Social Networks with Unobserved Links

Arthur Lewbel
Boston College

Xi Qu
Shanghai Jiao Tong University

Xun Tang
Rice University

Original: July 2019, revised February 2021

Abstract

We point identify and estimate linear social network models without observing any network links. The required data consist of many small networks of individuals, such as classrooms or villages, with individuals that are each only observed once. We apply our estimator to data from Tennessee's Student/Teacher Achievement Ratio (STAR) Project. Without observing the latent network in each classroom, we identify and estimate peer and contextual effects on students' performance in mathematics. We find that peer effects tend to be larger in bigger classes, and that increasing peer effects would significantly improve students' average test scores.

JEL classification: C31, I21, C51

Keywords: Social networks, Peer effects, Unobserved network, Classroom performance

1 Introduction

In many social and economic environments, an individual's behavior or outcome depends on both his own characteristics and on the behavior and characteristics of other individuals. Call such dependence between two individuals a *link*. A *network* consists of a group of individuals who are potentially linked or connected. Links between individuals can take either binary values indicating the presence or absence of a connection, or continuous values (*weights*) indicating the strength of the connection. We refer to linked individuals as *friends*. The structure of a social network is fully characterized by its *adjacency matrix*, which is a square matrix that lists all links (continuous or discrete) among the group members.

One goal of econometric network models is the estimation of various social effects based on observed outcomes and characteristics of network members. These structural parameters capture the effects on each individual's outcome of (i) the individual's own characteristics (*direct effects*) and group characteristics (*correlated effects*), (ii) the characteristics of friends (*contextual effects*) and (iii) the outcomes of friends (*peer effects*).

Existing methods of point identifying and estimating these structural parameters require either that the adjacency matrix be observed in the sample (as in, e.g., Bramoullé, Djebbari and Fortin (2009)), parameterized as in Rose (2017) or as the linear-in-means model described below, or that the reduced-form coefficients that correspond to a fixed, unknown network are already identified (as in Blume, Brock, Durlauf and Jayaraman (2015) and de Paula, Rasul and Souza (2018)). The usual way this latter requirement would be satisfied is by observing many repeated realizations of covariates and outcomes over a fixed unknown network.

We provide sufficient conditions to point identify and estimate the structural parameters in linear social network models when the adjacency matrix is unobserved, and where only a single realization of covariates and outcomes in each network is observed. Our identification assumes that we observe outcomes and covariates for individuals in many small networks such as classrooms or villages, but does not require any data on who is linked with whom within each network. Since most surveys do not include link data, our results have widespread potential applications.

For example, consider students who participated in the Student/Teacher Achievement Ratio (STAR) project. This data includes test scores and demographic information on each student, and reports what class each student is in, but does not provide any link data, such as which sets of children are friends or study partners within each class. Previous attempts to estimate peer effects with this data either assume a linear-in-means model where all classmates are assumed to be linked to each other with equal weights, e.g., Boozer and Cacciola (2001), or define links as functions of observed covariates as in Rose (2017).

While often assumed in practice, the linear-in-means assumption is very unlikely to hold

in many applications like classrooms, where peer and contextual effects are more likely to operate through actual friendships with varying strengths, instead of equal influence from all group members. We also show how to use our identification results to empirically test the linear-in-means assumption. We reject this assumption in the STAR data.

1.1. The Model. Let $y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^K$ denote the outcome and exogenous covariates, respectively, for an individual i. Each individual belongs to one of L groups, a.k.a. networks. Assume there are n_l individuals in each group $l \in \{1, ..., L\}$. Each group l has an unobserved $n_l \times n_l$ adjacency matrix G_l, whose (i, j)-th component is either binary (equals 1 if i is linked to j, and 0 otherwise), or is a generic number (a weight) indicating the strength of the link between i and j.[1]

The researcher only observes y_i and X_i for each individual i, and the identity of the group that each individual i belongs to. The researcher does *not* observe the adjacency matrices G_1,...G_L. For example, suppose each group is an elementary school class, and each G_l describes a network of friendships or study partners among the students in class l. The researcher observes each student i's test score y_i and the student's vector of demographic and other characteristics X_i. The researcher also observes which class (i.e., group) each student is in, but does not observe who is friends with whom, or who studies with whom, within each class. Instead of observing or modeling the adjacency matrices of each group (i.e., class), we only assume that there is an unknown distribution of latent adjacency matrices, from which each group's matrix G_l is drawn.

We assume a standard linear social interactions model:[2]

$$y_l = \alpha\iota + \lambda G_l y_l + X_l\beta + G_l X_l\gamma + \varepsilon_l, \tag{1}$$

where y_l and ε_l are $n_l \times 1$ vectors of outcomes and errors, respectively, ι an $n_l \times 1$ vector of ones, and X_l an $n_l \times K$ matrix of covariates. Assume for now that the errors ε_l are i.i.d. and uncorrelated with X_l (these conditions can be relaxed). Our asymptotics are that the number of members n_l of each network l is fixed, but the total number of networks L goes to infinity. Our goal is point identification and estimation of structural parameters consisting of the group effect coefficient $\alpha \in \mathbb{R}$, the peer effect $\lambda \in \mathbb{R}$, the vector of individual direct effects $\beta \in \mathbb{R}^K$, and the vector of contextual effects $\gamma \in \mathbb{R}^K$. We will later separate X_l into individual-level effects and group-level effects, with an additional vector δ of group-level coefficients.

[1] Links are typically assumed to be nonnegative in network models, but we do not need to impose that constraint.

[2] Note this linear model is far more general than linear-in-means. Linear-in-means is the very special case where every off-diagonal element of G_l is the same number.

If the adjacency matrix G_l were observed for each group l in the sample, then point identification and estimation of these parameters under general conditions would follow from existing methods in the literature. For example, one could use the linear instrumental variables estimator of Bramoullé, Djebbari and Fortin (2009), which uses data on friends of friends, i.e., $G_l^2 X_l$, as instruments for endogenous regressors $G_l y_l$.

1.2. Intuition for Identification and Estimation. To explain the intuition for our identification strategy, let us continue to use the example of students in a class. Begin by making the simplifying assumption that all classes are the same size, having n students per class (later, in Section 6.3, we describe multiple methods of generalizing our results to handle variation in group sizes).

Equation (1) says that each element of y_l (that is, each student's test score) is a linear function of the characteristics of that student, and of the test scores and characteristics of that student's friends. One could imagine trying to directly estimate these linear functions by linear regressions. However, we don't know who each student's friends are. Moreover, even if we did know, the test scores of friends are endogenous regressors. Without observing the adjacency matrices, we can't construct instruments for ones friend's test scores, as in Bramoullé, et. al. (2009).

So instead, consider estimating reduced-form regressions, where we solve equation (1) for y_l as a function of X_l and errors. In these regressions, each student's test score is regressed on all the characteristics of every child in that student's class. This means estimating the regression coefficients in a system of n linear equations (one equation for each class member), with each regression estimated using a sample of size L (the number of classes in the sample). The coefficients in these reduced-form regressions are all functions of the structural parameters, and of the underlying distribution of adjacency matrices across classes. More precisely, we show these reduced-form coefficients are all functions of the structural parameters α, β, and γ, and of $E(M_l)$ and $E(M_l G_l)$, where $M_l \equiv (I - \lambda G_l)^{-1}$, I denotes the identity matrix, and the expectations are over the unknown distribution of random matrices G_l across all classes.

We establish sufficient conditions for identifying the structural parameters from these reduced-form coefficients. These identification conditions are analogous to the traditional rank and order conditions for identifying structural parameters in classical linear simultaneous equation systems (e.g., systems of linear supply and demand equations). However, a complicating factor here, as compared to classical linear simultaneous systems, is the presence of many nuisance parameters, specifically, all the elements of the matrices $E(M_l)$ and $E(M_l G_l)$.

A key insight is that we do not need to observe or identify all of the adjacency matrices G_1,...G_L. For identifying the structural coefficients α, β, and γ, the only features of the

network that matter are $E(M_l)$ and $E(M_l G_l)$. What then makes this identification feasible is these matrices affect the reduced-form coefficients of each covariate in X_i in the same way. So having multiple covariates in the model provides identifying information regarding these matrices. As a result, from the reduced-form coefficients we can disentangle and identify the structural social effects, without observing the network, and without explicitly modeling either network structure or network formation.

An attractive feature of our identification strategy is that it is constructive, so the same steps used for identification can be replicated in data to obtain parameter estimates. Unlike traditional indirect least squares for linear simultaneous equations (recovering structural parameters from reduced-form estimates), our estimator requires a first step to estimate intermediate parameters. These intermediate parameters depend on the structural social effects but not the distribution of latent matrices.

Another attractive feature of our estimator is that, unlike other estimators that deal with unobserved networks, we do not need to either parameterize the networks, nor do we require repeated observations of the network. Moreover, since we identify and estimate functions of $E(M_l)$ and $E(M_l G_l)$, which are features of the distribution of adjacency matrices, we can use our estimates to test some models of link formation, such as testing if the linear-in-means model holds, or testing if links are determined randomly.

1.3. Classroom outcomes in Tennessee elementary schools. We apply our method without link data to estimate the impact of social networks on the test performance of elementary school students in the STAR data set mentioned above. For example, without observing any data on the links between students, we identify the peer effects coefficient λ, and estimate it to be 0.85 in small classes and 0.92 in large classes. Both estimates are statistically significant. We also find that, ceteris paribus, increasing the magnitude of peer effects would result in improved average test scores.

Would it be worthwhile to institute policies that encourage students to form additional links or friendships? Our results suggest that the impacts of such policies would be small, and could even have negative effects depending on class sizes. This is an example of a counterfactual exercise we can perform that would be difficult by other means with this data. We also test and reject alternative model specifications, including the linear-in-means model, and we also reject the random Poisson link formation model (also known as Erdős-Rényi (1959) networks).

The next section is a short literature review. It is followed by our formal model. We then present our new identification and estimation method for unobserved networks. Next, we provide the empirical application and conclusions. Proofs, derivations, and Monte Carlo simulations are in the appendix.

2 Literature Review

Standard estimators of social interactions models, like Lee (2007), Bramoullé, Djebbari and Fortin (2009), and Lin (2010) assume network links are reported in the data. One popular model that does not require observing the network is the "linear-in-means" model. This model simply assumes that everyone is equally linked to everyone else, either within groups, or in the entire network. So in that model, a simple network is assumed rather than observed.

An alternative to just assuming an unobserved network is to exploit alternative types of network information. For example, one may use spatial data to estimate adjacency matrices, assuming that G_l is a function of observed geographic distance or demographic difference. Examples are in Pinkse, Slade, and Brett (2002), LeSage and Pace (2009), Manresa (2016) and Rose (2018).

Another possibility is to assume a model of network formation, and estimate the resulting, possibly endogenous, network along with the structural model parameters. An example is an Erdős-Rényi (1959) network, which assumes that there is a fixed probability p that any element of G_l equals one versus zero. One might then estimate p along with structural parameters (we later show with our model that we can test the assumption of an Erdős-Rényi network, and we reject it in our application). Other endogenous network formation models are Hsieh and Lee (2016), Goldsmith-Pinkham and Imbens (2013), Hsieh, König, and Liu (2020), and Hsieh, Lee, and Boucher (2020). Many of the results in this endogenous network literature yield set rather than point identification, or are analyzed under the Bayesian framework.[3]

Another approach is to assume the researcher has additional information about the effects of the network, rather than additional information about its formation or structure. For example, if a survey directly asks questions related to the value of peers' outcomes and contextual effects, e.g., about $G_l y_l$ and $G_l X_l$, then the peer effects might be estimated without observing the network G_l itself. An example is Breza et al. (2020). Alternatively, Blume et al. (2015) provide identification results assuming that the reduced-form coefficients of individual characteristics on outcomes are already known to researchers. Obtaining these reduced-form coefficients would generally require many repeated observations of the same individuals in a fixed network, or observations of many groups, each of which was known to have the exact same network structure.[4]

[3] While we obtain point identification without making use of any specific model of network formation, we do require a relatively strong exogeneity condition regarding network formation versus outcomes. See Assumptions 2 and 3 below.

[4] Blume et al (2015) also consider a more general model where adjacency matrices for peer effects and for contextual effects are different. They show how to identify structural coefficients using partial knowledge

Perhaps the closest result to ours is de Paula, Rasul and Souza (2018), who identify and estimate a linear social network model where the network is completely unobserved, without additional information about networks or outcomes as above. They show identification assuming a panel data structure where researchers observe outcomes across multiple periods on a single fixed network. In their model, individual outcomes vary over time conditional on covariates, because they are generated by random draws of unobserved errors in each time period, while the unknown network structure is assumed constant over time. Given many time periods (or fewer time periods and some sparsity assumptions), they propose a consistent estimator for the social effects.

The assumptions we require to deal with unobserved networks are motivated by a different data structure than de Paula, Rasul and Souza (2018). While our methods could be applied to their data, unlike them we do not require a panel structure with the network fixed over time. Our method allows the unobserved network to vary across groups (e.g., classes or villages), and so could be applied in a cross-sectional setting where the network varies across groups within a single observed time period. Asymptotics in our case are defined in terms of the number of groups (each of which only needs to be observed once) going to infinity, rather than number of repeated observations of a single group.

Our identification argument also differs qualitatively from de Paula, Rasul and Souza (2018) in that ours is based on the relationship between the reduced-form impacts of multiple individual characteristics on outcomes. Also, our identification strategy is constructive, and thus leads to a simple two-stage estimator that has a closed form, is easy to compute, and attains standard parametric rate consistency and asymptotic normality.

Our empirical application looks at peer effects on students' academic performance. Other linear model studies of peer effects on student outcomes include Hauser et al. (2009), Calvó-Armengol et al. (2009), Lin, (2010), Lee et al.(2010), Patacchini and Zenou (2012), and Boucher et al. (2014).

A limitation of our model in equation (1) is that it assumes that peer effects λ and contextual effects γ operate through the same adjacency matrix G_l. This assumption is standard in the literature whenever both peer and contextual effects are included in a model; See, e.g., Lee (2007), Bramoullé et al (2009), and de Paula et al (2018). One paper that relaxes this assumption is Blume et al (2015). This assumption is generally imposed because it would be difficult to distinguish from data the extent to which any observed link applies to peer effects versus to contextual effects. We are not aware of any data sets where such information has been collected. However, since our identification is intended precisely to cover situations where link data is not, or cannot, be observed, it is possible that our methods

of both matrices (i.e., the complete set of individuals linked) and a priori restrictions on the cardinality of these links. Whether their model can be identified without such a priori restrictions is an open question.

could be extended to cover such models. We discuss the possibility of extending our method to cover this case of multiple adjacency matrices within each group in Appendix E.

