Social Networks with Unobserved Links

Arthur Lewbel — Boston College

Xi Qu — Shanghai Jiao Tong University

Xun Tang — Rice University

Original: July 2019, revised July 2022

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.

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 $l \in \{1, \ldots, L\}$. Each group l has an unobserved $n_l \quad n_l$ 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 $G_1,\ldots,G_L$. 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 network model:[2]

$$y_l = \quad + \quad G_l y_l + X_l \quad + G_l X_l \quad + \varepsilon_l, \tag{1}$$

where y_l and ε_l are $n_l \quad 1$ vectors of outcomes and errors, respectively, $\quad$ an $n_l \quad 1$ vector of ones, and X_l an $n_l \quad K$ 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

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

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.

e¤ects operate through the same adjacency matrix G_l. This assumption is standard in the literature whenever both peer and contextual e¤ects are included in a model. See, e.g., Lee (2007), Bramoullé, Djebbari and Fortin (2009), and de Paula Rasul and Souza (2020). One paper that relaxes this assumption is Blume , Brock, Durlauf and Jayaraman (2015). This assumption is generally imposed because it would be di¢ cult to distinguish from data the extent to which any observed link applies to peer e¤ects versus to contextual e¤ects. We are not aware of any data sets where such information has been collected. However, since our identi…cation is intended precisely to cover situations where link data is not, or cannot, be observed, it is possible that our methods 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 identi…cation of linear network models and identi…cation 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 (2020), which is equation (1), but simpli…ed by having $G_l = G$ and n_l = n416(9)]TJ/Fk6

The linear-in-means model, which corresponds to a G having all off-diagonal elements equal to $1/(n-1)$, 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 draws from some unknown distribution of possible networks. As explained below, our method requires these networks to be exogenous from the individual characteristics whose social effects are to be identified.

By convention in the literature, the diagonal entries in each G_l are all zeros, i.e., $G_{lii} = 0$ for $i = 1; \ldots; n_l$. 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 $G_{lij} = G_{lij} / \sum_{j'=1}^{n_l} G_{lij'}$

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

$$\begin{pmatrix} {}_k & {}_K \\ {}_k & {}_K \end{pmatrix}$$

has full rank. (ii) $_K \neq cI$ for any $c \in \mathbb{R}$, where $_K$ is a matrix of reduced-form coefficients for the K-th regressor as defined in equation (5).

Part (i) of Assumption 5 rules out the pathological case where some pair 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 $_K = 0$ (one of the regressors has no contextual effect) while $_K$, $_k$, and $_k$ 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

The first step of our identification strategy is to show how the reduced-form parameters relate to the structural components of our model. 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.

Lemma 1 Under Assumptions 1-4, the reduced-form parameters $_0$ and $_k$ for $1 \le k \le K$, defined in (5), are identified.

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

$$E(y_i \mid X) = {}_0 + e_i E(M) X + e_i E(MG) X \; ; \tag{6}$$

where e_i is a $1 \times n$ unit-vector whose i-th component is 1. Observe that the right-hand side of (6) 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 (6) also uses the fact that the row-normalization of G implies

$$M \quad = \quad \left[\sum_{s=0}^{\infty}(\; G)^s\right] \quad = \quad {}_0 \, . \tag{7}$$

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

In the reduced form of equation (6), the slope coefficient for the k-th regressor of individual j is $\;_k \; e_i E(M) e_j' \; + \; _k \; e_i E(MG) e_j'$. (Note that, for a generic $n \quad n$ 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 $\; _k$ for $k = 1, \ldots, K$.

Remark 1 The representation of $E(y \mid X)$ in (6) is consistent not only with the simultaneous social network model with complete information given by equation (1), but also with

size is moderately large.[7] Otherwise, the researcher needs to take measures to estimate the reduced-form coe¢ cients using limited data. For example, instead of requiring the sample size be large relative to the number of regressors in OLS, de Paula et al. (2020) impose a sparsity condition on the structural-form adjacency matrix, and then use a penalization approach to estimate the reduced-form interaction matrix. In contrast, we propose alternative ways to deal with such data de…ciency using anatomy of partitioned regressions in Section 6.2(f)7(y)-3894(nS(h)1
thrthrthrs9Td[()7Othr6c2(t7e(c)(o)11(n611(p)11(p)(t)8n)p12(c1(p)1r12(c1(p)12(t7e(c)1(p)(tc)r343(s-9Td

be a scalar multiple of I in order for (9) to hold for $(\tilde{a}_k; \tilde{b}_k)$. Case 3: $\tilde{a}_k \neq a_k$, $\tilde{b}_k \neq b_k$. Then (11) requires $_k = \frac{\tilde{b}_k \quad b_k}{\tilde{a}_k \quad a_k} \; _K$, which is a scalar multiple of $_K$. Again, this implies that in order for (9) to hold for $(\tilde{a}_k; \tilde{b}_k)$, $_K$ must be a scalar multiple of I. In each of these three cases, the implication of (11) contradicts part (ii) of Assumption 5. □

