Identification and Estimation of Social Interactions through Variation in Equilibrium Influence*

Mikko Packalen†

5 December 2010

Abstract

This paper presents a new method for estimating social interaction effects. The proposed approach is based on using network interaction structure induced variation in equilibrium influence to construct conditionally balanced interaction structures. As equilibrium influence is determined by the known interaction structure and the unknown endogenous social interaction parameter, interaction structures are constructed for different imputed values of the unknown parameter. Each constructed interaction structure is conditionally balanced in the sense that when it is combined with observations on the outcome variable to construct a new variable, the constructed variable is a valid instrumental variable for the endogenous social interaction regressor if the true and imputed parameter values are the same. Comparison of each imputed value with the associated instrumental variable estimate thus yields a confidence set estimate for the endogenous social interaction parameter as well as for other model parameters. We provide conditions for point identification and partial identification.

The contrast between the proposed and existing approaches is stark. In the existing approach instruments are constructed from observations on exogenous variables, whereas in the proposed approach instruments are constructed from observations on the outcome variable. Both approaches have advantages, and the two approaches complement one another. We demonstrate the feasibility of the proposed approach with analyses of the determinants of subjective college completion and income expectations among adolescents in the Add Health data and with Monte Carlo simulations of Erdös-Rényi and small-world networks.

Keywords: social interaction; spatial econometrics; networks; endogenous effect.
JEL Classification Codes: C31, C26.

*I thank Jay Bhattacharya and Tony Wirjanto for discussions. I'm grateful to Jay Bhattacharya for arranging access to the Add Health data and for providing the associated computational resources. 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 and funded by grant from National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations.

†Department of Economics, University of Waterloo, Canada. Email: packalen@uwaterloo.ca.

1 Introduction

External effects are central in economic theory and policy. External impacts can arise through various mechanisms, including social interaction, spatial competition and direct physical externalities, and can influence preferences, constraints, and expectations (Manski, 2000). Regardless of the influence mechanism, in many cases interaction has network structure: not all indirectly connected agents are directly connected and some connections are more important than others. Two prominent network structures are friendship networks and professional networks.

An important consequence of interaction through a network structure is that the strength of equilibrium influence between individuals varies within each network. There are many closely related sources for this variation. First, equilibrium influence between two individuals depends on whether the individuals are directly or only indirectly connected. Equilibrium influence that an individual's friends have on the individual is typically stronger than equilibrium influence that an individual's friends' friends have on the individual. Second, equilibrium influence between two directly connected individuals depends on how many connections the individuals have in common. Two friends who have many friends in common typically have a stronger equilibrium influence on one another than two friends who do not have friends in common. Third, equilibrium influence through each direct connection on an individual depends on how many connections the individual has. Equilibrium influence of a friend on an individual who has only one friend is typically stronger than equilibrium influence of a friend on an individual who has many friends. Additional sources for variation in equilibrium influence include variation in the strength of direct connections and directional variation in connections.

In this paper we show that when social interaction has network structure, variation in the strength of equilibrium influence between individuals within each network can be used to estimate models with endogenous and exogenous social interaction effects and correlated effects. Although we focus on network structures, the proposed estimation method can also be applied when the strength of equilibrium influence between individuals or other entities is governed by spatial distance. A prominent example of a spatial interaction structure is trade networks which structure is determined by geographical distance.

We use the variation in equilibrium influence between individuals to construct conditionally balanced interaction structures, which form the basis of our proposed estimation method. As equilibrium influence between individuals is determined by the known network structure and the unknown endogenous social interaction parameter, each conditionally balanced in-

teraction structure is constructed using the known network structure and an imputed value of the endogenous social interaction parameter. The constructed network structure and the observations on the endogenous variable are then combined to construct a new variable.

Each constructed network structure is conditionally balanced in the sense that if the associated imputed value and the true value of the endogenous social interaction parameter are the same then the constructed variable is a valid instrumental variable for the endogenous social interaction variable. Accordingly, we refer to the constructed variable as a potential instrumental variable. The potential validity of the constructed variable as an instrumental variable requires that each interaction structure is constructed in such a way that 1) the constructed variable resembles the original endogenous social interaction variable sufficiently–to satisfy the instrument relevance condition–and that 2) if the imputed and true values of the endogenous social interaction parameter are the same then the impact that the unobserved characteristics of an individual have on the network (or within-network) fixed effect demeaned value of the constructed variable for that individual is the same as the average impact that the unobserved characteristics of the individual have on the network (or within-network) fixed effect demeaned values the constructed variable across all individuals in the network–to potentially satisfy the instrument exogeneity condition.

The proposed estimation strategy yields a confidence set estimate for the endogenous social interaction parameter as well as for other model parameters. Each constructed potential instrumental variable is used to test the null hypothesis that the associated imputed value and the true value of the endogenous social effect parameter are the same. Those feasible values of the endogenous social interaction variable for which this null hypothesis is not rejected form the confidence set estimate for the endogenous social interaction parameter.

As has been argued by Moffitt (2001), the four key identification problems in social interaction analysis are the simultaneity problem, the correlated unobservables problem, the errors-in-variables problem, and the endogenous group/network membership problem.[1] Our proposed estimation method explicitly addresses the first three of these four identification problems. The simultaneity and correlated unobservables problems are addressed through the construction of the conditionally balanced interaction structures and the associated potential instrumental variables and through the inclusion of network (or within-network) fixed

[1]The simultaneity problem arises because outcomes of individuals whose outcomes directly influence the outcome of a particular individual are also influenced (directly or indirectly) by the outcome of that particular individual. The correlated unobservables problem arises because unobservable variables are potentially correlated for (directly or indirectly) connected individuals. It causes outcomes of directly connected individuals to be correlated even when the endogenous social effect is absent. Both problems introduce correlation between the error term and endogenous social interaction regressor.

effects. In an extension we show that measurement error in the dependent variable and a form of misspecification error can be addressed by imposing additional conditions in the construction of conditionally balanced interaction structures. To limit the scope of the analysis we assume throughout the paper that the network structure is exogenous. While important and relevant to most applications, systematic methods for accounting for the endogeneity of the network structure do not yet exist (Brock et al., 2010).

The proposed estimation method contributes to the literature on the estimation and identification of different forms of external effects. The closest related contributions are Bramoulle et al. (2009), Lin (2010), and Lee et al. (2010). These analyses are based on network structure induced variation in friends' friends' exogenous characteristics: the instrumental variable for the endogenous social interaction variable is constructed from observations on an individual's friends' friends' exogenous characteristics. In contrast, the approach proposed here is based on network structure induced variation in equilibrium influence: the instrumental variable for the endogenous social interaction variable is constructed from constructed conditionally balanced interaction structures and observations on the outcome variable. Hence, the contrast between our analysis and even the closest existing contributions is stark. While Bramoulle et al. (2009) emphasize the role of network fixed effects in capturing correlated effects and in their application consider the assumption of within-network independence of error terms reasonable, the approach developed and advanced in these related contributions is more general than ours in that it allows also for correlated effects that arise from within-network dependence in the error terms. An advantage of the proposed approach is that it does not rely on exclusion restrictions, which validity is difficult to determine, and which can yield too little identifying variation for the existing approach. The advantages of each approach render them complementary rather than competing.

These closest related contributions and our contribution build on two branches of literature– the social interaction literature and the spatial econometrics literature–which have otherwise remained largely separate.

Analysis of identification problems in social interaction settings by Manski (1993a) introduced the distinctions between endogenous (social) effects, contextual (social) effects, and correlated effects. In this categorization endogenous effects represent the influence of other individuals' outcomes on each individual's outcome. Contextual effects represent the influence of other individuals' exogenous observed characteristics on each individual's outcome and are interchangeably called as exogenous (social) effects. Correlated effects represent the tendency of individuals to have the similar outcomes because their unobserved characteristics are similar.

As was observed by Manski (1993a, 2000), the central identification challenge in social interaction settings is to distinguish between when group behavior affects individual behavior and when group behavior merely reflects the aggregate of individual behaviors. Manski (1993a) named this task as the reflection problem and emphasized that the reflection problem has several aspects. Distinguishing between endogenous and contextual effects and distinguishing between endogenous and correlated effects are the two central challenges. In a complementary analysis Moffitt (2001) distinguished between the four sources of endogeneity in identification of social interaction effects mentioned above and argued that the problem of correlated unobservables, which Moffitt (2001) modeled as group fixed effects, is at the core of the identification problem in social interaction settings.[2]

The model examined in the spatial econometrics literature (see e.g. Pinkse and Slade, 2010, and Anselin, 1988) differs from the model examined in the social interaction literature in two respects. First, while the social interactions literature has generally focused on modeling interaction in groups, the spatial econometrics model has spatial structure. In terms of the connections structure, the typical spatial econometrics model is therefore more flexible than the typical social interactions model. Second, the exogenous social effect is absent from the typical spatial econometrics model. In this sense the typical spatial econometrics model is less flexible than the typical social interactions model.

The analyses by Bramoulle et al. (2009), Lin (2010), Lee et al. (2010) and Liu and Lee (2009) and our analysis explicitly combine features from the social interactions and spatial econometrics literatures and share a common motivation: in many settings network structure offers both a more realistic description of interaction and a better basis for identification of endogenous influence than group structure.[3] For example, the academic achievement of a student is not necessarily equally affected by the performance and characteristics of all fellow students in the relevant group (such as grade or school). Instead, fellow students who are also the student's friends may have more influence on the student than other students in the same group. With this motivation in mind, these related analyses and our analysis adopt the network structure assumption instead of the group structure assumption employed in most contributions to the social interaction literature. Moreover, these related analyses and

[2]Manski (1993a) also stated that it is important for policy purposes to distinguish between the two types of social effects because only the endogenous social effects imply the existence of a social multiplier. Identification in social interaction models can be based on randomization (see e.g. Sacerdote, 2001), instrumental variables (see e.g. Ionnides and Zabel, 2003), the assumed absence of one of type of social effect (see e.g. Krauth, 2006, and Trogdon et al., 2008), variations in group sizes (Graham, 2008, and Lee, 2007), and non-linearities in discrete choice models (Brock and Durlauf, 2001 & 2007).

[3]The crucial role of network structure in our analysis stems from the fact that when social interaction has group structure, the strength of equilibrium influence between individuals does not vary within groups.

our analysis maintain the common assumption in the social interaction literature that both endogenous and contextual effects are present.

Bramoulle et al. (2009) formally show that network structure can facilitate identification of endogenous and contextual effects. In this existing approach network structure facilitates identification of the endogenous effect through variation in an individual's friends' friends' exogenous characteristics.[4] Bramoulle et al. (2009) and Lin (2010) apply the existing approach to estimate how club participation and academic achievement, respectively, are affected the friends' outcomes and exogenous characteristics. Lee et al. (2010) extend this approach to maximum likelihood estimation and the case in which also the error terms reflect network structure and show how this extension can improve estimation.[5] Liu and Lee (2009) extend the existing approach by using the number of connections of each individual to construct additional network-specific instruments for the endogenous social interaction variable.[6]

The paper is structured as follows. The model and the proposed estimation method are presented in Sections 2 and 3, respectively. Applications and Monte Carlo simulations are presented in Section 4. Identification is examined formally in Section 5. Extensions and directions for future research are discussed in Section 6. The final section concludes.

2 The Model

We assume that there are N independent observations on networks (or on a network). For expositional convenience we initially assume that the network structure is the same for all N observations. This assumption is relaxed in Section 6.3. An observation on a network is indexed by $k \in \{1, 2, ..., N\}$.

A network of size n is a collection of n individuals who are potentially affected by the endogenous and exogenous characteristics of other individuals in the same network. An observation on an individual within the observation k on a network is indexed interchangeably by ki and kj, where $i, j \in \{1, 2, ..., n\}$.

The strength of direct influence between individuals in a network is determined in part by the network structure, which is described by the interaction matrix G. Element G_{ij} of

[4] Also the approaches developed in Laschever (2009) and De Georgi (2010) are based on identification of the endogenous effect from peers' peers characteristics. Cohen-Cole (2006) and Cohen-Cole and Zanella (2008) examine identification in the presence of between-group contextual and endogenous effects.

[5] Keleija and Prucha (2010) and Lin and Lee (2010), however, show that because the maximum likelihood approach is only consistent under the assumption of homoskedasticity a Generalized Method of Moments can be a better approach to decrease the bias and increase precision in the estimates.

[6] We briefly discuss the asymptotics in Liu and Lee (2009) in Section 6.4.1 in the context of a potential extension to the proposed method based on many weak instrumental variables approaches.

the interaction matrix G depicts the strength of direct influence of individual j on individual i relative to the direct influence of other individuals in the same network on individual i. The influence may be directional, and thus the elements G_{ij} and G_{ji} may be different.[7]

The outcome variable is denoted by Y_k. The relationship between the outcome variable and the observed and unobserved exogenous variables is described by the regression equation

$$Y_k = \alpha_k + \beta G Y_k + \delta G X_k + \gamma X_k + \varepsilon_k, \tag{1}$$

where the variable X_k is an observed exogenous variable, the variable α_k is an unobserved network fixed effect, and the variable ε_k is the (unobserved) error term. The error term ε_k satisfies $E\left[\varepsilon_k \middle| X_k, \alpha_k\right] = 0$, and the elements ε_{ki} of the error term ε_k are independently distributed both across individuals in the same network and across networks. The unconditional variance of the element ε_{ki} of the error term ε_k is denoted by σ_i^2 and may differ across individuals i in the same network but is constant across networks. We assume that the observed and unobserved exogenous variables X_k and ε_k have finite fourth moments. The coefficient β on the endogenous social interaction variable GY_k represents the endogenous social interaction effect. The set of feasible values of the parameter β is denoted by $\Omega(\beta)$.[8] The coefficient δ on the exogenous social interaction variable GX_k represents the exogenous social interaction effect. The coefficient γ on the exogenous variable X_k, which is a scalar variable for expositional convenience, represents the effect of exogenous own characteristics.

To obtain an equation that describes the equilibrium influence of observed and unobserved exogenous variables on the outcome variable, we first rearrange equation (1) to collect all terms involving the outcome variable Y_k on the left-hand side and then multiply both sides of the resulting equation from the right by the generalized inverse $(I - \beta G)^{-1}$ of the matrix $(I - \beta G)$ to get the equation

$$Y_k = (I - \beta G)^{-1} \alpha_k + \delta (I - \beta G)^{-1} G X_k + \gamma (I - \beta G)^{-1} X_k + (I - \beta G)^{-1} \varepsilon_k. \tag{2}$$

This equation (2) describes the equilibrium influence of the exogenous variables α_k, X_k, and ε_k on the endogenous variable Y_k. This equilibrium influence is governed by the equilibrium influence matrix $E \equiv (I - \beta G)^{-1}$, which is an unknown matrix as it depends not only on

[7] In our applications and simulations we construct G such that $G_{ij} \in [0, 1]$ for all (i, j) and such that for all i either $\sum_{j=1}^{n} G_{ij} = 1$ or $G_{ij} = 0$ for all j. When for some i the property $G_{ij} = 0$ holds for all j, the outcome for individual i is not endogenous. To eliminate this additional source of identification, in our applications we consider also the case when $\sum_{j=1}^{n} G_{ij} = 1$ for all i (see footnote 13).

[8] The set of feasible values $\Omega(\beta)$ depends on the interaction structure G. The normalization $\sum_{j=1}^{n} G_{ij} = 1$ employed in our applications and simulations implies that the set of feasible values is $\Omega(\beta) = (-1, 1)$.

the known interaction structure G but also on the unknown endogenous social interaction parameter β. The introduced notation allows us to rewrite equation (2) as

$$Y_k = E\alpha_k + \delta EGX_k + \gamma EX_k + E\varepsilon_k. \tag{3}$$

The term $E\varepsilon_k$ on the right-hand side of this equation (3) shows that the element E_{ij} of matrix E represents the impact of the error term ε_{kj} on the outcome variable Y_{ki}. This implies that the element $(GE)_{ij}$ of matrix GE represents the impact of the error term ε_{kj} on the element $(GY_k)_i$ of the endogenous regressor GY_k in the original regression equation (1). The known matrix G and the unknown matrix E thus determine the relationship between the endogenous regressor GY_k and the error term ε_k. We make frequent use of this observation below.

3 Potential Instrumental Variable Estimation Method

The proposed estimation method produces a confidence set estimate for the endogenous social interaction parameter β as well as for other model parameters. This confidence set estimate is formed of all those feasible values $\tilde{\beta} \in \Omega(\beta)$ of the parameter β for which the null hypothesis H_0: $\beta = \tilde{\beta}$ is not rejected. In this section we first derive the test for each individual null hypothesis H_0: $\beta = \tilde{\beta}$. In the second subsection we then describe how the results of these tests are combined to form the confidence set estimate for the parameter β. The test against each individual null hypothesis H_0: $\beta = \tilde{\beta}$ is based on a constructed conditionally balanced interaction structure which construction is addressed in the third subsection.

3.1 The Potential Instrumental Variable Test

The construction of a test for each null hypothesis H_0: $\beta = \tilde{\beta}$, where $\tilde{\beta} \in \Omega(\beta)$, proceeds in 5 steps. Steps 1 and 2 consist of constructing a variable from the endogenous variable Y_k that is a valid instrumental variable for the endogenous regressor GY_k if $\beta = \tilde{\beta}$. Steps 3 through 5 consist of using the constructed variable as an instrumental variable for the endogenous regressor GY_k to obtain an estimate of the parameter β and the associated test statistic for the null hypothesis H_0: $\beta = \tilde{\beta}$.

3.1.1 STEP 1 of 5. Determine the Influence of Each Element ε_{kj} of the Error Term on Each Element Y_{ki} of the Endogenous Variable when $\beta = \tilde{\beta}$

In order to rely in part on the endogenous variable Y_k to construct a new variable that is a valid instrumental variable for the endogenous social interaction regressor GY_k when the null hypothesis H_0: $\beta = \tilde{\beta}$ holds, we first obtain a measure of the influence of each element ε_{kj} of the error term ε_k on each element Y_{ki} of the endogenous variable Y_k when $\beta = \tilde{\beta}$. To accomplish this, we impute the feasible value $\tilde{\beta}$ for β in the earlier definition $E \equiv (I - \beta G)^{-1}$ of the unknown matrix E to construct the known matrix $\tilde{E} \equiv (I - \tilde{\beta}G)^{-1}$. Using this definition of the matrix $\tilde{E}$, equation (3) can be rewritten as

$$Y_k = \tilde{E}\alpha_k + \delta\tilde{E}GX_k + \gamma\tilde{E}X_k + \tilde{E}\varepsilon_k \tag{4}$$

when $\beta = \tilde{\beta}$. The last term $\tilde{E}\varepsilon_k$ on the right-hand side of this equation (4) implies that the element $\tilde{E}_{ij}$ of the matrix $\tilde{E}$ is the impact of the error term ε_{kj} on the endogenous variable Y_{ki} if $\beta = \tilde{\beta}$. Correspondingly, if a variable WY_k is constructed from the endogenous variable Y_k using an arbitrary weight matrix W, then the element $(W\tilde{E})_{ij}$ of the matrix $W\tilde{E}$ is the impact of the error term ε_{kj} on the element $(WY_k)_i$ of the constructed variable WY_k.

