$

¹⁷ In this paper, we focus on the missing values of the dependent variable for two reasons. First, in our empirical application, the missing values of the dependent variable are due to the sampling design. Thus, the missing rate of the dependent variable is much higher than those of the covariates (see the last column of Table A1). Second, the missing values of the covariates can be imputed using some conventional methods (see, e.g. Little and Rubin, 2002). As explained later in this section, the consistency of the proposed NLS estimator relies on the condition that $E[\epsilon_r | G_r, X_r, \eta_r] = 0$. As long as this condition holds when the missing values of the covariates are imputed, the NLS estimator remains consistent.

¹⁸ The use of the core sample of the in-home survey is crucial because otherwise the sampled students in the in-home survey are not random.

where G_r^S is an $m_r \times n_r$ matrix of the first m_r rows of G_r and G_r^{SS} is an $m_r \times m_r$ matrix of the first m_r columns of G_r^S . Then, for the sampled individuals, we have

$$Y_r^S = \phi G_r^{SS} Y_r^S + X_r^S \beta + G_r^S X_r \gamma + \eta_r l_{m_r} + \epsilon_r^*, \quad (3.6)$$

where $Y_r^S = (y_{1,r}, \dots, y_{m_r,r})'$ denotes the $m_r \times 1$ vector of observations on the dependent variable of the sampled individuals, $X_r^S = (x_{1,r}, \dots, x_{m_r,r})'$ denotes the $m_r \times p$ matrix of observations on the covariates of the sampled individuals, and $\epsilon_r^* = \epsilon_r^S + \phi G_r^{SN} Y_r^N$ with $\epsilon_r^S = (\epsilon_{1,r}, \dots, \epsilon_{m_r,r})'$ and $Y_r^N = (y_{m_r+1,r}, \dots, y_{n_r,r})'$. As $E[\epsilon_r | G_r, X_r, \eta_r] = 0$, we have

$$E[\epsilon_r^* | G_r, X_r, \eta_r] = E[\epsilon_r^S + \phi G_r^{SN} Y_r^N | G_r, X_r, \eta_r] = \phi G_r^{SN} E[Y_r^N | G_r, X_r, \eta_r].$$

To obtain $E[Y_r^N | G_r, X_r, \eta_r]$, we need to inspect the reduced-form equation of the model. If $(I_{n_r} - \phi G_r)$ is nonsingular, then the reduced-form equation of (3.2) is given by

$$Y_r = (I_{n_r} - \phi G_r)^{-1} (X_r \beta + G_r X_r \gamma) + \frac{\eta_r}{1 - \phi} l_{n_r} + (I_{n_r} - \phi G_r)^{-1} \epsilon_r. \quad (3.7)$$

Let $D_r^N = [0_{(n_r - m_r) \times m_r}, I_{n_r - m_r}]$ denote an $(n_r - m_r) \times n_r$ matrix of the last $(n_r - m_r)$ rows of an $n_r \times n_r$ identity matrix. Then, it follows from (3.7) that

$$\begin{aligned} E[Y_r^N | G_r, X_r, \eta_r] &= D_r^N E[Y_r | G_r, X_r, \eta_r] \\ &= D_r^N (I_{n_r} - \phi G_r)^{-1} (X_r \beta + G_r X_r \gamma) + \frac{\eta_r}{1 - \phi} l_{n_r - m_r}. \end{aligned}$$

Therefore,

$$\begin{aligned} E[\epsilon_r^* | G_r, X_r, \eta_r] &= \phi G_r^{SN} E[Y_r^N | G_r, X_r, \eta_r] \\ &= \phi G_r^{SN} D_r^N (I_{n_r} - \phi G_r)^{-1} (X_r \beta + G_r X_r \gamma) + \frac{\phi \eta_r}{1 - \phi} G_r^{SN} l_{n_r - m_r}. \end{aligned} \quad (3.8)$$

If $\phi \neq 0$, then, in general, $E[\epsilon_r^* | G_r, X_r, \eta_r] \neq 0$. Furthermore, if $m_r/n_r \rightarrow c_r$ where $0 \leq c_r < 1$, then the entries of $E[\epsilon_r^* | G_r, X_r, \eta_r]$ may not converge to zero either. Hence, the 2SLS estimator given by (3.4) may not be consistent when the dependent variable has missing values.¹⁹

To overcome this missing data problem due to sampling, we consider the NLS approach suggested by Wang and Lee (2013a) based on the reduced-form equation (3.7). Let $D_r^S = [I_{m_r}, 0_{m_r \times (n_r - m_r)}]$ be an $m_r \times n_r$ matrix of the first m_r rows of an $n_r \times n_r$ identity matrix. Then,

$$Y_r^S = D_r^S Y_r = D_r^S (I_{n_r} - \phi G_r)^{-1} (X_r \beta + G_r X_r \gamma) + \frac{\eta_r}{1 - \phi} l_{m_r} + u_r, \quad (3.9)$$

where $u_r = D_r^S (I_{n_r} - \phi G_r)^{-1} \epsilon_r$. To eliminate the incidental parameters η_r , we apply a within transformation using the projector $J_r^S = I_{m_r} - (1/m_r) l_{m_r} l_{m_r}'$ so that (3.9) becomes

$$J_r^S Y_r^S = J_r^S D_r^S (I_{n_r} - \phi G_r)^{-1} (X_r \beta + G_r X_r \gamma) + J_r^S u_r.$$

¹⁹ It is worth noting that, if $m_r/n_r \rightarrow 1$, then the entries of $E[\epsilon_r^* | G_r, X_r, \eta_r]$ may converge to zero as $n_r \rightarrow \infty$. In this case, the 2SLS estimator given by (3.4) can still be consistent. However, this is certainly not the case in our empirical study, where the sampling rate of the in-home survey is about 13%.

The NLS estimator of $\theta = (\phi, \beta', \gamma')'$ is given by

$$\hat{\theta}_{\text{nls}} = \arg \min_{\theta} \sum_{r=1}^{\bar{r}} (Y_r^S - h_r(\theta))' J_r^S (Y_r^S - h_r(\theta)), \quad (3.10)$$

where $h_r(\theta) = D_r^S (I_{n_r} - \phi G_r)^{-1} (X_r \beta + G_r X_r \gamma)$. As

$$E[J_r^S u_r | G_r, X_r, \eta_r] = J_r^S D_r^S (I_{n_r} - \phi G_r)^{-1} E[\epsilon_r | G_r, X_r, \eta_r],$$

the NLS estimator is consistent as long as $E[\epsilon_r | G_r, X_r, \eta_r] = 0$, which is a common assumption in the social network literature; see, e.g. Bramoullé et al. (2009) and Lin (2010).