We conclude this literature review by noting a deep connection between identification of linear network models and identification of traditional structural systems of linear equations, going back to the rank and order conditions described by Koopmans (1949) and the Cowles foundation, and in more detail in Fisher (1966). First, consider the setting in de Paula, Rasul and Souza (2018), which is equation (1), but simplified by having $G_l = G$ and $n_l = n$, the same for all groups l (i.e. the number of members and the adjacency matrices are the same for all groups). Let $\widetilde{X}_l$ be a column vector that stacks all Kn elements in X_l and a constant term. We can write the model in (1) as

$$y_l = Ay_l + B\widetilde{X}_l + \varepsilon_l \tag{2}$$

where $A = \lambda G$ and $B\widetilde{X}_l = \alpha\iota + X_l\beta + GX_l\gamma$, so the elements of the matrix of coefficients B are functions of G, α, β, and γ. Equation (2) is a system of n linear equations. The reduced form (defined by solving for the endogenous y_l in terms of the exogenous covariates X_l) of equation (2) is

$$y_l = C\widetilde{X}_l + (I - A)^{-1}\varepsilon_l$$

where $C = (I - A)^{-1} B$. With L (the number of groups) large enough, one can identify and estimate the reduced-form matrix of coefficients C, by linearly regressing each element of y_l on the vector of regressors $\widetilde{X}_l$, assuming ε_l is uncorrelated with $\widetilde{X}_l$.

Some form of rank and order conditions are then needed to identify the structural coefficients A and B from C, and additional rank and order conditions would be needed to recover G, λ, α, β, and γ from A and B (or to just recover λ, α, β, and γ in standard models where G is known). By construction, A and B are functions of $n^2 + 2 + 2K$ structural parameters (G, λ, α, β, γ) while C consists of $n \times (Kn + 1)$ reduced-form coefficients. Thus it is straightforward to verify the order condition for identifying A and B from C for a given pair of n and K. The issues for identification here are rank conditions for identifying A and B, and the for recovering the structural parameters given A and B.

The linear-in-means model, which corresponds to a G having all off-diagonal elements equal to $1/n$, suffers from the "reflection problem" as pointed out by Manski (1993). The reflection problem is a failure to obtain identification because of a violation of the rank condition. As in ordinary linear simultaneous systems, the most common solution to the reflection problem is to regain identification by imposing exclusion assumptions, e.g., by assuming some contextual effects are zero as in Graham and Hahn (2005). In the above notation, this is equivalent to assuming some elements of γ equal zero, thereby restricting the matrix B and hence C to satisfy the rank condition. Both Blume et. al (2015) and

de Paula, Rasul and Souza (2018) can also be interpreted as providing rank conditions that suffice for identifying structural parameters from reduced form coefficients.

Our model of unobserved networks does not rule out linear-in-means networks as a special case, and so we also require exclusion assumptions for identification. Our model is more complicated than equation (2) in that we let the unobserved adjacency matrices G_l vary across groups $l \in \{1, ..., L\}$. So in our model equation (2) is replaced by

$$y_l = A_l y_l + B_l \widetilde{X}_l + \varepsilon_l.$$

Instead of the fixed matrices of coefficients A and B as in equation (2), variation in G_l across groups gives rise to matrices of random coefficients A_l and B_l. As a result we first identify and estimate a mean reduced-form matrix $C = E\left[(I - A_l)^{-1} B_l\right]$ where the expectation is over the distribution of random matrices A_l and B_l. Then, by making use of some exclusion (i.e. rank) restrictions, from this C we point identify the structural parameters λ, α, β, and γ, along with some features of the distribution of the random G_l matrices.

3 The Model

Let the data-generating process (DGP) be as specified in Section 1.1. The data consist of independent networks, or groups, indexed by $l = 1, 2, ..., L$. Examples of groups could be classrooms or villages. Each group l consists of n_l individual members, and has an n_l-by-n_l adjacency matrix G_l. These adjacency matrices vary across the groups, and are *not* reported in the data. What is observed are the outcomes and covariates of every member of each observed group l. Each group is only observed once.[5] We don't impose specific assumptions on how the latent, unobserved adjacency matrices are formed; instead we model them as independent draws from some unknown distribution of possible networks.

By convention in the literature, the diagonal entries in each G_l are all zeros, i.e., $G_{lii} = 0$ for $i = 1, ..., n_l$. The off-diagonal entries $G_{lij} \in \mathbb{R}$ measure the strength of the link between individuals i and j, with $G_{lij} = 0$ signifying the absence of a link. The unobserved adjacency matrices G_1, ..., G_L are assumed to be row normalized. That is, given a group adjacency matrix G_l^*, the (i, j)-th component in the row-normalized version G_l is $G_{lij} = G_{lij}^* / \left(\sum_{j'=1}^{n_l} G_{lij'}^*\right)$, where the sum in the denominator is positive almost surely. Row normalization imposes some strong behavioral restrictions, but is a commonly maintained assumption in the social networks literature.

[5] For identification and consistent estimation, there is no problem if outcomes and covariates of some or all groups are observed more than once. However, in that case, the asymptotic distribution of our estimator would need to account for the resulting correlation in errors and adjacency matrices across multiple observations of the same group.

For each individual i in the sample, it is assumed the group l that individual i belongs to is known. This is a sensible assumption in many applications, because groups are often defined by public information. Examples include geographic boundaries as in Banerjee et al (2017), where each l indexes a village, or registration/enrollment records such as class enrollment in the Add Health data set (see, e.g., Hunter et. al. 2008), where each l indexes a school-grade pair.

To fix ideas, for now let all groups in the data-generating process be of the same size $n_l = n$. Later we will relax this assumption by dividing the population into subgroups s, and allowing the group size (and some model coefficients) to vary by s. Another simplification we impose for now is to exclude any group-level variables from X_l. This means none of the columns in the matrix X_l consists of n identical entries. We can extend our method to accommodate such group-level variables; details are deferred to Section 6.1.

To save on notation, we suppress the subscript l while presenting identification results below.

Let X_{ck} denote the k-th column in X. That is, X_{ck} is an $n \times 1$ vector of the k-th regressor for all members in a group. The subscript c serves as a reminder that the index is for columns. Let $\tilde{X} \equiv (1, X_{c1}', X_{c2}', ..., X_{cK}')'$ denote a $(Kn+1) \times 1$ vector that stacks the regressors for all individuals in a group.

Assumption 1 *(Population model) The outcomes on the social network is determined by* $y = \alpha\iota + \lambda Gy + X\beta + GX\gamma + \varepsilon$*, where* y *and* ε *are* n*-by-*1*,* G *is* n*-by-*n*,* X *is* n*-by-*K*,* β *and* γ *are* K*-by-*1*, and* λ *is a nonzero scalar.* G *is row normalized, so the sum of the elements in every row of* G *equals one.*

Assumption 2 *(Exogenous networks)* $E(\varepsilon \mid G, X) = 0$*.*

Assumption 3 *(Independence)* G *is independent of* X*.*[6]

Assumption 4 *(Invertibility and no perfect collinearity) (i)* $E(\tilde{X}\tilde{X}')$ *exists and is non-singular. (ii)* $I - \lambda G$ *is invertible with probability one. (iii) All elements in* $E(M)$ and $E(MG)$ *are bounded above by a finite constant, where* $M \equiv (I - \lambda G)^{-1}$ *and* I *is the identity matrix.*

Assumption 5 *(Non-trivial effects) (i) For each* $k < K$*, the* 2*-by-*2 *matrix*

$$\begin{pmatrix} \beta_k & \beta_K \\ \gamma_k & \gamma_K \end{pmatrix}$$

has full rank. (ii) $\mu_K \neq cI$ *for any* $c \in \mathbb{R}$*, where* μ_K *is a matrix of reduced-form coefficients for the* K*-th regressor as defined in equation (4).*

[6] This condition can be replaced by "G^r is mean independent of X_l for all integers r". We later discuss how this condition can be further relaxed to allow dependence of G on some covariates.

Our method can be generalized to where Assumptions 2 and 3 hold conditional on other covariates. However, this extension adds notation and complicates the presentation, so we defer it to Appendix D, where it is discussed in detail. Nevertheless, even with this generalization, Assumptions 2 and 3 are strong restrictions, requiring that networks be conditionally exogenous. They rule out many kinds of potential endogeneity in group or link formation that could arise from unobserved heterogeneity.

In Assumption 4, invertibility of M is a common assumption in the literature. It holds, for example, if $|\lambda| < 1$, $\sum_{j \leq n} |\lambda G_{ij}| < 1$ for all $i \leq n$, and G is bounded in its norm. Row normalization, which we imposed in Assumption 1, is also often used in network models to facilitate invertibility. Given Assumption 4, we can obtain the reduced form of the population model $y = \alpha\iota + \lambda Gy + X\beta + GX\gamma + \varepsilon$ as

$$y = M\left(\alpha\iota + X\beta + GX\gamma + \varepsilon\right). \tag{3}$$

Part (i) of Assumption 5 rules out the pathological case where all pairs of regressors have proportional contextual and peer effects. As long as one regressor has contextual and peer coefficients that are not proportional to those of any other regressor, we can reorder the columns of X to make that regressor be the K-th regressor to satisfy part (i). A sufficient but not necessary condition for part (i) is if $\gamma_K = 0$ (one of the regressors has no contextual effect) while β_K, β_k, and γ_k are all nonzero for all $k < K$. Part (ii) of Assumption 5 rules out another pathological case, where the K-th regressor of each individual i has identical marginal effects on its own expected outcome, but no impact on that of any other group member.

In addition to Assumptions 1 to 5, to obtain identification we will require some exclusion restrictions, to satisfy a rank condition. These are discussed at length in Section 4.1.

4 Identification

Define *reduced-form parameters* to be the coefficients of a linear projection of y on $\tilde{X}$ in the population. As we show below, $E\left(y \mid \tilde{X}\right)$ is linear in $\tilde{X}$. Hence the reduced-form parameters can be alternatively defined as the coefficients of $\tilde{X}$ in this conditional expectation. The first step of our identification strategy is to show how these reduced-form parameters relate to the structural components of our model.

We arrange these reduced-form parameters into matrices μ_k for $k = 0, ..., K$, as follows. For each characteristic indexed by $k = 1, 2, ..., K$, let the (i, j)-th component in μ_k be the marginal effect of the k-th characteristic in X for individual j on the mean outcome of individual i. These marginal effects are just the coefficients of the elements of $\tilde{X}$ in the projections or conditional expectations described above.

Lemma 1 *Under Assumptions 1-4, the identified reduced-form parameters satisfy the following equations*

$$\begin{aligned} \mu_k &\equiv \beta_k E(M) + \gamma_k E(MG) \text{ for } k = 1, ..., K; \\ \mu_0 &\equiv \alpha/(1-\lambda). \end{aligned} \tag{4}$$

The proof of lemma 1 is Appendix A, but the intuition is as follows. Let y_i denote the outcome for individual i. By construction,

$$E(y_i \mid X) = \mu_0 + e_i E(M) X\beta + e_i E(MG) X\gamma, \tag{5}$$

where e_i is a $1 \times n$ unit-vector whose i-th component is 1. Observe that the right-hand side of (5) is linear in all Kn components of X, so $E\left(y \mid \tilde{X}\right)$ is linear in $\tilde{X}$. This equation holds because G and M are independent from X by Assumption 3, and $E(M\varepsilon \mid X) = E[ME(\varepsilon \mid X, G) \mid X] = 0$ by Assumption 2. The equality in (5) also uses the fact that the row normalization of G imply

$$\alpha M\iota = \alpha \left[\sum_{s=0}^{\infty} (\lambda G)^s\right] \iota = \mu_0 \iota. \tag{6}$$

The second equality here holds because, by row normalization, each row of M adds up to the same constant $1/(1-\lambda)$.

In the reduced form of equation (5), the slope coefficient for the k-th regressor of individual j is $\beta_k \left[e_i E(M) e_j'\right] + \gamma_k \left[e_i E(MG) e_j'\right]$. (Note that, for a generic $n \times n$ matrix Q, the product $e_i Q e_j'$ returns the (i,j)-th component in Q.) The full rank and the invertibility conditions in Assumption 4 guarantee the identification of these reduced-form coefficients. These identified vectors of regressor coefficients are then arranged into the K matrices of reduced-form coefficients μ_k for $k = 1, ..., K$.

Remark 1 *The representation of $E(y \mid X)$ in (5) is consistent not only with the simultaneous social network model with complete information given by equation (1), but also with an alternative model in which individuals have private information and rational expectations regarding peer outcomes:*

$$y = \alpha\iota + \lambda GE(y \mid G, X) + X\beta + GX\gamma + \varepsilon, \tag{7}$$

where the ε's are private shocks that are independent of other group members conditional on the commonly known G and the exogenous characteristics X. In equation (1), individuals have complete information about others in the same group and outcomes are simultaneously determined. In comparison, each group member in equation (7) has private shocks, and outcomes are determined through rational expectations of others' outcomes, conditional on

each individual's information set (G, X)*. Both models lead to the same representation of the conditional mean function*

$$E(y \mid G, X) = (I - \lambda G)^{-1}(\alpha\iota + X\beta + GX\gamma),$$

which in turn implies (5) under Assumption 3.

In the next lemma, we construct $2(K-1)$ intermediate parameters a_k and b_k for $k = 1, ..., K-1$ from the reduced-form coefficients μ_k. Later, the final step of the identification will recover the structural parameters λ, β, γ from these intermediate parameters a_k and b_k.