3.1.2 STEP 2 of 5. Construct the Potential Instrumental Variable

Next a matrix $\tilde{G}$ is constructed so that the constructed variable $\tilde{G}Y_k$ is a valid instrumental variable for the endogenous regressor GY_k if $\beta = \tilde{\beta}$. We refer to the constructed variable as a *potential instrumental variable*. The instrument relevance condition requires that the constructed variable $\tilde{G}Y_k$ and the endogenous regressor GY_k are correlated. For the instrument exogeneity condition to hold when $\beta = \tilde{\beta}$, it is sufficient that if $\beta = \tilde{\beta}$ then the impact of the element ε_{ki} of the error term ε_k on the element $(\tilde{G}Y_k)_i$ of the constructed variable $\tilde{G}Y_k$ is equal to the average impact of the same element ε_{ki} of the error term ε_k on all the elements $(\tilde{G}Y_k)_1, (\tilde{G}Y_k)_2, ..., (\tilde{G}Y_k)_n$ of the constructed variable $\tilde{G}Y_k$. Using the definition of the matrix $\tilde{E}$ constructed in Step 1, this instrument exogeneity condition can be written formally as

$$\left(\tilde{G}\tilde{E}\right)_{ii} = \frac{1}{n}\sum_{j=1}^{n}\left(\tilde{G}\tilde{E}\right)_{ji} \text{ for all } i \in \{1, ..., n\}\,. \tag{5}$$

In Step 4 we show that if the constructed matrix $\tilde{G}$ satisfies this instrument exogeneity condition (5) and $\beta = \tilde{\beta}$, then each element ε_{ki} of the error term ε_k has no impact on the network fixed effect demeaned value of corresponding element $(\tilde{G}Y_k)_i$ of the constructed

variable $\tilde{G}Y_k$. It is in this sense that the constructed matrix $\tilde{G}$ is conditionally balanced and, accordingly, we refer to the constructed matrix $\tilde{G}$ as a *conditionally balanced interaction structure*.[9] Construction of the matrix $\tilde{G}$ is addressed in Section 3.3. For now we simply assume the matrix $\tilde{G}$ and the associated potential instrumental variable $\tilde{G}Y_k$ have been constructed. In Step 4 we also show that when $\beta = \tilde{\beta}$ the constructed variable $\tilde{G}Y_k$ is a valid instrumental variable for the endogenous regressor GY_k.

3.1.3 STEP 3 of 5. Obtain First-Stage Estimates

Next the endogenous social interaction variable GY_k, exogenous regressors GX_k and X_k, dummy variables representing network fixed effects, and the constructed potential instrumental variable $\tilde{G}Y_k$ are used to estimate the first-stage regression equation

$$GY_k = \theta_k + \theta_{\tilde{G}Y}\tilde{G}Y_k + \theta_{GX}GX_k + \theta_X X_k + v_k, \tag{6}$$

where θ_k denote network fixed effects, v_k is the error term, and $\theta_{\tilde{G}Y}$, θ_{GX}, and θ_X are coefficients on the observed explanatory variables. We denote the Least Squares estimates of parameters θ_k, $\theta_{\tilde{G}Y}$, θ_{GX}, and θ_X by $\hat{\theta}_k$, $\hat{\theta}_{\tilde{G}Y}$, $\hat{\theta}_{GX}$, and $\hat{\theta}_X$, respectively.

Parameter estimates and the associated predicted values of the dependent variable GY_k from the first-stage regression (6) depend on the constructed matrix $\tilde{G}$. Consequently, we denote predicted values from the first-stage regression equation (6) by $\widehat{GY}_k^{\tilde{G}}$ and the associated residuals $GY_k - \widehat{GY}_k^{\tilde{G}}$ by $\hat{v}_k^{\tilde{G}}$. The predicted values are calculated as

$$\widehat{GY}_k^{\tilde{G}} = \hat{\theta}_k + \hat{\theta}_{\tilde{G}Y}\tilde{G}Y_k + \hat{\theta}_{GX}GX_k + \hat{\theta}_X X_k. \tag{7}$$

3.1.4 STEP 4 of 5. Obtain Second-Stage Estimates

Next a second-stage regression equation is estimated to obtain the potential instrumental variable estimate of the parameter β. Using the definition $\hat{v}_k^{\tilde{G}} \equiv GY_k - \widehat{GY}_k^{\tilde{G}}$, we can substitute $GY_k = \widehat{GY}_k^{\tilde{G}} + \hat{v}_k^{\tilde{G}}$ for GY_k in the original regression equation (1) to obtain the second-stage regression equation

$$Y_k = \alpha_k + \beta\widehat{GY}_k^{\tilde{G}} + \delta GX_k + \gamma X_k + \beta\hat{v}_k + \varepsilon_k. \tag{8}$$

[9] In contrast, the original interaction structure G is (almost always) unbalanced in this sense and thus the endogenous social interaction regressor GY_k is (almost always) correlated with the error term ε_k.

In this second-stage regression equation (8) the regressors are $\widehat{GY}_k^{\tilde{G}}$, GX_k, X_k, and the dummy variables representing the network fixed effects α_k. The error term is now $\beta\hat{v}_k + \varepsilon_k$. The Least Squares estimator of the coefficient β from the second-stage regression equation (8) is the potential instrumental variable estimator of the endogenous social interaction parameter β. This estimate of the parameter β depends in part on the imputed value $\tilde{\beta}$ used in constructing the conditionally balanced interaction structure $\tilde{G}$. Accordingly, we denote the potential instrumental variable estimator by $\hat{\beta}_{IV(\tilde{\beta})}$.

Before proceeding to Step 5, in which a test statistic for the null hypothesis $\beta = \tilde{\beta}$ is constructed from the potential instrumental variable estimate $\hat{\beta}_{IV(\tilde{\beta})}$, we examine the properties of the estimator $\hat{\beta}_{IV(\tilde{\beta})}$. Specifically, we now show that if $\beta = \tilde{\beta}$ then the estimator $\hat{\beta}_{IV(\tilde{\beta})}$ is a consistent estimator of the parameter β. It is of course important to also examine behavior of the estimator $\hat{\beta}_{IV(\tilde{\beta})}$ when $\beta \neq \tilde{\beta}$. This other side of identification is addressed computationally and analytically in Sections 4.2 and 5, respectively.

We first establish the following result.

Lemma 1. *The potential instrumental variable estimator* $\hat{\beta}_{IV(\tilde{\beta})}$ *satisfies*

$$\text{plim}_N \hat{\beta}_{IV(\tilde{\beta})} = \beta + \frac{\sum_i \left(\left(\tilde{G}E \right)_{ii} - \frac{1}{n} \sum_j \left(\tilde{G}E \right)_{ji} \right) \sigma_i^2}{R} \times \text{plim}_N \hat{\theta}_{\tilde{G}Y}, \tag{9}$$

where R *is a strictly positive constant.*

Proof. See Appendix 1.

We assume that the easily testable instrument relevance condition $\text{plim}_N \hat{\theta}_{\tilde{G}Y} \neq 0$ holds for the constructed potential instrumental variable $\tilde{G}Y_k$. Substituting the known matrix $\tilde{E} = (I - \tilde{\beta}G)^{-1}$ for the unknown matrix E in the above expression (9) yields the probability limit for the estimator $\hat{\beta}_{IV(\tilde{\beta})}$ when $\beta = \tilde{\beta}$. The result shows that if $\beta = \tilde{\beta}$ then the asymptotic bias $\text{plim}_N(\hat{\beta}_{IV(\tilde{\beta})} - \beta)$ of the estimator $\hat{\beta}_{IV(\tilde{\beta})}$ is zero if and only if

$$\sum_{i=1}^{n} \left[\left(\left(\tilde{G}\tilde{E} \right)_{ii} - \frac{1}{n} \sum_{j=1}^{n} \left(\tilde{G}\tilde{E} \right)_{ji} \right) \times \sigma_i^2 \right] = 0. \tag{10}$$

A sufficient condition for this condition (10) to hold regardless of how the unknown variances

σ_i^2 of the error terms ε_{ki} are distributed across individuals $i \in \{1, ..., n\}$ is that

$$\left(\tilde{G}\tilde{E}\right)_{ii} - \frac{1}{n}\sum_{j=1}^{n}\left(\tilde{G}\tilde{E}\right)_{ji} = 0 \text{ for all } i \in \{1, ..., n\}. \tag{11}$$

This condition (11) holds by construction as is the same condition as condition (5) that was used in the construction of the matrix $\tilde{G}$ in Step 2. Consequently, if $\beta = \tilde{\beta}$ the potential instrumental variable estimator $\hat{\beta}_{IV\left(\tilde{\beta}\right)}$ is a consistent estimator of the parameter β.

3.1.5 STEP 5 of 5. Calculate the Test Statistic

In the final step the test statistic $\frac{\hat{\beta}_{IV(\tilde{\beta})} - \tilde{\beta}}{\sqrt{\widehat{var}\left(\hat{\beta}_{IV(\hat{\beta})}\right)}}$, where $\widehat{var}(\hat{\beta}_{IV\left(\hat{\beta}\right)})$ is an estimate of the variance of the potential instrumental variable estimator $\hat{\beta}_{IV\left(\hat{\beta}\right)}$, is calculated and compared with the associated critical value to determine whether the null hypothesis H_0: $\beta = \tilde{\beta}$ is rejected. As the estimator $\hat{\beta}_{IV\left(\hat{\beta}\right)}$ is an asymptotically unbiased estimator of the parameter β when $\beta = \tilde{\beta}$ (see Step 4), the relevant critical values are obtained from the standard normal distribution provided that the constructed variable $\tilde{G}Y_k$ is relevant enough to serve as an instrumental variable i.e. has a sufficiently high first-stage F-statistic.

3.2 Construction of the Confidence Set Estimate for β

The $(1-p)\,\%$ confidence set estimate for the parameter β is constructed as the union of all those feasible values $\tilde{\beta} \in \Omega(\beta)$ of the parameter β for which the null hypothesis H_0: $\beta = \tilde{\beta}$ is not rejected at the $p\%$ level in Step 5. We denote the resulting $(1-p)\,\%$ confidence set estimate for the parameter β by $\Psi_{(1-p)}(\beta)$. By construction the test against each the null hypothesis H_0: $\beta = \tilde{\beta}$ in Step 5 has the nominal size p. Consequently, the probability that the true value of the endogenous social interaction parameter β is in the constructed confidence set estimate $\Psi_{(1-p)}(\beta)$ is arbitrarily close to $1-p$ when the number of observations N on networks is large enough. We now state this result as a formal proposition.

Proposition 1. *When a potential instrumental variable can be constructed for all feasible values of the endogenous social interaction parameter β, the confidence set estimate $\Psi_{(1-p)}(\beta)$ obtained using the potential instrumental variable estimation method satisfies the property*

$$\lim_{N\to\infty} P\left(\beta \in \Psi_{(1-p)}(\beta)\right) = 1 - p. \tag{12}$$

The other side of identification–what is the behavior of the estimator $\hat{\beta}_{IV(\tilde{\beta})}$ when $\tilde{\beta} \neq \beta$ and which feasible values will not lie in the constructed confidence set $\Psi_{(1-p)}(\beta)$–is addressed computationally and formally in Sections 4.2 and 5, respectively. A direct implication of Proposition 1 is that the potential instrumental variable estimation method yields a test for any null hypothesis H_0: $\beta = \beta_0$ with an asymptotically correct size: the result $\beta_0 \notin \Psi_{(1-p)}(\beta)$ indicates that the null hypothesis H_0: $\beta = \beta_0$ should be rejected. The qualifier in the beginning of Proposition 1 relates to the fact that for some network structures G a conditionally balanced interaction structure cannot be constructed for all feasible values of the parameter β. As this does not occur in our applications, discussion of the construction of the confidence set estimate in this case is postponed until Section 6.4.2.

3.3 Construction of Conditionally Balanced Interaction Structures

A constructed conditionally balanced interaction structure $\tilde{G}$ associated with feasible value $\tilde{\beta}$ must be such that the constructed variable $\tilde{G}Y_k$ is correlated with the endogenous social interaction regressor GY_k and that the constructed variable $\tilde{G}Y_k$ is conditionally balanced in the sense that when $\beta = \tilde{\beta}$ the constructed variable also satisfies the instrument exogeneity condition (5). Typically there are multiple such matrices $\tilde{G}$.[10]

To find an interaction structure $\tilde{G}$ that resembles the original network structure sufficiently, we construct an objective function which is increasing in the value of those elements $\tilde{G}_{ij}$ for which the corresponding element G_{ij} of the original interaction structure G is positive and decreasing in the value of those elements $\tilde{G}_{ij}$ for which the corresponding element G_{ij} of the original network structure G is zero. Formally, we find each conditionally balanced interaction structure $\tilde{G}$ as a solution to the constrained optimization problem

$$\tilde{G} \equiv \arg\max_{\tilde{G}} \sum_i \sum_j \left\{ \tilde{G}_{ij} G_{ij} - \tilde{G}_{ij} \chi_{[G_{ij}=0]} c \right\} \tag{13}$$

[10] Optimal $\tilde{G}$ would maximize correlation between network fixed effect demeaned values of the variables GY_k and $\tilde{G}Y_k$ subject to the instrument exogeneity condition (5). However, analytical tractability of such an optimization problem is questionable because the objective function would be nonlinear in the n^2 unknown parameters of matrix $\tilde{G}$ and involve taking the expectation of a nonlinear function of the error terms ε_k which have unknown distributions. Moreover, the objective function would contain an imputed equilibrium interaction matrix $\tilde{E}$, which would imply that when the null hypothesis $\beta = \tilde{\beta}$ does not hold the constructed potential instrumental variable $\tilde{G}Y_k$ might have a weak relationship with the instrumented variable GY_k. This would lead to weak power against alternative hypotheses and, consequently, the constructed confidence set estimates would be wide. For these reasons we favor the ad hoc approach in the text.

$$\text{s.t. } \left(\tilde{G}\tilde{E}\right)_{ii} - \frac{1}{n}\sum_{j=1}^{n}\left(\tilde{G}\tilde{E}\right)_{ji} = 0 \text{ for all } i \in \{1, ..., n\} \tag{14a}$$

$$\tilde{G}_{ij} \in [l_b, u_b] \text{ for all } i \in \{1, ..., n\} \text{ and for all } j \in \{1, ..., n\} \tag{14b}$$

$$\sum_i \sum_j \tilde{G}_{ij} = 1, \tag{14c}$$

where

$$\chi_{[G_{ij}=0]} = \begin{cases} 1 \text{ if } G_{ij} = 0 \\ 0 \text{ otherwise} \end{cases} \tag{15}$$

is an indicator function that indicates whether G_{ij} is zero, and where l_b, u_b and c are parameters set by the researcher. The term $\tilde{G}_{ij}G_{ij}$ (the term $\tilde{G}_{ij}\chi_{[G_{ij}=0]}c$) in the objective function (13) implies that the value of the objective function is increasing (decreasing, provided that $c > 0$) in the value of those elements $\tilde{G}_{ij}$ for which the corresponding element G_{ij} in the original network structure G is positive (zero). The first constraint (14a) is the instrument exogeneity condition (5). The second constraint (14b) reduces computational complexity and prevents potential solutions for which the values of a small number of elements in $\tilde{G}$ are either very high or very low.[11] The third constraint (14c) is a normalization.[12] In our applications and simulations we set $l_b = -1$, $u_b = 1$ and $c = 0.01$. To limit the scope of this paper, analyses of how the choice of parameters l_b, u_b and c and the choice of the objective function influence the properties of the proposed estimation method are left for future research.

4 Applications and Monte Carlo Simulations

In this section we first apply the proposed estimation method to the study of subjective income and college expectations. We then demonstrate the identification power of the proposed method with Monte Carlo simulations of Erdös-Rényi and small-world networks.

4.1 Applications with Add Health Data

We employ the National Longitudinal Study of Adolescent Health (Add Health) data and the proposed estimation method to estimate the impacts of variables such as gender, race,

[11] When the values of a small number of elements in $\tilde{G}$ are either very high or very low the value of the constructed variable $\tilde{G}Y_k$ is mostly determined by the value of the outcome variable Y_k for a relatively small number of individuals in each network and, consequently, the constructed variable $\tilde{G}Y_k$ is only a weak instrument for the endogenous regressor GY_k.

[12] The normalization implies that even when $c = 0$ there is a cost to increasing the value of any element $\tilde{G}_{ij}$ for which the corresponding element G_{ij} of the original network structure G is zero.

grade point average, and parents' education and occupational status on subjective income and college completion expectations among adolescents. To focus the main text on the methodology, we relegate to the Background Appendix a review of the literature on subjective expectations and discussion of the substantive motivation for analyses of income and college expectations. The Implementation Appendix in turn contains the details on the implementation of the proposed method; these details include the employed grid of feasible values for the parameter β, calculation of confidence set estimates for other model parameters, and how the analysis accounts for the fact that network structure is different across observations.

The main purpose of the applications is to demonstrate the feasibility of the proposed method. In this regard the results show that 1) conditionally balanced network interaction matrices can be constructed for interaction structures that occur in the real world (i.e. in the Add Health survey), 2) that in these real world applications the constructed potential instrumental variables are strong enough to satisfy the instrument relevance condition, and 3) that in these real world applications the proposed method has identification power in the sense that the obtained confidence set estimates are proper subsets of the set of feasible values. A secondary aim of the applications is the comparison of the obtained confidence set estimates with the confidence intervals obtained using the existing estimation method. In this regard the results demonstrate an advantage of the proposed method: it allows researchers to remain relatively agnostic about what exogenous effects should be included in the model as the results do not hinge on exclusion restrictions on variables constructed from observed exogenous variables.

4.1.1 Data

We employ the Add Health data because these data include friendship denomination data for most of the subjects in the study, which along with information on the grade and school of each respondent facilitates the construction of grade-school level network interaction matrices G. For this same reason the Add Health data were also employed in the three closest related studies by Bramoulle et al. (2009) Lin (2010) and Lee et al. (2010), which examined the determinants of club participation, grade point average, and hours spent on homework/watching tv, respectively. In part due to the ubiquity of applications that have utilized the Add Health data, we relegate to the Data Appendix the description of these data as well as the discussion of sample construction and associated descriptive statistics.

4.1.2 Caveats to Causal Interpretation

An important caveat to a causal interpretation of the estimates arises from the voluntary nature of friendship networks in the Add Health data. Among the many analyses of peer effects that have relied on Add Health data only a handful of papers, such as Calvo-Armengol et al. (2009) and Fletcher and Ross (2009), have incorporated analyses of the network formation problem and, as Brock et al. (2010) mention, systematic methods for the analysis of endogenous network formation do not yet exist. Another important caveat to a causal interpretation of the estimates arises from simultaneous determination of variables such as GPA, college expectations, and income expectations.

Both types of caveats apply also to the applications presented in each of the three closest related methodological contributions by Bramoulle et al. (2009), Lin (2010) and Lee et al. (2010). Moreover, these caveats do not interfere with the methodological objectives mentioned in the second paragraph of Section 4.1. Even if the estimates do not reflect causal effects the applications demonstrate that for network structures and variables observed in the real world the proposed method is feasible and the estimates obtained using the existing and proposed methods have certain relative properties.

4.1.3 Results for the Proposed Method

Results of applications of the proposed method to analyses of the determinants of subjective college completion and income expectations are shown in Tables 1 and 2, respectively. Confidence set estimates that do not include the value zero are indicated in bold. Columns 1-3 in each table show results for the fully networked sample in which every individual both nominated a friend and is nominated as a friend. Column 4 in each table shows results for the is/has friend sample in which every individual either nominated a friend or was nominated as a friend.[13] Before discussing these results, we discuss Figure 1 which illustrates how the confidence set estimate for the endogenous social interaction parameter β is derived. The sub-figures on the left and on the right in Figure 1 correspond to the specifications and results in Column 3 of the Table 1 and in Column 3 of Table 2, respectively.

The top sub-figures in Figure 1 depict the first-stage F-statistic as a function of the imputed value $\tilde{\beta}$. The first-stage F-statistic is high even when the imputed value $\tilde{\beta}$ is the

[13]See the Data Appendix for the construction of the two samples. In the fully networked sample the network-fixed effect demeaned value of the endogenous social interaction regressor GY_k is endogenous for every individual in every network. In contrast, in the is/has friend sample the endogenous social interaction regressor GY_k is exogenous for individuals who are not nominated as friend by any individual and for individuals who do not report any friends in the sample. The use of the fully networked sample demonstrates that the proposed estimation method can work even when these additional sources of identification are absent.

Table 1. 95% Confidence Set Estimates for **College Expectations**.

Dependent Variable: **y_College_Expectations**
Estimation Method: **Proposed (**Potential Instrumental Variable).