Let $J^S = \text{diag}\{J_r^S\}_{r=1}^{\bar{r}}$ and $D^S = \text{diag}\{D_r^S\}_{r=1}^{\bar{r}}$. Let $m = \sum_{r=1}^{\bar{r}} m_r$. If $m/n \rightarrow c$, where c is a finite positive constant, then it follows a similar argument as in Wang and Lee (2013a) that the NLS estimator $\hat{\theta}_{\text{nls}}$ is consistent with an asymptotic distribution

$$\sqrt{n}(\hat{\theta}_{\text{nls}} - \theta) \xrightarrow{d} N(0, \Sigma_{\text{nls}}),$$

where $\Sigma_{\text{nls}} = \lim_{n \rightarrow \infty} n(C' B' B C)^{-1} C' B' \Omega B C (C' B' B C)^{-1}$, with $B = J^S D^S (I - \phi G)^{-1}$, $C = [G(I - \phi G)^{-1} (X\beta + GX\gamma), X, GX]$ and $\Omega = \sigma^2 B B'$.

3.3. A simulation experiment

We conduct a Monte Carlo simulation in which we compare the finite sample performance of the 2SLS estimator given in (3.4) and the NLS estimator given in (3.10). The data-generating process follows model (3.1) with the adjacency matrices G_r and covariates X_r from the Add Health data.²⁰ We set $\phi = 0.6$ and $\beta_k = \gamma_k = 1$ for $k = 1, \dots, p$, and generate the network fixed effect η_r and the error term $u_{i,r}$ as i.i.d. standard normal random variables.

We conduct 2000 repetitions in the simulation. For each repetition, we draw random subsamples of the generated data with the sampling rates of 20%, 40%, 60% and 80%, respectively, and we estimate model (3.6) where only the dependent variable of the sampled individuals can be observed. As the empirical distribution of the 2SLS estimates has some outliers, we use robust measures of central tendency and dispersion for summary statistics of the 2SLS and NLS estimators, namely, the median bias (med. bias), the median of the absolute deviations (med. AD), the difference between the 0.1 and 0.9 quantile (dec. rge) in the empirical distribution and the coverage rate (cov. rate) of a nominal 95% confidence interval.

The simulation results are reported in Table 1.²¹ When the sampling rate is low, the 2SLS estimator has a substantial bias with high dispersion in its empirical distribution (i.e. large med. AD and dec. rge). The bias of the endogenous peer effect is negative. As the sampling rate increases, the bias and dispersion of the 2SLS estimator are reduced. However, the NLS estimator is essentially unbiased for all sampling rates considered. The dispersion of the NLS estimator is much lower than that of the 2SLS estimator and it is reduced as the sample size or sampling rate increases. The coverage rate of the NLS estimator is also closer to the nominal level of 95% than the 2SLS estimator.

²⁰ We use the same set of covariates as column (7) of Table 2 in the empirical study and all real networks with size between 10 and 300. We do not consider large networks (i.e. networks with sizes between 300 and 400) for ease of computation.

²¹ To save space, we only report the estimates of ϕ, β_1, γ_1 in Table 1.

Table 1. Simulation results.

Sampling rate		2SLS	NLS
20%	ϕ	−0.588 (0.588) [0.191] 0.000	0.002 (0.019) [0.071] 0.927
	β_1	0.378 (0.383) [0.631] 0.661	−0.000 (0.094) [0.357] 0.939
	γ_1	0.980 (0.985) [1.263] 0.331	−0.002 (0.163) [0.600] 0.943
40%	ϕ	−0.567 (0.567) [0.194] 0.000	0.001 (0.014) [0.052] 0.951
	β_1	0.378 (0.379) [0.516] 0.486	−0.003 (0.061) [0.233] 0.950
	γ_1	0.994 (0.997) [1.123] 0.232	−0.001 (0.112) [0.448] 0.935
60%	ϕ	−0.536 (0.536) [0.207] 0.002	0.001 (0.012) [0.045] 0.950
	β_1	0.368 (0.370) [0.487] 0.450	−0.005 (0.048) [0.188] 0.955
	γ_1	0.973 (0.976) [1.155] 0.223	−0.001 (0.091) [0.355] 0.943
80%	ϕ	−0.452 (0.452) [0.228] 0.007	0.001 (0.011) [0.040] 0.952
	β_1	0.326 (0.332) [0.573] 0.594	−0.001 (0.042) [0.161] 0.953
	γ_1	0.878 (0.889) [1.435] 0.353	−0.008 (0.080) [0.317] 0.949

Note: Columns 3 and 4 show med. bias (med. AD) [dec. rge] cov. rate. Number of replications = 2000. Sample size = 1961. Network sizes are between 10 and 300.

4. EMPIRICAL RESULTS

We estimate model (3.6) using both NLS and 2SLS, with increasing sets of controls. The estimation results are reported in Table 2. The dependent variable is the time a student goes to bed on weeknights during the school year.²² Column 1 of Table 2 gives the NLS estimates controlling for individual characteristics (age, gender and race) and family background (household size, parents' education and parents' occupation).²³ It appears that the estimated endogenous peer effect $\hat{\phi}$ is positive and statistically significant. If we ignore, for the moment, the feedback effect, then a one-hour delay in the average bedtime of the peers translates into a roughly 53-minute (0.877×60) delay in an individual's bedtime, holding the covariates constant.

However, similar behaviour between an individual and their peers may be due to the similarity in characteristics (i.e. the contextual effect), rather than to the endogenous peer effect. The uniqueness of our data, in which both respondents and friends are interviewed, allows us to control for peers' characteristics and thus to disentangle the endogenous peer effect from the contextual effect. When we control for peers' characteristics (column 2 of Table 2), the estimated endogenous peer effect decreases, showing that indeed some of the effect attributed to peers' behaviour is in fact due to the similarity in peers' characteristics.²⁴ Nevertheless,

²² If all schools were to start at the same time, then by looking at the time students go to sleep we could recover their sleeping duration. This is not the case in the US, where school districts set school starting time keeping in mind the minimization of busing costs. Middle/junior high schools (grades 7 and 8) and high schools (grades 9–12) typically have different starting times. However, because all the friend nominations in our data are within a school (typically within a grade), the effect of school starting time is captured by the network fixed effect.