The reduced-form coe¢ cients $_0$ and $_k$ are identi…ed by Lemma 1. Therefore, for each $k \quad K \quad 1;$

For general cases with $K > 3$, the linear system in (14) is generalized to:

$$\underbrace{\begin{pmatrix} 0_{(K-1)\times 1} & H & 0_{(K-1)\times K} \\ {}_{(K-1)\times 1} & 0_{(K-1)\times K} & H \\ m & I & I \end{pmatrix}}_{} \underbrace{\begin{pmatrix} \\ \\ \\ \end{pmatrix}}_{} = \underbrace{\begin{pmatrix} {}_{(K-1)\times 1} \\ 0_{(K-1)\times 1} \\ m \end{pmatrix}}_{}, \tag{15}$$

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

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

The rank of the matrix is generically $2K-1$. It cannot be greater than $2K-1$ by construction, and is strictly less than $2K-1$ only if the DGP 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 finite number of environments in the population. We allow all model parameters and group sizes to vary across environments, and so all can be given an s superscript. Within each environment, the structural parameters are fixed. For example, environment can be defined by classroom size as in our empirical study. Notice $S = L$ is ruled out because S is finite and $L \to \infty$. 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).

Because structural parameters $^{(s)}$ $(^{(s)}, ^{(s)\prime}, ^{(s)\prime})' \in \mathbb{R}^{2K+1}$ and the distribution of (G, X, ε) vary by environment in general, we index them with superscripts s, $(G^{(s)}, X^{(s)}, \varepsilon^{(s)})$, to emphasize that they are allowed to be drawn from different distributions across different environments. For example, for two groups l and k from the same environment s, their adjacency matrices G_l and G_k differ but are drawn from the same distribution indexed by s; in comparison, for two groups l and k' from different environments s and s', the adjacency matrices G_l and $G_{k'}$ are drawn from two different distributions, indexed by s and s' respectively. Now identification of the model requires that we identify $^{(s)}$ for each environment s.

Suppose $^{(s)}$ and the distribution of $(G^{(s)}, X^{(s)}, \varepsilon^{(s)})$ satisfy the restrictions in Assumpwhere $\boldsymbol{\theta}$ and d are column vectors that stack $^{(s)}$ and $^{(s)}$ respectively for $s = 1; :::; S$; and where all o¤-diagonal elements of $G^{(s)}$ equal $1=(n^{(s)} \quad 1)$. The re‡ection problem shows that in this model, even if $G^{(s)}$ were known, the structural parameters would not be identi…ed 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 identi…cation.[9]

There are two types of rank restrictions that are most natural to impose. The …rst type are exclusion restrictions, which consist of assuming that some elements of either or 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.[10] To illustrate, suppose $K = 3$ and $S = 1$. In this case it su¢ ces to assume that one regressor X_k has no contextual e¤ect ($_k^{(1)} = 0$) and a non-zero direct e¤ect ($_k^{(1)} \neq 0$), while another regressor $X_{k'}$ has no direct e¤ect ($_{k'}^{(1)} = 0$) and a non-zero contextual e¤ect ($_{k'}^{(1)} \neq 0$). More generally, with $K \quad 3$, has full rank generically if R is de…ned by the exclusion restrictions that there exist $k, k^0 < K$ with $_k = 0$, $_{k'} = 0$ and $_k \neq 0;$ $_{k'} \neq 0$. So essentially, we get identi…cation if one regressor has no contextual e¤ects and another has no direct e¤ects. In contrast, restricting two regressors to both have no contextual e¤ects but nonzero individual e¤ects would not su¢ ce to make full rank (this turns out to be a case where the order condition would be satis…ed but the rank condition is not).

Since it would be unusual for covariates to have context.na no sutoooould nae05.2790Td[(.)8-17(e)-416(

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) equals zero, and we impose the constraint that $^{(1)} \neq {}^{(2)}$, then that provides enough restrictions to generically satisfy Theorem 1.

Note that the requirement that $^{(1)} \neq {}^{(2)}$ can be tested in this case, since, by equation (16), $^{(1)} \neq {}^{(2)}$ if and only if $m^{(1)} \neq m^{(2)}$.