	(1)	(2)	(3)	(4)
Endogenous Effect				
$Gy\_College\_Expectations$	$[\mathbf{0.62}, \mathbf{0.68}]$	$[\mathbf{0.37}, \mathbf{0.46}]$	$[\mathbf{0.38}, \mathbf{0.49}]$	$[\mathbf{0.59}, \mathbf{0.73}]$
Network Fixed Effect	yes	yes	yes	yes
Own Characteristics				
x_GPA		$[\mathbf{0.48}, \mathbf{0.58}]$	$[\mathbf{0.51}, \mathbf{0.58}]$	$[\mathbf{0.50}, \mathbf{0.58}]$
x_Age		$[\mathbf{-0.19}, \mathbf{-0.10}]$	$[\mathbf{-0.19}, \mathbf{-0.10}]$	$[\mathbf{-0.20}, \mathbf{-0.12}]$
x_Female		$[\mathbf{0.20}, \mathbf{0.27}]$	$[\mathbf{0.24}, \mathbf{0.33}]$	$[\mathbf{0.29}, \mathbf{0.38}]$
x_Asian		$[-0.00, 0.14]$	$[-0.03, 0.15]$	$[\mathbf{0.04}, \mathbf{0.19}]$
x_Black		$[\mathbf{0.19}, \mathbf{0.31}]$	$[\mathbf{0.07}, \mathbf{0.28}]$	$[\mathbf{0.23}, \mathbf{0.39}]$
$x\_Hispanic$		$[-0.07, 0.10]$	$[-0.08, 0.10]$	$[-0.09, 0.08]$
$x\_Mom\_College$		$[\mathbf{0.16}, \mathbf{0.25}]$	$[\mathbf{0.17}, \mathbf{0.25}]$	$[\mathbf{0.22}, \mathbf{0.30}]$
$x\_Dad\_College$		$[\mathbf{0.17}, \mathbf{0.27}]$	$[\mathbf{0.18}, \mathbf{0.27}]$	$[\mathbf{0.21}, \mathbf{0.29}]$
$x\_Mom\_Professional$		$[\mathbf{0.01}, \mathbf{0.09}]$	$[\mathbf{0.01}, \mathbf{0.10}]$	$[\mathbf{0.21}, \mathbf{0.29}]$
$x\_Dad\_Professional$		$[-0.05, 0.06]$	$[-0.05, 0.06]$	$[-0.05, 0.06]$
$x\_Mom\_White\_Collar$		$[\mathbf{0.09}, \mathbf{0.20}]$	$[\mathbf{0.09}, \mathbf{0.20}]$	$[\mathbf{0.09}, \mathbf{0.19}]$
$x\_Dad\_White\_Collar$		$[\mathbf{0.04}, \mathbf{0.16}]$	$[\mathbf{0.04}, \mathbf{0.16}]$	$[\mathbf{0.06}, \mathbf{0.17}]$
$x\_Parent\_Homemaker$		$[-0.05, 0.07]$	$[-0.05, 0.07]$	$[-0.04, 0.07]$
$x\_Parent\_Military$		$[\mathbf{0.03}, \mathbf{0.20}]$	$[\mathbf{0.02}, \mathbf{0.20}]$	$[-0.01, 0.17]$
Exogenous/Contextual Effects				
Gx_GPA			$[-0.24, 0.02]$	$[\mathbf{-0.53}, \mathbf{-0.06}]$
Gx_Age			$[-0.08, 0.07]$	$[\mathbf{-0.24}, \mathbf{-0.10}]$
Gx_Female			$[\mathbf{-0.26}, \mathbf{-0.08}]$	$[\mathbf{-0.43}, \mathbf{-0.14}]$
Gx_Asian			$[-0.09, 0.18]$	$[-0.25, 0.04]$
Gx_Black			$[-0.05, 0.23]$	$[-0.38, 0.01]$
$Gx\_Hispanic$			$[-0.11, 0.14]$	$[-0.15, 0.07]$
$Gx\_Mom\_College$			$[-0.17, 0.03]$	$[-0.24, 0.02]$
$Gx\_Dad\_College$			$[\mathbf{-0.22}, \mathbf{-0.00}]$	$[\mathbf{-0.34}, \mathbf{-0.04}]$
$Gx\_Mom\_Professional$			$[-0.13, 0.03]$	$[-0.12, 0.04]$
$Gx\_Dad\_Professional$			$[-0.08, 0.12]$	$[-0.09, 0.13]$
$Gx\_Mom\_White\_Collar$			$[-0.04, 0.17]$	$[-0.18, 0.11]$
$Gx\_Dad\_White\_Collar$			$[-0.03, 0.18]$	$[-0.11, 0.16]$
$Gx\_Parent\_Homemaker$			$[-0.06, 0.13]$	$[-0.15, 0.07]$
$Gx\_Parent\_Military$			$[-0.28, 0.04]$	$[-0.23, 0.10]$
Sample	Fully Networked	Fully Networked	Fully Networked	Has/Is Friend
Number of Networks (N)	486	486	486	489
Observations ($\sum_{k=1}^{N} n_k$)	42,827	42,827	42,827	60,495
Network Parameters ($\sum_{k=1}^{N} n_k^2$)	6,287,061	6,287,061	6,287,061	12,298,399

Table 2. 95% Confidence Set Estimates for **Income Expectations**.

Dependent Variable: **y_Income_Expectations**
Estimation Method: **Proposed (**Potential Instrumental Variable).

	(1)	(2)	(3)	(4)
Endogenous Effect				
Gy_Income_Expectations	[**0.29**, **0.40**]	[**0.04**, **0.15**]	[**0.03**, **0.16**]	[**0.07**, **0.22**]
Network Fixed Effect	yes	yes	yes	yes
Own Characteristics				
x_College_Expectations		[**0.26**, **0.29**]	[**0.26**, **0.29**]	[**0.26**, **0.29**]
x_GPA		[**0.15**, **0.23**]	[**0.16**, **0.22**]	[**0.14**, **0.19**]
x_Age		[**−0.09**, **−0.00**]	[**−0.08**, **−0.00**]	[**−0.09**, **−0.02**]
x_Female		[**−0.22**, **−0.12**]	[**−0.21**, **−0.10**]	[**−0.20**, **−0.11**]
x_Asian		[**−0.32**, **−0.10**]	[**−0.26**, **−0.02**]	[**−0.29**, **−0.09**]
x_Black		[−0.18, 0.01]	[−0.23, 0.02]	[−0.14, 0.04]
x_Hispanic		[**−0.26**, **−0.07**]	[**−0.23**, **−0.06**]	[**−0.23**, **−0.09**]
x_Mom_College		[**0.02**, **0.14**]	[**0.03**, **0.14**]	[**0.00**, **0.10**]
x_Dad_College		[−0.05, 0.08]	[−0.04, 0.08]	[−0.01, 0.09]
x_Mom_Professional		[−0.00, 0.11]	[−0.00, 0.11]	[**0.01**, **0.10**]
x_Dad_Professional		[**0.04**, **0.19**]	[**0.04**, **0.19**]	[**0.07**, **0.20**]
x_Mom_White_Collar		[**0.03**, **0.15**]	[**0.03**, **0.14**]	[**0.04**, **0.14**]
x_Dad_White_Collar		[**0.01**, **0.15**]	[**0.01**, **0.15**]	[**0.02**, **0.14**]
x_Parent_Homemaker		[**0.08**, **0.20**]	[**0.08**, **0.20**]	[**0.07**, **0.17**]
x_Parent_Military		[−0.10, 0.09]	[−0.10, 0.09]	[−0.04, 0.13]
Exogenous/Contextual Effects				
Gx_College_Expectations			[−0.08, 0.05]	[**0.07**, **0.17**]
Gx_GPA			[−0.06, 0.08]	[−0.08, 0.08]
Gx_Age			[−0.06, 0.06]	[−0.07, 0.02]
Gx_Female			[−0.13, 0.04]	[−0.10, 0.05]
Gx_Asian			[**−0.35**, **−0.06**]	[−0.30, 0.04]
Gx_Black			[−0.12, 0.16]	[−0.16, 0.06]
Gx_Hispanic			[−0.22, 0.07]	[−0.23, 0.02]
Gx_Mom_College			[−0.03, 0.16]	[−0.11, 0.07]
Gx_Dad_College			[−0.18, 0.01]	[−0.07, 0.09]
Gx_Mom_Professional			[−0.10, 0.08]	[−0.13, 0.06]
Gx_Dad_Professional			[−0.15, 0.11]	[−0.09, 0.16]
Gx_Mom_White_Collar			[−0.04, 0.15]	[**0.01**, **0.20**]
Gx_Dad_White_Collar			[−0.09, 0.16]	[−0.15, 0.09]
Gx_Parent_Homemaker			[−0.12, 0.09]	[−0.07, 0.13]
Gx_Parent_Military			[−0.06, 0.27]	[−0.10, 0.18]
Sample	Fully Networked	Fully Networked	Fully Networked	Has/Is Friend
Number of Networks (N)	486	486	486	489
Observations ($\sum_{k=1}^{N} n_k$)	42,827	42,827	42,827	60,495
Network Parameters ($\sum_{k=1}^{N} n_k^2$)	6,287,061	6,287,061	6,287,061	12,298,399

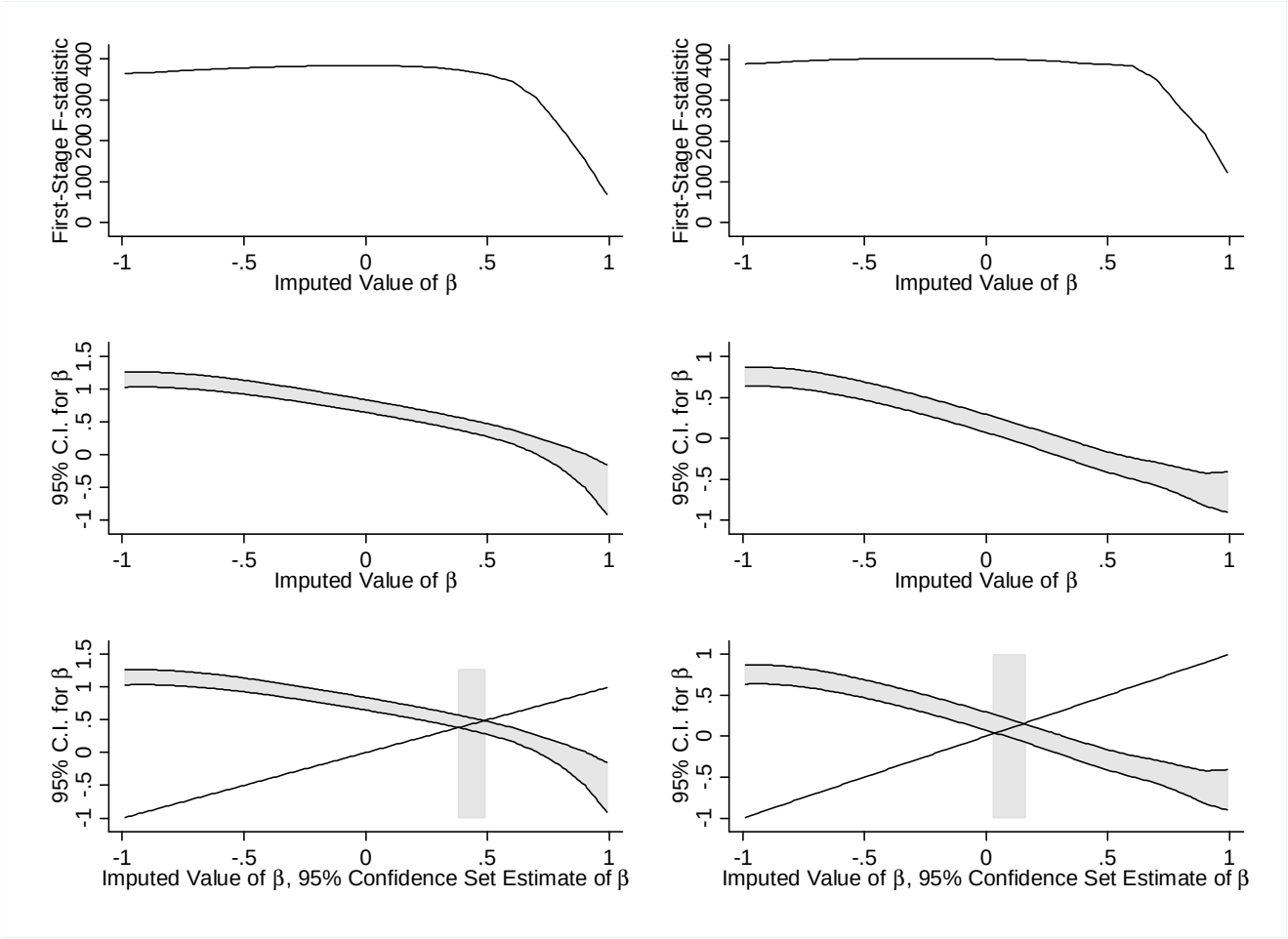

Figure 1: Two illustrations of the derivation of the confidence set estimate.

highest value (0.99) in the selected grid of feasible values for the parameter β. Moreover, the first-stage F-statistic decreases as $\tilde{\beta}$ increases. This feature is as expected in the Add Health application. The underlying reason, as we discuss next, is that there is considerable variation in the number of friends across individuals within networks in the Add Health data.

The influence that the outcome of friends have on an individual is proportional to β, whereas the influence that the outcome of individuals who are k connections removed from an individual have on the individual is proportional to β^k. When β is small, the relative magnitudes of β and β^k are very different and, consequently, variation in equilibrium influence within a given network is relatively high. In contrast, when β is close to 1 the relative magnitudes of β and β^k are very similar. Consequently, when β is very close to 1 the equilibrium influence of an individual on other individuals in the network is almost the same across

the network. Therefore, for any imputed value $\tilde{\beta}$ close to 1, an associated conditionally balanced interaction structure $\tilde{G}_k$ must have approximately as many connections for every individual for the constructed interaction structure be conditionally balanced.[14] However, as there is within network variation in the number of friends in the Add Health data, for any imputed value $\tilde{\beta}$ close to 1 constructed conditionally balanced interaction structures $\tilde{G}_k$ will then be quite different from original interaction structures G_k, which in turn implies that the relationship between the endogenous regressor G_kY_k and the constructed variable $\tilde{G}_kY_k$ is weaker when the imputed value $\tilde{\beta}$ is closer to 1. In contrast, when each individual has the same number of friends, conditionally balanced interaction structures $\tilde{G}_k$ associated with imputed value $\tilde{\beta}$ close to 1 will be very similar to the corresponding original interaction structures G_k.

The middle sub-figures in Figure 1 depict the 95% confidence interval indicated by each potential instrumental variable estimate $\hat{\beta}_{IV(\tilde{\beta})}$ and the associated standard error as a function of the associated imputed value $\tilde{\beta}$. The width of the confidence interval for each imputed value $\tilde{\beta}$ reflects the associated value of the first-stage F-statistic. The upper and lower bounds of the confidence interval are both roughly linear functions of $\tilde{\beta}$ (except near extreme feasible values of parameter β), which justifies the use of linear interpolation to determine the confidence interval for feasible values of parameter β that are not in the selected grid of feasible values. Moreover, the upper and lower bounds of the confidence interval decrease as a function of the imputed value $\tilde{\beta}$. This is a desirable feature from the perspective of identification as it implies a fixed point property that yields point identification of the parameter β (see Section 5).

The bottom sub-figures in Figure 1 depict the confidence intervals against the 45-degree line $\beta = \tilde{\beta}$. The 95% confidence set estimate $\Psi_{0.95}(\beta)$ of parameter β is constructed as the union of all those feasible values β for which the 95% confidence interval contains the feasible value $\beta = \tilde{\beta}$, where $\tilde{\beta}$ is the imputed value used in the construction of the 95% confidence interval. Accordingly, in each bottom sub-figure in Figure 1, the 95% confidence set estimate is constructed as the union of all those feasible values β for which the 45-degree line $\beta = \tilde{\beta}$ is between the confidence interval mapping. The shaded vertical bar in each bottom sub-figure in Figure 1 indicates this 95% confidence set estimate.

From a methodological perspective the results in Tables 1 and 2 demonstrate that the proposed estimation method works. The constructed conditionally balanced interaction struc-

[14] If for an imputed value $\tilde{\beta}$ close to 1 the constructed interaction structure $\tilde{G}_k$ for network k had more connections for individual i in the network than others, the influence of the error term ε_{ki} on the element $(\tilde{G}_kY_k)_i$ of the constructed variable would be higher than the average influence of the ε_{ki} is on the constructed variable $\tilde{G}_kY_k$. This would violate the condition that any constructed $\tilde{G}_k$ must be conditionally balanced.

tures $\tilde{G}_k$ are similar enough with the original network structures G_k that the associated constructed potential instrumental variables $\tilde{G}_k Y_k$ are similar enough with the endogenous social interaction variable $G_k Y_k$ to satisfy the instrument relevance condition. Moreover, while we have postponed formal analysis of the the other side of identification–behavior of the potential instrumental variable estimator $\hat{\beta}_{IV(\tilde{\beta})}$ when $\tilde{\beta} \neq \beta$–until Section 5, the results here indicate that the proposed method has identification power. For both outcome variables the obtained 95% confidence set estimate of the endogenous social interaction parameter β is quite narrow across the different specifications. The different specifications illustrate that the proposed method can be applied in the absence of any explanatory variables (Column 1), in the presence of only endogenous social interaction effects (Column 2), and in the presence of both exogenous and endogenous social interaction effects (Columns 3 and 4).

Substantively the results in Tables 1 and 2 suggest that the endogenous social interaction effect is positive and quite large for subjective college expectations and smaller but still positive for subjective income expectations. With respect to own characteristics, the results are as expected.[15] With respect to exogenous social interaction effects, the results show a negative association between an individual's friends' parents' college education and the individual's college expectations. One potential causal explanation is that the influence that friends' college expectations have on an individual's college expectations is heterogenous in the individual's friends' parents' education: a decision to go to college by a friend whose parents are not college educated might be more informative for an individual about the importance of college than the same decision by a friend whose parents are college educated. Similar reasoning can be applied to support a causal interpretation of the result that when college expectations is the outcome variable the coefficient on friend being female is negative.

4.1.4 Comparison with Results for the Existing Estimation Method

Results obtained using the estimation method proposed by Bramoulle et al. (2009) and Lin (2010) are shown in Table 3. In Panel A friends' friends' characteristics ($G_k G_k x\_^*$) are used as instruments for the endogenous social interaction variable $G_k Y_k$. In Panel B the exogenous effects ($G_k x\_$) are not included in the model and thus also friends' characteristics ($G_k x\_^*$)

[15] GPA is positively associated with both college and income expectations, and college expectations are positively associated with income expectations. Being female is positively associated with college expectations but negatively associated with income expectations. Age is negatively associated with both college and income expectations, reflecting the fact that the oldest individuals in each grade are those that have repeated a grade. Consistent with the results in Jacob and Wilder (2010), college expectations are positively associated with being black. In contrast, but still as expected, being in any racial minority is negatively associated with income expectations. Also as expected is the finding that many of the variables measuring parents' educational and occupational status are positively associated with college and income expectations.

Table 3. 95% Confidence Intervals for **College and Income Expectations**.

Estimation Method: **Existing (**Instruments in Panel A: $GGx\_^*$ [Panel B: $Gx\_^*$ and $GGx\_^*$] of included $\_x^*$).