²³ In the Add Health data, less than 0.6% of the fathers and less than 3% of the mothers are unemployed. None of the parents of the individuals in our final sample is unemployed. However, values for 13% of the observations are missing. We use a missing value dummy that takes value 1 if the value is missing, and 0 otherwise, to minimize the loss of information.

²⁴ Observe that in our model we include both individual characteristics and average peer characteristics. Hence, we control for the similarity of an agent with their peers on average (i.e. $x_{ik,r} - \sum_{j=1}^{n_r} g_{ij,r} x_{jk,r} = \sum_{j=1}^{n_r} g_{ij,r} (x_{ik,r} - x_{jk,r})$).

Table 2. Peer effect estimation: increasing set of controls (dependent variable is bedtime).

	NLS				2SLS				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Endogenous peer effect	0.877*** (0.019)	0.825*** (0.003)	0.798*** (0.103)	0.782*** (0.141)	0.737*** (0.154)	0.658*** (0.126)	0.602*** (0.132)	0.025 (0.021)	0.016 (0.023)
Female	-0.211 (5.162)	-11.373*** (3.162)	-7.616 (5.228)	-6.715 (4.842)	-7.429 (4.679)	-8.629 (5.550)	-12.346** (5.831)	-12.429*** (4.509)	-15.282*** (5.622)
Grade	1.856*** (0.627)	22.592*** (1.955)	32.125*** (6.600)	32.182*** (5.726)	30.761*** (5.284)	31.188*** (5.147)	29.932*** (4.894)	27.940*** (2.303)	24.772*** (2.255)
Black	-1.006 (2.861)	23.618*** (6.436)	30.005* (15.748)	29.686** (13.591)	34.593*** (12.659)	35.595*** (12.075)	34.433*** (11.703)	21.280** (8.668)	29.150*** (8.794)
Asian	7.413 (5.792)	26.292*** (9.020)	22.680 (17.390)	23.355 (14.776)	25.368* (14.190)	23.099 (14.197)	22.124 (14.248)	26.193*** (12.889)	26.294** (12.863)
Household size	1.797 (2.003)	-0.083 (1.190)	-0.643 (1.974)	-0.665 (1.996)	0.694 (1.987)	0.322 (1.925)	0.170 (1.945)	-1.639 (2.084)	-1.175 (2.074)
Parental education	5.713* (3.361)	3.650 (2.722)	6.033* (3.080)	6.467** (3.223)	7.853*** (3.210)	7.748** (3.171)	7.512** (3.171)	6.460* (3.380)	7.388*** (3.369)
Father's occ. military	6.253 (15.817)	-19.224** (8.737)	-25.997*** (11.850)	-25.319** (12.050)	-20.809* (12.196)	-22.308* (12.045)	-22.937* (12.197)	-21.889 (13.870)	-18.587 (13.691)
Father's occ. farmer	3.508 (13.212)	-17.951*** (8.826)	-18.232 (12.279)	-18.443 (12.295)	-18.707 (11.935)	-20.547* (11.770)	-21.562* (11.593)	-23.228** (11.601)	-23.300*** (11.473)
School performance				-3.412 (3.320)	2.056 (3.312)	2.255 (3.298)	1.652 (3.311)		2.876 (3.324)
Risky behaviour					14.643*** (2.141)	14.726*** (2.116)	14.752*** (2.0839)		14.293*** (2.041)
Parental occupation dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Extracurricular sport	No	No	No	No	No	Yes	Yes	No	Yes
Extracurricular other	No	No	No	No	No	No	Yes	No	Yes
Contextual effects	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Network fixed effects	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: 1,740 sampled individuals over 7,916 individuals in 33 networks. Robust standard errors in parentheses. *, **, and *** denote significant at the 10%, 5% and 1% levels, respectively. Parental occupation dummies are described in Table A1. We report the effects of parental occupation that are statistically significant. Extracurricular sport includes dummies for the participation in teams of football, soccer, baseball/softball, basketball, swimming, volleyball, track, tennis and other sports. Extracurricular other includes dummies for the participation in band, cheer/dance, chorus, orchestra, drama and other activities. See Table A1 for the definition of these variables.

the estimated endogenous peer effect remains positive and statistically different from zero.

A remaining concern relates to the presence of unobserved factors. The observed characteristics of peers may not capture all the nuances of the social environment. There may be two types of unobservables: (a) unobservables that are common to all individuals in a (broadly defined) social circle and/or (b) unobservables that are individual-specific. The bi-dimensional nature of network data (we observe individuals over networks) allows us to control for the presence of unobserved factors of type (a) by including network fixed effects. By doing so, we purge our estimates from the effects of unobserved factors that are common among directly and indirectly related individuals. Column 3 of Table 2 reports the estimation results when network fixed effects are included in the model. The estimated endogenous peer effect decreases further, but it retains its statistical significance. The presence of type (b) unobservables is probably the most difficult empirical challenge in the identification of peer effects with network data. Unobservable student characteristics may affect friendship formation, thus making the network structure endogenous; they may arise from unobservable characteristics of the family environment or, more broadly, of the social environment. We address this problem in the following way. First, the richness of our data allows us to use indicators of a set of activities that may be related to bedtime decisions as additional controls. Second, we perform a set of robustness checks in Section 5.

The results of the first exercise are contained in columns 4–7 of Table 2. In column 4, we introduce a school performance indicator and, in column 5, an indicator of risky behaviour.²⁵ Next, we add participation in several sports (column 6) and participation in other extracurricular activities (column 7) as further controls. The estimated endogenous peer effect decreases substantially across columns, but it remains statistically different from zero. In the specification with the more extensive set of controls (column 7), having peers who go to bed one hour later on average translates into a delay of approximately 36 minutes (0.602×60) in one's own bedtime. The effects of the covariates are in line with previous results in medical research. Indeed we find that, on average, females go to bed about 10 minutes earlier than males, black people go to bed more than 20 minutes after white people and students in higher grades and those with risky behaviour delay bedtime; see, e.g. Lauderdale et al. (2008) and National Sleep Foundation (2010, 2014). Interestingly, we find that children whose fathers have a military career or work in the agricultural sector go to bed about 18 minutes earlier than children with fathers who have other occupations. Having highly educated parents is associated with going to bed later. Observe that our evidence on the existence of peer effects in sleeping behaviour suggests that these quantifications would be different. Indeed, if $\phi \neq 0$ (in model (3.5)), then the marginal effect of the k th covariate would be $(I_{n_r} - \phi G_r)^{-1}(I_{n_r}\beta_k + G_r\gamma_k)$, which is an $n_r \times n_r$ matrix with its (i, j) th element representing the effect of a change in $x_{jk,r}$ on $y_{i,r}$. The marginal effects are heterogeneous across individuals, as they depend on the individual's position in the network. We show in panel (b) of Table 3 the mean, standard deviation, minimum and maximum of the main diagonal elements of the matrix. This can be considered as a summary measure of the own-partial derivatives (labelled as direct effects); see Pace and LeSage (2009). Panel (c) reports the mean, standard deviation, minimum and maximum of the cumulative sum of off-diagonal elements. This reflects the cross-partial derivatives and provides a summary measure of spillovers (labelled