The assumption that and do not vary by environment in this example can be relaxed. For example, if the direct effects are the same across groups but the contextual effects vary, so $^{(1)} \neq {}^{(2)}$, 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 $^{(1)}$ and $^{(2)}$ 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 allow to vary by class size while fixing and . This generalizes the models using class size variation to estimate constant peer effects (e.g., Boozer and Cacciola (2001) and Graham (2008)). 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 corresponding to days of absence is set to zero. This exclusion restriction is motivated by the fact that common specifications of student outcomes in the empirical literature typically do further rank restrictions. The two approaches proposed in this section could precisely serve this purpose. For example, if the model imposes no contextual effects, i.e., $\ _k = 0$ for $k = 1; 2; 3$, we can uniquely solve for $(\ ;\ _1;\ _2;\ _3)$ from the linear system (14) provided the coefficient matrix, after dropping the last three rows, has full rank (four). Alternatively, we can accommodate contextual effects but exploit the presence of multiple environments to add rank restrictions by adopting the second approach proposed above. We note that these additional required rank restrictions may in practice impose strong additional assumptions on the model.

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 ability, one could simply randomly label individuals from 1 to n in each group. However,

5 Estimation

To estimate the structural parameters of our model, we use a sample of outcomes and regressors over random networks $(y_l, X_l)_{l=1,2,\ldots,L}$. Assume that across $l = 1, \ldots, L$, $(y_l, G_l, X_l, \epsilon_l)$

where $\hat{}$ is the coe¢ cient matrix formed by stacking (12) and (13) along with the exclusion restrictions $R\boldsymbol{\theta} = c$, as in Theorem 1.

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

$$\hat{} \left(\begin{array}{ccccccc} 0 & \hat{a}_1 & 0 & \hat{b}_1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & \hat{a}_1 & 0 & \hat{b}_1 \\ 0 & 0 & \hat{a}_2 & \hat{b}_2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & \hat{a}_2 & \hat{b}_2 \\ \hat{m} & & & \vert & & & \vert \\ & & & R & & & \end{array} \right); \hat{c} \left(\begin{array}{c} 1 \\ 0 \\ 1 \\ 0 \\ \hat{m} \\ c \end{array} \right);$$

with $R = c$ representing equalities describing the exclusion restrictions, such as some of the contextual and direct e¤ects being set to zero. Finally, the remaining structural parameter is estimated by $b = (1 \quad ^{b})b_0$.

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

$$\hat{\boldsymbol{\theta}} \quad \arg\min_{2} (\hat{}\boldsymbol{\theta} \quad \hat{c})' \Xi^{\ 1} (\hat{}\boldsymbol{\theta} \quad \hat{c}),$$

where denotes the feasible parameter space and ^{1} is a chosen weight matrix that is function in (17) depends on $\hat{}_k^{(s)}$ smoothly. As $L \to \infty$, this objective function converges in probability, uniformly over the parameter space, to its limit where $\hat{}_k^{(s)}$ is replaced by $_k^{(s)}$. 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 and v consist of known constants, $a_k^{(s)}$, $b_k^{(s)}$, and $m_k^{(s)}$ for $k \quad K$ and $s \quad S$. It then follows from the Slutsky Theorem that $\hat{\theta}$ is consistent for θ.

In Appendix A, we also explain why $\hat{\theta}$ is $\sqrt{L}$-convergent and asymptotically normal. Essentially, this result comes from the parametric convergence of OLS 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 $z_l \in R^P$. 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

$$y_l = \quad + \quad G_l y_l + \; z_l \; + X_l \; + G_l X_l \; + \varepsilon_l,$$

with $\in R^P$ being a column vector of additional coefficients. One could interpret 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 (i) of Assumption 4 hold with $\tilde{X}_l \quad (1; z_l\, X$

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 $^{(s)}_{0};\ ^{(s)}_{k}\ _{k\ K}$ for s S and $R\boldsymbol{\theta} = c$ as before, and estimate each $\hat{}^{(s)}$ using $\hat{}^{(s)} = \hat{}^{(s)}(1 \quad \hat{}^{(s)})$.

Finally, this procedure can be further extended to accommodate unobserved group-level fixed effects (denoted $\$_l$). 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 $y \quad y$, and following the same steps as before to estimate the structural parameters $\boldsymbol{\theta}$. Then, we can recover the remaining parameters and by plugging the estimates for $\boldsymbol{\theta}$ into the non-demeaned reduced form in (18), and applying an exogeneity and location normalization assumption that $E(\$_l \mid z_l; X_l; G_l) = 0$. 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 matrices $_k$ for k K. Then, given these $_k$ 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 $Kn^{(s)}$. We adopt the second approach to estimate reduced form coefficients in our application.