	(1)	(2)	(3)	(4)
Dependent Variable	**y_College_Exp.**	**y_College_Exp.**	**y_Income_Exp.**	**y_Income_Exp.**
Sample	Fully Networked	Is/Has Friend	Fully Networked	Is/Has Friend
Network Fixed Effect	yes	yes	yes	yes
PANEL A				
Endogenous Effect	[**0.59**, **1.00**]	[**0.34**, **0.56**]	[−0.15, 0.74]	[**0.03**, **0.61**]
Own Characteristics				
x_College_Expectations			[**0.26**, **0.29**]	[**0.26**, **0.29**]
x_GPA	[**0.49**, **0.56**]	[**0.52**, **0.58**]	[**0.16**, **0.22**]	[**0.13**, **0.19**]
x_Age	[−**0.17**, −**0.09**]	[−**0.20**, −**0.14**]	[−**0.08**, −**0.00**]	[−**0.08**, −**0.02**]
x_Female	[**0.24**, **0.33**]	[**0.29**, **0.36**]	[−**0.20**, −**0.10**]	[−**0.19**, −**0.11**]
x_Asian	[−0.04, 0.16]	[**0.05**, **0.18**]	[−0.24, 0.00]	[−**0.27**, −**0.08**]
x_Black	[**0.05**, **0.28**]	[**0.23**, **0.37**]	[−**0.22**, −**0.05**]	[−0.14, 0.04]
x_Hispanic	[−0.08, 0.11]	[−0.09, 0.06]	[−**0.22**, −**0.05**]	[−**0.22**, −**0.08**]
x_Mom_College	[**0.16**, **0.24**]	[**0.23**, **0.30**]	[**0.03**, **0.13**]	[**0.00**, **0.10**]
x_Dad_College	[**0.18**, **0.27**]	[**0.22**, **0.30**]	[−0.04, 0.08]	[−0.01, 0.09]
x_Mom_Professional	[**0.02**, **0.10**]	[**0.01**, **0.09**]	[−0.00, 0.11]	[**0.01**, **0.10**]
x_Dad_Professional	[−0.06, 0.06]	[−0.04, 0.06]	[**0.04**, **0.18**]	[**0.07**, **0.19**]
x_Mom_White_Collar	[**0.07**, **0.17**]	[**0.11**, **0.19**]	[**0.03**, **0.14**]	[**0.04**, **0.13**]
x_Dad_White_Collar	[**0.02**, **0.14**]	[**0.07**, **0.17**]	[**0.02**, **0.15**]	[**0.02**, **0.14**]
x_Parent_Homemaker	[−0.06, 0.07]	[−0.04, 0.07]	[**0.08**, **0.20**]	[**0.07**, **0.17**]
x_Parent_Military	[**0.01**, **0.19**]	[**0.00**, **0.17**]	[−0.10, 0.09]	[−0.04, 0.13]
Exogenous/Contextual Effects				
Gx_College_Expectations			[−0.20, 0.06]	[−0.18, 0.01]
Gx_GPA	[−**0.49**, −**0.18**]	[−**0.27**, −**0.07**]	[−0.13, 0.07]	[−0.12, 0.04]
Gx_Age	[−0.02, 0.14]	[−**0.16**, −**0.10**]	[−0.05, 0.07]	[−**0.11**, −**0.01**]
Gx_Female	[−**0.34**, −**0.16**]	[−**0.30**, −**0.14**]	[−0.12, 0.11]	[−0.09, 0.06]
Gx_Asian	[−0.15, 0.13]	[−0.17, 0.04]	[−0.32, 0.01]	[−0.23, 0.08]
Gx_Black	[−0.16, 0.13]	[−0.21, 0.00]	[−0.09, 0.19]	[−0.14, 0.07]
Gx_Hispanic	[−0.09, 0.17]	[−0.13, 0.06]	[−0.20, 0.14]	[−0.18, 0.06]
Gx_Mom_College	[−**0.25**, −**0.06**]	[−0.13, 0.02]	[−0.05, 0.14]	[−0.12, 0.05]
Gx_Dad_College	[−**0.32**, −**0.12**]	[−**0.21**, −**0.05**]	[−0.11, 0.07]	[−0.08, 0.09]
Gx_Mom_Professional	[−0.16, 0.02]	[−0.11, 0.03]	[−0.11, 0.07]	[−0.14, 0.04]
Gx_Dad_Professional	[−0.10, 0.12]	[−0.09, 0.10]	[−0.18, 0.09]	[−0.10, 0.13]
Gx_Mom_White_Collar	[−0.12, 0.11]	[−0.06, 0.12]	[−0.06, 0.14]	[−0.00, 0.17]
Gx_Dad_White_Collar	[−0.08, 0.15]	[−0.03, 0.17]	[−0.11, 0.14]	[−0.16, 0.06]
Gx_Parent_Homemaker	[−0.08, 0.14]	[−0.10, 0.08]	[−0.14, 0.08]	[−0.08, 0.11]
Gx_Parent_Military	[−0.31, 0.06]	[−0.17, 0.10]	[−0.06, 0.27]	[−0.11, 0.16]
First-Stage F-statistic	10.18	21.49	1.80	3.40
PANEL B (w/out Exog. Effects)				
Endogenous Effect	[**0.30**, **0.37**]	[**0.05**, **0.06**]	[**0.04**, **0.13**]	[**0.01**, **0.03**]
First-Stage F-statistic	69.55	4119.68	60.56	4343.63
Overidentification test, p-value	0.000	0.000	0.001	0.018

are employed as instruments for the endogenous social interaction variable G_kY_k.

Results in Table 3 are qualitatively similar to the corresponding results in Tables 1 and 2. From a substantive perspective such robustness of results to the choice of the estimation approach is naturally a desired feature in any application. The results in Panel A of Table 3 come with the caveat that for these analyses the first-stage F-statistic is low. This aspect of the analysis highlights the main disadvantage of the existing estimation method. When the included exogenous effects are weak, the existing estimation approach does not have good identification power because a low value of the coefficient on the variable $G_kx\_some\_variable_k$ implies that the variables $G_kG_kx\_some\_variable_k$ and $G_kG_kG_kx\_some\_variable_k$ are weak instruments. The existing approach is similarly disadvantaged when exogenous effects are not included in the model but own characteristics have only a weak impact on the outcome variable. While variables $G_kx\_^*$ can then be used as an instruments for the endogenous social interaction variable, provided of course that the exclusion restrictions are correct, a low value of the coefficient on $x\_some\_variable_k$ implies that the variable $G_kx\_some\_variable_k$ is a weak instrument. A related problem with using $G_kG_kx\_some\_variable_k$ and $G_kG_kG_kx\_some\_variable_k$ as instruments is that variation in such variables is often limited for reasons discussed at the end of the Data Appendix.

As expected, the value of the first-stage F-statistic is higher when exogenous effects are excluded (see Panel B of Table 3). However, a standard overidentification test (Hansen's J-statistic) indicates that not all of the exclusion restrictions are valid. While in each case there may exist a specification in which the exclusion restrictions are valid and instruments are relevant enough, such a specification search comes with the cost of additional potential complications arising from issues such as nested hypothesis testing and power of overidentification tests. In comparison, when the proposed approach is applied the researcher can remain relatively agnostic about which exogenous effects should be included in the specification as identification in the proposed approach is not based on instruments constructed from the observed exogenous variables and the associated exclusion restrictions.

4.2 Monte Carlo Simulations

We now present results from Monte Carlo simulations of Erdös-Rényi and small-world networks. In an Erdös-Rényi network links are i.i.d. and any two nodes in the network are connected with probability p (Erdös and Rényi, 1959). A small-world network of size N is generated by starting from $\frac{N}{k}$ disjoint sub-groups of size k. Initially all pairs of nodes within each sub-groups are connected, while no pairs of nodes in different sub-networks are

connected. Then, with probability p each link in the initial network structure is reconfigured at random (Watts and Strogatz, 1998). Previously, Bramoulle et al. (2009) have employed these network structures to examine the performance of the existing estimator. We employ these network structures to demonstrate that the proposed estimation method has identification power and to demonstrate that an important determinant of the relative properties of the two estimation methods is the strength of the exogenous effect.

We set the network size at $n = 100$ and the number of observations on the network at $N = 50$. For small-world networks we set the sub-group size at $k = 10$. Observations are generated using model (1) with one exogenous variable X_k. We set $X_{ki} \sim N(0, 2)$, $\varepsilon_{ki} \sim N(0, 1)$ and $\alpha_k \sim N(0, 1)$, where both X_{ki} and ε_{ki} are independently distributed both within and across networks. We set coefficients on variables X_k and GX_k at $\gamma = 1$ and $\delta = 0.5$, respectively. We show four graphs that depict coverage probabilities for $\beta = -0.5$, $\beta = 0$ and $\beta = 0.5$, and the power function against the null hypothesis H_0: $\beta = 0$. The coverage probabilities and power functions are calculated using the 95% confidence set estimate for the proposed method and the 95% confidence interval for the existing method.[16] In the existing method the variables GGX_k and $GGGX_k$ are used as instruments for the endogenous regressor GY_k.[17]

Figures 2 and 3 depict results for undirected Erdös-Rényi and small-world networks, respectively, as a function of the network structure parameter p. Coverage probabilities for the proposed method (solid line) indicate that the width of the associated confidence set estimates is generally increasing in the parameter p. A similar pattern holds for the existing method (dashed line), except in the case $\beta = -0.5$. When $\beta = -0.5$, the chosen parameters satisfy $\beta\gamma + \delta = 0$ and, as was shown by Bramoulle et al. (2009), the existing method then has no identification power. For an Erdös-Rényi network with $p = 0.9$ the number of connections in the network is very high and there is very little variation in the regressors GY_k and GX_k because the set of friends is then almost the same for individuals in the network and because the variables GY_k and GX_k are then constructed as averages of so many random variables that variation in them is small. Consequently, for an Erdös-Rényi network with $p = 0.9$ neither method has much (if any) identification power. The proposed method has identification power in all other cases. And more specifically, the results suggest that in most

[16] We only show results for undirected networks; results for directed versions of these networks are similar. We set the grid of feasible values of the parameter β as $[-0.99, -0.9, -0.8, -0.7, ..., 0.9, 0.99]$. Each estimate is based on 400 replications. The network structure G is the same across the replications. This limits computational demands and yields estimates for a given network instead of the average estimate across networks generated with the same parameters. Bramoulle et al. (2009) apply the same approach.

[17] Network fixed effects are of course included in both estimation approaches. Efficiency of both approaches can be improved by using the estimators developed in Keleijan and Prucha (2010) and Lin and Lee (2010).

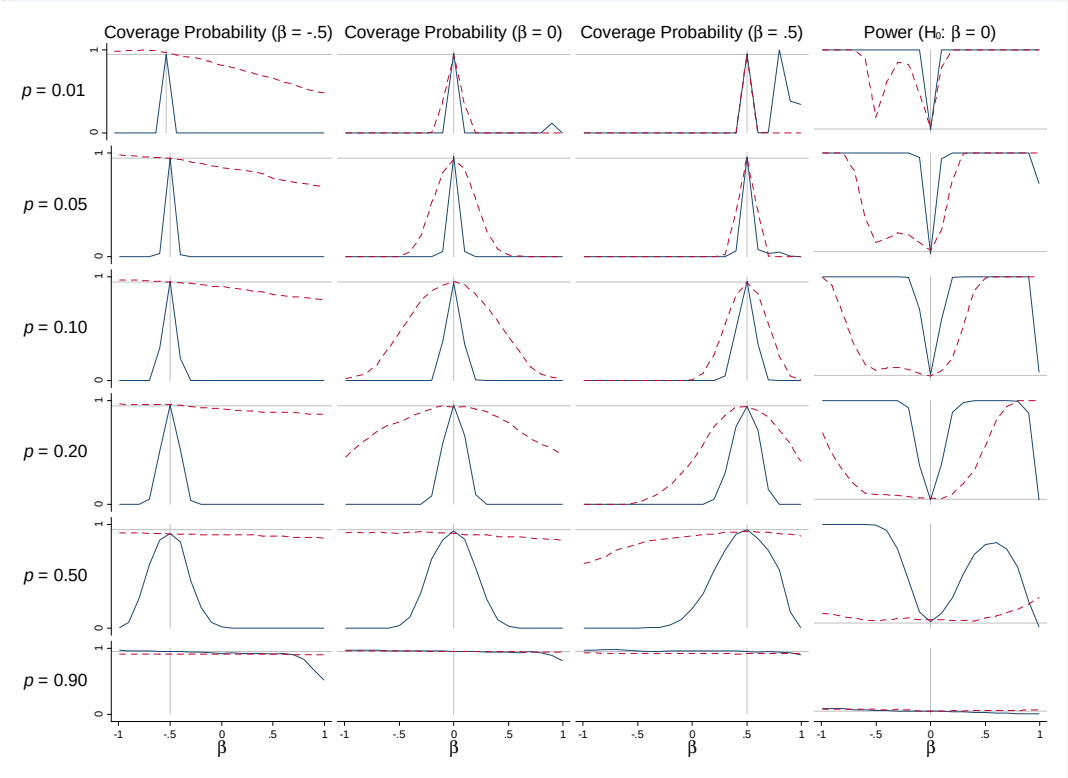

Figure 2: Proposed (solid line) and existing (dashed line) methods in Erdös-Rényi networks.

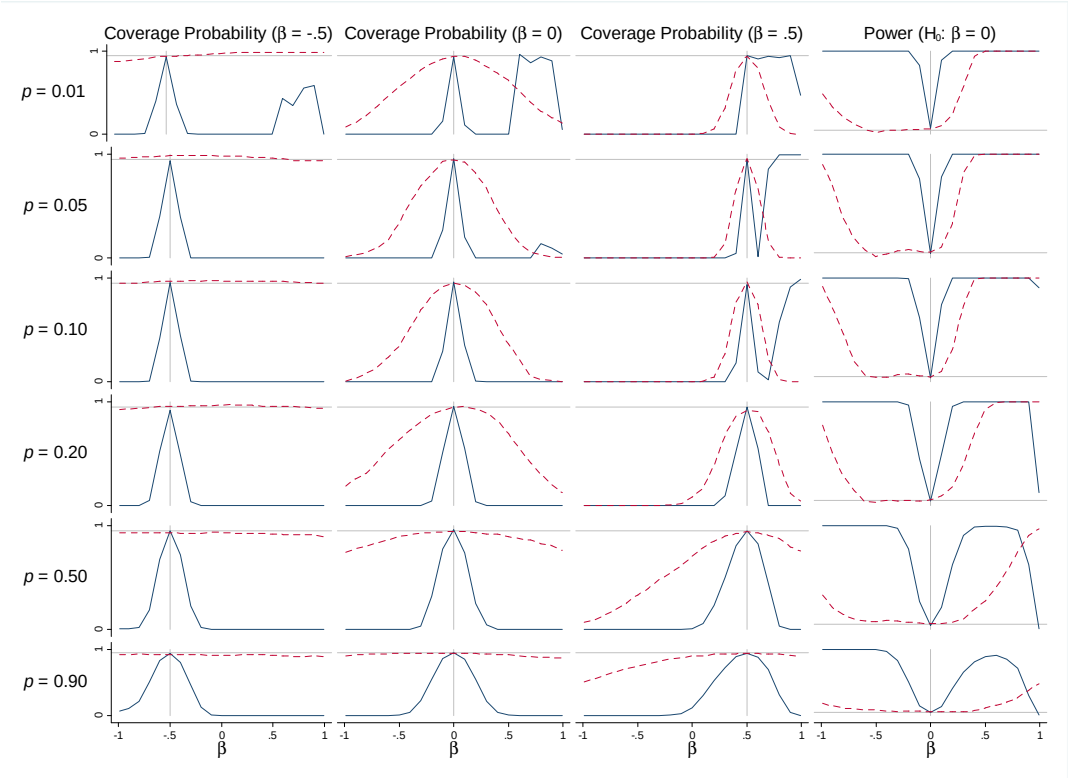

Figure 3: Proposed (solid line) and existing (dashed line) methods in small-world networks.

of these cases the proposed method yields point identification, although results for small-world networks with $p \leq 0.1$ and $\beta = .5$ demonstrate that the confidence set estimate is not always continuous. The results also show that the power of the proposed method can be poor against the null hypothesis H_0: $\beta = 0$ when the true parameter value is $\beta = 0.99$. This is because $\beta = 0.99$ implies that there is very little variation in the endogenous regressor GY_k. Comparison of the results for the proposed and existing methods show that the proposed estimation method can yield much more precise estimates than the existing method.

Figure 4 depicts results for Erdös-Rényi networks with $p = 0.05$ as a function of the exogenous effect δ. Results for the case $\beta = 0$ demonstrate the general feature that the lower is the absolute value of the exogenous effect δ the better is the performance of the proposed method relative to the existing method as a low absolute value of the coefficient δ on the variable GX_k implies that the variables GGX_k and $GGGX_k$ are only weak instruments. Results for the cases $\beta = -0.5$ and $\beta = 0.5$ are driven by the fact that, as discussed above, the existing method has little identification power when $\beta\gamma + \delta \approx 0$

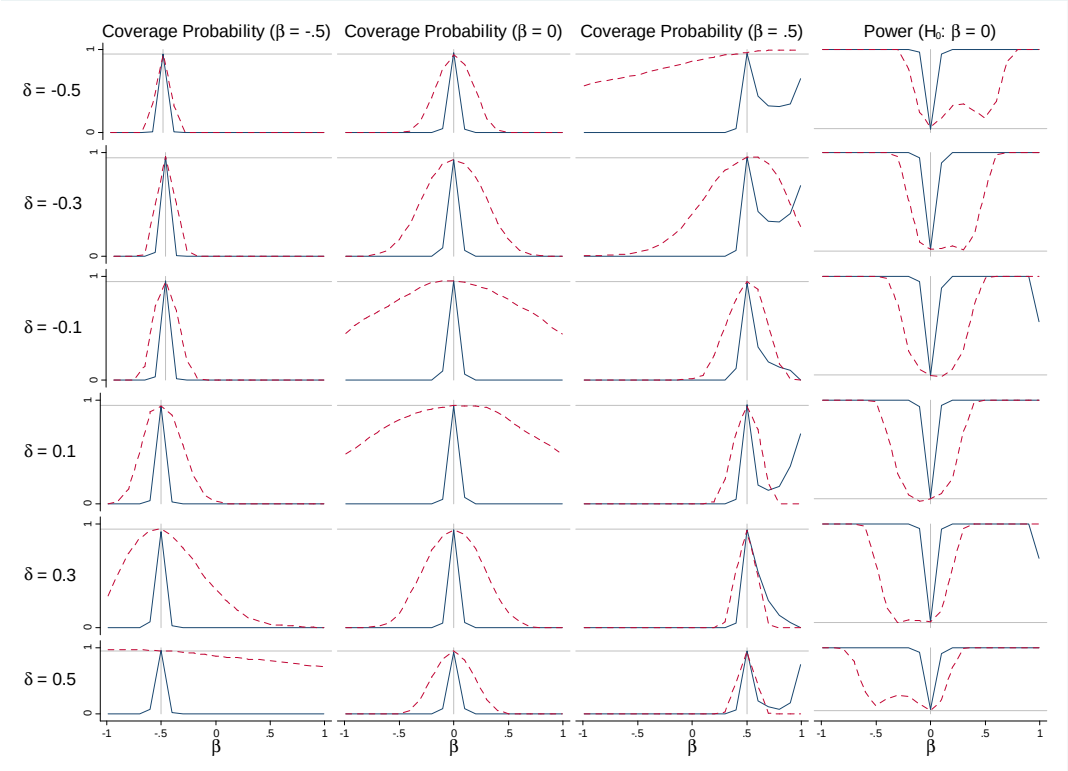

Figure 4: Proposed (solid line) and existing (dashed line) methods in Erdös-Rényi networks as a function of the exogenous effect δ.

5 Identification

Proposition 1 shows that the confidence set estimate $\Psi_{1-p}(\beta)$ contains the true value of the endogenous social interaction parameter β at the chosen probability $1-p$. We now consider the other side of identification formally: we examine behavior of the confidence set estimate $\Psi_{1-p}(\beta)$ in relation to feasible values $\tilde{\beta} \neq \beta$ when the number of observations N on the network increases without limit. It is of course desirable that the probability that the confidence set estimate $\Psi_{1-p}(\beta)$ contains any such feasible value $\tilde{\beta} \neq \beta$ is as small as possible.

We provide sufficient conditions for point identification and for two types of partial identification. Of course, either point or partial identification may occur even when these sufficient conditions do not hold. The conditions for partial identification imply that the parameter β is point identified for a connected subset of the set of feasible values $\Omega(\beta)$ and possibly only set identified for each of the one or two remaining connected subsets of the set of feasible values $\Omega(\beta)$. Accordingly, we refer to the conditions for partial identification as conditions for "partial point identification". These conditions are most useful when the feasible value $\beta = 0$ lies in the point identified subset as the data then provide at least qualitative information on the presence and direction of the endogenous effect β (i.e. whether $\beta < 0$, $\beta = 0$ or $\beta > 0$). Throughout the analysis we assume that the estimator $\hat{\beta}_{IV(\tilde{\beta})}$ can be constructed for each feasible value of the endogenous social interaction parameter β. This condition is easily verifiable and holds in our Add Health applications. For expositional convenience, we assume that the set of feasible of values for parameter β is $\Omega(\beta) = (-1, 1)$.