²⁵ The indicator of risky behaviour is the score of a factor analysis run on alcohol consumption, cigarette smoking and general health.

Table 3. Direct and indirect effects of exogenous variables.

	Panel (a)		Panel (b)			Panel (c)					
	β		Mean	Std	Direct effects	Max	Mean	Std	Indirect effects	Min	Max
Female (minutes)	-12.346 (-7)	-13.651 (-8)	1.063 (1)		-19.346 (-12)	-12.580 (-8)	-0.057 (0)	0.321 (0)		-11.644 (-7)	-0.000 (0)
Grade (minutes)	29.932 (18)	33.148 (20)	2.581 (2)		30.547 (18)	46.977 (28)	0.140 (0)	0.780 (0)		0.000 (0)	28.275 (16)
Black (minutes)	34.433 (21)	38.105 (23)	2.967 (2)		35.116 (21)	54.003 (32)	0.161 (0)	0.897 (1)		0.000 (0)	32.504 (19)
Parental education (minutes)	7.512 (5)	8.311 (5)	0.647 (0)		7.659 (5)	11.778 (7)	0.035 (0)	0.195 (0)		0.000 (0)	7.089 (4)
Father's occ. military (minutes)	-22.937 (-14)	-25.382 (-15)	1.976 (1)		-35.971 (-22)	-23.390 (-14)	-0.107 (0)	0.597 (0)		-21.650 (-12)	-0.000 (0)
Father's occ. farmer (minutes)	-20.547 (-12)	-23.866 (-14)	1.858 (1)		-33.824 (-20)	-21.994 (-13)	-0.101 (0)	0.562 (0)		-20.358 (-12)	-0.000 (0)

Note: Marginal effects are reported (minutes in parentheses). Direct effects for variable k are computed using the average of the main diagonal elements of the matrix $(I_{nr} - \phi G_r)^{-1}(I_{nr}\beta_k + G_r\gamma_k)$. Indirect effects are the averages of the cumulative sum of off-diagonal elements. Direct and indirect effects are computed within each network. Only statistically significant effects are reported.

as indirect effects); see Pace and LeSage (2009).²⁶ Panel (a) of Table 3 reports the estimates of Table 2 (column 7) for comparison. Table 3 reveals that the direct effects of the exogenous variables in panel (b) are about 10% higher than those in panel (a). The indirect effects are tiny, but are highly heterogeneous across individuals.

Before moving to further robustness checks, it is interesting to compare our estimates with those that would be obtained if we had neglected the sampling issue. That is, if we estimate model (3.6) using 2SLS. Columns 8 and 9 of Table 2 collect the estimates that are obtained using the same set of controls as in columns 3 and 7, respectively. In line with the simulation results, without taking into account the sampling issue, the downward bias leads the 2SLS estimate of the endogenous peer effect to be statistically insignificant.

5. ROBUSTNESS CHECKS

5.1. Endogenous network formation

If the variables that drive the process of selection into friendship networks are not fully observable, then potential correlations between (unobserved) network-specific factors and the adjacency matrix G_r could lead to biased NLS estimates. The network fixed effect is a remedy for the selection bias that originates from the possible sorting of individuals with similar unobserved characteristics into a network. The underlying assumption is that such unobserved characteristics are common to the individuals within each network. This assumption is reasonable when the networks are relatively small. However, if there is unobservable individual heterogeneity that drives both network formation and behavioural choice, then G_r is likely to be endogenous even after controlling for the network fixed effect. Here, we provide a robustness check that helps to alleviate this concern.

We use a dyadic friendship formation model that is common in the empirical literature on network formation; see, e.g. De Weerd (2002), Udry and Conley (2004), Fafchamps and Gubert (2007), Mayer and Puller (2008), Mihaly (2009), Santos and Barrett (2010) and Graham (2014). It is an homophily model, in which the variables that explain $g_{ij,r}^*$ are distances in terms of observed and unobserved characteristics between students i and j :²⁷

$$d_{ij,r} = \alpha + \sum_{k=1}^p \delta_k x_{ij,k,r} + \theta v_{ij,r} + \eta_r + u_{ij,r}. \quad (5.1)$$

In (5.1), $d_{ij,r}$ is the latent dependent variable such that $g_{ij,r}^* = \mathbf{1}(d_{ij,r} > 0)$, where $\mathbf{1}(\cdot)$ is an indicator function. Here, $x_{ij,k,r} = \mathbf{1}(x_{ik,r} = x_{jk,r})$ if $x_{ik,r}$ is a discrete variable, while $x_{ij,k,r} = 1/|x_{ik,r} - x_{jk,r}|$ if $x_{ik,r}$ is a continuous variable. $v_{ij,r}$ is equal to $|v_{i,r} - v_{j,r}|$, where $v_{i,r}$ and

²⁶ See LeSage (2014) for further details.

²⁷ Homophily, the tendency for people to associate with others similar to themselves, has been widely recognized as a pervasive feature of social and economic networks; see, e.g. McPherson et al. (2001) and Golub and Jackson (2012b). In particular, there is ample evidence that inbreeding (i.e. the decision of agents to connect with others of the same type) holds true for the nationally representative sample of US teenagers covered by the Add Health dataset; see, e.g. Currarini et al. (2009, 2010) and Golub and Jackson (2012a). The specific homophily model adopted in this paper has been used by Hsieh and Lee (2014) and Goldsmith-Pinkham and Imbens (2013) to explain friendship formation among students in the Add Health dataset. Our test is informative under the assumption that this homophily model is a reasonable approximation of network formation.

Table 4. Robustness check: endogenous network formation.