Lemma 2 *Suppose Assumptions 1-5 hold. Then for each* $k < K$*, the equation*

$$a_k\mu_k + b_k\mu_K = I \tag{8}$$

has a unique solution $(a_k, b_k) \in \mathbb{R}^2$*, where*

$$\begin{pmatrix} a_k \\ b_k \end{pmatrix} = \begin{pmatrix} \beta_k & \beta_K \\ \gamma_k & \gamma_K \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ -\lambda \end{pmatrix}. \tag{9}$$

Proof of Lemma 2. For any $k = 1, ..., K-1$, the inverted matrix on the right-hand side of (9) has full rank under condition (i) in Assumption 5. Hence the solution (a_k, b_k) is well-defined, and $(a_k, b_k) \neq (0, 0)$. By construction, $a_k\beta_k + b_k\beta_K = 1$ and $a_k\gamma_k + b_k\gamma_K = -\lambda$. Therefore,

$$a_k\mu_k + b_k\mu_K = E[M(a_k\beta_k I + a_k\gamma_k G + b_k\beta_K I + b_k\gamma_K G)] = E[M(I - \lambda G)] = I.$$

Next, we show that for each k, (a_k, b_k) as defined in (9) is the *unique* solution for (8). That is, there exists no $(\tilde{a}_k, \tilde{b}_k) \neq (a_k, b_k)$ such that

$$(\tilde{a}_k - a_k)\mu_k + (\tilde{b}_k - b_k)\mu_K = \mathbf{0}. \tag{10}$$

Consider three mutually exclusive cases. Case 1: $\tilde{a}_k = a_k$, $\tilde{b}_k \neq b_k$. Then (10) requires $\mu_K = \mathbf{0}$. Case 2: $\tilde{a}_k \neq a_k$, $\tilde{b}_k = b_k$. Then (10) requires $\mu_k = \mathbf{0}$. This in turn implies μ_K must be a scalar multiple of I in order for (8) to hold for $(\tilde{a}_k, \tilde{b}_k)$. Case 3: $\tilde{a}_k \neq a_k$, $\tilde{b}_k \neq b_k$. Then (10) requires $\mu_k = -\frac{\tilde{b}_k - b_k}{\tilde{a}_k - a_k}\mu_K$, which is a scalar multiple of μ_K. Again, this implies that in order for (8) to hold for $(\tilde{a}_k, \tilde{b}_k)$, μ_K must be a scalar multiple of I. In each of these three cases, the implication of (10) contradicts part (ii) of Assumption 5. □

The reduced-form coefficients μ_0 and μ_k are identified by Lemma 1. Therefore, for each $k \leq K-1$, (a_k, b_k) can be recovered as the unique solution to equation (8). For each k, this matrix equation yields n^2 equalities, namely, $a_k\mu_{k,ij} + b_k\mu_{K,ij} = 0$ for all $i \neq j$ and $a_k\mu_{k,ii} + b_k\mu_{K,ii} = 1$ for all i, where i and j go from 1 to n. In Section 5, we construct an

estimator for each pair (a_k, b_k) by minimizing the L_2-distance between $a_k\mu_k + b_k\mu_K$ and the identity matrix.

Now consider identification of the structural parameters (λ, β, γ) given a_k and b_k. Lemma 2 provides the linear equations

$$\begin{pmatrix} \beta_k & \beta_K \\ \gamma_k & \gamma_K \end{pmatrix} \begin{pmatrix} a_k \\ b_k \end{pmatrix} = \begin{pmatrix} 1 \\ -\lambda \end{pmatrix} \text{ for } k = 1, ..., K-1. \tag{11}$$

And, by the row normalization of G in Assumption 1, we get the additional equations

$$m_k \equiv (\iota'\mu_k\iota)/n = \frac{\beta_k + \gamma_k}{1-\lambda} \text{ for } k = 1, ..., K, \tag{12}$$

where m_k is the sum of components in μ_k divided by n, which is identified due to Lemma 1.

Combining equations (11) and (12) yields a system of $2(K-1)+K$ linear equations for $2K+1$ parameters $\theta \equiv (\lambda, \beta', \gamma')'$ with $\beta \equiv (\beta_1, \beta_2, ..., \beta_K)'$ and $\gamma \equiv (\gamma_1, \gamma_2, ..., \gamma_K)'$. The rank of the matrix of coefficients for θ in this linear system is at most $2K-1$, because $a_k m_k + b_k m_K = 1$ for all $k < K$ by construction.

To illustrate, the system of linear equations we obtain from equations (11) and (12) when $K = 3$ is:

$$\begin{pmatrix} 0 & a_1 & 0 & b_1 & 0 & 0 & 0 \\ 0 & 0 & a_2 & b_2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & a_1 & 0 & b_1 \\ 1 & 0 & 0 & 0 & 0 & a_2 & b_2 \\ m_1 & 1 & 0 & 0 & 1 & 0 & 0 \\ m_2 & 0 & 1 & 0 & 0 & 1 & 0 \\ m_3 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ m_1 \\ m_2 \\ m_3 \end{pmatrix}. \tag{13}$$

Here $\theta = (\lambda, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3)'$ which has seven elements, while the rank of the matrix that multiplies θ in equation (13) is bounded above by five.[7] For general cases with $K > 3$, the linear system is:

$$\underbrace{\begin{pmatrix} 0_{(K-1)\times 1} & H & 0_{(K-1)\times K} \\ \iota_{(K-1)\times 1} & 0_{(K-1)\times K} & H \\ m & I & I \end{pmatrix}}_{\pi} \underbrace{\begin{pmatrix} \lambda \\ \beta \\ \gamma \end{pmatrix}}_{\theta} = \underbrace{\begin{pmatrix} \iota_{(K-1)\times 1} \\ 0_{(K-1)\times 1} \\ m \end{pmatrix}}_{\tau}, \tag{14}$$

with $m \equiv (m_1, m_2, ..., m_K)'$, I is a $K \times K$ identity matrix, and H is a $(K-1)$-by-K matrix constructed from $(a_k, b_k)_{k=1,...,K-1}$ as follows:

$$H \equiv [diag(a_1, ..., a_{K-1}),\ (b_1, b_2, ..., b_{K-1})'].$$

[7] To see this, note that the sum of the first and the third row equals a weighted sum of the fifth and the last row (as $a_1 m_1 + b_1 m_3 = 1$ by construction). Likewise, the sum of the second and fourth rows equals a weighted sum of the last two rows.

The rank of the π matrix is generically $2K-1$. It cannot be greater than $2K-1$ by construction, and is strictly less than $2K-1$ only if the data generating process generates one or more pathological equality constraint coincidences among the a_k, b_k, and m_k terms.

Next, we define what we call an *environment*. An environment s is a subpopulation of groups, defined by observable information, that satisfies Assumptions 1 to 5. Each group lies in one and only one environment. Let S denote the number of environments in the population. We allow all model parameters and group sizes to vary by environment, and so all can be given an s superscript. For example, to accommodate data that has groups of different sizes, we can assume a different environment s for each possible group size $n^{(s)}$ (additional ways to deal with varying group sizes are discussed later). Structural parameters can vary by environment, so identification of the model requires that we identify $\theta^{(s)} \equiv (\lambda^{(s)}, \beta^{(s)}, \gamma^{(s)}) \in \mathbb{R}^{2K+1}$ for each environment s.

Repeating the construction of equation (14) for each environment $s = 1, ..., S$, we obtain S linear systems

$$\pi^{(s)}\theta^{(s)} = \tau^{(s)} \text{ for } s = 1, 2, ..., S,$$

We then stack these S systems to get

$$\Pi\boldsymbol{\theta} = d,$$

where $\boldsymbol{\theta}$ and d are column vectors that stack $\theta^{(s)}$ and $\tau^{(s)}$ respectively for $s = 1, ..., S$; and Π is a block-diagonal matrix with diagonal blocks $\pi^{(s)}$.

Finally, suppose there are additional known linear equality constraints that hold among the elements of $\boldsymbol{\theta}$. For example, some structural parameters might take the same value in different environments, or one or more structural parameters might be known to equal zero (i.e., exclusion restrictions). Denote these additional restrictions by $R\theta = c$ where R and c are known a priori (see the next subsection for details). Let $\Psi \equiv [\Pi; R]$ denote a combined coefficient matrix constructed by stacking Π on top of R, and define the vector $v \equiv (d', c')'$. We can then summarize all these equality constraints by the linear system

$$\Psi\boldsymbol{\theta} = v.$$

Theorem 1 *Assume the population consists of S environments for some fixed constant S. Let Assumptions 1-5 hold for each environment $s = 1, ..., S$. Assume that Ψ has full rank. Then $\lambda^{(s)}$, $\beta^{(s)}$, $\gamma^{(s)}$,and $\alpha^{(s)}$ for $s = 1, ..., S$ are all identified.*

To prove Theorem 1, we first get identification of $\boldsymbol{\theta}$, and hence of $\lambda^{(s)}$, $\beta^{(s)}$, and $\gamma^{(s)}$ for all s, by $\boldsymbol{\theta} = (\Psi'\Psi)^{-1}\Psi'v$. Then, using lemma 1, $\alpha^{(s)}$ is identified by $\alpha^{(s)} = (1 - \lambda^{(s)})\mu_0^{(s)}$.

Once the structural parameters are identified, equation (4) (which can now vary by environment s) provides equality constraints that the matrices $E(M^{(s)})$ and $E(M^{(s)}G^{(s)})$

must satisfy. These constraints are not sufficient to identify moments of the distribution of the adjacency matrix itself, but they provide restrictions that we will later use to test hypotheses about the networks, such as whether they are linear-in-means or not.

We discuss restrictions that suffice to give Ψ full rank, as required by Theorem 1, in the next section.

4.1 Rank restrictions

To satisfy the rank condition in Theorem 1, we require restrictions of the form $R\boldsymbol{\theta} = c$. The number of rows in R must at least equal the number of required restrictions on the coefficients $\boldsymbol{\theta}$ to satisfy the rank condition in Theorem 1. This number depends on both the number of regressors K and the number of environments S. For example, with $K = 3$ and $S = 1$, we require two additional linear restrictions on $\boldsymbol{\theta}$ to make Ψ full rank.

One way to see why rank restrictions like $R\boldsymbol{\theta} = c$ are needed is to consider Manski's (1993) reflection problem again. Manski's linear-in-means social interactions model is a special case of our model where $G^{(s)}$ is the same for all groups in each environment s, and where all off-diagonal elements of $G^{(s)}$ equal $1/n^{(s)}$. The reflection problem shows that in this model, even if $G^{(s)}$ were known, the structural parameters would not be identified without additional restrictions. Since our model includes this linear-in-means model as a special case, we must require at least as many additional restrictions for identification.[8]

There are two types of rank restrictions that are most natural to impose. The first type are exclusion restrictions, which consist of assuming that some elements of either β or γ equal zero (like the exclusion restrictions commonly used to identify linear simultaneous systems of equations). Graham and Hahn (2005) use such exclusion restrictions to identify the linear-in-means model.[9] To illustrate, suppose $K = 3$ and $S = 1$. In this case it suffices to assume that one regressor X_k has no contextual effect ($\gamma_k^{(1)} = 0$) and a non-zero direct effect ($\beta_k^{(1)} \neq 0$), while another regressor $X_{k'}$ has no direct effect ($\beta_{k'}^{(1)} = 0$) and a non-zero contextual effect ($\gamma_{k'}^{(1)} \neq 0$). More generally, with $K \geq 3$, Ψ has full rank generically if R is defined by the exclusion restrictions that there exist k, $k' < K$ with $\gamma_k = 0$, $\beta_{k'} = 0$ and $\beta_k \neq 0, \gamma_{k'} \neq 0$. So essentially, we get identification if one regressor has no contextual effects and another has no direct effects. In contrast, restricting two regressors to both have no contextual effects but nonzero individual effects would not suffice to make Ψ full rank (this turns out to be a case where the order condition would be satisfied but the rank condition

[8] It is not sufficient to rule out the linear-in-means model to eliminate this problem, since there exist many other models in our framework that are also not identified without additional restrictions.

[9] Graham and Hahn (2005) also use instruments from outside the model to obtain identification. In contrast we only consider restrictions on coefficients to gain identification.

is not).

Since it would be unusual for covariates to have contextual but not direct effects, we consider a second type of rank restriction, which exploits the presence of multiple environments s. These restrictions are that some structural parameters do not vary by environment. To illustrate, suppose we have two different environments, so $S = 2$, and we assume that peer effects vary by environment, but direct and contextual effects do not. Then the restrictions $R\boldsymbol{\theta} = c$ will include the equations $\beta^{(1)} - \beta^{(2)} = 0$ and $\gamma^{(1)} - \gamma^{(2)} = 0$. In this case $\Psi\boldsymbol{\theta} = v$ simplifies to

$$\begin{pmatrix} 0 & 0 & H^{(1)} & 0 \\ \iota & 0 & 0 & H^{(1)} \\ m^{(1)} & 0 & I & I \\ 0 & 0 & H^{(2)} & 0 \\ 0 & \iota & 0 & H^{(2)} \\ 0 & m^{(2)} & I & I \\ & R & & \end{pmatrix} \begin{pmatrix} \lambda^{(1)} \\ \lambda^{(2)} \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} \iota \\ 0 \\ m^{(1)} \\ \iota \\ 0 \\ m^{(2)} \\ c \end{pmatrix}, \tag{15}$$

where we have let $\beta = \beta^{(1)} = \beta^{(2)}$ and similarly for γ. Inspection of equation (15) shows this still does not provide enough restrictions for identification (note that increasing S from 1 to 2 increased the number of required restrictions). However, if we impose one exclusion restriction, such as assuming that one contextual effect (i.e., one element of γ) equals zero, and we impose the constraint that $\lambda^{(1)} \neq \lambda^{(2)}$, then that provides enough restrictions to generically satisfy Theorem 1.

Note that the requirement that $\lambda^{(1)} \neq \lambda^{(2)}$ can be tested in this case, since, by equation (15), $\lambda^{(1)} \neq \lambda^{(2)}$ if and only if $m^{(1)} \neq m^{(2)}$, and we estimate the reduced-form parameters $m^{(s)}$.

The assumption that β and γ do not vary by environment in this example can be relaxed. For example, if the direct effects β are the same across groups but the contextual effects vary, so $\gamma^{(1)} \neq \gamma^{(2)}$, then the full rank condition required for identification will still hold generically if one of the regressors has no contextual effect in either environment, that is, if one element in $\gamma^{(1)}$ and $\gamma^{(2)}$ equals zero.

For our empirical application in Section 7, we analyze students' math test scores. In that application, we assume two environments corresponding to small ($s = 1$) and large ($s = 2$) class sizes. For identification we impose the above restriction that λ varies by class size while β and γ do not. We then need one additional exclusion restriction. For this we assume that a student's number of days of absence from school has an impact on his own test score but not on those of other classmates, so the element of γ corresponding to days of absence is set to zero.

4.2 Individual labels

Define the *label* of an individual in a group l to be the row of Y_l and X_l where that individual's data appears, and hence is also the row of G_l that contains that individual's links. When we refer to individual members $i = 1, ..., n$ of a group l, any given member's value of i is that member's label.

The ordering, or labeling, of individuals in a group l determines the ordering of the rows of that group's adjacency matrix G_l. Therefore, the labeling of individuals in each group affects the distribution of adjacency matrices. As a result, the validity of our assumptions may depend in part on how individuals in each group are labeled. In particular, our assumptions require that, for the chosen labeling of individuals, every group's random array $(X_l, G_l, \varepsilon_l)$ is drawn from the same underlying joint distribution of group arrays, and that distribution satisfies the properties given in Assumptions 2, 3, and 4.

Analogous labeling requirements exist in other papers that identify social network models from reduced-form coefficients, including de Paula et al (2018) and Blume et al (2015). Similar requirements apply to the labeling of players in many empirical game models. For example, to infer private values from bids in auctions, it is assumed that bidders who share the same label across different auctions be independent draws from the same underlying distribution of private values. See, e.g., Section 3.2.2 and 4.1 in Athey and Haile (2005). Another example is the labeling of firms in matching markets. For example, Fox, Yang, and Hsu (2015) recover unobserved complementarities from matching patterns across many markets in a sample. Their method requires either that the labels of firms on both sides have common meaning across markets in the data, or that the distribution of unobserved characteristics is fully exchangeable in firm labels.

