# Social Networks with Misclassified or Unobserved Links

Arthur Lewbel
Boston College

Xi Qu
Shanghai Jiao Tong University

Xun Tang
Rice University

June 25, 2019

**Abstract**

We identify and estimate social network models when network links are either misclassified or unobserved. We first derive conditions under which some misclassification of links does not interfere with the consistency or asymptotic properties of standard instrumental variable estimators of social effects. Second, we construct a consistent estimator of social effects in a model where network links are not observed at all. Our method does not require repeated observations of individual network members. 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; C51

*Keywords:* Social networks; Peer effects; Misclassification; Unobserved network

# 1 Introduction

In many social and economic environments, an individual's behavior or outcome (e.g. a consumption choice or a test score) depends not only on his own characteristics, but also on the behavior and characteristics of other individuals. We refer to such dependence between two individuals as a *link* and individuals with such links as *neighbors*. A *social network* consists of a group of linked individuals. In general, the set of neighbors varies across individuals within the group, and each individual may assign heterogenous weights to his neighbors. The structure of a social network is fully characterized by a square matrix which lists all links (with possibly heterogenous weights) among the individuals in the group, known as the *adjacency matrix*.

Much of the econometric literature on social networks focuses on disentangling and estimating various social effects based on observed outcomes and characteristics of network members. These structural parameters include the effects on each individual's outcome of (i) the individual's own and group characteristics (*direct effects*), (ii) the characteristics of his neighbors (*contextual effects*) and (iii) the outcomes of his neighbors (*peer effects*).

Existing methods of estimating these structural network effect parameters require either that the adjacency matrix of links among individuals in the sample be observed (as in, e.g., Bramoullé, Djebbari and Fortin (2009)), or that we observe many observations of outcomes of the same individuals (as in, e.g., Blume, Brock, Durlauf and Jayaraman (2015) or De Paula, Rasul and Souza (2018)).[1]

**1.1. Our contribution.** In this paper we relax these data requirements. We first consider the case where some network links are either misreported or missing. Here we provide good news for empirical researchers; we show that standard instrumental variables estimation and quasi-maximum likelihood estimation of network models (consisting of either many separate networks or a single growing network) remain consistent even in the presence of misclassified or unreported links, as long as the number of such links does not grow too quickly with the sample size.

We next consider point identification and estimation of structural social effects parameters when the adjacency matrix is not observed at all, and where individuals are each observed only once. Since many surveys do not include link data, these results have very many potential applications.

In this case where the adjacency matrix is unobserved, we assume the data consists of individuals in many separate (or almost separate) networks, such as students in many different schools or residents of many different villages. In this data generating process, we know the group or network each individual belongs to (e.g., which school or village each person is in), but we do not observe any information about the links between individuals within each group. Instead, we assume each group network is a draw from some underlying distribution of possible networks, and we identify some features of that underlying distribution along with the structural parameters of the model. The

[1] Blume et al (2015) do not explicitly assume many observations of the same individuals. Rather, they assume that the reduced-form coefficients implied by a fixed unknown network is identified, which would presumably require some kind of repeated observations in practice.

first and second parts of the paper are unified by applying our earlier results to these unobserved networks, thereby allowing some links to exist between groups.

We illustrate these results by identifying and estimating the magnitude of peer and contextual effects of student outcomes in classrooms, using a data set that contains no information about the social links among the students, and where each student in each class is only observed once.

**1.2. Motivation.** There are many reasons why network links may be mismeasured or unreported. Typical surveys in economics only deal with individuals and not connections, and so provide no link data at all. Other surveys collect data on proximity or similarity of individuals (e.g., geographic location) from which links might be imputed, but any such imputation will likely entail errors, such as not linking people who happen to be friends despite living far apart.

Misclassification of links may also arise because links that are observed in one context may be irrelevant for outcomes under study, e.g., two people who are observed as linked on a social media platform may be connected there for business or political reasons, and have no effect on each other's personal outcomes.

Even in data sets where observed links are directly relevant for observed outcomes, link data may suffer from a variety of reporting or recording errors. For example, many surveys limit the number of links (e.g., the number of friends) one can report, leading to missing links for popular individuals. Studies that measure links within groups such as classrooms or villages may not observe links across these groups, (e.g. friendships across classrooms or villages). Also, in some surveys on networks with undirected links, an individual A could claim to be friends with B, but B does not report being friends with A. This leaves the status of their undirected link uncertain.