	Full set of controls		Grade omitted	
Dependent variable: probability of forming a link (g_{ij})	OLS	LOGIT	OLS	LOGIT
Residuals	−0.117 (0.091)	−0.919 (2.275)	−0.361*** (0.078)	−9.364*** (1.853)
Female	0.007*** (0.002)	0.187*** (0.069)	0.007*** (0.002)	0.217*** (0.039)
Grade	0.121*** (0.012)	3.185*** (0.097)		
Black	0.025*** (0.007)	0.543*** (0.168)	0.022*** (0.005)	1.029*** (0.103)
Asian	0.008* (0.004)	0.385*** (0.114)	0.010*** (0.003)	0.418*** (0.123)
Household size	−0.003 (0.015)	0.021 (0.370)	−0.003 (0.014)	0.077 (0.351)
Parental education	0.002 (0.002)	0.203 (0.244)	0.002 (0.002)	0.315 (0.272)
Father's occ. military	0.002 (0.004)	0.256 (0.265)	0.002 (0.003)	0.283 (0.312)
Father's occ. farmer	0.007*** (0.003)	0.294*** (0.103)	0.006*** (0.003)	0.398*** (0.153)
School performance	0.007*** (0.002)	0.312*** (0.078)	0.012*** (0.002)	0.260*** (0.051)
Risky behaviour	0.016*** (0.005)	0.491*** (0.117)	0.019*** (0.004)	0.418*** (0.062)
Parental occupation dummies	Yes	Yes	Yes	Yes
Extracurricular sport	Yes	Yes	Yes	Yes
Extracurricular other	Yes	Yes	Yes	Yes
Network fixed effects	Yes	Yes	Yes	Yes
Observations	123,319	123,319	123,319	123,319

Note: 1,740 sampled individuals over 7,916 individuals in 33 networks. Robust standard errors in parentheses. *, ** and *** denote significant at the 10%, 5% and 1% levels, respectively. The difference in residuals is divided by 100,000 and the difference in household size is divided by 10.

$v_{j,r}$ are the residuals from the outcome equation (3.9) for individual i and j , respectively. η_r is a network-specific parameter, $u_{ij,r}$ is the error term and θ is the parameter of interest, which captures how differences in unobserved individual characteristics affect the probability to be friends. Intuitively, the test evaluates whether homophily in unobserved factors that drive similar behavioural decisions also helps to explain friendship formation decisions. A statistically significant estimate of θ would suggest that the adjacency matrix G_r is endogenous and the NLS estimator is inconsistent.

OLS and logit estimates of the network formation model (5.1) are presented in the first and second columns of Table 4. The results show no sign of correlation between differences in unobserved individual characteristics and link formation. It should also be noted that, because there are several thousand individual-pair observations in a given regression, the power to detect small departures from zero is quite high.

Table 5. Robustness check: simulated peers.

	Matching criterion				
	Parental education	Household size	Father's occupation	Mother's occupation	Grade
Peer effect	−0.001 (0.004)	−0.004 (0.004)	−0.003 (0.004)	0.003 (0.004)	−0.001 (0.004)
Female	−13.054** (5.862)	−13.012** (5.857)	−13.049** (5.858)	−13.120** (5.858)	−13.045** (5.860)
Grade	27.045*** (4.466)	26.952*** (4.464)	26.954*** (4.467)	27.029*** (4.467)	27.087*** (4.467)
Black	29.774*** (10.175)	29.755*** (10.170)	29.613*** (10.172)	29.767*** (10.172)	29.821*** (10.179)
Asian	26.497 (15.543)	26.656 (15.545)	26.498 (15.537)	26.503 (15.538)	26.439 (15.543)
Household size	−0.753 (2.054)	−0.720 (2.052)	−0.718 (2.053)	−0.687 (2.054)	−0.763 (2.054)
Parental education	6.452* (3.390)	6.580* (3.391)	6.497* (3.389)	6.498* (3.389)	6.488* (3.396)
Father's occ. military	−22.862* (13.616)	−22.401* (13.622)	−23.059* (13.616)	−23.141* (13.612)	−22.996* (13.623)
Father's occ. farmer	−26.823** (11.427)	−26.556** (11.427)	−26.854** (11.421)	−26.404** (11.430)	−26.891** (11.426)
School performance	3.553 (3.361)	3.552 (3.356)	3.500 (3.355)	3.667 (3.359)	3.561 (3.358)
Risky behaviour	14.122*** (2.109)	14.087*** (2.107)	14.133*** (2.108)	14.169*** (2.109)	14.136*** (2.109)
Parental occupation dummies	Yes	Yes	Yes	Yes	Yes
Extracurricular other	Yes	Yes	Yes	Yes	Yes
Extracurricular sport	Yes	Yes	Yes	Yes	Yes
Contextual effects	Yes	Yes	Yes	Yes	Yes
Network fixed effects	Yes	Yes	Yes	Yes	Yes

Note: Dependent variable is bedtime. There are 1,740 sampled individuals over 7,916 individuals in 33 networks. For each individual, we draw at random a number of friends equal to the nominated one of a specific type. As a matching criterion, parental education is recoded as a dummy taking a value equal to 1 if the parent has graduated from college and above, and 0 otherwise. Household size is recoded as a dummy taking a value equal to 1 if the number of household members is greater than 4, and 0 otherwise. Father's and mother's occupations are recoded as dummies taking a value equal to 1 if the occupation is manager, sales or technical, and 0 otherwise. Grade is coded as a dummy taking a value equal to 1 if the student is in grades 9–12 (high school), and 0 otherwise (middle school). *, ** and *** denote significant at the 10%, 5% and 1% levels, respectively.

To increase confidence in our exercise, we perform the following experiment. We deliberately leave out one individual characteristic, which will then act as an unobserved factor (to the econometrician). We exclude the grade, which is relevant to both the link formation process and bedtime decisions. If our exercise detects this problem, then we should obtain a negative and statistically significant $\hat{\theta}$. The last two columns of Table 4 report the results. We can see that $\hat{\theta}$ is indeed different from zero, which suggests that the proposed test has a reasonable power.