There are two methods we can use to deal with this labeling issue. One is to assume that the joint distribution of $(X_l, G_l, \varepsilon_l)$ is exchangeable in individual labels. In this case, how the individuals are labeled would have no impact on the identification strategy or on the asymptotic properties of the estimator we propose in the next section. Under exchangeability, one could simply randomly label individuals from 1 to n in each group. However, exchangeability is a strong symmetry restriction that in many ways resembles (though is still less restrictive than) the linear-in-means model.[10] Note that it would be sufficient for our results to only assume exchangeability within environments, not across environments.

An alternative to assuming exchangeability is to order, and hence label, individuals in

[10] In the linear-in-means model, M is constant and therefore identical to $E(M)$, and has a simple form where the ratio between any diagonal and an any off-diagonal element of M is a known function of n and λ. In contrast, in our model, even under exchangeability, that ratio is jointly determined by the distribution of network links in addition to n and λ. So even with exchangeability, our model is more general than the linear-in-means model.

each group based on some observable characteristics that may affect link formation, but are otherwise exogenous (and so are not included in X_l). In our empirical application, we order students in classrooms based on their date of birth. Within classrooms, students' dates of birth are typically not included as an element of X_l in models of test score outcomes. See, e.g., Krueger and Whitmore (2001).[11] On the other hand, date of birth may have non-trivial impacts on link formation, if a child is less likely to consider a much younger or older classmate as a friend than one with a closer birthday. Finally, note that if exchangeability is satisfied, then labeling based on observed characteristics is harmless.

5 Estimation

To estimate the structural parameters of our model, we use a sample of outcomes and regressors over random networks $(y_l, X_l)_{l=1,2,...,L}$. Assume that across $l = 1, .., L$, $(y_l, G_l, X_l, \varepsilon_l)$ are independent draws from the population model. Our estimator is based on sample analogs of the moments and steps used for identification. The estimator is analogous to indirect least squares, in that we first estimate reduced-form coefficients, and then use them to recover the structural coefficients.

To fix ideas, we first consider the case of a single environment ($S = 1$), so the required rank restrictions $R\boldsymbol{\theta} = c$ are all exclusion restrictions.

Step 1: For each $i \in \{1, ..., n\}$, linearly regress the outcome $(y_{l,i})_{l=1,...,L}$ on a constant and on $(X_l)_{l=1,...,L}$, yielding $Kn + 1$ slope coefficients for each i. Note that each of these regressions uses L observations. These regressions correspond to equation (5). The constant term in each regression should be the same μ_0, so these regressions can be estimated jointly, imposing the constraint that the estimated intercept in each regression be the same $\hat{\mu}_0 \in \mathbb{R}$.[12]

After running these regressions, the resulting coefficients are then arranged into matrices $\hat{\mu}_k \in \mathbb{R}^{n \times n}$ for $k = 1, 2, ..., K$, as described immediately before and after Lemma 1. Also, construct $\hat{m}_k \equiv (\iota' \hat{\mu}_k \iota)/n$ for $k = 1, 2, ..., K$. Note at this stage one could test the condition of non-trivial marginal effects required by part (ii) of Assumption 5, using these estimates and their associated standard errors.

Step 2: For each $k = 1, 2, ..., K - 1$, estimate the solution to equation (8) using the

[11] Some papers, such as Angrist and Krueger (1991), showed that the dates of birth could affect broader, longer term outcomes. However, since we use essentially the same STAR data as Krueger and Whitmore (2001), we follow them to exclude dates of birth as a direct regressor. Still, our partitioning of classes into more vs less disbursed dates of birth does allow for some indirect effects on outcomes via differences in link formation.

[12] Alternatively, we may first demean the data, estimate these regressions separately, each without an intercept, and then recover an estimate of the common intercept $\hat{\mu}_0$.

extremum estimator

$$(\hat{a}_k, \hat{b}_k) \equiv \arg \min_{a_k, b_k \in \mathbb{R}} \sum_{i,j} \left[e_i (a_k \hat{\mu}_k + b_k \hat{\mu}_K - I) e_j' \right]^2 . \tag{16}$$

This step entails a numerical search. A potentially less efficient, but closed-form alternative would be to use a subset of the information in (8) to construct a smaller linear system that could then be solved for (a_k, b_k) by matrix inversion. An example of such a system would be just the equalities that the diagonal entries in $a_k \hat{\mu}_k + b_k \hat{\mu}_K$ sum to n and the off-diagonal entries add up to 0. These closed-form estimates could be used as starting values for the extremum estimation above.[13]

Step 3: Given the estimates from Step 2, calculate the closed-form estimator of $\hat{\theta} \equiv (\hat{\lambda}, \hat{\beta}_1, ..., \hat{\beta}_K, \hat{\gamma}_1, ..., \hat{\gamma}_K)'$ using

$$\hat{\theta} \equiv \left(\hat{\Psi}' \hat{\Psi} \right)^{-1} \hat{\Psi} \hat{v},$$

where $\hat{\Psi}$ is the coefficient matrix formed by stacking (11) and (12) along with the exclusion restrictions $R\boldsymbol{\theta} = c$, as in Theorem 1.

For example, in the case with $K = 3$ above:

$$\hat{\Psi} \equiv \begin{pmatrix} 0 & \hat{a}_1 & 0 & \hat{b}_1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & \hat{a}_1 & 0 & \hat{b}_1 \\ 0 & 0 & \hat{a}_2 & \hat{b}_2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & \hat{a}_2 & \hat{b}_2 \\ \hat{m} & & & I & & & I \\ & & & R & & & \end{pmatrix}; \hat{v} \equiv \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ \hat{m} \\ c \end{pmatrix};$$

with $R\theta = c$ representing equalities describing the exclusion restrictions, such as some of the contextual and direct effects being set to zero. Finally, the remaining structural parameter α is estimated by $\widehat{\alpha} = (1 - \widehat{\lambda})\widehat{\mu}_0$.

Now consider how this procedure can be generalized to handle multiple environments, so $S \geq 2$. To do so, first implement steps 1 and 2 separately for each environment s, to get estimates $\hat{a}_k^{(s)}, \hat{b}_k^{(s)}, \hat{m}_k^{(s)}$, $s \leq S$. Then, for step 3, stack the estimated matrices $\hat{\Pi}$ with R, and the estimated vector $\hat{d}$ with c as in the preceding subsection, to obtain $\hat{\Psi}$ and $\hat{v}$. Then $\boldsymbol{\theta}$ is estimated by a classical minimum distance method:

$$\hat{\boldsymbol{\theta}} \equiv \arg \min_{\theta \in \Theta} (\hat{\Psi}\boldsymbol{\theta} - \hat{v})' \boldsymbol{\Xi}^{-1} (\hat{\Psi}\boldsymbol{\theta} - \hat{v}),$$

[13] Many alternative smaller linear systems could be constructed for inefficient closed-form estimation, each using a different subset of the equalities in equation (8). One could then estimate $(\hat{a}_k, \hat{b}_k)$, or obtain consistent starting values for extremum estimation of these coefficients, by taking a (possibly weighted) average of these many closed-form estimates.

where Θ denotes the feasible parameter space and Ξ^{-1} is a chosen weight matrix that is symmetric and positive definite. The first-order condition for this minimization yields the estimator

$$\hat{\boldsymbol{\theta}} = \left(\hat{\Psi}'\boldsymbol{\Xi}^{-1}\hat{\Psi}\right)^{-1}\left(\hat{\Psi}'\boldsymbol{\Xi}^{-1}\hat{v}\right)$$

The following theorem shows consistency of this estimator under standard conditions.

Assumption 6. *For each $s \leq S$, the parameter space for $(a_k^{(s)}, b_k^{(s)})$ defined in (9) is compact for all $k \leq K$.*

Theorem 2 *Suppose Assumptions 1-6 hold for each $s \leq S$, and Ψ has full rank. Then $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}$ as $L \to \infty$.*

In Appendix A, we provide the proof of Theorem 2. To see intuition for the consistency of $\hat{\boldsymbol{\theta}}$, recall that for each environment s and each $k \leq K$, $\hat{\mu}_k^{(s)}$ consists of ordinary least squares coefficient estimates from linear regressions with L observations, and $\hat{m}_k^{(s)}$ is a simple linear function of all the elements in $\hat{\mu}_k^{(s)}$. Hence these estimators are consistent for the actual $\mu_k^{(s)}$ and $m_k^{(s)}$ in the data-generating process. In addition, $(\hat{a}_k^{(s)}, \hat{b}_k^{(s)})$ are two-step extremum estimators, whose objective function in (16) depends on $\hat{\mu}_k^{(s)}$ smoothly. As $L \to \infty$, this objective function converges in probability, uniformly over the parameter space, to its limit where $\hat{\mu}_k^{(s)}$ is replaced by $\mu_k^{(s)}$. Lemma 2 implies this limit is uniquely minimized at the actual $(a_k^{(s)}, b_k^{(s)})$. By a standard argument for the consistency of extremum estimators, $(\hat{a}_k^{(s)}, \hat{b}_k^{(s)})$ converges in probability to $(a_k^{(s)}, b_k^{(s)})$ for each s and k. Note that Ψ and v consist of known constants and $a_k^{(s)}$, $b_k^{(s)}$, and $m_k^{(s)}$ for $k \leq K$ and $s \leq S$. It then follows from the Slutsky Theorem that $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$.

In Appendix A, we also explain why $\hat{\boldsymbol{\theta}}$ is $\sqrt{L}$-convergent and asymptotically normal. Essentially, this result comes from the parametric convergence of ordinary least squares regression coefficients, and application of the delta method.

6 Extensions

6.1 Group-level variables and group fixed effects

The identification and estimation methods in Sections 4 and 5 can be readily extended to accommodate group-level regressors. Suppose each group l has a row vector of group-level characteristics $z_l \in \mathbb{R}^P$. For example these could be attributes of the teacher when each group is an elementary school class.

For the moment, consider just a single environment, so $S = 1$ and the s superscript is omitted. Including group level effects the structural model becomes

$$y_l = \alpha\iota + \lambda G_l y_l + \iota z_l \rho + X_l \beta + G_l X_l \gamma + \varepsilon_l,$$

with $\rho \in \mathbb{R}^P$ being a column vector of additional coefficients. One could interpret ρ as a source of "correlated effects". Let Assumption 1, 2 and 3 hold with X_l replaced by (X_l, z_l), and let part (ii) of Assumption 4 hold with $\tilde{X}_l \equiv (1, z_l, X'_{l,c1}, X'_{l,c2}, ..., X'_{l,cK})$. The reduced form is now

$$E(y_l \mid X_l, z_l) = \mu_0 + E(M_l)\iota z_l \rho + E(M_l) X_l \beta + E(M_l G_l) X_l \gamma. \tag{17}$$

The first step linear regressions identify μ_0 and $(\mu_k)_{k \leq K}$ as before. But in addition, these regressions now include z_l. Denote the reduced form coefficients for z_l as $\nu \in \mathbb{R}^P$. The p-th component of ν, denoted ν_p, satisfies $\nu_p \equiv \rho_p/(1-\lambda)$. This equality follows from equation (17) and the row normalization in Assumption 1, which as noted earlier implies that each row in M_l adds up to the same constant $1/(1-\lambda)$. We used this same relationship earlier to obtain $\mu_0 = \alpha/(1-\lambda)$. Applying Theorem 1, we identify $\lambda, \beta, \gamma, \alpha$ from $\mu_0, (\mu_k)_{k \leq K}$, and $R\boldsymbol{\theta} = c$. Finally, the parameters ρ can then be identified by $\rho = \upsilon(1-\lambda)$. Correspondingly, for estimation let $\hat{\rho} = \hat{\nu}(1-\hat{\lambda})$, where $\hat{\nu}$ are the ordinary least squares estimates for the slope coefficients of z_l in the reduced-form regression in equation (17).

Now if we have multiple environments, then run the above reduced-form regressions separately for each environment s as before, but now including z_l as additional regressors. We may then identify and estimate $\boldsymbol{\theta}$ from $\mu_0^{(s)}$, $\left(\mu_k^{(s)}\right)_{k \leq K}$ for $s \leq S$ and $R\boldsymbol{\theta} = c$ as before, and estimate each $\hat{\rho}^{(s)}$ using $\hat{\rho}^{(s)} = \hat{\nu}^{(s)}(1 - \hat{\lambda}^{(s)})$.

Finally, this procedure can be further extended to accommodate unobserved group-level fixed effects (denoted ϖ_l). Essentially, we can remove these fixed effects by applying group-level demeaning of the outcomes to the reduced form, prior to recovering the structural parameters. Specifically, the method consists of replacing the dependent variables y in the first stage reduced-form regressions with demeaned outcomes $y - \bar{y}$, and following essentially the same steps as before to estimate the structural $\boldsymbol{\theta}$ parameters. Then, we can recover the remaining parameters ρ and α by plugging the estimated $\boldsymbol{\theta}$ parameters into the non-demeaned reduced form in (17), and applying an exogeneity and location normalization assumption that $E(\varpi_l \mid z_l, X_l, G_l) = 0$. Details of this procedure are provided in Appendix F.

6.2 Dimension reduction

Again, begin by considering the case of only one environment, so s superscripts can be dropped. In the first-step regressions of $y_{l,i}$ on X_l for each $i \leq n$, we need the number of

groups L in the sample to be large relative to the dimension of regressors Kn (where n is the group size and K is the number of individual characteristics in X). Other network estimation papers have similar data requirements on number of groups, including Blume et al (2015) and de Paula et al (2018). However, in some applications, L might not be large relative to Kn.

One possible way to deal with this issue could be to apply sparsity related methods like Tibshirani's (1996) LASSO to these first step reduced-form regressions, with the caveat that setting small elements of μ_k equal to zero could have effects of unknown magnitude on the resulting structural model parameters. See, e.g., de Paula et al (2018) for a similar use of sparsity assumptions in reduced form network equations.

Alternatively, by making an additional uncorrelatedness assumption regarding characteristics, our method can be implemented using sequential steps that involve just K dimensional regressions. Suppose for each individual i that the vector of characteristics $x_{l,i} \in \mathbb{R}^K$ is uncorrelated with those of other group members $(x_{l,j})_{j \neq i}$. This may occur if, e.g., members are randomly assigned to groups. We may then transform all observed variables into mean deviation form: $\Delta y_{l,i} \equiv y_{l,i} - \bar{y}_i$ and $\Delta x_{l,i} \equiv x_{l,i} - \bar{x}_i$ for $i = 1, ..., n$ where $\bar{y}_i \equiv \frac{1}{L} \sum_{l' \leq L} y_{l',i}$, $\bar{x}_i \equiv \frac{1}{L} \sum_{l' \leq L} x_{l',i}$. Now, for each i and j from 1 to n, separately regress $\Delta y_{l,i}$ on $\Delta x_{l,j}$. This gives a total of n^2 regressions, each having K regressors and L observations. The resulting coefficients from these regressions can then be assembled into the reduced form coefficient matrices μ_k for $k \leq K$. Then, given these μ_k matrices, one can proceed as before to estimate the model.