For point identification we provide sufficient conditions under which in the limit the probability that the constructed confidence set estimate $\Psi_{1-p}(\beta)$ contains any feasible value $\tilde{\beta} \neq \beta$ is zero. Formally, we provide sufficient conditions under which the property

$$\lim_{N\to\infty} P\left(\tilde{\beta} \in \Psi_{1-p}(\beta) \mid \tilde{\beta} \neq \beta\right) = 0 \text{ for all } \beta \in \Omega(\beta) \text{ and for all } \tilde{\beta} \in \Omega(\beta) \qquad (16)$$

holds. Together with Proposition 1 this condition (16) implies that in the limit the probability that a feasible value $\tilde{\beta} \in \Omega(\beta)$ is contained in the confidence set estimate $\Psi_{1-p}(\beta)$ is non-zero if and only if the feasible value $\tilde{\beta}$ is the true value β.

Under the first set of conditions for partial identification point identification occurs for all feasible values $\beta \in \Omega_1(\beta)$, where $\Omega_1(\beta) = (-1, \bar{\beta}]$, and set identification occurs for all feasible values $\beta \in \Omega_2(\beta)$, where $\Omega_2(\beta) = (\bar{\beta}, 1)$. The sets $\Omega_1(\beta)$ and $\Omega_2(\beta)$ are thus mutually disjoint connected subsets of the set $\Omega(\beta)$ and satisfy the property $\Omega_1(\beta) \cup \Omega_2(\beta) = \Omega(\beta)$.

Formally, we provide sufficient conditions under which the properties

$$\lim_{N\to\infty} P\left(\tilde{\beta} \in \Psi_{1-p}(\beta) \mid \tilde{\beta} \neq \beta\right) = 0 \text{ for all } \beta \in \Omega(\beta) \text{ and for all } \tilde{\beta} \in \Omega_1(\beta) \tag{17}$$

and

$$\lim_{N\to\infty} P\left(\tilde{\beta} \in \Psi_{1-p}(\beta) \mid \tilde{\beta} \neq \beta\right) = 0 \text{ for all } \beta \notin \Omega_2(\beta) \text{ and for all } \tilde{\beta} \in \Omega_2(\beta) \tag{18}$$

hold. Together with Proposition 1 these conditions (17) and (18) imply that in the limit the probability that a feasible value $\tilde{\beta} \in \Omega_1(\beta)$ is contained in the confidence set estimate $\Psi_{1-p}(\beta)$ is non-zero if and only if the feasible value $\tilde{\beta}$ is the true value β, and that in the limit the probability that a feasible value $\tilde{\beta} \in \Omega_2(\beta)$ is contained in the confidence set estimate $\Psi_{1-p}(\beta)$ is positive only if the true value β is in the same subset of feasible values $\Omega_2(\beta)$.

Under the second set of conditions for partial identification point identification occurs for all feasible values $\beta \in \Omega_1(\beta)$, where $\Omega_1(\beta) = [-\underline{\beta}, \bar{\beta}]$, and set identification occurs for all $\beta \in \Omega_0(\beta)$, where $\Omega_0(\beta) = (-1, \underline{\beta})$, and also for all $\beta \in \Omega_2(\beta)$, where $\Omega_2(\beta) = (\bar{\beta}, 1)$. The sets $\Omega_0(\beta)$, $\Omega_1(\beta)$ and $\Omega_2(\beta)$ are thus mutually disjoint connected subsets of the set $\Omega(\beta)$ and satisfy the property $\Omega_0(\beta) \cup \Omega_1(\beta) \cup \Omega_2(\beta) = \Omega(\beta)$. Formally, we provide sufficient conditions under which properties (17), (18) and

$$\lim_{N\to\infty} P\left(\tilde{\beta} \in \Psi_{1-p}(\beta) \mid \tilde{\beta} \neq \beta\right) = 0 \text{ for all } \beta \notin \Omega_0(\beta) \text{ and for all } \tilde{\beta} \in \Omega_0(\beta) \tag{19}$$

hold. Together with Proposition 1 these conditions (17), (18) and (19) imply that in the limit the probability that a feasible value $\tilde{\beta} \in \Omega_1(\beta)$ is contained in the confidence set estimate $\Psi_{1-p}(\beta)$ is positive if and only if the feasible value $\tilde{\beta}$ is the same as the true value β, and that in the limit the probability that a feasible value $\tilde{\beta} \in \Omega_0(\beta)$ (feasible value $\tilde{\beta} \in \Omega_2(\beta)$) is contained in the confidence set estimate $\Psi_{1-p}(\beta)$ is positive only if the true value β is in the same subset of feasible values $\Omega_0(\beta)$ (subset of feasible values $\Omega_2(\beta)$).

Given that the estimator $\hat{\beta}_{IV(\tilde{\beta})}$ determines whether feasible value $\tilde{\beta}$ is included in the confidence set estimate $\Psi_{1-p}(\beta)$, a necessary and sufficient condition for the condition (16) for point identification to hold is

$$\text{plim}_N \hat{\beta}_{IV(\tilde{\beta})} \neq \beta \text{ for all } \beta \neq \tilde{\beta} \text{ and for all } \tilde{\beta} \in \Omega(\beta). \tag{20}$$

A set of sufficient conditions for this condition (20) to hold are

$$\text{plim}_N \hat{\beta}_{IV(\tilde{\beta})} \leq \beta \text{ for all } \beta < \tilde{\beta} \text{ and for all } \tilde{\beta} \in \Omega(\beta) \tag{21}$$

and

$$\text{plim}_N \hat{\beta}_{IV(\tilde{\beta})} \geq \beta \text{ for all } \beta > \tilde{\beta} \text{ and for all } \tilde{\beta} \in \Omega(\beta). \tag{22}$$

Condition (22) implies that for all imputed values $\tilde{\beta}$ lower than the true value β, in the limit the estimate will be above the true value. Condition (21) in turn implies that for all imputed values $\tilde{\beta}$ higher than the true value β, in the limit the estimate will be below the true value. Together the conditions (21) and (22) thus imply a fixed point property that yields point identification.

Similarly, necessary and sufficient conditions for the conditions (17) and (18) for the first type of partial point identification to hold are

$$\text{plim}_N \hat{\beta}_{IV(\tilde{\beta})} \neq \beta \text{ for all } \beta \neq \tilde{\beta} \text{ and for all } \tilde{\beta} \in \Omega_1(\beta) \tag{23}$$

and

$$\text{plim}_N \hat{\beta}_{IV(\tilde{\beta})} \neq \beta \text{ for all } \beta \notin \Omega_2(\beta) \text{ and for all } \tilde{\beta} \in \Omega_2(\beta), \tag{24}$$

respectively. A set of sufficient conditions for these conditions (23) and (24) to hold are

$$\text{plim}_N \hat{\beta}_{IV(\tilde{\beta})} \leq \beta \text{ for all } \beta < \tilde{\beta} \text{ and for all } \tilde{\beta} \in \Omega_1(\beta) \tag{25}$$

$$\text{plim}_N \hat{\beta}_{IV(\tilde{\beta})} \geq \beta \text{ for all } \beta > \tilde{\beta} \text{ and for all } \tilde{\beta} \in \Omega_1(\beta), \tag{26}$$

and

$$\text{plim}_N \hat{\beta}_{IV(\tilde{\beta})} \leq \beta \text{ for all } \beta \notin \Omega_2(\beta) \text{ and for all } \tilde{\beta} \in \Omega_2(\beta). \tag{27}$$

For all $\beta \in \Omega_1(\beta)$ these conditions (25), (26) and (27) imply a fixed point property–similar to the fixed point property implied by the conditions (21) and (22) for point identification–which implies point identification for all $\beta \in \Omega_1(\beta)$. For all $\beta \in \Omega_2(\beta)$ the condition (26) implies that for imputed values $\tilde{\beta} \in \Omega_1(\beta)$, which are lower than any true value $\beta \in \Omega_2(\beta)$, in the limit the estimate is higher than the true value, and thus in the limit the confidence set estimate does not include any imputed value $\tilde{\beta} \in \Omega_1(\beta)$ implying a form of partial identification for all $\beta \in \Omega_2(\beta)$.

Similarly, necessary and sufficient conditions for the conditions (17), (18) and (19) for

the second type of partial point identification to hold are conditions (23), (24) and

$$\text{plim}_N \hat{\beta}_{IV\left(\tilde{\beta}\right)} \neq \beta \text{ for all } \beta \notin \Omega_0\left(\beta\right) \text{ and for all } \tilde{\beta} \in \Omega_0\left(\beta\right), \tag{28}$$

respectively. A set of sufficient conditions for these conditions (23), (24) and (28) to hold are conditions (25), (26), (27) and

$$\text{plim}_N \hat{\beta}_{IV\left(\tilde{\beta}\right)} \geq \beta \text{ for all } \beta \notin \Omega_0\left(\beta\right) \text{ and for all } \tilde{\beta} \in \Omega_0\left(\beta\right). \tag{29}$$

Given the expression (74) for the probability limit of the estimator $\hat{\beta}_{IV\left(\tilde{\beta}\right)}$, a sufficient condition for the first condition (21) for point identification to hold is

$$\left(\tilde{G}E\right)_{ii} - \frac{1}{n}\sum_{j=1}^{n}\left(\tilde{G}E\right)_{ji} \leq 0 \text{ for all } i \in \{1,...,n\} \text{ and for all } \beta < \tilde{\beta} \text{ and for all } \tilde{\beta} \in \Omega\left(\beta\right), \tag{30}$$

and a sufficient condition for the second condition (22) for point identification to hold is

$$\left(\tilde{G}E\right)_{ii} - \frac{1}{n}\sum_{j=1}^{n}\left(\tilde{G}E\right)_{ji} \geq 0 \text{ for all } i \in \{1,...,n\} \text{ and for all } \beta > \tilde{\beta} \text{ and for all } \tilde{\beta} \in \Omega\left(\beta\right). \tag{31}$$

Whether conditions (30) and (31) hold in a given application depends on the constructed conditionally balanced interaction structures $\tilde{G}$ and on the equilibrium interaction matrix E, which in turn depends on the known interaction structure G and the unknown parameter β. While the parameter β is unknown, it is straightforward to verify whether conditions (30) and (31) hold for all feasible values in the selected grid of feasible values of the parameter β.

In our experience the sufficient conditions (30) and (31) rarely hold, which is why we place more emphasis on slightly more stringent conditions for point identification and the corresponding conditions for partial point identification. More specifically, we establish conditions for point and partial identification when the variances σ_i^2 of the error terms ε_{ki} within each network satisfy the property

$$\sigma_{MAX}^2 \leq c \times \sigma_{MIN}^2, \tag{32}$$

where σ_{MAX}^2 and σ_{MIN}^2, respectively, are the highest and lowest variances among the unconditional error term variances σ_i for the observations $i \in \{1,2,...,n\}$ within a network and $c \geq 1$ is a constant. While the variance ratio restriction parameter c is unknown, its interpretation is straightforward.

When the variance ratio assumption (32) holds, the expression (74) for the probability limit of the estimator $\hat{\beta}_{IV(\tilde{\beta})}$ implies that sufficient conditions for the conditions (21) and (22) for point identification to hold are

$$\sum_i \left\{ (I_{\Delta_i<0} + c \times I_{\Delta_i>0}) \left[\left(\tilde{G}E\right)_{ii} - \frac{1}{n}\sum_{j=1}^{n} \left(\tilde{G}E\right)_{ji} \right] \right\} \leq 0 \text{ for all } \beta < \tilde{\beta} \text{ and for all } \tilde{\beta} \in \Omega(\beta) \tag{33}$$

and

$$\sum_i \left\{ (c \times I_{\Delta_i<0} + I_{\Delta_i>0}) \left[\left(\tilde{G}E\right)_{ii} - \frac{1}{n}\sum_{j=1}^{n} \left(\tilde{G}E\right)_{ji} \right] \right\} \geq 0 \text{ for all } \beta > \tilde{\beta} \text{ and for all } \tilde{\beta} \in \Omega(\beta), \tag{34}$$

where indicator functions $I_{\Delta_i<0}$ and $I_{\Delta_i>0}$ are defined as $I_{\Delta_i<0} = 1$ if $(\tilde{G}E)_{ii} - \frac{1}{n}\sum_{j=1}^{n}(\tilde{G}E)_{ji} < 0$ and $I_{\Delta_i<0} = 0$ otherwise, and $I_{\Delta_i>0} = 1$ if $(\tilde{G}E)_{ii} - \frac{1}{n}\sum_{j=1}^{n}(\tilde{G}E)_{ji} > 0$ and $I_{\Delta_i>0} = 0$ otherwise. Condition (33) implies that even if $\sigma_i^2 = \sigma_{MIN}^2$ for all i for which $(\tilde{G}E)_{ii} - \frac{1}{n}\sum_{j=1}^{n}(\tilde{G}E)_{ji} < 0$ and $\sigma_i^2 = \sigma_{MAX}^2$ for all i for which $(\tilde{G}E)_{ii} - \frac{1}{n}\sum_{j=1}^{n}(\tilde{G}E)_{ji} > 0$, the sum $\sum_i((\tilde{G}E)_{ii} - \frac{1}{n}\sum_j(\tilde{G}E)_{ji})\sigma_i^2$, which determines the sign of the bias for the estimator $\hat{\beta}_{IV(\tilde{\beta})}$, is still negative. Interpretation of condition (34) is analogous and omitted.

Corresponding conditions for partial identification are derived similarly. When the variance ratio assumption (32) holds, the expression (74) for the probability limit of the estimator $\hat{\beta}_{IV(\tilde{\beta})}$ implies that sufficient conditions for the conditions (25), (26) and (27) for the first type of partial point identification to hold are

$$\sum_i \left\{ (I_{\Delta_i<0} + c \times I_{\Delta_i>0}) \left[\left(\tilde{G}E\right)_{ii} - \frac{1}{n}\sum_{j=1}^{n} \left(\tilde{G}E\right)_{ji} \right] \right\} \leq 0 \text{ for all } \beta < \tilde{\beta} \text{ and for all } \tilde{\beta} \in \Omega_1(\beta), \tag{35}$$

$$\sum_i \left\{ (c \times I_{\Delta_i<0} + I_{\Delta_i>0}) \left[\left(\tilde{G}E\right)_{ii} - \frac{1}{n}\sum_{j=1}^{n} \left(\tilde{G}E\right)_{ji} \right] \right\} \geq 0 \text{ for all } \beta > \tilde{\beta} \text{ and for all } \tilde{\beta} \in \Omega_1(\beta) \tag{36}$$

and

$$\sum_i \left\{ (I_{\Delta_i<0} + c \times I_{\Delta_i>0}) \left[\left(\tilde{G}E\right)_{ii} - \frac{1}{n}\sum_{j=1}^{n} \left(\tilde{G}E\right)_{ji} \right] \right\} \leq 0 \text{ for all } \beta \notin \Omega_2(\beta) \text{ and for all } \tilde{\beta} \in \Omega_2(\beta), \tag{37}$$

and that sufficient conditions for the conditions (25), (26), (27) and (29) for the second type

of partial point identification are conditions (35), (36), (37) and

$$\sum_i \left\{ \left(I_{\Delta_i<0} + c \times I_{\Delta_i>0} \right) \left[\left(\tilde{G}E \right)_{ii} - \frac{1}{n} \sum_{j=1}^{n} \left(\tilde{G}E \right)_{ji} \right] \right\} \geq 0 \text{ for all } \beta \notin \Omega_0(\beta) \text{ and for all } \tilde{\beta} \in \Omega_0(\beta). \tag{38}$$

We now summarize the above results in a formal proposition.

Proposition 2. *When a potential instrumental variable can be constructed for all feasible values of the endogenous social interaction parameter β and the variance ratio assumption (32) holds, the confidence set estimate $\Psi_{1-p}(\beta)$ obtained using the potential instrumental variable estimation method yields point identification under conditions (33) and (34), the first type of partial point identification under conditions (35), (36) and (37), and the second type of partial point identification under conditions (35), (36), (37) and (38).*

Table 4 shows to what extent the three sets of identification conditions in Proposition 2 hold in the Add Health data as a function of the parameter c in the variance ratio condition (32). When the variances of error terms within a given network are the same (unrestricted) $c = 1$ ($c = \infty$). In calculating these frequencies we only consider those cases of partial identification for which point identification occurs at $\beta = 0$ i.e. we require that $\underline{\beta} < 0$ ($\underline{\beta} < 0$ and $\bar{\beta} > 0$) for type 1 (type 2) partial identification to occur. The results show that while the conditions for point identification rarely hold regardless of the variance ratio parameter c, the conditions for the second type of partial identification hold for 73% of the 489 networks even when the variance ratio parameter c is as high as 10. Moreover, this percentage increases from 73% to 94% when the 200 smallest networks are excluded from the analysis. The sufficient conditions for identification established in Proposition 2 are thus relevant for network structures observed in the real world. We reiterate that because the conditions established in Proposition 2 are only sufficient conditions, point or partial identification may occur even when these conditions do not hold.

	Point	Partial, Type 1	Partial, Type 2
$c = 1$	18%	71% [median $\bar{\beta}$: 0.9]	94% [median $\underline{\beta}, \bar{\beta} : -1, 0.8$]
$c = 5$	4%	19% [median $\bar{\beta}$: 0.8]	81% [median $\underline{\beta}, \bar{\beta} : -0.8, 0.6$]
$c = 10$	2%	6% [median $\bar{\beta}$: 0.8]	73% [median $\underline{\beta}, \bar{\beta} : -0.6, 0.5$]
$c = 100$	0 %	0% [median $\bar{\beta}$: 0.8]	8% [median $\underline{\beta}, \bar{\beta} : -0.2, 0.3$]
$c = \infty$	0 %	0% [median $\bar{\beta}$: -]	0% [median $\underline{\beta}, \bar{\beta}$: -]

Table 4. Percentage of networks in the is/has friend sample of Add Health data for which identification conditions in Proposition 2 hold.

6 Extensions and Directions for Future Research

6.1 Within-Network Fixed Effects

Within-network fixed effects capture unobserved shocks that are common to individuals in each mutually exclusive subset of individuals in a given network and which may be correlated with observed exogenous variables. For example, when network structure represents friendships between students in the same school, grade-specific within-network fixed effects can capture the influence of unobserved shocks that are common to all students within each grade in a given school.[18]

Extending the proposed estimation method to allow for within-network fixed effects is straightforward, with within-network fixed effects formally defined as variables that are constant across all observations in each mutually exclusive subset of observations in each network. Let S_i index the within-network subset of observation i in a network. Define n_{S_i} as the number of observations in the within-network subset S_i. Define the characteristic variable

$$\chi_{[S_j=S_i]} = \begin{cases} 1 \text{ if } S_j = S_i \\ 0 \text{ otherwise} \end{cases} \tag{39}$$

to capture whether observation j in the network belongs in the same within-network subset as observation i. A conditionally balanced network structure $\tilde{G}$ associated with the feasible value $\tilde{\beta}$ is now constructed in such a way that if $\tilde{\beta} = \beta$ then the influence of error term ε_{ki} on the constructed variable $(\tilde{G}Y_k)_i$ is the same as the average influence of the error term ε_{ki} is for all observations in the within-network subset in which observation i belongs. Formally, conditionally balanced interaction structures $\tilde{G}$ are constructed using condition

$$\left(\tilde{G}\tilde{E}\right)_{ii} = \frac{1}{n_{S_i}} \sum_{j=1}^{n} \left\{ \left(\tilde{G}\tilde{E}\right)_{ji} \times \chi_{[S_j=S_i]} \right\} \text{ for all } i \in \{1, ..., n\}. \tag{40}$$

instead of condition (5). Using condition (40) the comparison group for each observation is all observations in the same within-network subset in the same network. Within-network fixed effects can thus vary across networks.