To conclude, after controlling the (unusually) large set of individual characteristics provided by Add Health, peer characteristics and network fixed effects, we find no evidence of network

endogeneity that could bias the NLS estimator, given the homophily network formation model assumed.²⁸

5.2. *Simulated peers*

Along the lines of Bifulco et al. (2011), we run placebo tests in which we replace the actual peers with simulated peers. We draw, at random, peers with the same family size, or parental education, or parental occupation, or the cohort of actual peers. More specifically, for each individual, we draw, at random, a number of friends equal to the nominated friends of a given type (i.e. belonging to a given social circle as defined by parental education, occupation, etc.). If our estimates simply capture unobserved social circle characteristics, then these regressions should continue to show a statistically significant peer effect. However, if our strategy is valid, then we should not find any effect of simulated peers' behaviour on one's own behaviour in these placebo regressions. The results are contained in Table 5. No evidence of a significant endogenous peer effect is revealed. Thus, this evidence provides further confirmation that our strategy, which is based on a large set of controls about individuals and their peers as well as network fixed effects, is able to cope with sorting issues that could confound our estimates.

6. CONCLUSIONS

Our study brings two main contributions to the literature. First, we extend the NLS estimator in Wang and Lee (2013a) to estimate social network models with sampled observations on the dependent variable. Second, we analyse peer effects in bedtime decisions using a representative sample of US teenagers, finding non-negligible endogenous peer effects. That is, besides the impact of individual and friend characteristics, we show that the sleeping behaviour of friends helps to shape own sleeping behaviour. Our findings are consistent with a behavioural model of peer effects with conformist preferences in which bedtime decisions among peers are partly made following the peer group norm. However, there are a variety of utility functions (or a variety of social processes) that can be consistent with our evidence. To discriminate between the different mechanisms empirically is an extremely difficult exercise, which requires (at least) better data. It is worth noting, however, that our methodology applies to any SAR model, irrespective of its microfoundation.

ACKNOWLEDGEMENTS

This research uses data from Add Health, a programme 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 to 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 web site

²⁸ Patacchini and Venanzoni (2014) use a similar strategy to demonstrate the importance of network fixed effects in identifying peer effects in the demand for housing quality.

(<http://www.cpc.unc.edu/addhealth>). No direct support was received from grant P01-HD31921 for this analysis. The views expressed here do not necessarily reflect those of the Bank of Italy or the Eurosystem.

REFERENCES

- Akerlof, G. A. (1980). A theory of social custom, of which unemployment may be one consequence. *Quarterly Journal of Economics* 94, 749–75.
- Akerlof, G. A. (1997). Social distance and social decisions. *Econometrica* 65, 1005–27.
- Benhabib, J., A. Bisin and M. O. Jackson (Eds.) (2011). *Handbook of Social Economics, Volume 1A*. Amsterdam: North-Holland.
- Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy* 102, 841–77.
- Biddle, J. E. and D. S. Hamermesh (1990). Sleep and the allocation of time. *Journal of Political Economy* 98, 922–43.
- Bifulco, R., J. M. Fletcher and S. L. Ross (2011). The effect of classmate characteristics on post-secondary outcomes: evidence from the Add Health. *American Economic Journal: Economic Policy* 3, 25–53.
- Bramoullé, Y., H. Djebbari and B. Fortin (2009). Identification of peer effects through social networks. *Journal of Econometrics* 150, 41–55.
- Calvó-Armengol, A., E. Patacchini and Y. Zenou (2009). Peer effects and social networks in education. *Review of Economic Studies* 76, 1239–67.
- Chandrasekhar, A. and R. Lewis (2011). Econometrics of sampled networks. Working paper, MIT.
- Currarini, S., M. O. Jackson and P. Pin (2009). An economic model of friendship: homophily, minorities, and segregation. *Econometrica* 77, 1003–45.
- Currarini, S., M. O. Jackson and P. Pin (2010). Identifying the roles of race-based choice and chance in high school friendship network formation. *Proceedings of the National Academy of Sciences of the United States of America* 107, 4857–61.
- De Weerd, J. (2002). Risk-sharing and endogenous network formation. Research Paper DP2002/57, United Nations University World Institute for Development Economics Research (UNU-WIDER).
- Fafchamps, M. and F. Gubert (2007). Risk sharing and network formation. *American Economic Review* 97(2), 75–79.
- Fershtman, C. and Y. Weiss (1998). Social rewards, externalities and stable preferences. *Journal of Public Economics* 70, 53–73.
- Goldsmith-Pinkham, P. and G. W. Imbens (2013). Social networks and the identification of peer effects. *Journal of Business and Economic Statistics* 31, 253–64.
- Golub, B. and M. O. Jackson (2012a). Does homophily predict consensus times? Testing a model of network structure via a dynamic process. *Review of Network Economics* 11, Issue 3, Article 9.
- Golub, B. and M. O. Jackson (2012b). How homophily affects the speed of learning and best-response dynamics. *Quarterly Journal of Economics* 127, 1287–338.
- Graham, B. S. (2014). An econometric model of link formation with degree heterogeneity. NBER Working Paper 20341, National Bureau of Economic Research.
- Hsieh, C.-S. and L. F. Lee (2014). A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics* 31, 1099–255.
- Kandel, E. and E. P. Lazear (1992). Peer pressure and partnerships. *Journal of Political Economy* 100, 801–17.
- Kelejian, H. H. and I. Prucha (2010). Spatial models with spatially lagged dependent variables and incomplete data. *Journal of Geographical Systems* 12, 241–57.