With multiple environments ($S > 1$), the above regressions would be run separately in each environment, before proceeding to the later steps of identification and estimation as before. Either of the above dimension reduction methods may be especially useful in applications with multiple environments, where the number of groups in some environments s could be small relative to $Kn^{(s)}$.

6.3 Variation in group sizes

Our identification and estimation method assumes that all groups within each environment s have the same group size $n^{(s)}$. But with K individual characteristics in X, this requires observing enough groups of size $n^{(s)}$ (meaning that $L^{(s)}$, the number of groups in environment s, is large enough) to estimate first-stage reduced-form regressions consisting of $Kn^{(s)}$ coefficients in each environment s. However, in some samples we may not observe enough groups of each size to implement these regressions. We propose two ways to resolve such data deficiencies. One requires some additional uncorrelatedness assumptions, while the other imposes some restrictions on structural coefficients across groups of different sizes.

The first approach exploits the dimension-reduction methods in Section 6.2. To fix ideas, first suppose individual characteristics $x_{l,i} \in \mathbb{R}^K$ are uncorrelated across group members (as would happen if, e.g., individuals were randomly assigned to groups with different sizes). Then, as explained in Section 6.2, one can estimate the reduced-form coefficients for each i via a sequence of lower-dimension regressions, each involving only K instead of $Kn^{(s)}$ regressors. In this case, one can account for variation in group sizes in each of these lower-dimension regressions by including dummy variables for group sizes and interacting them with the slope coefficients. This method can be generalized to allow for correlated individual characteristics, by instead applying the partitioned regressions to estimate reduced-form coefficients, and again including group size dummies (and their interactions with slope coefficients) in these regressions.

The second approach we propose can be used even if the sample has very few observations of some group sizes. This second approach pools groups of different sizes into a single environment, and so requires that the structural parameters $\lambda, \beta, \gamma, \alpha$ be the same among all the different sized groups being pooled. This second approach takes smaller groups, and augments them with additional simulated "pseudo-individuals" to artificially increase their size, to match the size of other, larger groups. Under certain conditions, the resulting pooled regressions then consistently estimate a weighted average of reduced-form coefficient matrices for groups of different sizes, yielding consistent estimates of the structural parameters. Details are in Appendix C.

In our empirical application, we apply the second method. We define $S = 2$ environments: "small class size" and "large class size." Small classes pool classes (groups) having 15 to 20 students, while large classes pool classes of 21 to 25 students.

7 Peer Effects in Tennessee Elementary Schools

We apply our method to analyze the social effects among elementary school students who participated in the Student/Teacher Achievement Ratio (STAR) Project in the U.S. State of Tennessee. The STAR project was a four-year longitudinal study funded by the Tennessee General Assembly and conducted by the Tennessee State Department of Education. The goal of the project was to assess the impact of class sizes on students' academic performance through randomized experiments.[14] The STAR sample data does not report any measure of links among students, and so is a candidate for applying our method of estimation.

The typical method of evaluating potential peer effects in a model without link data is to assume a linear-in-means specification. In classroom applications, this is equivalent

[14] A general survey of influences on learning and associated outcomes is Heckman and Mosso (2014).

to assuming every class has an adjacency matrix where each student in the class is linked to all others in the class, with equal weights. Examples of papers that use this method include estimates of contextual effects of student-teacher races in Dee (2004), gender ratios in Whitmore (2005), and a composite of peer characteristics in Graham (2008) and Sojourner (2013). Boozer and Cacciola (2001) apply a linear-in-means specification to the STAR data, using experimental variation in class quality (fraction of students exposed in the previous year to small classes) as an instrument to identify peer effects.

Instead of assuming each student in a class is linked to all the others with equal weights, our estimator makes no assumption about what the within-class unobserved links actually are, and allows these links to vary across classes. We also do not require an instrument, although we do require exclusion assumptions. We nevertheless identify both peer and contextual effects. We also use our results to test some hypotheses about these effects, and about the link formation process, and we use our structural model estimates to perform some counterfactual calculations.

7.1 Data description

We observe a cohort of students who were in kindergarten in 1985-1986. Seventy-nine public schools were selected to participate in the project, representing various geographic locations (inner city, urban, suburban or rural). Students and teachers were randomly assigned to classes with varying sizes of 13 to 25 students.[15] Note that our estimator neither requires nor directly exploits this random assignment; however, random assignment does make some of our assumptions more plausible. An example is the dimension reduction discussed in Section 6.2.

Our sample consists of 258 classes that had at least 15 but no more than 25 students each. The total number of students in the sample is 5,189. We partition the classes in the sample into $S = 2$ environments: smaller classes with 15-20 students, and larger classes with 21-25 students according to the original design of the project. In each class, we order the students by their dates of birth, and use this ordering to label individual students. Table 7.1 reports summary statistics of the students' mathematics test scores in the second and third grade ($t2$ and $t3$) and other individual-level or class-level variables to be used in our empirical analysis. These include a student's number days of absence from school (abs), students' self-reported motivation scores (mot), and a discretized measure of teachers' years of experience (tec). We standardize the math scores in the second grade $t2$ using the overall mean and standard deviation of raw scores of all classes in the sample.

[15] Students who joined the cohort at STAR schools after 1985-1986 were also included in the experiment throughout later years

Table 7.1. Summary Statistics

	Small class size (122 classes)				Large class size (136 classes)			
	mean	median	std dev	range	mean	median	std dev	range
t3	620.7	618.0	40.88	[487.0, 774.0]	616.6	616.0	40.15	[510.0, 774.0]
t2	0.077	0.287	0.936	[-5.902, 1.042]	-0.029	0.287	1.023	[-6.355, 1.042]
abs	6.743	5.000	6.643	[0, 59]	6.902	5.000	6.429	[0, 55]
mot	49.29	50.00	3.990	[17, 59]	49.14	50.00	4.013	[18, 60]
tec	13.30	13.00	8.416	[0, 36]	14.19	14.00	9.079	[0, 38]

Notes: *t3*: raw scores for 3rd grade math; *t2*: standardized scores for 2rd grade math (using overall mean and std dev across *all* classes); *abs*: days of absence; *mot*: self-reported motivation score; *tec*: teacher experience (in # yrs).

Table 7.1 reports that the average math score in the third grade is 620.7 for small classes, and 616.6 for large classes. In addition, Table 7.2 shows that a t-test for the null hypothesis of equal mean scores in small and large classes (allowing for unequal variances) rejects the null at the 1% level. The sign of this difference is consistent with findings in Krueger (1999), which reports in a bigger sample that on average Grade K-3 test scores in smaller classes are about 5 percentage points (or 0.2 standard deviations) higher than in larger classes. Other papers that report similar patterns include Hanushek (1999) and Krueger and Whitmore (2001).

Table 7.2. Test of Equal Means

(small vs. large classes)

	p-value		p-value
t3	0.001	*abs*	0.401
t2	<0.001	*mot*	0.161
		tec	0.420

Table 7.2 also reports the p-values for testing the equality of means of demographic variables in small versus large classes. Unlike the test scores, we fail to reject the null of equal means for each of the demographic variables. This provides some support for the assumption that the assignment of students and teachers to classes is independent of these demographic variables. On the other hand, Table 7.2 suggests that the small classes have a higher average for Grade 2 scores than large classes, and the difference is highly statistically significant at the 1% level. One explanation, which is reconcilable with earlier findings in the literature, is that the students enrolled in smaller classes had already developed better math skills than their peers in larger classes before the beginning of the third grade.

7.2 Econometric specification

Our model, corresponding to equation (1), is

$$\begin{aligned} t3_{l,i} &= \alpha^{(s)} + \lambda^{(s)} \textstyle\sum_j G^{(s)}_{lij} t3_{l,j} + \beta_1 abs_{l,i} + \beta_2 mot_{l,i} + \beta_3 t2_{l,i} + \delta^{(s)} tec_l \\ &\quad + \gamma_1 \textstyle\sum_j G^{(s)}_{lij} mot_{l,j} + \gamma_2 \textstyle\sum_j G^{(s)}_{lij} t2_{l,j} + \varepsilon_{l,i}, \end{aligned}$$

where i and j are indices (labels) for individual students, l is an index for class, and (s) is the environment index. Each summation $\sum_j$ is over all students in the same class l as student i. For each pair i and j, $G^{(s)}_{lij}$ is the row-normalized unobserved zero or nonzero link between the members labeled i and j in class l, in environment s. The coefficients to be estimated are peer effects $\lambda^{(s)}$, direct effects $(\beta_1, \beta_2, \beta_3)$, contextual effects (γ_1, γ_2), intercepts $\alpha^{(s)}$, and correlated effects $\delta^{(s)}$ (this last is the marginal impact of teacher experience, a group-level covariate).

The rank restrictions we have imposed for identification are as follows. First, this specification assumes *abs* has a direct effect ($\beta_1 \neq 0$) but no contextual effects. That is, a student's absence from school affects his own test scores, but has no impact on his classmates other than through peer effects. This is an exclusion restriction. Other covariates *mot* (self-reported motivation score) and $t2$ (Grade 2 math score) are not restricted, and so can have both direct and contextual effects. Our second rank restriction is that we assume the individual effects β and contextual effects γ are the same in the two environments, small and large class sizes (which is why β and γ do not have s superscripts above). All other structural parameters, i.e., the intercept $\alpha^{(s)}$, the peer effect $\lambda^{(s)}$, and the correlated effect $\delta^{(s)}$, are permitted to differ between small ($s = 1$) vs large ($s = 2$) classes. These constraints result in more rank restrictions than are required to satisfy Theorem 1. Our model is therefore over-identified, which we will exploit by providing some model specification tests.

Our methodology does not require explicit modeling or parametrization of the network formation process. However, as discussed earlier and in Appendix D, we do require conditional independence between the random adjacency matrices and the ε errors in our model. To control for possible dependence of the network on student demographics (other than characteristics included in the regressors X), we partition the classes (and hence partition each of the environments) in our sample into those with higher versus lower dispersion in birthdays.[16] In the notation of Appendix D, X^a_l is a dummy indicating low versus high birthday dispersion, and X^e_l is the set of other covariates in the model. We then report estimates for social effects that are sample-size-weighted averages of estimates obtained across these partitions.

[16] For each class we calculate the standard deviation of students' birthdays. We label a class as having "high birthday dispersion" if the standard deviation exceeds six months.

7.3 Estimation results

Table 7.3 reports our structural coefficient estimates. Standard errors are calculated using $B = 1000$ bootstrap samples, each of which is constructed by drawing classes from the original sample with replacement.

Table 7.3: Estimates of Social Effects

		Small Class		Large Class	
Effects	Coef.	est.	(s.e.)	est.	(s.e.)
Peer	λ	0.8478***	(0.0189)	0.9208***	(0.0215)
Group	δ	0.0709	(0.2885)	0.2032	(0.2609)
Constant	α	94.543***	(26.221)	48.126***	(14.450)
		est.		(s.e.)	
Direct	β_1	-0.3639**		(0.1611)	
	β_2	0.0384		(0.0653)	
	β_3	23.356***		(5.3011)	
Context	γ_1	-0.0118		(0.0742)	
	γ_2	13.129**		(5.8605)	

Notes: Standard errors are computed using $B = 1000$ bootstrap samples. ***: significant at 1%; **: significant at 5%.

Estimates of peer effects are statistically significantly positive in both small and large classes, with the estimated coefficient λ being 0.85 and 0.92 respectively. A t-test for the equality of peer effects in small and large classes rejects the null of equality at the 1% level. The magnitudes of our λ estimates are comparable to earlier findings that used the same data but very different methodologies. For example, using a linear-in-means specification (with average class size of students in the previous year as an instrument) Boozer and Cacciola (2001) estimate the peer effects to be 0.86 for the second grade and 0.92 for the third grade. Defining links to be a simple function of measured social distance and employing some variance restrictions, Rose (2017) estimates the peer effects of 0.90. Graham (2008) reports a peer effect of 0.86 for normalized math scores in a linear-in-means social interaction model. The estimated magnitudes of peer effects are quite similar across these different papers and modeling strategies, though the implications and hence implied counterfactuals differ somewhat by specification. Moreover, we later test and reject the linear-in-means specification, and we obtain estimates of both direct and contextual effects in addition to peer effects.

Unlike these previous papers, our peer effect estimates differ in small vs large classes. The bigger value of λ in larger classes could be due to students having more options to form links (like friendships or study partners) in larger classes. This could on average lead to better matches, and hence be conducive to more productive relationships.

Our estimates also show that the number of days absent from school has a small but statistically significant direct effect on a student's test performance. We find that self-reported motivation scores have no significant direct or contextual effects. In contrast, students' performance in the second grade ($t2$) have both direct and contextual effects that are positive and statistically significant. A unit (one standard deviation) increase in a student's score in the second grade improves his own raw score in the third grade by 23.36 points, and increases his friend's third grade scores by 13.13 points.

We infer that the higher average Grade 3 score in small classes should be mostly attributed to better Grade 2 preparation in small classes, as demonstrated in Tables 7.1 and 7.2. While Table 7.3 shows that positive peer effects are bigger in large classes, this effect is not sufficient to counteract the trajectory of higher Grade 2 preparation in small classes. Note that the structural intercept α is also higher in smaller classes. This also contributes to the higher average Grade 3 performance in small classes.

7.4 Specification tests

In this section we report results from a general specification test of our model, and tests of some specific adjacency matrix specifications.

We first exploit the fact that we imposed sufficient rank restrictions to over-identify our model. Recall in Theorem 1, $\boldsymbol{\theta}$ is identified from $\Psi\boldsymbol{\theta} = v$. In our empirical application, $\boldsymbol{\theta}$ has seven elements while Ψ has fifteen linearly independent rows, yielding eight degrees of over-identification. Our estimator minimizes a measure of distance between $\Psi\boldsymbol{\theta}$ and v, so under the null of correct specification, the minimized objective function is asymptotically zero. To test this, exploiting the over-identification, we use $B = 1000$ bootstrap samples to estimate the sampling distribution of the minimized objective function and calculate p-values under the null.[17] Recall that, to control for possible network dependence, we partition the data by low versus high dispersion in the date of birth, and minimize separately for each. We therefore obtain separate test statistics for each partition, and find a p-value of 0.569 for less

[17]Note that our estimator does not lend itself to the use of classical J-tests of over-identified models in Generalized Method of Moments. This is because the coefficient matrix in the last step of estimation is constructed from the estimates of reduced-form coefficients in earlier steps, analogous to indirect least squares. Once these reduced-form coefficient estimates are calculated, the linear system used in the last step is deterministic.

dispersed classes and 0.358 for more dispersed classes. We therefore fail to reject the null of correct model specification.