[18] Another example is the case when socioeconomic status is captured by a categorical variable (e.g. "neither parent is white collar; one parent is white-collar; two parents are white-collar") and the associated within-network fixed effects are employed to capture the possibility that the impact socioeconomic status on the outcome variable may vary across networks. An additional benefit of employing a within-network fixed effects approach is that interacting the endogenous social interaction regressor with dummy variables that represent the characteristic captured with within-network fixed effects (e.g. grade or sosioeconomic status) facilitates inspection of whether the endogenous effect is heterogenous in this characteristic.

6.2 Measurement and Misspecification Error

As was mentioned in the introduction, Moffitt (2001) lists measurement error as one of four key problems in the identification of social interaction effects. To our knowledge neither the existing nor the proposed approach is robust to the type of measurement error in explanatory variables modeled by Moffitt (2001). The proposed estimator can be modified to become robust to classical measurement error in the dependent variable. Suppose that the true model is still (1) but observations are on the variable Y_k^{MM}, which is defined as

$$Y_k^{MM} = Y_k + e_k, \tag{41}$$

where e_k represents i.i.d. measurement error or i.i.d. misspecification error due to, for example, optimization errors by individuals. Combining model (1) and expression (41) yields

$$Y_k^{MM} - e_k = \alpha_k + \beta G\left(Y_k^{MM} - e_k\right) + \delta G x_k + \gamma x_k + \varepsilon_k, \tag{42}$$

which can be rearranged as

$$Y_k^{MM} = (I - \beta G)^{-1}\alpha_k + \delta\,(I - \beta G)^{-1} G x_k + \gamma\,(I - \beta G)^{-1} x_k + e_k + (I - \beta G)^{-1}\varepsilon_k. \tag{43}$$

Expression (43) shows that constructed interaction structures $\tilde{G}$ must now be such that if $\tilde{\beta} = \beta$ then also influence of the element e_{ki} of the measurement/misspecification error term e_k on the element $(\tilde{G}Y_k)_i$ of the constructed variable $\tilde{G}Y_k$ is the same as the average influence of the same error term e_{ki} is on all observations of the constructed variable $\tilde{G}Y_k$ in the same network. Formally, each constructed interaction structure $\tilde{G}$ must satisfy also condition

$$\tilde{G}_{ii} - \frac{1}{n}\sum_{j=1}^{n}\tilde{G}_{ji} = 0 \text{ for all } i \in \{1, ..., n\}. \tag{44}$$

Figure 5 depicts such measurement/misspecification error robust confidence set estimates for the proposed method in directed (solid line) and undirected (dashed line) Erdös-Rényi networks with $p = 0.05$ as a function of the variance σ^2 of error terms ε_{ki}. Other parameters are set as in Section 4.2. The results show that the proposed method continues to have identification power. Especially power against the null hypothesis H_0: $\beta = 0$ continues to be high. This occurs because for imputed value $\tilde{\beta} = 0$ the additional condition (44) and the original condition (5) are the same (for $\tilde{\beta} = 0$ the associated matrix $\tilde{E}$ in condition (5) is an identity matrix). Imposing the additional condition (44) in the construction of conditionally

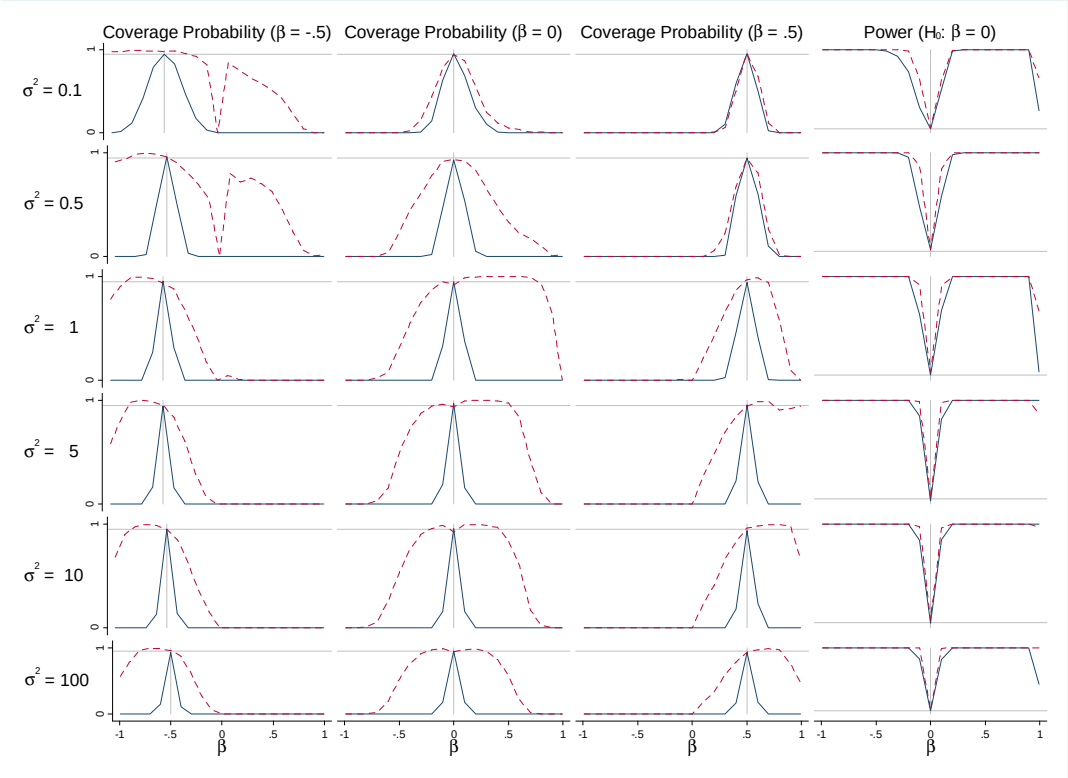

Figure 5: Measurement/misspecification error robust proposed method in directed (solid line) and undirected (dashed line) Erdös-Rényi networks as a function of error term variance.

balanced interaction structures $\tilde{G}$ therefore does not impact power of the proposed method against the null hypothesis H_0: $\beta = 0$.

In contrast, for feasible values $\tilde{\beta}$ for which $\tilde{\beta} \neq 0$ and $\tilde{\beta} \neq \beta$ imposing the additional condition (44) can markedly increase the probability that the feasible value $\tilde{\beta}$ is included in the confidence set estimate. Coverage probabilities shown in Figure 5 indicate that confidence set estimates are quite narrow for directed Erdös-Rényi networks but much wider for undirected Erdös-Rényi networks. Moreover, coverage probabilities for the proposed method in an undirected Erdös-Rényi network in row 3 of Figure 5 are much higher than coverage probabilities for the proposed method in row 2 of Figure 2 (calculated for the same parameterization but without condition (44)). Results in Figure 5 also demonstrate the general feature of the proposed method that coverage probabilities are not necessarily monotonic in variance of the error terms. This occurs because an increase in variance of the error terms also increases in variation in the endogenous regressor GY_k and in variation in the constructed instrument $\tilde{G}Y_k$ which form the basis for identification in the proposed method.

6.3 Observations on Networks with Different Network Structures

We now show that the analysis extends to the case when the k observations on networks do not necessarily have the same network structure. For this purpose, denote the network structure for observation k by G_k, the constructed network structure for observation k associated with the imputed feasible value $\tilde{\beta}$ by $\tilde{G}_k$, and the equilibrium influence matrices for observation k associated with the true and imputed parameter values β and $\tilde{\beta}$ by E_k and $\tilde{E}_k$, respectively. For observation k, the condition (5) for a constructed network structure to be conditionally balanced is now written as

$$\left(\tilde{G}_k\tilde{E}_k\right)_{ii} = \frac{1}{n}\sum_j \left(\tilde{G}_k\tilde{E}_k\right)_{ji} \text{ for all } i \in \{1,...,n\}, \tag{45}$$

which must hold for all observations k. Substituting $\tilde{G}_k$ for $\tilde{G}$ and E_k for E in expression (71) for the probability limit of the estimator $\hat{\beta}_{IV\left(\tilde{\beta}\right)}$ yields

$$\text{plim}_N \hat{\beta}_{IV\left(\tilde{\beta}\right)} = \beta + \frac{\text{plim}_N \frac{1}{N}\sum_k \sum_i \left(\left(\tilde{G}_k E_k\right)_{ii} - \frac{1}{n}\sum_j \left(\tilde{G}_k E_k\right)_{ji}\right)\varepsilon_{ki}^2}{R} \times \text{plim}_N \hat{\theta}_{\tilde{G}Y}. \tag{46}$$

Consider first the case $\tilde{\beta} = \beta$. Substituting $\tilde{E}_k$ for E_k in the above expression (46) and then employing condition (45) yields

$$\text{plim}_N \hat{\beta}_{IV\left(\tilde{\beta}\right)} = \beta + \left[\frac{\text{plim}_N \frac{1}{N}\sum_k \sum_i (0)\left(\varepsilon_{ki}\right)^2 + 0}{R}\right] \times \text{plim}_N \hat{\theta}_{\tilde{G}Y}. \tag{47}$$

Hence, provided that the constructed interaction structures $\tilde{G}_k$ resemble the original interaction structures G_k sufficiently to satisfy the easily testable instrument relevance condition $\text{plim}_N \hat{\theta}_{\tilde{G}Y} \neq 0$, the property $\text{plim}_N \hat{\beta}_{IV\left(\tilde{\beta}\right)} = \beta$ again holds if $\tilde{\beta} = \beta$. Thus, the confidence set estimate $\Psi_{1-p}(\beta)$ will again include the true value β at the chosen probability $1-p$.

Consider now the case $\tilde{\beta} \neq \beta$. The terms in the numerator of the expression (46) for the probability limit of the estimator $\hat{\beta}_{IV\left(\tilde{\beta}\right)}$ each have expectation

$$\left(\left(\tilde{G}_k E_k\right)_{ii} - \frac{1}{n}\sum_j \left(\tilde{G}_k E_k\right)_{ji}\right)\sigma_{k,i}^2, \tag{48}$$

where $\sigma_{k,i}^2$ denotes the variance of the error term ε_{ki} in network k. When the variance ratio assumption (32) and conditions (33) and (34) for point identification hold for every network

$k \in \{1, ..., N\}$ the expectation of every term in the numerator in expression (46) is non-positive for all $\beta < \tilde{\beta}$ and non-negative for all $\beta > \tilde{\beta}$. A version of the law of large numbers due to Chebyshev (1867), which allows expectations and variances to vary across the terms in the sum as long as they are both bounded, then implies that the probability limit of the sum in expression (46) has the same sign as the expectations of the individual terms in the sum.[19] This result in turn implies that $\text{plim}_N \hat{\beta}_{IV(\tilde{\beta})} \leq \beta$ for all $\beta < \tilde{\beta}$ and $\text{plim}_N \hat{\beta}_{IV(\tilde{\beta})} \geq \beta$ for all $\beta > \tilde{\beta}$. Hence, when the variance ratio assumption (32) and conditions (33) and (34) hold for all observed networks, point identification occurs even when the network structure is different across observations.

The same line of argument shows that if the variance ratio assumption (32) and either the set of conditions (35), (36) and (37) or the set of conditions (35), (36), (37) and (38) for partial identification hold for every network then the type of partial identification occurs even when network structure is different across observations.

6.4 Directions for Future Research

6.4.1 Many Instruments

When network structure is different across observations constructed conditionally balanced structures $\tilde{G}_k$ will resemble original interaction structures G_k more for some observations than others. Accordingly, correlation between those parts of $\tilde{G}_k Y_k$ and $G_k Y_k$ that are not explained by fixed effects or observed exogenous variables will be stronger for some observations than others. Interacting the variable $\tilde{G}_k Y_k$ with N network dummy variables, and using the resulting N variables as potential instrumental variables, may then yield a better basis for estimation than using the variable $\tilde{G}_k Y_k$ as the lone instrumental variable. Implementation of this approach in applications presented here is hampered by computational limitations. The relevant estimators, for which Chao et al. (2010) provide the asymptotic analysis, are jackknife instrumental variables estimators.[20] As a result, even when the recursive residual based approach in Chao et al. (2010) is employed, computation of the estimates and the

[19] The same theorem implies that even when network structure is different across observations a probability limit exists for the parameter $\hat{\theta}_{\tilde{G}Y}$ in the first-stage regression as well as for expression (65) for R in the derivation of Lemma 1.

[20] Network-specific instrumental variables constructed from centrality measures are key to the estimation approach developed in Liu and Lee (2010). The asymptotics in Liu and Lee (2010) invoke the assumption that also the size n of networks increases without limit. When this assumption is employed, also the estimators examined in Hausman et al. (2009), which unlike Liu and Lee (2010) allow (do not allow) for heteroskedasticity (within-network correlation), can be employed in the present context in an analysis with network-specific potential instrumental variables.

associated variance covariance matrix is computationally very demanding. The computation involves a triple summation over the sample space, which is large in the Add Health data as is the variable space when network-specific potential instrumental variables are present.

6.4.2 Using Sign of Bias in Constructing Confidence Set Estimates

Let $\hat{\beta}_{LS}$ denote the Least Squares estimator of the parameter β in the original regression equation (1). The estimator $\hat{\beta}_{LS}$ is biased for almost all $\beta \in \Omega(\beta)$. However, as we discuss here briefly, network structure can enable researchers sign the bias of the estimator $\hat{\beta}_{LS}$ when $\beta = \breve{\beta}$ for a given feasible value $\breve{\beta} \in \Omega(\beta)$ and thereby use the estimator $\hat{\beta}_{LS}$ to determine whether the feasible value $\breve{\beta}$ belongs in the confidence set estimate $\Psi_{1-p}(\beta)$. This approach is potentially beneficial when the original interaction structure G is such that for some feasible values $\tilde{\beta}$ an associated conditionally balanced interaction structure $\tilde{G}$ cannot be constructed and when a constructed potential instrumental variable $\tilde{G}Y_k$ can be constructed but is too weak to satisfy the instrument relevance condition. Similarly, also the ability to determine the sign of the bias of a potential instrumental variable estimator $\hat{\beta}_{IV(\tilde{\beta})}$ when $\beta = \breve{\beta}$, where $\breve{\beta} \neq \tilde{\beta}$, can be used to determine whether the feasible value $\breve{\beta}$ belongs in the confidence set estimate $\Psi_{1-p}(\beta)$.

Following the steps in the derivation of the probability limit for the potential instrumental variable estimator $\hat{\beta}_{IV(\tilde{\beta})}$ in Appendix 1, it is straightforward to show that the probability limit of the Least Squares estimator $\hat{\beta}_{LS}$ is

$$\text{plim}_N \hat{\beta}_{LS} = \beta + \frac{\sum_i \left((GE)_{ii} - \frac{1}{n} \sum_j (GE)_{ji} \right) \sigma_i^2}{R}, \tag{49}$$

where R again denotes a strictly positive constant, and σ_i^2 is the variance of the error term ε_{ki}. In expression (49) the equilibrium influence matrix $E = (I - \beta G)^{-1}$ is unknown but for $\beta = \breve{\beta}$ can be expressed as $\breve{E} = (I - \breve{\beta} G)^{-1}$. Therefore, if $\beta = \breve{\beta}$ the sign of the bias of the estimator $\hat{\beta}_{LS}$ is the same as the sign of the expression

$$\sum_{i=1}^{n} \left[\left(\left(G\breve{E} \right)_{ii} - \frac{1}{n} \sum_{j=1}^{n} \left(G\breve{E} \right)_{ji} \right) \times \sigma_i^2 \right]. \tag{50}$$

Hence, the sign of the bias of the estimator $\hat{\beta}_{LS}$ is positive (negative) when $\beta = \breve{\beta}$ if condition

$$\left(G\breve{E} \right)_{ii} - \frac{1}{n} \sum_{j=1}^{n} \left(G\breve{E} \right)_{ji} > 0 \text{ for all } i \in \{1, ..., n\} \ \ (< 0 \text{ for all } i \in \{1, ..., n\}) \tag{51}$$

holds. Suppose that for some feasible value $\breve{\beta}$ condition (51) holds with the positive inequality and that the feasible value $\breve{\beta}$ is above the confidence interval computed using the Least Squares estimate $\hat{\beta}_{LS}$. Expression (50) for the sign of the bias and the confidence interval then enable researchers to reject the null hypothesis H_0: $\beta = \breve{\beta}$ and exclude the feasible value $\breve{\beta}$ from the confidence set estimate $\Psi_{1-p}(\beta)$ for parameter β. The case when condition (51) holds with the negative inequality is treated analogously.

Such a sign of bias based approach can only yield partial identification. Moreover, in many cases condition (51) holds with neither positive nor negative inequality. However, it can still be possible to determine the sign of the bias provided that one is willing to make additional assumptions about how the variances σ_i^2 of the error terms ε_{ki} vary across individuals in a network, analogously to the variance ratio restriction (32) employed in Section 5.

6.4.3 Combining Existing and Proposed Approaches

The distinct advantages of the proposed and existing approaches render the approaches complementary. A parallel application of the two approaches provides researchers an opportunity to examine the robustness of results to the choice of estimation method and the associated assumptions. A sequential application of the two approaches may in turn be employed to improve the precision of the estimates. Estimates from the existing approach can be used to first perform a (Cochrane-Orcutt) transformation to account for (spatial) correlation among unobservables not captured by network or within network fixed effects. The proposed approach, which as our Monte Carlo simulations have shown can have much better finite-sample performance than the existing method, can then be applied to the transformed variables.

7 Conclusion

The method proposed in this paper allows researchers to take advantage of interaction structure induced variation in equilibrium influence to estimate endogenous and exogenous social interaction effects. Variation in equilibrium influence is present in network structure based models of social interaction and in spatial interaction models.

In implementing the proposed method, conditionally balanced interaction structures are first constructed using the original interaction structure and an imputed value of the endogenous social interaction parameter. Each constructed interaction structure is then combined with observations on the outcome variable to construct a potential instrumental variable for the endogenous social interaction variable. Comparison of each potential instrumental variable estimate with the associated imputed value of the endogenous social interaction

parameter yields a confidence set estimate for the endogenous social interaction parameter as well as for other parameters in the model.

Implementation of the proposed method is straightforward. And while the method is computationally relatively demanding, recent advances in computation render the method feasible. We have demonstrated this with analyses of the determinants of subjective income and college completion expectations among adolescents. Implementation of these applications (using the is/has friend Add Health sample) involved solving for each considered feasible value of the endogenous social interaction parameter 489 constrained optimization problems with over 12 million unknown interaction structure parameters.

Comparison of the proposed approach with the existing network structure based estimation approach is sharp. In the proposed approach instruments are constructed from the outcome variable whereas in the existing approach excluded exogenous variables are used as instruments. The disadvantage of the proposed method is that while it allows for correlated effects that can be represented by network fixed effects or within-network fixed effects, the existing method allows an unrestricted within-network correlation structure for unobserved variables. The advantage of the proposed method is that it does not rely on excluded exogenous variables for identification. An important step in implementing the existing method is the determination of whether the relevant exclusion restrictions are valid, which is complicated by issues related to nested hypothesis testing and power of overidentification tests. Moreover, when exclusion restrictions concern peers' peers' characteristics and peers' peers' peers' characteristics, as has been the case in the applications offered in the literature, identifying variation is limited as we have argued in this paper. The proposed approach, in contrast, does not rely on such variation and associated exclusion restrictions for identification. Accordingly, the proposed approach allows researchers to remain relatively agnostic about which exogenous variables can be excluded from the model.

The advantages of each approach make the approaches complementary rather than competing. Application of the approaches in parallel allows researchers to examine the robustness of results to the choice of estimation method and the associated assumptions. In addition, researchers may benefit from a sequential application of the two approaches. Analyses of this and other potential extensions discussed in Section 6.4 are left for future research.

References

Anselin, L., 1988, *Spatial Econometrics*. Kluwer Academic Publishers, Dordrecht.

Blau, F. and M. Ferber, 1991, "Career Plans and Expectations of Young Women and Men," *Journal of Human Resources*, vol. 26, pp. 581-607.

Blume, L. E., Brock, W. A., Durlauf, S. N. and Y. M. Ioannides, 2010, "Identification of Social Interactions," in: Benhabib, J., Bisin, A., and M. Jackson (eds.) *Handbook of Social Economics*, forthcoming.