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

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.[16] 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' math 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 of days of absence from school (abs), students' self-reported motivation scores (mot 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.

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 7.2. Test of Equal Means
(small vs. large classes)

	p-value		p-value
t3	0.001	abs	0.402
t2	< 0.001	motm3161r161-387̄2-387955Tfd[(0)10(.)7(4)10e0.40.4	

7.2 Econometric specification

Our model, corresponding to equation (1), is

$$
\begin{aligned}
t3_{l;i} = {}& {}^{(s)} + {}^{(s)} \sum_j G^{(s)}_{lij} t3_{l;j} + {}_1 abs_{l;i} + {}_2 mot_{l;i} + {}_3 t2_{l;i} + {}^{(s)} tec_l \\
& + {}_2 \sum_j G^{(s)}_{lij} mot_{l;j} + {}_3 \sum_j G^{(s)}_{lij} t2_{l;j} + "_{l;i},
\end{aligned}
$$

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 ${}^{(s)}$, direct effects $({}_1; {}_2; {}_3)$, contextual effects $({}_2; {}_3)$, intercepts ${}^{(s)}$, and correlated effects ${}^{(s)}$ (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 allows abs to have a direct effect $({}_1 \neq 0)$ but no contextual effects $({}_1 = 0)$. 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 and contextual effects are the same in the two environments, small and large class sizes (which is why and do not have s superscripts above). All other structural parameters, i.e., the intercept ${}^{(s)}$, the peer effect ${}^{(s)}$, and the correlated effect ${}^{(s)}$, are permitted to differ between small $(s = 1)$ vs large $(s = 2)$ classes. These con-

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.

Estimates of peer effects are statistically significant and positive in both small and large classes, with the estimated coefficient 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 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, uden8(t)8(i)6(v)3(s)8(t)0(0)11(1)11rst0mates. St8(029(1)11(%)-300(0)6)-30

7.4 Speci…cation tests

Table 7.4: Tests for Over-identification

	p-values
low disp.	0:569
high disp.	0:358

Table 7.5: Wald Test Statistics 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)

Table 7.6: CMD Test Statistics 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)

Table 7.7: Differences in Test Scores under the Linear-in-Means Network

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

Table 7.8: Impact of Counterfactual Peer Effects

	Est. mean	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 : 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.

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

$$_k \quad (I \quad G)^{1}(\,_k I + \,_k G) = I + \frac{}{1} G\ (\,_k I + \,_k G).$$

This in turn means that all the o¤-diagonal components in $_k$ must be identical. We calculate Wald test statistics using a 6 6 leading principal minor of the reduced form co a large number of simulated draws r) of the simulated model-implied marginal effects $(I$ $\hat{}G_r(p))$ $^1(\hat{}_k I + \hat{}_k G_r(p))$. 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

Appendix

A. Proofs

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

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

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

$${}_{l,i} \quad [\ {}_0; (\ {}_1 M_{l,ri} + \ {}_1 M_{l,ri} G_l); \ldots; (\ {}_K M_{l,ri} + \ {}_K M_{l,ri} G_l)]'$$

with ${}_k; {}_k$ being the k-th components in $\;;\;$. Recall that the joint distribution of $(y_l; X_l)$ is directly identified in the data-generating process (DGP) under Assumption 1. By construction, for each individual i,

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

where the second equality holds because of the exogeneity of $(G; X)$ in Assumption 2, and the independence between G and X in Assumption 3. Under the non-singularity of $E \ \tilde{X}_l \tilde{X}_l'$ in Assumption 4-(i), we can recover $E(\ {}_{l,i})$ from the joint distribution of $(y_l; X_l)$ as

$$E(\ {}_{l,i}) = \left[E \ \tilde{X}_l \tilde{X}_l' \right]^{-1} E \ \tilde{X}_l y_{l,i}$$

for each $i = 1; 2; \ldots; n$. Rearranging the components in $E(\ {}_{l,i})$, we identify ${}_0 \quad =(1 \quad)$ and ${}_k \quad E[M_l(\ {}_k I + \ {}_k G_l)]$ for each $k = 1; \ldots; K$. □