Next we turn to tests of network structure. We cannot identify and estimate individual adjacency matrices $G_l^{(s)}$. However, we do identify the expected value of some functions of these matrices. Since our model only imposes regularity assumptions on the distribution of adjacency matrices, we can use these identified functions to test some models of network specification against arbitrary regular alternatives. In particular, we consider two different null hypotheses: the linear-in-means specification, and a Poisson random network, i.e., the Erdős-Rényi (1959) network, where links are drawn independently from a heterogenous Bernoulli distribution. Both of these network specifications imply restrictions on the reduced-form coefficient matrices μ_k that we use to construct tests.

In the linear-in-means specification, for every group l in each environment s, the adjacency matrix $G_l^{(s)}$ is constant (the same for all l) with all off-diagonal elements taking the exact same value. With the s superscript dropped for simplicity, this implies that, for each individual characteristic k,

$$\mu_k \equiv (I - \lambda G)^{-1}(\beta_k I + \gamma_k G) = \left(I + \tfrac{\lambda}{1-\lambda} G\right)(\beta_k I + \gamma_k G).$$

This in turn means that all the off-diagonal components in μ_k must be identical. We calculate Wald test statistics using a 6×6 leading principal minor of the reduced-form coefficient for $t2$ (standardized Grade 2 score) in each of the partitions (defined by birthday dispersion) for each of the environments (defined by class size). We choose $t2$ as the characteristic k to base the test on, because its coefficients were the most precisely estimated. The resulting test statistics are reported in Table 7.4:

Table 7.4: Wald Tests for Linear-in-Means (d.f.=29)

	small class (p-val)	large class (p-val)
low disp.	79.915 (<.001)	63.874 (<.001)
high disp.	45.112 (.028)	61.061 (<.001)

The number of restrictions, which equals the degrees of freedom, of each test is $d.f. = 6 \times 6 - 6 - 1 = 29$. Based on the p-values reported in Table 7.4, we reject the hypothesis that the data were generated by the linear-in-means model. Note there could exist other models that also imply identical off-diagonal components in μ_k, in which case those models would also be rejected.

To provide a sense of the magnitude of the difference between our estimates and a linear-in-means model, we can compare our estimate of $E(M)$ to the constant linear-in-means M. These matrices are large, but to summarize, consider just small groups with low birthday

dispersion. For this environment and partition, the average of the n estimates for diagonal entries in $E(M)$ is 1.0611, with a standard deviation of 0.0562, whereas the average of the $n(n-1)$ off-diagonal entries is 0.0801 with a standard deviation of 0.1147. These values differ substantially from the linear-in-means M matrix, which has all diagonal entries equal to 1.2785 and all off-diagonal entries equal to 0.2785 (based on our estimated peer effect of 0.8478).

Next, we construct classical minimum distance (CMD) tests for the null hypothesis of Poisson random network formation, again controlling for class size and birthday dispersion. Specifically, this null hypothesis posits a random link formation process where, before row normalization, each element of each group's adjacency matrix equals one with some success probability, and equals zero with one minus that probability, independent of all the other elements of the adjacency matrix (and of the model error). We allow the success probability to take one of three values, depending on the difference between the birthdays of the two students' being potentially linked. Let p be the vector of these three success probabilities.

We construct the CMD objective function for estimating p by simulation. That is, for any given value of p, we simulate a large number of networks by drawing independently from a Bernoulli distribution with success probabilities given by the vector p. Let $G_r(p)$ denote the simulated adjacency matrix in the r-th draw. Define the objective function $\hat{Q}(p)$ as the distance between the estimated reduced-form coefficients $\hat{\mu}_k$ and the average (over a large number of simulated draws r) of the simulated model-implied marginal effects $(I - \hat{\lambda}G_r(p))^{-1}(\hat{\beta}_k I + \hat{\gamma}_k G_r(p))$. We define the distance between these two matrices as a weighted sum of the differences in average diagonal and off-diagonal components, respectively. We estimate p by minimizing $\hat{Q}(p)$. This objective function would asymptotically converge to zero if the Poisson network specification is correct, so our test statistic is just the minimized value of $\hat{Q}(p)$, with a standard error obtained by bootstrapping. The degrees of freedom of the limit distribution under the null is 3.[18] As before, we implement this procedure and test separately for the two partitions defined by birthday dispersion and the two environments defined by class size. Results are reported in Table 7.5. We strongly reject the null of Poisson random network formation.

Table 7.5: CMD Tests for Poisson Random Network (d.f.=3)

	small class (p-val)	large class (p-val)
low disp.	49.880 (<.001)	171.327 (<.001)
high disp.	36.954 (<.001)	101.636 (<.001)

[18] This is because the restrictions (number of links between reduced-form coefficients and model implied marginal effects) used in the CMD objective function is $2K = 6$, and the number of structural parameters is $\dim(p) = 3$.

In conclusion of this section, we do not reject our general model, and we do reject both the simple linear-in-means model and the simple model of independently drawn random links.

7.5 Counterfactuals

Our first counterfactual exercise is to ask how test scores would change if the unobserved networks that generated our data were replaced with linear-in-means networks, holding our estimated parameters fixed. This can be interpreted as measuring the potential benefits or costs of encouraging more links (i.e., more friendships or other connections) among students.

For each class, we calculate the within-class average difference in test scores between those observed in the data, and those obtained if every class's adjacency matrix G_l (which we do not observe) were replaced with linear-in-means adjacency matrices, holding all our parameter estimates fixed. The counterfactual in this setting is equivalent to redistributing some link weight onto classmates who previously were not friends. This could impact a student's score in either direction, depending on whether the counterfactual "new friends" would have a positive or negative impact on a student's test performance, relative to their actual friends.

Table 7.6 reports the resulting difference in average test scores across the classes in each sub-population defined by class size and birthday dispersion. Table 7.6 indicates that the effects of encouraging more links among students would be modest. For comparison, test scores have a standard deviation of 40 in the raw data (see Table 7.1). These results should be interpreted cautiously given their lack of statistical significance, but they suggest that having more friends would slightly increase test scores in small classes, and decrease them in large classes.

Table 7.6: Differences in Test Scores under the Linear-in-Means Network

	Est. mean Δ	p-val
small, low disp	6.054	0.105
large, low disp	-9.596	0.060
small, high disp	5.810	0.184
large, high disp	-6.405	0.239

Notes: Est. mean Δ: average difference in class means of grade three math scores in a network with equal weights on all friends.

In the next counterfactual exercise, we maintain the linear-in-means counterfactual, and consider alternative magnitudes of peer effects. Specifically, we swap the estimated peer

effects between small and large classes (i.e., increase λ to 0.9208 in small classes and decrease λ to 0.8478 in large classes). The goal of this exercise is to assess how peer effect magnitudes interact with the contextual and other differences between small and large classes.

Table 7.7 reports the average changes in class means within each subpopulation again. The table shows that increasing peer effects in small classes would lead to significantly better test scores, and reducing peer effects in large classes would yield worse performance. These effects are larger and highly statistically significant for the low dispersion partition.

Table 7.7: Impact of Counterfactual Peer Effects

	Est. mean Δ	p-val
small, low disp	16.198	0.003
large, low disp	-11.637	0.001
small, high disp	2.954	0.620
large, high disp	-5.301	0.187

Notes: Est. mean Δ: average difference in class means of grade three math scores when peer effects in small and large classes are swapped in a network with equal weights on all friends.

8 Conclusions

We provide an original method for identifying and estimating social interaction effects on many small networks, when the networks are not observed. We propose a two-step estimator, and apply our method to estimate direct, contextual and peer effects among elementary school students. Among other results, we find that the peer effects are larger in bigger classes, that encouraging more links/friendships among students might not significantly improve outcomes (and could make them worse), and we can reject the usual linear-in-means specification of network links.

One limitation of our model is that it requires network formation to be exogenous, after conditioning on covariates. Relaxing this constraint, perhaps with some model of the joint determination of network links and outcomes, would be a useful area for future research.

Appendix

A. Proofs

Proof of Lemma 1. The outcome of each individual i in group l is

$$y_{l,i} = \tilde{X}_l' \kappa_{l,i} + \tilde{\varepsilon}_{l,i},$$

where $\tilde{\varepsilon}_{l,i} \equiv M_{l,ri}\varepsilon_l$ with $M_{l,ri}$ being the i-th row in M_l, and $\kappa_{l,i}$ is a $(Kn+1)$-by-1 random vector:

$$\kappa_{l,i} \equiv \left[\mu_0, \left(\beta_1 M_{l,ri} + \gamma_{01} M_{l,ri} G_l\right), ..., \left(\beta_K M_{l,ri} + \gamma_{0K} M_{l,ri} G_l\right)\right]'$$

with β_{0k}, γ_{0k} being the k-th components in β, γ. Regressing $(y_{l,i})_{l \le L}$ on $(\tilde{X}_l)_{l \le L}$ gives:

$$\begin{aligned} & \left(\sum_l \tilde{X}_l \tilde{X}_l'\right)^{-1} \left(\sum_l \tilde{X}_l y_{l,i}\right) \\ = & \underbrace{\left(\frac{1}{L}\sum_l \tilde{X}_l \tilde{X}_l'\right)^{-1}}_{A_L} \underbrace{\left(\frac{1}{L}\sum_l \tilde{X}_l \tilde{X}_l' \kappa_{l,i}\right)}_{B_L} + \left(\frac{1}{L}\sum_l \tilde{X}_l \tilde{X}_l'\right)^{-1} \underbrace{\left(\frac{1}{L}\sum_l \tilde{X}_l \tilde{\varepsilon}_{l,i}\right)}_{C_L}. \end{aligned}$$

As $L \to \infty$, $A_L \xrightarrow{p} E\left(\tilde{X}_l \tilde{X}_l'\right)^{-1}$ and $C_L \xrightarrow{p} E\left(\tilde{X}_l \tilde{\varepsilon}_{l,i}\right) = 0$ because of the weak law of large numbers and Assumptions 1, 2 and 4-(i). Furthermore,

$$B_L \xrightarrow{p} E\left(\tilde{X}_l \tilde{X}_l' \kappa_{l,i}\right) = E\left(\tilde{X}_l \tilde{X}_l'\right) E\left(\kappa_{l,i}\right),$$

where the equality follows from Assumption 3. This implies

$$\left(\sum_l \tilde{X}_l \tilde{X}_l'\right)^{-1} \left(\sum_l \tilde{X}_l y_{l,i}\right) \xrightarrow{p} E(\kappa_{l,i}).$$

Thus $E(\kappa_{l,i})$ is identified for $i = 1, ..., n$ under maintained assumptions. By rearranging the components in $E(\kappa_{l,i})$, we identify $\mu_0 \equiv \alpha/(1-\lambda)$ and $\mu_k \equiv E[M_l(\beta_k I + \gamma_k G_l)]$ for each $k = 1, ..., K$. $\square$

Proof of Theorem 2. The estimators for reduced-form coefficients in Step 1 are ordinary least squares estimators for slope coefficients in a regression. Thus under Assumptions 1-3 and 4-(i), $\hat{\mu}_k \xrightarrow{p} \mu_k, \hat{m}_k \xrightarrow{p} m_k$ for all $k \le K$. Next, For each $k = 1, ..., K-1$,

$$\begin{aligned} & \left|\sum_{i,j}\left[e_i(a_k\hat{\mu}_k + b_k\hat{\mu}_K - I)e_j'\right]^2 - \sum_{i,j}\left[e_i(a_k\mu_k + b_k\mu_K - I)e_j'\right]^2\right| \\ = & \left|\sum_{i,j}\left\{e_i\left[a_k(\hat{\mu}_k + \mu_k) + b_k(\hat{\mu}_K + \mu_K) - 2I\right]e_j'\right\}\left\{e_i\left[a_k(\hat{\mu}_k - \mu_k) + b_k(\hat{\mu}_K - \mu_K)\right]e_j'\right\}\right| \\ \le & \max_{i,j}\left|e_i\left[a_k(\hat{\mu}_k + \mu_k) + b_k(\hat{\mu}_K + \mu_K) - 2I\right]e_j'\right| \times \left\{\sum_{i,j}\left|e_i\left[a_k(\hat{\mu}_k - \mu_k) + b_k(\hat{\mu}_K - \mu_K)\right]e_j'\right|\right\} \\ \le & \left[\left(|a_k| + |b_k|\right)\max_{i,j,k'}\left|e_i(\hat{\mu}_{k'} + \mu_{k'})e_j'\right| + 2\right] \times n^2 \times \left(|a_k| + |b_k|\right)\max_{i,j,k'}\left|e_i(\hat{\mu}_{k'} - \mu_{k'})e_j'\right|. \end{aligned}$$

where the inequalities are due to the triangular and Cauchy-Schwarz inequalities. By the consistency of $\hat{\mu}_k$ and the Continuous Mapping Theorem, the first term on the right-hand side of the last inequality is bounded in probability, and the last term is $o_p(1)$. Therefore, due to compact parameter space in Assumption 6, the objective function of the extremum estimator in Step 2 converges in probably to its population counterpart uniformly over (a_k, b_k). That is, for all $k \leq K$,

$$\sup_{a_k, b_k} \left| \sum_{i,j} \left[e_i (a_k \hat{\mu}_k + b_k \hat{\mu}_K - I) e_j' \right]^2 - \sum_{i,j} \left[e_i (a_k \mu_k + b_k \mu_K - I) e_j' \right]^2 \right| \xrightarrow{p} 0.$$

By Lemma 2, the limit function $\sum_{i,j} \left[e_i (a_k \mu_k + b_k \mu_K - I) e_j' \right]^2$ is uniquely minimized at the solution of (a_k, b_k) in (9). By Theorem 2.1 in Newey and McFadden (1994), $(\hat{a}_k, \hat{b}_k) \xrightarrow{p} (a_k, b_k)$ for all $k \leq K$. Because Ψ has full rank and the weight matrix $\boldsymbol{\Xi}^{-1}$ is positive definite, $\Psi' \boldsymbol{\Xi}^{-1} \Psi$ is invertible. The consistency of $\hat{\boldsymbol{\theta}}$ then follows from the Slutsky Theorem. $\square$

The estimator $\hat{\boldsymbol{\theta}}$ is $\sqrt{L}$-convergent and asymptotically normal under standard regularity conditions. To see this, note that for each $k \leq K$, $\hat{\mu}_k$ consists of slope coefficient estimates from a regression. Besides, $\left(\hat{a}_k, \hat{b}_k \right)$ are two-step extremum estimators whose objective function depends on $\hat{\mu}_k$, and $\hat{m}_k$ is a linear function of $\hat{\mu}_k$ (i.e. the sum of all components in $\hat{\mu}_k$ divided by the group size n). By a standard argument of two-step estimators similar to Section 6.1 of Newey and McFadden (1994) or Chapter 12.4 in Wooldridge (2010), one can show that $(\hat{a}_k, \hat{b}_k, \hat{m}_k)_{k=1,\ldots,K}$ are jointly $\sqrt{L}$-convergent and asymptotic normal, with a limit covariance that takes account of first-stage estimation error in $\hat{\mu}_k$'s. Next, recall that our estimator has a closed form $\hat{\boldsymbol{\theta}} \equiv \left(\hat{\Psi}' \boldsymbol{\Xi}^{-1} \hat{\Psi} \right)^{-1} \left(\hat{\Psi}' \boldsymbol{\Xi}^{-1} \hat{v} \right)$, with $\hat{\Psi}, \hat{v}$ being a matrix and a vector that consist of elements in $(\hat{a}_k, \hat{b}_k, \hat{m}_k)_{k=1,\ldots,K}$. Also note that $\hat{\Psi}' \boldsymbol{\Xi}^{-1} \hat{\Psi}$ converges in probability to an invertible matrix, because Ψ has full rank and $\boldsymbol{\Xi}^{-1}$ is symmetric and positive definite. Hence one can apply the delta method to show that $\hat{\boldsymbol{\theta}}$ is $\sqrt{L}$-convergent and asymptotically normal.