Bramoulle, Y., Djebbari, H. and B. Fortin, 2009, "Identification of Peer Effects through Social Networks," *Journal of Econometrics*, vol. 150, pp. 41-55.

Brock, W. and S. Durlauf, 2001, "Interaction-Based Models," in: Heckman, J. and E. Leamer (eds.), *Handbook of Econometrics*, vol. 5. North-Holland, Amsterdam.

Brock, W. and S. Durlauf, 2007, "Identification of Binary Choice Models with Social Interactions," *Journal of Econometrics*, vol. 140, pp. 57-75.

Brooks-Gunn, J., Duncan, G. J., Klebanov, P. K. and N. Sealand, 1993, "Do Neighborhoods Influence Child and Adolescent Development?" *The American Journal of Sociology*, vol. 99, pp. 353-95.

Calvó-Armengol, A., Patacchini, E. and Y. Zenou, 2009, "Peer Effects and Social Networks in Education," *Review of Economic Studies*, vol. 76, pp. 1239-67.

Chao, J. C., Swanson, N. R., Hausman, J. A., Newey, W. K. and T. Woutersen, 2010, "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments," Mimeo.

Chebyshev, L. P., 1867, "Des valeurs moyennes (on mean values)", *Journal de Mathématiques Pures et Appliquées*, vol. 12, pp. 177-84.

Cohen-Cole, E., 2006, "Multiple Groups Identification in the Linear-in-Means Model," *Economics Letters*, vol. 92, pp. 157-62.

Cohen-Cole, E. and G. Zanella, 2008, "Unpacking Social Interactions," *Economic Inquiry*, vol. 46, pp. 19-24.

Davies, M. and D. B. Kandel, 1981, "Parental and Peer Influences on Adolescents' Educational Plans: Some Further Evidence," *American Journal of Sociology*, vol. 87, 363-87.

Delavande, A., 2008, "Pill, Patch or Shot? Subjective Expectations and Birth Control Choice," *International Economic Review*, vol. 49, pp. 999-1042.

Dominitz, J., and C. Manski, 1996, "Eliciting Student Expectations of the Returns to Schooling," *Journal of Human Resources*, vol. 31, pp. 1-26.

Dominitz, J., and C. Manski, 1997a, "Using Expectations Data to Study Subjective Income Expectations," *Journal of the American Statistical Association*, vol. 92, pp. 855-67.

Dominitz, J., and C. Manski, 1997b, "Perceptions of Economic Insecurity: Evidence from the Survey of Economic Expectations," *Public Opinion Quarterly*, vol. 61, pp. 261-87.

Erdös, P. and A. Rényi, 1959, "On random graphs," *Publicationes Mathematicae*, vol. 6, pp. 290-7.

Fischhoff, B, Parker, A., Bruine De Bruin, W., Downs, J., Palmgren, C. Dawes, R. and C. Manski, 2000, "Teen Expectations for Significant Life Events," *Public Opinion Quarterly*, vol. 64. pp. 189-205.

Fletcher, J. M. and S. L. Ross, 2009, "Estimating the Effects of Friendship Networks on Health Behaviors of Adolescents," Mimeo.

De Giorgi, G., M. Pellizari and S. Redaelli, 2010, "Identification of Social Interactions through Partially Overlapping Groups," *American Economic Journal: Applied Economics*, vol. 2, pp. 241-75.

Graham, B. S., 2008, "Identifying Social Interactions through Conditional Variance Restrictions," *Econometrica*, vol. 76, pp. 643-60.

Guiso, L., Jappelli, A. and D. Terlizzese, 1992, "Earnings Uncertainty and Precautionary Saving," *Journal of Monetary Economics*, vol. 30, pp. 307-37.

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C. and N. R. Swanson, 2010, "Instrumental Variable Estimation with Heteroskedasticity and Many Instruments," Mimeo.

Hurd, M. and K. McGarry, "Evaluation of the Subjective Probabilities of Survival in the Health and Retirement Study," *Journal of Human Resources*, vol. 30, pp. S268-92

Hurd, M., Smith, J. and J. Zissimopoulos, "The Effects of Subjective Survival on Retirement and Social Security Claiming," *Journal of Applied Econometrics*, vol. 19, pp. 761-75.

Ioannides, Y. and A. Zabel, 2003, "Neighborhood Effects and Housing Demand," *Journal of Applied Econometrics*, vol. 18, pp. 563-84.

Jacob, B. and T. Wilder, 2010, "Educational Expectations and Attainment," National Bureau of Economic Research working paper No. 15683.

Juster, T. and R. Suzman, 1995, "An Overview of the Health and Retirement Study," *Journal of Human Resources*, vol. 30, pp. S7-S56.

Keleija, H. H. and I. R. Prucha, 2010, "Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances," *Journal of Econometrics*, forthcoming.

Kiuru, N., Aunola, K., Vuori, J. and J.-E. Nurmi, 2007, "The Role of Peer Groups in Adolescents' Educational Expectations and Adjustment," *Journal of Youth and Adolescence*, vol. 36, pp. 995-1009.

Krauth, B., 2006, "Simulation-Based Estimation of Peer Effects," *Journal of Econometrics*, vol. 133, pp. 243-71.

Laschever, R., 2009, "The Doughboys Network: Social Interactions and the Employment of World War I Veterans," Mimeo.

Lee, L. F., 2003, "Best Spatial Two-Stage Least Squares Estimators for a Spatial Autoregressive Model with Autoregressive Disturbances," *Econometric Reviews*, vol. 22, pp. 307-35.

Lee, L. F., 2007, Identification and Estimation of Econometric Models with Group Interactions, Contextual Factors and Fixed Effects," *Journal of Econometrics*, vol. 140, pp. 333-74.

Lee, L. F., Liu, X. and X. Lin, 2010, "Specification and Estimation of Social Interaction Models with Network Structure, Contextual Factors, Correlation and Fixed Effects," *Econometrics Journal*, vol. 24, pp. 257-281.

Lin, X., 2010, "Identifying Peer Effects in Student Academic Achievement by Spatial Autoregressive Models with Group Unobservables," *Journal of Labor Economics*, vol. 28, pp. 825-60.

Lin, X. and L. F. Lee, 2010, "GMM Estimation of Spatial Autoregressive Models with Unknown Heteroskedasticity," *Journal of Econometrics*, forthcoming.

Liu, X. and L-f. Lee, 2009, "GMM Estimation of Social Interaction Models with Centrality," Mimeo.

Lochner, L., 2004, "Individual Perceptions of the Criminal Justice System," *American Economic Review*, vol. 97, pp. 444-60.

Manski C. F., 1993a, "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, vol. 60, pp. 531-42.

Manski, C. F., 1993b, "Adolescent Econometricians: How Do Youth Infer the Returns to Schooling?" in: Clotfelter, C. and M. Rothschild (eds.) *Studies of Supply and Demand in Higher Education*. University of Chicago Press, Chicago.

Manski, C. F., 1995, *Identification Problems in the Social Sciences*. Harvard University. Press, Cambridge.

Manski, C. F., 2000, "Economic Analysis of Social Interactions," *Journal of Economic Perspectives*, vol. 14, pp. 115-36.

Manski, C. F., 2004, "Measuring Expectations," *Econometrica*, vol. 72, pp. 1329-76.

Moffitt, R., 2001, "Policy Interventions, Low-Level Equilibria, and Social Interactions," In: Durlauf, S. and P. Young (Eds.), *Social Dynamics*. MIT Press.

Nicholson, S. and N. Souleles, 2001, "Physician Income Expectations and Specialty Choice," National Bureau of Economic Research working paper No. 8536.

Pinkse, J. and M. E. Slade, 2010, "The Future of Spatial Econometrics," *Journal of Regional Science*, vol. 50, pp. 103-17.

Quadrel, M. Fischhoff, B. and W. Davis, 1993, "Adolescent (In)vulnerability," *American Psychologist*, vol. 48, pp. 102-16.

Sacerdote, B., 2001, "Peer Effects with Random Assignment: Results for Darmouth Roommates," *Quarterly Journal of Economics*, vol. 116, pp. 681-704.

Smith, H. and B. Powell, 1990, "Great Expectations: Variations in Income Expectations Among College Seniors," *Sociology of Education*, vol. 63, pp. 194-207.

Souleles, N., 2004, "Expectations, Heterogeneous Forecast Errors, and Consumption: Micro Evidence from the Michigan Consumer Sentiment Surveys," *Journal of Money, Credit, and Banking*, vol. 36, pp. 39-72.

Stock, J. H. and M. W. Watson, 2008, "Heteroskedasticity–Robust Standard Errors for Fixed Effects Panel Data Regression," *Econometrica*, vol. 76. pp. 155-74.

Trogdon, J., Nonnemaker, J. and J. Pais, 2008, "Peer Effects in Adolescent Overweight," *Journal of Health Economics*, vol. 27 , pp. 1388-99.

Watts, D. J. and S. H. Strogatz, 1998, "Collective dynamics of small-world networks," *Nature*, vol. 393, pp. 440-2.

Appendix 1: Proof of Lemma 1 (Probability Limit of the Potential Instrumental Variable Estimator)

The estimator $\hat{\beta}_{IV(\tilde{\beta})}$ can be expressed as

$$\hat{\beta}_{IV(\tilde{\beta})} = \frac{\sum_k \sum_i \hat{r}_{i1} \left(Y_k\right)_i}{\sum_k \sum_i \hat{r}_{i1}^2}, \tag{52}$$

where $\hat{r}_{i1}$ are the residuals

$$\hat{r}_{i1} \equiv \left(\widehat{GY_k} - \sum_k \hat{f}_{D_k} D_k - \hat{f}_{GX} GX_k - \hat{f}_X X_k \right)_i \tag{53}$$

from the regression of the variable $\widehat{GY_k}$ on the variables D_k, GX_k, and X_k. This auxiliary regression equation is formally written as

$$\widehat{GY_k} = \sum_k f_{D_k} D_k + f_{GX} GX_k + f_X X_k + r_k. \tag{54}$$

Substituting the second-stage regression equation (8) for Y_k in expression (52) yields

$$\hat{\beta}_{IV(\tilde{\beta})} = \frac{\sum_k \sum_i \hat{r}_{i1} \left(\beta \widehat{GY_k} + \sum_k \alpha_{D_k} D_k + \delta GX_k + \gamma X_k + \beta \hat{v}_k + \varepsilon_k \right)_i}{\sum_k \sum_i \hat{r}_{i1}^2}. \tag{55}$$

Variables D_k, GX_k and δX_k are among the regressors in the regression equation (54) that yields the residuals $\hat{r}_{i1}$ and are thus orthogonal to the residuals $\hat{r}_{i1}$. Using this observation expression (55) for $\hat{\beta}_{IV(\tilde{\beta})}$ can be rewritten as

$$\hat{\beta}_{IV(\tilde{\beta})} = \frac{\sum_k \sum_i \hat{r}_{i1} \left(\beta \widehat{GY_k} + \beta \hat{v}_k + \varepsilon_k \right)_i}{\sum_k \sum_i \hat{r}_{i1}^2}, \tag{56}$$

which can be reorganized as

$$\hat{\beta}_{IV(\tilde{\beta})} = \beta \frac{\sum_k \sum_i \hat{r}_{i1} \left(\widehat{GY_k} \right)_i}{\sum_k \sum_i \hat{r}_{i1}^2} + \frac{\sum_k \sum_i \hat{r}_{i1} \left(\beta \hat{v}_k + \varepsilon_k \right)_i}{\sum_k \sum_i \hat{r}_{i1}^2}. \tag{57}$$

Using again the observation that the variables D_k, GX_k and δX_k are orthogonal to the residuals $\hat{r}_{i1}$, the above expression can be rewritten as

$$\hat{\beta}_{IV(\tilde{\beta})} = \beta \frac{\sum_k \sum_i \hat{r}_{i1} \left(\widehat{GY_k} - \sum_k \hat{f}_{D_k} D_k - \hat{f}_{GX} GX_k - \hat{f}_X X_k \right)_i}{\sum_k \sum_i \hat{r}_{i1}^2} + \frac{\sum_k \sum_i \hat{r}_{i1} \left(\beta \hat{v}_k + \varepsilon_k \right)_i}{\sum_k \sum_i \hat{r}_{i1}^2}. \tag{58}$$

Using definition (53) of $\hat{r}_{i1}$ we can substitute $\hat{r}_{i1}$ for $\left(\widehat{GY_k} - \sum_k \hat{f}_{D_k} D_k - \hat{f}_{GX} GX_k - \hat{f}_X X_k \right)_i$ in the numerator of the first term to get

$$\hat{\beta}_{IV(\tilde{\beta})} = \beta + \frac{\sum_k \sum_i \hat{r}_{i1} \left(\beta \hat{v}_k + \varepsilon_k \right)_i}{\sum_k \sum_i \hat{r}_{i1}^2}. \tag{59}$$

Using definition (53) of $\hat{r}_{i1}$ we now substitute $\left(\widehat{GY_k} - \sum_k \hat{f}_{D_k} D_{ki} - \hat{f}_{GX} GX_k - \hat{f}_X X_k \right)_i$ for $\hat{r}_{i1}$ in the numerator of the second term to get

$$\hat{\beta}_{IV(\tilde{\beta})} = \beta + \frac{\sum_k \sum_i \left(\widehat{GY_k} - \sum_k \hat{f}_{D_k} D_k - \hat{f}_{GX} GX_k - \hat{f}_X X_k \right)_i \left(\beta \hat{v}_k + \varepsilon_k \right)_i}{\sum_k \sum_i \hat{r}_{i1}^2}. \tag{60}$$

Reorganizing the terms in the numerator yields

$$\hat{\beta}_{IV(\tilde{\beta})} = \beta + \frac{\sum_k \sum_i \left(\widehat{GY_k} \right)_i \left(\beta \hat{v}_k + \varepsilon_k \right)_i - \sum_k \sum_i \left(\sum_k \hat{f}_{D_k} D_k + \hat{f}_{GX} GX_k + \hat{f}_X X_k \right)_i \left(\beta \hat{v}_k + \varepsilon_k \right)_i}{\sum_k \sum_i \hat{r}_{i1}^2}. \tag{61}$$

Regressors D_k, GX_k and δX_k in the first-stage regression (6) and predicted values $\left(\widehat{GY_k} \right)_i$ from the first-stage regression are orthogonal to the residuals $\left(\hat{v}_k \right)_i$ from the same regression. Using this observation yields

$$\hat{\beta}_{IV(\tilde{\beta})} = \beta + \frac{\sum_k \sum_i \left(\widehat{GY_k} \right)_i \left(\varepsilon_k \right)_i - \sum_k \sum_i \left(\sum_k \hat{f}_{D_k} D_k + \hat{f}_{GX} GX_k + \hat{f}_X X_k \right)_i \left(\varepsilon_k \right)_i}{\sum_k \sum_i \hat{r}_{i1}^2}. \tag{62}$$

Using definition (7) of the first-stage predicted values $\widehat{GY_k}$ allows us to rewrite this as

$$\begin{aligned} \hat{\beta}_{IV(\tilde{\beta})} \;=\; & \beta + \frac{\sum_k \sum_i \left(\sum_k \hat{\theta}_{Dk} D_k + \hat{\theta}_{\tilde{G}Y} \tilde{GY}_k + \hat{\theta}_{GX} GX_k + \hat{\theta}_X X_k \right)_i \left(\varepsilon_k \right)_i}{\sum_k \sum_i \hat{r}_{i1}^2} \\ & - \frac{\sum_k \sum_i \left(\sum_k \hat{f}_{D_k} D_k + \hat{f}_{GX} GX_k + \hat{f}_X X_k \right)_i \left(\varepsilon_k \right)_i}{\sum_k \sum_i \hat{r}_{i1}^2}. \end{aligned} \tag{63}$$

Substituting the network fixed-effect demeaned versions of the variables $\tilde{G}Y_k$, GX_k, and X_k for the variables $\tilde{G}Y_k$, GX_k, and X_k yields

$$\begin{aligned}\hat{\beta}_{IV(\tilde{\beta})} &= \beta + \frac{\hat{\theta}_{\tilde{G}Y}\sum_k\sum_i\left[\left(\tilde{G}Y_k\right)_i - \frac{1}{n}\sum_j\left(\tilde{G}Y_k\right)_j\right](\varepsilon_k)_i}{\sum_k\sum_i\hat{r}_{i1}^2} \\ &+\frac{\hat{\theta}_{GX}\sum_k\sum_i\left[(GX_k)_i - \frac{1}{n}\sum_j(GX_k)_j\right](\varepsilon_k)_i + \hat{\theta}_X\sum_k\sum_i\left[(X_k)_i - \frac{1}{n}\sum_j(X_k)_j\right](\varepsilon_k)_i}{\sum_k\sum_i\hat{r}_{i1}^2} \\ &-\frac{\sum_k\sum_i\left(\hat{f}_{GX}\left(GX_k - \frac{1}{n}\sum_j(GX_k)_j\right) + \hat{f}_X\left(X_k - \frac{1}{n}\sum_j(X_k)_j\right)\right)_i(\varepsilon_k)_i}{\sum_k\sum_i\hat{r}_{i1}^2}.\end{aligned} \tag{64}$$

Using definition

$$R \equiv \text{plim}_N\frac{\sum_k\sum_i\hat{r}_{i1}^2}{N} \tag{65}$$

and the assumption $E\left[\varepsilon_k\middle|\, X_k, \alpha_k\right] = 0$ we have

$$\text{plim}_N\hat{\beta}_{IV(\tilde{\beta})} = \beta + \frac{\text{plim}_N\frac{1}{N}\sum_k\sum_i\left(\left(\tilde{G}Y_k\right)_i - \frac{1}{n}\sum_j\left(\tilde{G}Y_k\right)_j\right)(\varepsilon_k)_i + 0}{R} \times \text{plim}_N\hat{\theta}_{\tilde{G}Y}. \tag{66}$$

Using expression (3) for the relationship between the outcome variable Y_k and the observed and unobserved exogenous variables allows us to rewrite this as

$$\begin{aligned}\text{plim}_N\hat{\beta}_{IV(\tilde{\beta})} &= \beta + \left[\frac{\text{plim}_N\frac{1}{N}\sum_k\sum_i\left(\left(\tilde{G}\left[\tau\alpha_k + \delta EGX_k + \gamma EX_k + E\varepsilon_k\right]\right)_i\right)(\varepsilon_k)_i}{R}\right. \\ &\left.-\frac{\text{plim}_N\frac{1}{N}\sum_k\sum_i\left(\frac{1}{n}\sum_i\left(\tilde{G}\left[\tau\alpha_k + \delta EGX_k + \gamma EX_k + E\varepsilon_k\right]\right)_i\right)(\varepsilon_k)_i}{R}\right] \\ &\times\text{plim}_N\hat{\theta}_{\tilde{G}Y}.\end{aligned} \tag{67}$$

Using the assumption $E\left[\varepsilon_k\middle|\, X_k, \alpha_k\right] = 0$ again allows us to rewrite this as

$$\text{plim}_N\hat{\beta}_{IV(\tilde{\beta})} = \beta + \frac{\text{plim}_N\frac{1}{N}\sum_k\sum_i\left(\left(\tilde{G}E\varepsilon_k\right)_i - \frac{1}{n}\sum_j\left(\tilde{G}E\varepsilon_k\right)_j\right)(\varepsilon_k)_i + 0}{R} \times \text{plim}_N\hat{\theta}_{\tilde{G}Y}. \tag{68}$$