- Lauderdale, D., K. L. Knutson, L. L. Yan, K. Liu and P. J. Rathouz (2008). Sleep duration: how well do self-reports reflect objective measures? The CARDIA sleep study. *Epidemiology* 19, 838–45.
- Lee, L. F., X. Liu and X. Lin (2010). Specification and estimation of social interaction models with network structures. *Econometrics Journal* 13, 145–76.
- LeSage, J. (2014). What regional scientists need to know about spatial econometrics. *Review of Regional Studies* 44, 13–32.
- Lin, X. (2010). Identifying peer effects in student academic achievement by a spatial autoregressive model with group unobservables. *Journal of Labor Economics* 28, 825–60.
- Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. New York, NY: Wiley.
- McPherson, M., L. Smith-Lovin and J. M. Cook (2001). Birds of a feather: homophily in social networks. *Annual Review of Sociology* 27, 415–44.
- Manski, C. F. (1993). Identification of endogenous social effects: the reflection problem. *Review of Economic Studies* 60, 531–42.
- Mayer, A. and S. L. Puller (2008). The old boy (and girl) network: social network formation on university campuses. *Journal of Public Economics* 92, 329–47.
- Mihaly, K. (2009). Do more friends mean better grades? Student popularity and academic achievement. RAND Working Paper WR-678, RAND Corporation.
- National Sleep Foundation (2006). NSF's 2006 Sleep in America poll: highlights and key findings. <https://sleepfoundation.org/sleep-polls-data/sleep-in-america-poll/2006-teens-and-sleep>.
- National Sleep Foundation (2010). NSF's 2010 Sleep in America poll: highlights and key findings. See <https://sleepfoundation.org/sleep-polls-data/sleep-in-america-poll/2010-sleep-and-ethnicity>.
- National Sleep Foundation (2014). NSF's 2014 Sleep in America poll: highlights and key findings. See <https://sleepfoundation.org/sleep-polls-data/sleep-in-america-poll/2014-sleep-and-family>.
- Pace, R. K. and J. LeSage (2009). *Introduction to Spatial Econometrics*. Boca Raton, FL: Chapman & Hall/CRC.
- Patacchini, E. and G. Venanzoni (2014). Peer effects in the demand for housing quality. *Journal of Urban Economics* 83, 6–17.
- Patacchini, E. and Y. Zenou (2008). The strength of weak ties in crime. *European Economic Review* 52, 209–36.
- Patacchini, E. and Y. Zenou (2012). Juvenile delinquency and conformism. *Journal of Law, Economics, and Organization* 28, 1–31.
- Rao, G., F. Schillbach and H. Schofield (2015). Sleepless in Chennai: the economic effects of sleep deprivation among the poor. Working paper, MIT.
- Santos, P. and C. B. Barrett (2010). Identity, interest and information search in a dynamic rural economy. *World Development* 38, 1788–96.
- Sojourner, A. (2013). Identification of peer effects with missing peer data: evidence from project STAR. *Economic Journal* 123, 574–605.
- Udry, C. R. and T. G. Conley (2004). Social networks in Ghana. Yale University Economic Growth Center Discussion Paper No. 888, Yale University.
- Wang, W. and L. F. Lee (2013a). Estimation of spatial autoregressive models with randomly missing data in the dependent variable. *Econometrics Journal* 16, 73–102.
- Wang, W. and L. F. Lee (2013b). Estimation of spatial panel data models with randomly missing data in the dependent variable. *Regional Science and Urban Economics* 43, 521–38.
- Wolfson, A. R. and M. A. Carskadon (2003). Understanding adolescent's sleep patterns and school performance: a critical appraisal. *Sleep Medicine Reviews* 7, 491–506.

APPENDIX A: DATA DESCRIPTION

In Table A.1, we give the sample summary statistics. The descriptions of the explanatory variables are as follows.

The sampled variable is bedtime. This is the answer to the question: ‘During the school year, what time do you usually go to bed on week nights?’. An hour is rescaled in 100 units, and half an hour is thus 50 units. The mean is thus 10.37 p.m. and the standard deviation is about one hour.

Table A.1. Summary statistics.

Variable	Mean	Std dev.	Min	Max	No. obs.	Missing rate
Sampled variable						
Bedtime	1060.236	100.254	800	1,800	1,740	93%
Demographic characteristics						
Female	0.535	0.499	0	1	7,916	1%
Grade	9.598	1.610	7	12	7,916	1%
Black	0.123	0.328	0	1	7,916	0%
Asian	0.092	0.289	0	1	7,916	0%
Family characteristics						
Family size	4.466	0.993	1	6	7,916	4%
Mother’s occupation						
Prof. tech.	0.271	0.445	0	1	7,916	0%
Manager	0.062	0.241	0	1	7,916	0%
Sales	0.244	0.430	0	1	7,916	0%
Manual	0.114	0.318	0	1	7,916	0%
Military	0.004	0.064	0	1	7,916	0%
Farmer	0.037	0.190	0	1	7,916	0%
Father’s occupation						
Prof. tech.	0.213	0.410	0	1	7,916	0%
Manager	0.148	0.355	0	1	7,916	0%
Sales	0.077	0.267	0	1	7,916	0%
Manual	0.366	0.482	0	1	7,916	0%
Military	0.041	0.198	0	1	7,916	0%
Farmer	0.054	0.225	0	1	7,916	0%
Other	0.023	0.149	0	1	7,916	0%
Parental education	2.440	0.828	0	4	7,916	9%
Activities						
Drama	0.086	0.280	0	1	7,916	0%
Band	0.161	0.367	0	1	7,916	0%
Cheer/dance	0.103	0.304	0	1	7,916	0%
Chorus	0.140	0.347	0	1	7,916	0%
Orchestra	0.025	0.156	0	1	7,916	0%
Other club	0.244	0.430	0	1	7,916	0%
Baseball/softball	0.222	0.416	0	1	7,916	0%
Basketball	0.244	0.430	0	1	7,916	0%

Continued

Table A.1. Continued.

Variable	Mean	Std dev.	Min	Max	No. obs.	Missing rate
Field hockey	0.016	0.124	0	1	7,916	0%
Football	0.138	0.345	0	1	7,916	0%
Hockey	0.024	0.152	0	1	7,916	0%
Soccer	0.100	0.300	0	1	7,916	0%
Swimming	0.063	0.243	0	1	7,916	0%
Tennis	0.072	0.258	0	1	7,916	0%
Track	0.155	0.362	0	1	7,916	0%
Volleyball	0.104	0.306	0	1	7,916	0%
Wrestling	0.039	0.195	0	1	7,916	0%
Other sport	0.127	0.333	0	1	7,916	0%
School performance	2.798	0.811	1	4	7,916	13%
Risk behaviour	1.177	1.160	0	6.227	7,916	7%

Note: There are 7,916 individuals over 33 networks, with network sizes between 10 and 400 individuals. The missing rate is the ratio between the number of observations with missing values and the entire sample size in the original sample. Missing value dummies are used for parental occupations.

In the demographic characteristics, female is a dummy variable taking a value of 1 if the respondent is female, grade is the grade of the student in the current year, and black and asian are race dummies, with white being the reference group.

In the family characteristics, family size is the number of people living in the household, where category '6' includes six or more people, and the mother's and father's occupation dummies are the closest descriptions of the jobs of the (biological or non-biological) parents who are living with the child. Parental education is the schooling level of the (biological or non-biological) parents who are living with the child, distinguishing between 'never went to school', 'did not graduate from high school', 'high school graduate', 'graduated from college or a university', 'professional training beyond a four-year college', coded as 0, 1, 2, 3 and 4. We consider the maximum education across the parents who are present in the household. If the information on one parent is missing, then we use the education of the other parent.