B. Monte carlo simulation

We provide a simulation study of the finite sample performance of our estimator. We simulate 200 samples, each of which consists of L independent groups. Each group involves n individuals, where n is a fixed small integer.

The structural equation in our data-generating process is $y = \alpha + \lambda G y + X\beta + GX\gamma + \varepsilon$, where X is an $n \times 3$ matrix that consists of three characteristics. The parameter values are: $\alpha = 1$; $\lambda = 0.7$; $\beta = (1.5, 2, 0)'$ and $\gamma = (0.9, 0, 0.6)'$. For each observation $i = 1, .., n$, the error terms ε_i is independently drawn from a standard normal distribution. The elements in the first column of X are independently drawn from a multinomial distribution with equal

probability mass over $\{-1, 1, 2\}$, the second from a standard normal $N(0, 1)$, and the third from a normal $N(1, 2)$. The three characteristics are uncorrelated with each other. The links in the latent adjacency matrix G^* (of which G is a row normalization) are each independently drawn from Bernoulli with success probability 0.5.

Table B.1. Finite-sample Performances of the Estimator with Unobserved Links

(Group size: $n = 10$)

	$L = 60$			$L = 120$			$L = 240$			$L = 480$		
	m.s.e.	bias	std	m.s.e.	bias	std	m.s.e.	bias	std	m.s.e.	bias	std
λ	0.0197	-0.0305	0.1374	0.0044	-0.0162	0.0648	0.0017	-0.0061	0.0409	0.0010	-0.0069	0.0314
β_1	0.7232	0.0288	0.8521	0.0143	0.0133	0.1190	0.0047	0.0123	0.0677	0.0024	0.0086	0.0487
β_2	0.6762	0.0590	0.8223	0.0078	0.0130	0.0876	0.0031	0.0072	0.0553	0.0018	0.0074	0.0416
γ_1	1.3511	0.2260	1.1430	0.2911	0.0808	0.5347	0.1009	0.0399	0.3159	0.0760	0.0357	0.2740
γ_3	0.1192	0.0370	0.3441	0.0484	0.0151	0.2200	0.0225	-0.0016	0.1505	0.0125	0.0061	0.1119
α	0.5919	0.1020	0.7645	0.2349	0.0955	0.4763	0.0956	0.0336	0.3082	0.0495	0.0382	0.2198

Note: m.s.e., bias and std are calculated from empirical distribution of coefficient estimates in 200 simulated samples.

Table B.2. Finite-sample Performances of the Estimator with Unobserved Links

(Group size: $n = 20$)

	$L = 60$			$L = 120$			$L = 240$			$L = 480$		
	m.s.e.	bias	std	m.s.e.	bias	std	m.s.e.	bias	std	m.s.e.	bias	std
λ	0.0181	-0.0340	0.1305	0.0037	-0.0086	0.0603	0.0017	-0.0037	0.0417	0.0007	-0.0059	0.0258
β_1	0.0151	0.0199	0.1216	0.0031	0.0024	0.0556	0.0015	0.0051	0.0389	0.0006	-0.0020	0.0238
β_2	0.0118	0.0184	0.1071	0.0022	0.0044	0.0463	0.0008	0.0028	0.0283	0.0004	-0.0017	0.0207
γ_1	1.4307	0.2101	1.1805	0.2747	0.0443	0.5236	0.1233	0.0255	0.3510	0.0546	0.0279	0.2326
γ_3	0.1422	0.0448	0.3753	0.0373	0.0006	0.1937	0.0209	0.0016	0.1448	0.0105	0.0184	0.1010
α	0.5534	0.1597	0.7284	0.1794	0.0582	0.4206	0.1041	0.0213	0.3228	0.0495	0.0268	0.2215

Note: m.s.e., bias and std are calculated from empirical distribution of coefficient estimates in 200 simulated samples.

We estimate the model using the method in Section 5. In the first step, we use the first dimension-reduction algorithm (when regressors are uncorrelated across group members) to estimate the reduced-form coefficients, as explained in Section 6.2. Table B.1 and B.2 report the mean-squared error (m.s.e.), the bias and the standard deviation of the estimators for group sizes $n = 10$ and 20, using the empirical distribution of estimates from 200 simulate samples. We increase the sample size L, i.e. the number of groups in each sample, from $L = 60$ to $L = 480$.

The results show that our estimator is reasonably accurate even when the sample is moderately small with $L = 60$. Furthermore, the mean-squared errors diminish at the parametric rate, i.e. the same rate as the increase in sample sizes. In fact the reduction in m.s.e. between $L = 60$ and $L = 120$ is faster than the increase in sample size. This is because the first-step estimation of reduced-form coefficients consists of $n \times n$ regressions on

$K = 3$ characteristics. The reduction in estimation error in such a low-dimension regression is substantial as the number of observations increases from $L = 60$ to $L = 120$.

It is worth noting that the difference in m.s.e. is small between the DGP with small group size $n = 10$ versus the larger $n = 20$. This illustrates a desirable feature of our estimator: The precision of the estimator depends primarily on the accuracy of the first-step reduced-form coefficients. Once the constants a_k, b_k are recovered from the reduced-form coefficients, the second step does not introduce additional sampling error. A useful result for practitioners is that the first-stage estimation precision can be enhanced using the dimension-reduction methods explained in 6.2. For example, in the current simulation example, the dimension-reduction method replaces $n = 10$ regressions on $n \times K = 30$ explanatory variables with $n \times n = 100$ regressions on $K = 3$ characteristics. This dimension-reduction helps obtain the encouraging performance results reported in Tables B.1 and B.2.

C. Pooling groups with different sizes

In this appendix, we explain how to impute smaller groups with simulated "pseudo-individuals". Doing so allows us to run a pooled regression with balanced group sizes, and consistently estimate a weighted average of reduced-form coefficient matrices.

To fix ideas, let there be two group sizes $n_l \in \{\underline{n}, \bar{n}\}$ in the data-generating process only, and suppose the assumptions in Section 4 hold conditional on either group size. For each group l with $n_l = \underline{n}$, define an $\bar{n} \times K$ matrix X_l^* by stacking the observed matrix X_l (i.e., the $\underline{n} \times K$ matrix of regressors for group l in the sample) with an $(\bar{n} - \underline{n}) \times K$ matrix of draws simulated from the distribution of regressors of the other $(\bar{n} - \underline{n})$ individuals in groups with $\bar{n}$ members. By construction, X_l^* can be considered as a draw from the distribution of $X_{l'}$ when $n_{l'} = \bar{n}$. Define a $(\bar{n}K + 1)$-dimensional column vector:

$$\tilde{X}_l \equiv \begin{cases} \left(1, X'_{l,c1}, ..., X'_{l,cK}\right)' & \text{if } n_l = \bar{n} \\ \left(1, X^{*\prime}_{l,c1}, ..., X^{*\prime}_{l,cK}\right)' & \text{if } n_l = \underline{n} \end{cases},$$

with $X_{l,ck}$ denoting the k-th column in X_l as before. By construction, $E(\tilde{X}_l \tilde{X}'_l)$ is invariant to group sizes.

For a large group l with $n_l = \bar{n}$ and all $i \leq n_l$, we have $E(\tilde{X}_l y_{l,i} | n_l = \bar{n}) = E(\tilde{X}_l \tilde{X}'_l)\Phi_i(\bar{n})$, where

$$\Phi_i(\bar{n}) \equiv \left(\mu_0, \mu_{1,ri}(\bar{n}), ..., \mu_{K,ri}(\bar{n})\right)'$$

and $\mu_{k,ri}(\bar{n})$ denotes the i-th row of the $\bar{n} \times \bar{n}$ matrix of reduced-form coefficients $\mu_k(\bar{n})$ defined in Lemma 1. (Note that we now write μ_k as a function of n_l in order to emphasize its dependence on group sizes.) Likewise, for any small group l with $n_l = \underline{n}$ and all $i \leq n_l$,

we have $E(\tilde{X}_l y_{l,i}|n_l = \underline{n}) = E(\tilde{X}_l \tilde{X}_l')\Phi_i(\underline{n})$, where

$$\Phi_i(\underline{n}) \equiv \left(\mu_0, \mu_{1,ri}(\underline{n}), \mathbf{0}, \mu_{2,ri}(\underline{n}), \mathbf{0}, ..., \mu_{K,ri}(\underline{n}), \mathbf{0}\right)'$$

and $\mu_{k,ri}(\underline{n})$ denotes the i-th row of the $\underline{n} \times \underline{n}$ matrix $\mu_k(\underline{n})$ and $\mathbf{0}$ a row vector of $(\bar{n} - \underline{n})$ zeros.

Let $p(\cdot)$ denote the probability mass for n_l in the population. It then follows that for all $i = 1, ..., \underline{n}$,

$$\begin{aligned} & E(\tilde{X}_l y_{l,i}) = E(\tilde{X}_l \tilde{X}_l') \left[p(\bar{n})\Phi_i(\bar{n}) + p(\underline{n})\Phi_i(\underline{n})\right] \\ \Rightarrow \quad & E[\Phi_i(n_l)] = \left[E(\tilde{X}_l \tilde{X}_l')\right]^{-1} E(\tilde{X}_l y_{l,i}). \end{aligned}$$

Thus $E[\mu_k(n_l)]$, with n_l integrated out as a random variable, are identified and consistently estimable for $k = 1, 2, ..., K$. Assuming $\lambda, \beta, \gamma, \alpha$ are the same for small and large classes, one can then proceed and apply the method in Section 4 to estimate the structural parameters of social effects. We use this method to balance group sizes within the environments of small or large classes in our application.

D. Dependent networks

In practice, the formation of links on a network may depend on individual characteristics in the data. We now discuss how to generalize our estimator to deal with this dependence.

Begin by considering a single environment s, where all groups within the environment have the same size n, and we omit the environment superscript. This procedure can be applied separately for each environment in the data to obtain reduced-form coefficients, which would then be combined to obtain the structural parameters as in Theorems 1 and 2. Partition individual characteristics into two parts $X_l = (X_l^a, X_l^e)$. Let X_l^e denote an $n \times K_e$ matrix of excluded characteristics, i.e., covariates that affect outcomes but not link formation; let X_l^a denote an n-by-K_a matrix that affect individuals' outcomes, link formation decisions, or both. For example, in our empirical application, we let X_l^e be students' days of absence from school and test scores from previous years. This assumes friendships are independent of test scores conditional on observed demographics such as proximity of age. If we observe all variables that jointly determine network formation and outcomes, then our method can be applied after conditioning on X_l^a.

There is large and growing literature on network formation. To just name a few, Graham (2017), Hsieh, König, and Liu (2020), Hsieh, Lee, and Boucher (2020), Leung (2015), Leung (2020), and Sheng (2020) explicitly model how the links are formed as an equilibrium outcome. As stated in Graham (2019), "Ultimately, of course, the goal is to study the formation of networks and their consequences jointly, but such an integrated treatment remains

largely aspirational at this stage". Our focus in this paper is on peer effects with unobserved links, so we simply adopt the conditional independence to deal with potential endogeneity in network formation.

Suppose network formation is given by $G_l = \zeta(X_l^a, u_l)$, which does not involve X_l^e. The reduced form is:

$$E(y_l|X_l) = \int \left[\sum_{k=1}^{K} M_l(\beta_k I + \gamma_k G_l) X_{l,ck} + M_l E(\varepsilon_l|X_l, G_l)\right] dF(G_l|X_l), \tag{18}$$

where $X_{l,ck}$ denotes the k-th column in X_l as before. Assume (i) ε_l is independent of X_l^e conditional on (X_l^a, u_l) and (ii) u_l is independent of X_l^e conditional on X_l^a. These conditions allow the unobserved errors ε_l and u_l to be correlated conditional on X_l^a. Under these assumptions, $E(M_l|X_l)$ and $E(M_l G_l|X_l)$ is a function of X_l^a but not X_l^e, and

$$\int M_l E(\varepsilon_l|X_l, G_l) dF(G_l|X_l) = \int M_l E(\varepsilon_l|X_l^a, G_l) dF(G_l|X_l^a) \equiv \phi(X_l^a).$$

Conditional on X_l^a, the reduced-form coefficients for X_l in (18) are:

$$\mu_k(X_l^a) \equiv \beta_k E(M_l|X_l^a) + \gamma_k E(M_l G_l|X_l^a) \text{ for all } k \leq K.$$

With an abuse of notation, let K_a and K_e denote the set of indices for characteristics in X_l^a, X_l^e respectively so that K_a and K_e partition $\{1, 2, ..., K\}$. We can write (18) as

$$E(y_l|X_l) = \sum_{k \in K_e} \mu_k(X_l^a) X_{l,ck} + \underbrace{\sum_{k' \in K_a} \mu_{k'}(X_l^a) X_{l,ck'} + \phi(X_l^a)}_{\psi(X_l^a)},$$

which is linear in X_l^e conditional on X_l^a.

We can identify the model by the following steps. First, recover $\psi(X_l^a)$ and $\mu_k(X_l^a)$ for all $k \in K_e$ for a given realization of X_l^a. In practice, this can be estimated using reduced-form methods such as kernel estimation of an average derivative $E\left[\partial E(y_l|X_l)/\partial X_l^e\right]$, or, exploiting the structure of $E(y_l|X_l)$, using sieve regressions that replace the $\mu_k(X_l^a)$ and $\phi(X_l^a)$ functions with sieve expansions, or by linearly regressing y_l on X_l^e conditioning on discrete values of X_l^a. Then, for all $k \in K_e$, identify $\lambda, \beta_k, \gamma_k$ from $\mu_k(X_l^a)$, using the methods in Section 4. We can also back out $E(M_l|X_l^a)$ and $E(M_l G_l|X_l^a)$ from $\mu_k(X_l^a)$, $k \in K_e$, using β_k, γ_k, $k \in K_e$.[19]

E. Multiple adjacency matrices

In this part of the appendix, we discuss how our method might be extended to allow peer effects and contextual effects to operate through different adjacency matrices. The reasons

[19] One could also recover the model elements $\phi(\cdot)$ and $\mu_k(\cdot)$ for $k \in K_a$ from $\psi(\cdot)$ by making additional functional form assumptions, e.g., assuming index sufficiency in $\phi(\cdot)$ and $\mu_k(\cdot)$ for $k \in K_a$.

why one might be interested in this extension, and citations to previous literature on the subject, are in Section 2.