Finally, as noted above, existing results that identify social interactions without link data require many repeated observations of the same network, which are often not available either because individuals may only be observed once (or a small number of times), or because the underlying network could change over time.

**1.3. The Model.** Let $y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^K$ denote the outcome and exogenous covariates for individual $i$ respectively. The sample data includes observations of $y_i$ and $X_i$ for $i = 1, ..., n$. The asymptotics are that $n$ goes to infinity. Let $\overline{y}_i$ and $\overline{X}_i$ denote the average outcome and average covariates among all individuals linked with individual $i$.

Consider the model

$$y_i = \alpha + \lambda \overline{y}_i + X_i'\beta + \overline{X}_i'\gamma + \varepsilon_i,$$

where $\varepsilon_i$ is the i.i.d. error term. The structural parameters of interest include the intercept $\alpha \in \mathbb{R}$, the endogenous peer effect $\lambda \in \mathbb{R}$, the vector of individual effects $\beta \in \mathbb{R}^K$, and the vector of contextual effects $\gamma \in \mathbb{R}^K$.

The adjacency matrix, denoted by $G^*$, is an $n$-by-$n$ matrix whose $(i, j)$-th component equals one if $i$ is linked to $j$ and zero otherwise. Constructing $\overline{y}_i$ and $\overline{X}_i$ requires knowing who is linked to individual $i$, because only those people are included in these averages. So construction of $\overline{y}_i$ and $\overline{X}_i$ for every individual $i$ in the sample requires that $G^*$ be observed.

A popular estimator for the structural coefficients $\alpha$, $\lambda$, $\beta$, and $\gamma$ is the instrumental variable (IV) estimator which uses the average covariates for all individuals $j$ who are friends of friends of $i$ as an instrument for $\overline{y}_i$ (see Bramoullé, Djebbari and Fortin (2009)). This estimator assumes the data includes perfect measurement of all the links in $G^*$, not only to construct $\overline{y}_i$ and $\overline{X}_i$, but also to construct the instruments for $\overline{y}_i$. Another possible estimator is quasi-maximum likelihood estimation (QMLE), which is based on parameterizing the distribution of $\varepsilon_i$'s and maximizing the associated likelihood function. This estimator also assumes $G^*$ is correctly observed.

**1.4. Estimation with misclassified links.** Our first set of results characterizes the impact of misclassification of links in the data on the asymptotic properties of IV estimators of $\alpha$, $\lambda$, $\beta$, and $\gamma$. For these results, instead of observing the true adjacency matrix $G^*$, we observe $H^*$, which is a noisy measure of $G^*$. The difference between $H^*$ and $G^*$ is the matrix of measurement errors in the network in our data.

We first investigate the asymptotic properties of the IV estimator when the misclassified links matrix $H^*$ is used instead of the true $G^*$ for constructing $\overline{y}_i$, $\overline{X}_i$ and the instruments for $\overline{y}_i$. As noted above, the results are good news for researchers. If the expected number of misclassified links grows at a rate slower than $\sqrt{n}$, then the IV estimator remains $\sqrt{n}$-consistent and asymptotically normal under regularity assumptions, and the usual formulas for estimating standard errors remain consistent. Therefore, under these conditions researchers can safely ignore the presence of these misclassified links, because both the estimator and its standard errors based on $H^*$ remain valid.[2] We show the same result also holds for parameters that are estimated using QMLE.

These results can be applied to many of the above listed examples of measurement errors in link data. For example, consider the common modeling environment where the data come from many independent small groups, such as schools or villages. We can think of the groups as blocks along the diagonal of a single growing network $G^*$. Models using such data often ignore or assume away links between individuals across villages or schools (i.e., links between one block and another), either for theoretical convenience or because no data are collected on such links. Under the general framework we consider, this is equivalent to misclassifying all links that exist outside the diagonal blocks in $G^*$. Our results show that IV estimators (and QMLE) using such a block-diagonal $H^*$ as a proxy for the true $G^*$ remain consistent, and the usual standard error formulas remains valid, provided $G^*$ is sufficiently sparse outside the diagonal blocks.