The activities are extracurricular activity dummies. The dummy variable takes a value of 1 if the respondent participates (or plans to participate in the current school year) in the listed activity.

School performance is the average across available test scores in Mathematics, English, Science and History/Social Science at the most recent grading period. Test scores are coded as A = 4, B = 3, C = 2 and D = 1.

Risk behaviour is the total score of a factor analysis run on alcohol consumption, cigarette smoking and health. Alcohol consumption is obtained using the response to the question: 'How often did you drink beer, wine or liquor?', coded as 0 = never, 1 = once or twice, 2 = once a month or less, 3 = two or three days a month, 4 = once a week, 5 = three to five days a week and 6 = nearly every day. Information on cigarette smoking is obtained using the response to the question: 'How often did you smoke cigarettes?', coded as 0 = never, 1 = once or twice, 2 = once a month or less, 3 = two or three days a month, 4 = once a week, 5 = three to five days a week and 6 = nearly every day. Information on respondent's health status is obtained using the question: 'In general, how is your health?', coded as 5 = excellent, 4 = very good, 3 = good, 2 = fair and 1 = poor.

APPENDIX B: MICROFOUNDATION

Following Patacchini and Zenou (2012), we present a social network model of peer effects with conformity preferences for the demand for sleep.

B.1. The network

Suppose there are n individuals in the economy partitioned in $\bar{r}$ networks. Let n_r be the number of individuals in the r th network, so that $n = \sum_{r=1}^{\bar{r}} n_r$. The adjacency matrix $G_r^* = [g_{ij,r}^*]$ of the r th network keeps track of the direct connections in this network. Here, $g_{ij,r}^* = 1$ if two players i and j are directly connected (i.e. best friends), and $g_{ij,r}^* = 0$ otherwise. We also set $g_{ii,r}^* = 0$. The set of individual i 's best friends (direct connections) is $N_{i,r} = \{j \neq i \mid g_{ij,r}^* = 1\}$, which is of size $g_{i,r}^*$ (i.e. $g_{i,r}^* = \sum_{j=1}^n g_{ij,r}^*$ is the number of direct links of individual i). This means in particular that if i and j are best friends, then in general $N_{i,r} \neq N_{j,r}$ unless the network is complete (i.e. each individual is friends with everybody in the network). This also implies that groups of friends may overlap if individuals have common best friends. To summarize, the reference group of each individual i is $N_{i,r}$ (i.e. the set of their best friends that does not include them).

B.2. Preference

We denote by $y_{i,r}$ the bedtime of individual i in network r and by $Y_r = (y_{1,r}, \dots, y_{n,r})'$ the population bedtime profile in network r . Denote by $\bar{y}_{i,r}$ the average bedtime of individual i 's best friends, given by

$$\bar{y}_{i,r} = \sum_{j=1}^{n_r} g_{ij,r} y_{j,r}, \quad (\text{B.1})$$

where $g_{ij,r} = g_{ij,r}^* / g_{i,r}^*$. Each agent i in network r decides on bedtime $y_{i,r} \geq 0$, and obtains a payoff $u_{i,r}(Y_r, G_r)$ that depends on the profile Y_r and on the underlying network G_r , in the following way:

$$u_{i,r}(Y_r, G_r) = a_{i,r} y_{i,r} - \frac{1}{2} y_{i,r}^2 - \frac{d}{2} (y_{i,r} - \bar{y}_{i,r})^2 \quad (\text{B.2})$$

where $d > 0$. The benefit part of this utility function is given by $a_{i,r} y_{i,r}$ while the cost is $(1/2) y_{i,r}^2$; both are increasing in own time $y_{i,r}$. In this part, $a_{i,r}$ denotes the agent's ex-ante idiosyncratic heterogeneity, which is assumed to be perfectly observable by all individuals in the network. The second part of the utility function $(d/2)(y_{i,r} - \bar{y}_{i,r})^2$ reflects the influence of friends' behaviour on own actions. It is such that each individual wants to minimize the social distance between them and their reference group, where d is the parameter describing the taste for conformity. Here, the individual loses utility $(d/2)(y_{i,r} - \bar{y}_{i,r})^2$ from failing to conform to others.²⁹ In the context of adolescents' bedtime decisions, a taste for conformity captures the idea that adolescents tend to make similar decisions to those of their friends. For example, if, in a given friendship circle, there is a widespread idea that going early to bed early is 'not cool' or childish, then all members of that group will tend to go to bed later to conform to the social norm.

Observe that the social norm here captures the differences between individuals due to network effects. This means that individuals have different types of friends and thus different reference groups $\bar{y}_{i,r}$. As a result, the social norm each individual i faces is endogenous and depends on their location in the network as well as the structure of the network.

B.3. Nash equilibrium and econometric model

In this game where agents choose $y_{i,r} \geq 0$ simultaneously, there exists a unique Nash equilibrium (Patacchini and Zenou, 2012) with the best response function given by

$$y_{i,r}^* = \phi \sum_{j=1}^{n_r} g_{ij,r} y_{j,r}^* + (1 - \phi) a_{i,r}, \quad (\text{B.3})$$

²⁹ This is the standard way economists have modelled conformity; see, e.g. Akerlof (1980, 1997), Bernheim (1994), Fershtman and Weiss (1998), Kandel and Lazear (1992) and Patacchini and Zenou (2012).

where $\phi = d/(1 + d)$. The equilibrium bedtime $y_{i,r}^*$ depends on the individual ex-ante heterogeneity and on the average bedtime of the reference group.

The econometric model considered in this paper follows the equilibrium best response function (B.3) by assuming

$$(1 - \phi)a_{i,r} = \sum_{k=1}^p \beta_k x_{ik,r} + \sum_{k=1}^p \gamma_k \sum_{j=1}^{n_r} g_{ij,r} x_{jk,r} + \eta_r + \varepsilon_{i,r}. \quad (\text{B.4})$$

Although we assume that $a_{i,r}$ is perfectly observable by all individuals in the network, we allow some components of $a_{i,r}$ to be unobservable to the researcher. In (B.4), $x_{ik,r}$ corresponds to the observable (to the researcher) characteristics of individual i (e.g. gender, race, age, parental education), η_r denotes the unobservable (to the researcher) network characteristics and $\varepsilon_{i,r}$ represents the unobservable (to the researcher) characteristics of individual i .

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Replication files