Again we start with the case of a single environment where all groups have identical size n, and we suppress the group subscript l throughout this section to simplify notation. Let G and W be two possibly different n-by-n adjacency matrices. For each group, peer effects and contextual effects operate through two different adjacency matrices G and W, respectively. Divide the set of individual characteristic regressors into two matrices: V is an n-by-J matrix of regressors that have both direct and contextual effects, while X is an n-by-K matrix of regressors that only have direct effects. The structural equation for the outcome is now

$$y = \lambda Gy + X\beta_X + V\beta_V + WV\gamma + \varepsilon,\ E(\varepsilon|G, X, V) = 0$$

where y and ε are n-by-1 vectors. Let $\Pr\{G \neq W\} > 0$. The structural parameters are $\lambda, \beta_X, \beta_V, \gamma, \delta$.

Now consider the same steps we used before to achieve identification. The reduced form is now

$$\begin{aligned} E(y|X,V) &= E[\underbrace{(I-\lambda G)^{-1}}_{M}(X\beta_X + V\beta_V + WV\gamma + \varepsilon)|X,V] \\ &= E\left(MX\beta_X + MV\beta_V + MWV\gamma|X,V\right) \\ &= \sum_k \underbrace{E(\beta_{X,k}M)}_{\varphi_k} X_{ck} + \sum_j \underbrace{E\left[M(\beta_{V,j}I + \gamma_j W)\right]}_{\mu_j} V_{cj} \end{aligned}$$

where X_{ck}, V_{cj} are n-by-1 vectors of single characteristics (the k-th column in X and j-th column in V). The last equality assumes $(G,W) \perp (X,V)$. Note that φ_k and μ_j are each n-by-n matrices of reduced-form coefficients. Maintain the following two conditions on the structural parameters:

$$\beta_{X,k} \neq 0\ \forall k \leq K;\ rank\begin{pmatrix} \beta_{V,j} & \beta_{V,J} \\ \gamma_j & \gamma_J \end{pmatrix} = 2\ \forall j < J, \tag{19}$$

where the second condition rules out proportional social effects as well as the special case with no contextual effects ($\gamma_j = 0$ for all $j \leq J$). We also assume

$$\begin{aligned} & E(M) \neq 0,\ E(MW) \neq 0; \\ \nexists(\Delta_1, \Delta_2) \neq\ & (0,0) \text{ s.t. } \Delta_1 E(M) + \Delta_2 E(MW) = 0. \end{aligned} \tag{20}$$

This second condition in (20) rules out the pathological case where the $n \times n$ entries in $E(M)$ are proportional to those in $E(MW)$.

Lemma E.1. For each $j \leq J-1$ and $k \leq K$, the system of equations

$$a_{jk}\mu_j + b_{jk}\mu_J = \varphi_k \tag{21}$$

has a unique solution

$$\begin{pmatrix} a_{jk} \\ b_{jk} \end{pmatrix} = \begin{pmatrix} \beta_{V,j} & \beta_{V,J} \\ \gamma_j & \gamma_J \end{pmatrix}^{-1} \begin{pmatrix} \beta_{X,k} \\ 0 \end{pmatrix}. \tag{22}$$

Proof of Lemma E.1. It is straightforward to check that (a_{jk}, b_{jk}) defined in (22) solves (21). To see that this is a unique solution, suppose there exists $(\tilde{a}_{jk}, \tilde{b}_{jk}) \neq (a_{jk}, b_{jk})$ such that (21) holds with (a_{jk}, b_{jk}) replaced by $(\tilde{a}_{jk}, \tilde{b}_{jk})$, and

$$\begin{pmatrix} \beta_{V,j} & \beta_{V,J} \\ \gamma_j & \gamma_J \end{pmatrix} \begin{pmatrix} \tilde{a}_{jk} - a_{jk} \\ \tilde{b}_{jk} - b_{jk} \end{pmatrix} = \begin{pmatrix} \Delta_1 \\ \Delta_2 \end{pmatrix} \neq 0,$$

where the inequality follows from the rank condition in (19). It then follows that

$$(\tilde{a}_{jk} - a_{jk})\mu_j + \left(\tilde{b}_{jk} - b_{jk}\right)\mu_J = E\left(\Delta_1 M + \Delta_2 MW\right) = 0. \tag{23}$$

The last equality is ruled out by (20). □

Lemma E.1 provides an analog to Lemma (1). It may then be possible to combine these equality constraints with rank restrictions like exclusions and multiple environments to construct a corresponding extension of Theorem 1 to attain identification of this extended model.

F. Group-level fixed effects

Our identification strategy can be extended to allow for group-level unobserved heterogeneity, i.e., group-level fixed effects. First, we note that if the group-level unobserved heterogeneity is mean independent from the group and individual-level covariates in (z, X) (corresponding to the usual assumption in random effects models), then the estimation method described in Section 6.1 can be directly applied, because in this case the conditional mean of y given (z, X) is as specified in equation (17).

Now, consider instead the more general fixed effects model. We now have the reduced-form

$$y = M\left(X\beta + GX\gamma + \varepsilon\right) + \frac{\alpha}{1-\lambda}\iota + \frac{z\rho}{1-\lambda}\iota + \frac{\varpi}{1-\lambda}\iota,$$

where α is still the intercept, z are observed group characteristics and ϖ is the unobserved group heterogeneity (fixed effects). Let $D = I - C$, where C is an n-by-n matrix of identical

entries $1/n$, so that Dy returns the within transformation of y. Then under the assumptions that $E(\varepsilon|X,G)=0$ and $G\perp X$, a within transformation leads to

$$Dy = DM(X\beta + GX\gamma + \varepsilon) \Rightarrow E(Dy|X) = E(DM)X\beta + E(DMG)X\gamma.$$

Thus we can write the reduced-form coefficients for the k-th characteristic from a regression using the within transformation as

$$\tilde{\mu}_k \equiv E\left(\beta_k DM + \gamma_k DMG\right) = DE[M(\beta_k I + \gamma_k G)].$$

Assume the rank condition in Assumption 5-(i) holds and that

$$\tilde{\mu}_K \neq cD \text{ for any } c \in \mathbb{R}. \tag{24}$$

This condition can in principle be checked directly using the identifiable $\tilde{\mu}_K$. It can then be established that the following system

$$a_k\tilde{\mu}_k + b_k\tilde{\mu}_K = D$$

admits a unique solution

$$\begin{pmatrix} a_k \\ b_k \end{pmatrix} = \begin{pmatrix} \beta_k & \beta_K \\ \gamma_k & \gamma_K \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ -\lambda \end{pmatrix}. \tag{25}$$

The proof is almost the same as that of Lemma 1, except that the condition "$\mu_k \neq cI$ for any $c \in \mathbb{R}$" in Assumption 5 is replaced by (24). It is worth emphasizing that the first stage reduced-form regressions now consist of regressing demeaned outcomes Dy on the *undemeaned* characteristics X to get reduced-form coefficients. Given (25), we can then apply the constructions of Theorem 1 to identify λ, β, γ.

This "fixed-effect" approach does not immediately identify the coefficient for group-level variables ρ or the intercept α, in the same way that the "within-transformation" does not identify the coefficients of variables that only vary by time in linear panel data models. We can identify these remaining parameters from undemeaned reduced-form $E(y_l|X_l, z_l) = \mu_0 + E(M_l)\iota z_l\rho + E(M_l)X_l\beta + E(M_lG_l)X_l\gamma$ by imposing the exogeneity and location normalization condition $E(\varpi_l|z_l, X_l, G_l) = 0$.

References

Angrist, J.,and A. Krueger. 1991. Does Compulsory School Attendance Affect Schooling and Earnings? The Quarterly Journal of Economics, 106(4), 979-1014.

Athey, S. and P. Haile. 2007. Nonparametric Approaches to Auctions. In J. Heckman and E. Leamer, eds., Handbook of Econometrics, Vol. 6A, Elsevier, ch60, 3847-3965.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton. 2017. From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. Journal of Economic Perspectives, 31(4), 73-102.

Blume, L., W. Brock, S. Durlauf, and R. Tayaraman. 2015. Linear social interactions models. Journal of Political Economy 123(2), 444-496.

Boozer, M A., and S E. Cacciola. 2001, Inside the 'Black Box' of Project Star: Estimation of Peer Effects Using Experimental Data, working paper.

Boucher, V, Y. Bramoullé, H. Djebbari, and B. Fortin. 2014. Do peers affect student achievement? Evidence from Canada using group size variation. Journal of Applied Econometrics 29, 91–109.

Bramoullé, Y, H. Djebbari, and B. Fortin. 2009. Identification of peer effects through social networks. Journal of Econometrics 150, 41–55.

Breza, E., A. G. Chandrasekhar, T. H. McCormick, and M. Pan. 2020. Using Aggregated Relational Data to Feasibly Identify Network Structure without Network Data. American Economic Review, 110(8), 2454-84.

Calvó-Armengol A, E. Patacchini, and Y. Zenou. 2009. Peer effects and social networks in education. Review of Economic Studies 76, 1239–1267.

Dee, T.S. 2004, Teachers race and student achievement in a randomized experiment, Review of Economics and Statistics, 86(1), 195–210.

De Paula Á., I. Rasul, and P. CL Souza, 2018. Recovering social networks from panel data: identification, simulations and an application, CeMMAP working papers CWP58/18.

Erdős, P. and A. Rényi, 1959. On Random Graphs. Publicationes Mathematicae 6, 290–297.

Fisher, F. 1966. The Identification Problem in Econometrics. Huntington, N.Y.: Krieger Publishing Company.

Fox, J., C. Yang, and D. Hsu. 2015. Unobserved Heterogeneity in Matching Games Journal of Political Economy, 126(4), 1339-1373.

Graham, B. S. 2008, Identifying social interactions through conditional variance restrictions. Econometrica, 76(3), 643–60.

Graham, B. S., and J. Hahn. 2005. Identification and estimation of the linear-in-means model of social interactions. Economics Letters, 88(1), 1-6.

Graham, B. S. 2017, An econometric model of network formation with degree heterogeneity. Econometrica, 85(4), 1033–1063.

Graham, B. S. 2019, Network data, NBER Working paper 26577.

Griffith, A. 2021. Name Your Friends, But Only Five? The Importance of Censoring in Peer Effects Estimates Using Social Network Data. Working paper, The University of Washington.

Goldsmith-Pinkham, P., and GW. Imbens. 2013. Social networks and the identification of peer effects. Journal of Business and Economic Statistics 31, 253–264.

Hauser, C., M. Pfaffermayr, G. Tappeiner, and J. Walde. 2009. Social capital formation and intra familial correlation: A social panel perspective. Singapore Economic Review 54, 473–488.

Hanushek, E., 1999, Some Findings From an Independent Investigation of the Tennessee STAR Experiment and From Other Investigations of Class Size Effects, Educational Evaluation and Policy Analysis, 21(2), 143-164.

Heckman, J. J. and S. Mosso, 2014, The Economics of Human Development and Social Mobility, Annual Review of Economics, 6, 689-733.

Hsieh, C. and L. Lee. 2016. A Social Interactions Model with Endogenous Friendship Formation and Selectivity, Journal of Applied Econometrics 31, 301-319.

Hsieh, C, M. D. König, and X. Liu. 2020. A Structural Model for the Coevolution of Networks and Behavior, The Review of Economics and Statistics, forthcoming.

Hsieh, C, L. Lee, and V. Boucher. 2020. Specification and estimation of network formation and network interaction models with the exponential probability distribution, Quantitative Economics, 11(4), 1349-1390.

Hunter, D. R., S. M. Goodreau, and M. S. Handcock. 2008. Goodness of Fit of Social Network Models, Journal of the American Statistical Association, 103:481, 248-258.

Koopmans, T. C. (1949), Identification problems in economic model construction, Econometrica, 17, 125-144.

Krueger, A. B., 1999. Experimental Estimates of Education Production Functions, Quarterly Journal of Economics, 114(2), 497-532.

Krueger, A. B., and D. M. Whitmore, 2001, The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR, The Economic Journal, 111(1), 1-28.

Lee, L. 2007. Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. Journal of Econometrics 140: 333–374.

Lee, L., X. Liu, and X. Lin. 2010. Specification and estimation of social interaction models with network structures. Econometrics Journal 13, 145–176.

LeSage and Pace 2009. Introduction to spatial econometrics. Taylor Francis/CRC Press,

Boca Raton.

Leung, M. P. 2015. Two-Step Estimation of Network-Formation Models With Incomplete Information, Journal of Econometrics, 188(1), 182–195.

Leung, M. P. 2020. Equilibrium computation in discrete network games, Quantitative Economics, 11(4), 1325-1347.

Lin, X. 2010. Identifying peer effects in student academic achievement by spatial autoregressive models with group unobservables. Journal of Labor Economics 28, 825–860.

Manresa, E., 2016. Estimating the Structure of Social Interactions Using Panel Data, working paper.

Manski, C F. 1993. Identification of Endogenous Social Effects: The Reflection Problem. Review of Economic Studies 60(3): 531–42.

Newey, W K. and D. McFadden. 1994. Large sample estimation and hypothesis testing. Handbook of Econometrics. IV. Elsevier Science, 2111–2245.

Patacchini E, and Y. Zenou. 2012. Juvenile delinquency and conformism. Journal of Law, Economics, and Organization 28: 1–31.

Pinkse, J., ME. Slade, and C. Brett, 2002. Spatial price competition: A semiparametric approach, Econometrica 70, 1111–1153.

Rose, C. 2017. Identification of peer effects through social networks using variance restrictions. Econometrics Journal 20, S47–S60.

Rose, C. 2018. Identification of Spillover Effects using Panel Data, working paper.

Sheng, S. 2020. A structural econometric analysis of network formation games. Econometrica, 88(5) 1829–1858.

Sojourner, A. 2013. Inference on peer effects with missing peer data: evidence from project STAR, Economic Journal, 123(569), 574–605.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 58(1), 267-288.

Whitmore, D. 2005. Resource and peer impacts on girls academic achievement: evidence from a randomized experiment, American Economic Review, 95(2), 199–203.

Wooldridge, J. 2010. Econometric Analysis of Cross Section and Panel Data. 2nd Ed. MIT Press.