Proof of Theorem 2. The estimators for reduced form coe¢ cients in Step 1 are OLS estimators for slope coe¢ cients in a regression. Thus under Assumptions 1-3 and 4-(i), $\hat{}_k \overset{p}{\to} {}_k$, $\hat{m}_k \overset{p}{\to} m_k$ for all $k \quad K$. Next, for each $k = 1; \ldots; K \quad 1$

in Step 2 converges in probably to its population counterpart uniformly over $(a_k; b_k)$. That is, for all $k \le K$,

$$\sup_{a_k;b_k} \left| \sum_{i;j} \left(e_i(a_k \hat{\Sigma}_k + b_k \hat{\Sigma}_K - I)e_j' \right)^2 - \sum_{i;j} \left(e_i(a_k \Sigma_k + b_k \Sigma_K - I)e_j' \right)^2 \right| \xrightarrow{p} 0:$$

By Lemma 2, the limit function $\sum_{i;j} \left(e_i(a_k \Sigma_k + b_k \Sigma_K - I)e_j' \right)^2$

second step does not introduce additional sampling errors. A useful result for practitioners is that the first-step 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 $n \quad K = 30$ explanatory variables with $n \quad n = 100$ 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 “pseudoand $_{k;ri}(\underline{n})$ denotes the i-th row of the $\underline{n} \times \underline{n}$ matrix $_k(\underline{n})$ and $\mathbf{0}$ a row vector of $(n - \underline{n})$ zeros.

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

$$E(\tilde{X}_l y_{l;i}) = E(\tilde{X}_l \tilde{X}_l')\,[p(n)\ _i(n) + p(\underline{n})\ _i(\underline{n})]$$
$$\Rightarrow \quad E[\ _i(n_l)] = \left[E(\tilde{X}_l \tilde{X}_l')\right]^{-1} E(\tilde{X}_l y_{l;i}).$$

Thus $E[\ _k(n_l)]$, with n_l integrated out as a random variable, are identified and consistently estimable for $k = 1, 2, \ldots, K$. Assuming $\ ;\ ;\ ;\ $ 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 a 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 $G_l = (X_l^a, u_l)$, which does not involve X_l^e. The reduced form is:

$$E(y_l|X_l) = \int \Big[\sum_{k=1}^{K} M_l({}_k I + {}_k G_l) X_{l,ck} + M_l E(\epsilon_l|X_l, G_l) \Big] dF(G_l|X_l), \qquad (19)$$

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

$$\int M_l E(\epsilon_l|X_l, G_l) dF(G_l|X_l) = \int M_l E(\epsilon_l|X^{a}_{l}$$

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

has a unique solution

$$\begin{pmatrix} a_{jk} \\ b_{jk} \end{pmatrix} = \begin{pmatrix} V_{;j} & V_{;J} \\ j & J \end{pmatrix}^{1} \begin{pmatrix} X_{;k} \\ 0 \end{pmatrix}. \tag{23}$$

Proof of Lemma E.1. It is straightforward to check that $(a_{jk}; b_{jk})$ defined in (23) solves (22). To see that this is a unique solution, suppose there exists $(\tilde{a}_{jk}; \tilde{b}_{jk}) \neq (a_{jk}; b_{jk})$ such that (22) holds with $(a_{jk}; b_{jk})$ replaced by $(\tilde{a}_{jk}; \tilde{b}_{jk})$, and

$$\begin{pmatrix} V_{;j} & V_{;J} \\ j & J \end{pmatrix} \begin{pmatrix} \tilde{a}_{jk} \quad a_{jk} \\ \tilde{b}_{jk} \quad b_{jk} \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \neq 0,$$

where the inequality follows from the rank condition in (20). It then follows that

$$(\tilde{a}_{jk} \quad a_{jk}) \; {}_j + \; \tilde{b}_{jk} \quad b_{jk} \; {}_J = E(\; {}_1M + \; {}_2MW) = 0. \tag{24}$$

The last equality is ruled out by (21). □

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 (18).

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

$$y = M(X \quad + GX \quad + ") + \frac{}{1} + \frac{z}{1} + \frac{\$}{1},$$

where is still the intercept, z are observed group characteristics and $\$$ is the unobserved group heterogeneity (fixed effects). Let $D = I \quad C$, 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 $E("|X; G) = 0$ and $G ? X$, a within transformation leads to

$$Dy = DM(X \quad + GX \quad + ") \;) \; E(Dy|X) = E(DM)X \quad + E(DMG)X \; .$$

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

$$\sim_k \quad E({}_k DM + {}_k DMG) = DE[M({}_k I + {}_k G)].$$

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

$$\sim_K \neq cD \text{ for any } c \in R. \tag{25}$$

This condition can in principle be checked directly using the identifiable $\sim_K$. It can then be established that the following system

$$a_k \sim_k + b \quad {}_b Ka$$

Fisher, F. 1966. The identification problem in econometrics. Huntington, N.Y.: Krieger Publishing Company.

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.