Reorganizing yields

$$\begin{aligned}\text{plim}_N\hat{\beta}_{IV(\tilde{\beta})} &= \beta + \left[\frac{\text{plim}_N \frac{1}{N}\sum_k\sum_i\left(\sum_h\left(\tilde{G}E\right)_{ih}(\varepsilon_k)_h - \frac{1}{n}\sum_j\sum_h\left(\tilde{G}E\right)_{jh}(\varepsilon_k)_h\right)(\varepsilon_k)_i}{R}\right] \\ &\times \text{plim}_N\hat{\theta}_{\tilde{G}Y}, \end{aligned} \tag{69}$$

which can be rewritten as

$$\begin{aligned}\text{plim}_N\hat{\beta}_{IV(\tilde{\beta})} &= \beta + \left[\frac{\text{plim}_N \frac{1}{N}\sum_k\sum_i\left(\left(\tilde{G}E\right)_{ii}(\varepsilon_k)_i - \frac{1}{n}\sum_j\left(\tilde{G}E\right)_{ji}(\varepsilon_k)_i\right)(\varepsilon_k)_i}{R}\right. \\ &\left. + \frac{\text{plim}_N \frac{1}{N}\sum_k\sum_i\left(\sum_{h\neq i}\left(\tilde{G}E\right)_{ih}(\varepsilon_k)_h - \frac{1}{n}\sum_j\left(\tilde{G}E\right)_{ji}(\varepsilon_k)_i\right)(\varepsilon_k)_i}{R}\right] \\ &\times \text{plim}_N\hat{\theta}_{\tilde{G}Y}. \end{aligned} \tag{70}$$

Using independence of the unobserved error terms ε_k within and across networks yields

$$\text{plim}_N\hat{\beta}_{IV(\tilde{\beta})} = \beta + \frac{\text{plim}_N \frac{1}{N}\sum_k\sum_i\left(\left(\tilde{G}E\right)_{ii} - \frac{1}{n}\sum_j\left(\tilde{G}E\right)_{ji}\right)\varepsilon_{ki}^2}{R} \times \text{plim}_N\hat{\theta}_{\tilde{G}Y}. \tag{71}$$

Reorganizing yields

$$\text{plim}_N\hat{\beta}_{IV(\tilde{\beta})} = \beta + \frac{\sum_i\left(\left(\tilde{G}E\right)_{ii} - \frac{1}{n}\sum_j\left(\tilde{G}E\right)_{ji}\right)\text{plim}_N \frac{1}{N}\sum_k(\varepsilon_{ki})^2}{R} \times \text{plim}_N\hat{\theta}_{\tilde{G}Y}. \tag{72}$$

Using the result

$$\text{plim}_N \frac{\frac{1}{N}\sum_k(\varepsilon_{ki})^2}{N} = \sigma_i^2 \tag{73}$$

yields

$$\text{plim}_N\hat{\beta}_{IV(\tilde{\beta})} = \beta + \frac{\sum_i\left(\left(\tilde{G}E\right)_{ii} - \frac{1}{n}\sum_j\left(\tilde{G}E\right)_{ji}\right)\sigma_i^2}{R} \times \text{plim}_N\hat{\theta}_{\tilde{G}Y}. \tag{74}$$

Data Appendix

The National Longitudinal Study of Adolescent Health (Add Health) is a nationally representative school-based study of adolescents. In wave I of this study an in-school questionnaire of this study was administered in 132 schools to all students on grades 7-12 in September 1994 - April 1995. More detailed in-home interviews were conducted in four waves to a much smaller sample. We employ only data from the wave I in-school questionnaire. The in-school data contain information on respondents' demographics, family characteristics, academic performance, expectations, club participation and psychological as well as physical well-being. As was mentioned in the main text, the Add Health data also contain information on which other students in the same school each respondent nominates as their friend.

In constructing the sample we initially follow Lin (2010) to facilitate ease of replication across related studies. Networks are constructed at the school-grade level. Each grade in each school thus forms its own network. Lin (2010) provides a detailed documentation of how elimination of observations with missing or invalid information on respondent ID, age, grade, gender, race, number of years in school, family structure and academic performance affects the sample size and descriptive statistics variable by variable. The cumulative impact of these eliminations is that sample size is reduced from $90,118$ to $70,639$. As is noted in Lin (2010), only $60,495$ students in this sample of $70,639$ students either nominate as their friend another student in the sample or are nominated as friend by another student in the sample. This sample of $60,495$ students forms our "is/has friend sample". To construct our "fully networked sample" we recursively eliminate students who do not nominate as friend a student in the remaining sample and students who are not nominated as friend by a student in the remaining sample. This recursive elimination leaves a sample of $42,827$ observations.[21] Descriptive statistics for networks in the two employed samples are shown in Table A.1.

	Fully Networked Sample	Is/Has Friend Sample
Number of Networks (N)	486	489
Observations ($\sum_{k=1}^{N} n_k$)	$42,827$	$60,495$
Network Parameters ($\sum_{k=1}^{N} n_k^2$)	$6,287,061$	$12,298,399$
Network Sizes: Median [Min, Max]	75 [2, 410]	103 [2, 484]
Nominated Friends: Median [Min, Max]	3 [1, 10]	3 [0, 10]
Nominated as Friend: Median [Min, Max]	3 [1, 32]	3 [0, 35]

Table A.1. Descriptive Statistics: Networks.

[21] The "Network Sample" with $49,559$ observations in Lin (2010) is constructed by recursively elimating only students who do not nominate as friend a student in the remaining sample.

The employed measure of income expectations $y\_Income\_Expectations$ is constructed from responses to the question "*On a scale from "No chance" to "It will happen" what do you think are the chances you will: Have a middle-class family income by age 30?*" The employed measure of college completion expectations $y\_College\_Expectations$ is constructed from responses to the question "*On a scale from "No chance" to "It will happen" what do you think are the chances you will: Graduate from college?*" The distributions for these variables are shown in Table A.2. For both variables the scale indicated in the questionnaire consists of integers 0 through 8 with values 0, 2, 4, 6 and 8 labeled as shown in Table A.2. For observations with missing/invalid information on either variable we set the value of the variable equal to the mean value of the variable in the data.

	$y\_Income\_Expectations$		$y\_College\_Expectations$	
	Fully Networked Sample	Is/Has Friend Sample	Fully Networked Sample	Is/Has Friend Sample
0 - "*No chance*"	3.62%	4.30%	2.03%	2.62%
1	1.91%	2.03%	0.74%	0.92%
2 - "*Some chance*"	8.26%	8.61%	3.45%	4.06%
3	2.76%	2.80%	0.85%	0.98%
4 - "*About 50-50*"	19.10%	18.97%	6.35%	6.99%
5	5.70%	5.35%	1.84%	1.93%
6 - "*Pretty likely*"	23.81%	22.65%	12.52%	12.54%
7	11.78%	10.87%	10.55%	9.83%
8 - "*It will happen*"	19.06%	18.76%	48.50%	45.80%
Missing/invalid	4.02%	5.66%	13.18%	14.32%

Table A.2. Descriptive Statistics: Distributions of dependent variables.

Construction of the employed explanatory variables from the raw Add Health data is self-explanatory except for dummy variable $x\_Mom\_College$ [$x\_Dad\_College$], which is set equal to 1 if respondent indicates that resident mother [resident father] graduated from college/university or has professional training beyond a 4-year college, for dummy variable $x\_Mom\_Professional$ [$x\_Mom\_Professional$], which is set equal to 1 if respondent indicates that resident mother [resident father] is a professional of type 1 (doctor/lawyer/scientist), professional of type 2 (teacher/librarian/nurse), or a manager/executive, and for dummy variable $x\_Mom\_White\_Collar$ [$x\_Dad\_White\_Collar$], which is set

equal to 1 if respondent indicates that resident mother [resident father] a professional of type 1 (doctor/lawyer/scientist), professional of type 2 (teacher/librarian/nurse), manager/executive, technical/computer specialist/radiologist, office worker/book keeper/clerk/secretary, or sales worker/insurance agent/store clerk. The mean value of each explanatory variable is shown in Table A.3.

	Fully Networked Sample	Is/Has Friend Sample
x_GPA	2.90	2.84
x_Age	14.92	15.00
x_Female	0.56	0.53
x_Asian	0.06	0.06
x_Black	0.15	0.17
$x\_Hispanic$	0.11	0.13
$x\_Mom\_College$	0.29	0.29
$x\_Dad\_College$	0.28	0.27
$x\_Mom\_Professional$	0.29	0.27
$x\_Dad\_Professional$	0.23	0.22
$x\_Mom\_White\_Collar$	0.54	0.52
$x\_Dad\_White\_Collar$	0.34	0.32
$x\_Parent\_Homemaker$	0.16	0.16
$x\_Parent\_Military$	0.04	0.04

Table A.3. Descriptive Statistics: Mean values of explanatory variables.

Standard deviations and correlations for selected variables are shown in Tables A.4 and A.5. The results illustrate the issue mentioned in Section 4.1.4 that one problem with using variables $GGx\_some\_variable_k$ and $GGGx\_some\_variable_k$ constructed from exogenous variables as instrumental variables for the endogenous social interaction regressor is that identifying variation in these variables is limited. As Table A.4 demonstrates, variation in variable $G^hx\_some\_variable_k$ decreases as h increases. This occurs because the set of friends' friends' friends is more similar across observations than the set of friends or the set of friends' friends. Table A.5 in turn demonstrates that in the is/has friend sample correlation between variables $G^hx\_some\_variable$ and $G^hx\_some\_other\_variable$ increases as h increases. This occurs because the values of $G^hx\_some\_variable_{ki}$ and $G^hx\_some\_other\_variable_{ki}$, where $h \geq 2$, are in large part determined by whether the friends, friends' friends, and friends' friends' friends of individual ki have any friends themselves. If many of the friends, friends' friends, and friends' friends' friends of individual ki do not have any friends themselves, the value of both $G^hx\_some\_variable_{ki}$ and $G^hx\_some\_other\_variable_{ki}$, where $h \geq 2$, is close to zero for the individual ki.

	Fully Networked Sample	Is/Has Friend Sample
$x\_Mom\_College$	0.43	0.42
$Gx\_Mom\_College$	0.28	0.28
$GGx\_Mom\_College$	0.21	0.21
$GGGx\_Mom\_College$	0.18	0.18

Table A.4. Descriptive Statistics: Standard deviations for selected variables.

	$x\_Dad\_College$	$Gx\_Dad\_C.$	$GGx\_Dad\_C.$	$GGGx\_Dad\_C.$
$x\_Mom\_College$	0.38			
$Gx\_Mom\_Coll.$		0.52		
$GGx\_Mom\_Coll.$			0.66	
$GGGx\_Mom\_Coll.$				0.73

Table A.5. Descriptive Statistics: Correlations for selected variables in the is/has friend sample.

Implementation Appendix

Interaction structure G_k for each network is constructed from the friendship denomination data. An auxiliary interaction matrix W_k is first constructed by setting $(W_k)_{ij} = 1$ if student i in network k has nominated student j in the same network k as friend and by setting $(W_k)_{ij} = 0$ otherwise. The actual interaction structure G_k is then constructed using the normalization $(G_k)_{ij} = (W_k)_{ij} / \sum_{j=1}^{n} (W_k)_{ij}$ for all (i, j) for which $(W_k)_{ij} = 1$ and by setting $(G_k)_{ij} = 0$ for all (i, j) for which $(W_k)_{ij} = 0$.

We set $\{-0.99, -0.9, -0.8, ..., 0.8, 0.9, 0.99\}$ as the grid of feasible values of parameter β. For each feasible value $\tilde{\beta}$ in this grid, whether the feasible value $\tilde{\beta}$ is included in the confidence set estimate $\Psi_{0.95}(\beta)$ is determined in 5 steps (corresponding to the 5 steps in Section 3.1):

Step 1. Calculate equilibrium influence matrix $\tilde{E}_k = \left(I - \tilde{\beta} G_k\right)^{-1}$ for each observed network G_k (there are 486 networks in the fully networked sample and 489 networks in the is/has friend sample).

Step 2. First solve the constrained linear optimization problem (13) for each observed network G_k to obtain a conditionally balanced interaction structure $\tilde{G}_k$ for each observed network G_k. Parameters l_b, u_b and c are set as mentioned at the end of Section 3.3. Then combine the constructed interaction structures $\tilde{G}_k$ and observations on the endogenous outcome variable Y_k to construct variable $\tilde{G}_k Y_k$.

Steps 3-4. Use the constructed variable $\tilde{G}_k Y_k$ as an instrumental variable for the endogenous regressor $G_k Y_k$ to obtain a potential instrumental variable estimate $\hat{\beta}_{IV(\tilde{\beta})}$ of parameter β in equation (1).

Step 5. Use the potential instrumental variable estimate $\hat{\beta}_{IV(\tilde{\beta})}$ and the associated standard error $S.E.(\hat{\beta}_{IV(\tilde{\beta})})$ to construct the confidence interval $(\hat{\beta}_{IV(\tilde{\beta})} - 1.96 \times S.E.(\hat{\beta}_{IV(\tilde{\beta})}),$ $\hat{\beta}_{IV(\tilde{\beta})} + 1.96 \times S.E.(\hat{\beta}_{IV(\tilde{\beta})}))$. Given the inconsistency of standard heteroskedasticity-robust standard errors in the presence of fixed effects (Stock and Watson, 2008), we employ cluster-robust standard errors with clustering at the network level. The imputed feasible value $\tilde{\beta}$ is included in the confidence set estimate $\Psi_{0.95}(\beta)$ if the constructed confidence interval contains the imputed value $\tilde{\beta}$ i.e. if $\tilde{\beta} \in (\hat{\beta}_{IV(\tilde{\beta})} - 1.96 \times S.E.(\hat{\beta}_{IV(\tilde{\beta})}),\ \hat{\beta}_{IV(\tilde{\beta})} + 1.96 \times S.E.(\hat{\beta}_{IV(\tilde{\beta})}))$,

For any feasible value β that falls between two adjacent feasible values $\tilde{\beta}_{LOWER}$ and $\tilde{\beta}_{HIGHER}$ in the selected grid of feasible values of parameter β, we use linear interpola-

tion to determine whether the feasible value is included in the confidence set estimate. Let $\tilde{\beta}^{LOWER}_{CI\_L\_BOUND}$ and $\tilde{\beta}^{LOWER}_{CI\_U\_BOUND}$ ($\tilde{\beta}^{HIGHER}_{CI\_L\_BOUND}$ and $\tilde{\beta}^{HIGHER}_{CI\_U\_BOUND}$) denote the lower and upper bounds of the confidence interval implied by the potential instrumental variable estimate computed using imputed value $\tilde{\beta}_{LOWER}$ ($\tilde{\beta}_{HIGHER}$). The lower and upper confidence interval bounds $\beta_{CI\_L\_BOUND}$ and $\beta_{CI\_U\_BOUND}$ associated with any parameter value β that falls between the adjacent values $\tilde{\beta}_{LOWER}$ nor $\tilde{\beta}_{HIGHER}$ in the selected grid of feasible values of parameter β are then constructed using expressions

$$\begin{aligned}\beta_{CI\_L\_BOUND} &= \tilde{\beta}^{LOWER}_{CI\_L\_BOUND} \\ &\quad +(\tilde{\beta}^{HIGHER}_{CI\_L\_BOUND} - \tilde{\beta}^{LOWER}_{CI\_L\_BOUND})\frac{\beta - \tilde{\beta}_{LOWER}}{\tilde{\beta}_{HIGHER} - \tilde{\beta}_{LOWER}} \end{aligned} \tag{75}$$

and

$$\begin{aligned}\beta_{CI\_U\_BOUND} &= \tilde{\beta}^{LOWER}_{CI\_U\_BOUND} \\ &\quad +(\tilde{\beta}^{HIGHER}_{CI\_U\_BOUND} - \tilde{\beta}^{LOWER}_{CI\_U\_BOUND})\frac{\beta - \tilde{\beta}_{LOWER}}{\tilde{\beta}_{HIGHER} - \tilde{\beta}_{LOWER}}. \end{aligned} \tag{76}$$

The feasible value $\beta \in (\tilde{\beta}_{LOWER}, \tilde{\beta}_{HIGHER})$ is included in the confidence set estimate $\Psi_{0.95}(\beta)$ if the interpolated confidence interval contains the feasible value β i.e. if $\beta \in (\beta_{CI\_L\_BOUND}, \beta_{CI\_U\_BOUND})$.

For other model parameters the confidence set estimate is constructed as a union of all those confidence intervals for each parameter that are implied for the parameter by any potential instrumental variable estimate calculated using an imputed value $\tilde{\beta}$ that is either included in the confidence set estimate $\Psi_{0.95}(\beta)$ or is adjacent to an imputed value that is included in the confidence set estimate $\Psi_{0.95}(\beta)$. Hence, we do not use linear interpolation in construction of confidence set estimates for other model parameters.

Background Appendix: Motivation and Related Literature for Analyses of Subjective Expectations

The rationale for the analysis and measurement of subjective expectations arises because observed choices are consistent with different combinations of preferences and expectations, and because solving this conundrum merely by assuming a specific form for expectations–such as rational expectations–is often implausible (Manski, 2004). With this motivation in mind, the literature on measurement of subjective expectations and on the impact of subjective expectations on outcomes has grown considerably during the last two decades.[22]

An important motivation for the study of subjective income expectations more specifically is to attain better analyses of consumption/savings decisions. For example, Guiso et al. (1992) use data on subjective income expectations to find that income uncertainty is not as important a determinant of precautionary saving than had been implied by prior studies based on indirect measures of income risk. Regarding the study of college expectations, an important rationale is that educational expectations may influence later educational outcomes. For example, Jacob and Wilder (2010) find a positive relationship between educational expectations and attainment among adolescents although they caution against interpreting either their own or related prior findings as evidence of causal effects.

The literature on the determinants of subjective income expectations include Dominitz and Manski (1996, 1997a), Blau and Ferber (1991), and Smith and Powell (1990), and Nicholson (2004). To our knowledge none of the contributions to this literature has examined the role of peer influence in the formation of income expectations. Contributions that examine the formation of subjective college completion or attendance expectations include Davies and Kandel (1981), Fischhoff et al. (2000), Kiuru et al. (2009) and Jacob and Wilder (2010). While Jacob and Wilder (2010) do not examine peer effects, the other three papers find that an individual's college expectations are correlated with the individual's peers' college expectations. An important aspect in many of these studies and in our study is that the subjects are adolescents. While ex-ante one might not expect adolescents to form sensible subjective expectations, prior studies have generally found that with the exception of subjective ex-

[22] Early contributions–such as Juster and Suzman (1995), Hurd and McGarry (1995), Dominitz and Manski (1997a, 1997b) and Manski (1993a)–during this surge of analyses focused mostly on methodological questions surrounding the measurement of expectations. Consistent with the long-term objective of this research program (as stated by Manski, 2004), some recent applications, including the analyses of retirement, crime, and contraception by Hurd et al. (2002), Lochner (2004) and Delavande (2008), respectively, have built on the earlier contributions to study the impact of expectations on outcomes. Yet another important strand of the expectations literature is the study of the determinants of the differences between subjective expectations and outcomes (see e.g. Souleles, 2004, and Jacob and Wilder, 2010).

pectations about mortality, the adolescents' subjective expectations are largely sensible (see Manski, 2004, Dominitz and Manski, 1996, and Fischhoff et al., 2000). Moreover, Quadrel et al. (1993) compare the risk perceptions of adolescents and their parents and conclude that their differences in cognitive decision-making processes are small, and Jacob and Wilder (2010) find support for the conclusion that adolescents form informed college expectations as they find that expectations are revised based on factors such as changes in grades and having a child.

The rationale for the inclusion of peer effects as possible determinants of subjective income and college expectations among adolescents is rooted in various theories that predict that an adolescent's choices are influenced by the choices and characteristics of the individual's peers, corresponding to the endogenous and exogenous social interaction effects, respectively. These include contagion based theories of problem behavior, role model and collective monitoring based theories, competition based theories, and relative status based theories (see e.g. Brooks-Gunn et al., 1993, and the references therein). Underlying our analysis of peer effects in the formation of subjective income or college completion expectations is a conjecture that some of these same mechanisms or yet another possibly information or social complementarity based mechanism may also induce also an adolescent's subjective expectations to be influenced by the individual's friends' subjective expectations and characteristics.[23]

[23]For example, if role models or information obtained from peers are important in the formation of subjective income expectations, an adolescent whose friends' parents are teachers, doctors or other professionals may be more inclined to expect to attain middle-income status than an adolescent whose friends' parents are unemployed. Similarly, if information obtained from peers or social complementarities are important in the formation of educational expectations, an adolescent whose friends expect to complete college may be more inclined to expect to complete college than an adolescent whose friends do not expect to complete college.