**1.5. Identification and Estimation without link data.** Our second set of results consists of a new constructive point identification strategy and associated closed form estimator for the case where the data contain no link information at all, so not even an $H^*$ is observed. For these results we assume observed individuals are in finite-sized groups indexed by $l = 1, .., L$ with $L \to \infty$. In data these groups might correspond to classrooms or villages. This is equivalent to assuming the true unobserved adjacency matrix $G^*$ is block diagonal. Suppose we know what block each

[2] We also find that if the expected number of misclassified links grows at a rate faster than $\sqrt{n}$ but slower than $n$, then the IV estimator is still consistent, but in this case the rate of convergence of the coefficients will be less than $\sqrt{n}$ and the usual standard error formulas no longer apply.

individual belongs to (e.g., what classroom or village), but do not have any information regarding the links within each block. We also consider the extension to allowing additional unobserved sparse links outside these diagonal blocks, as above.

Our identification strategy makes use of the fact that the same unobserved $G^*$ that determines $\overline{y}_i$ also determines $\overline{X}_i$, and information about $\overline{X}_i$ (on average) can be recovered from the reduced form coefficients obtained by regressing each $y_i$ on the $X_j$'s for all individuals $j$ in the same group as $i$ (analogous to indirect least squares estimation of simultaneous systems of equations). Observation of many groups is what allows us to identify these reduced form coefficients.

One attractive feature of this identification strategy is that it is constructive, so the same steps used for identification can be replicated in data to obtain parameter estimates. Another feature is that it does not require us to model how links are formed, e.g., no assumptions are needed regarding how link probabilities are determined.

**1.6. Classroom outcomes in Tennessee elementary schools.** We apply our estimator without link data to estimate the impact of social networks on the test performance of elementary school students from a data set collected in Tennessee, USA. For example, without observing any data on the links between students, we identify the peer effects coefficient $\lambda$, and estimate it to equal 0.85 in small classes and 0.92 in large classes. Both estimates are statistically significant. These estimates are roughly similar to those obtained by other researchers who use a linear-in-means specification for identification with this data, but we find we can reject that model (and other simple models of link formation). 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 size. This is an example of a counterfactual exercise we can perform that would be difficult by other means with this data. We can also test various alternative models of link formation.

The next section is a short literature review. This is followed by our formal model. We then present our results for mismeasured networks, followed by our new identification and estimation method for unobserved networks. We then present some simulation results, followed by our empirical application and conclusions. Proofs and derivations are in the Appendix.

## 2 Literature Review

A standard social interaction model links an individual's outcome to his/her own characteristics, contextual influences from her peers' characteristics, and endogenous effects from her peers' outcomes. The traditional linear-in-means model suffers the "reflection problem" as pointed out by Manski (1993). This issue can be overcome in models with more complicated social interaction structures. Lee (2007) uses conditional maximum likelihood and instrumental variable methods to estimate peer and contextual effects in a spatial autoregressive social interaction model, assuming

links are perfectly observed in the data. Bramoullé, Djebbari and Fortin (2009) and Lin (2010) provide specific conditions on observed network structure in order to identify peer effects in social interaction models.

Given results like these, the model described in the introduction has been widely used to estimate peer effects in a variety of settings. Examples are studies of peers influence on students' academic performance, sport and club activities, and delinquent behaviors (Hauser et al., 2009; Calvó-Armengol et al., 2009; Lin, 2010; Lee et al., 2010; Liu et al., 2014; Boucher et al., 2014; Patacchini and Zenou, 2012). Two key assumptions maintained in these models are that the network structure is exogenously given and is correctly measured in the data.

The issue of potentially misclassified links is acknowledged and discussed in Patacchini and Venanzoni (2014), Liu et al. (2014), and Lin (2015) among others. But these papers do not provide a formal analysis of the asymptotic impact of mismeasured links on the performance of standard estimators. Our results in Section 4.1 fill this void.[3]

There are several publications that investige identification when network links are unobserved. Blume, Brock, Durlauf and Jayaraman (2015) provide identification results in a setting where the network structure is a fixed, unobserved model element that need to be recovered jointly with the social effects. Their results assume that the reduced-form coefficients in front of individual characteristics *for a given, fixed network structure* are already known to researchers. In the setting of cross-sectional data, this essentially requires the latent network structure be identical across a large number of cross-sectional units (e.g., groups such as classes or villages).

De Paula, Rasul and Souza (2018) identify and estimate a linear social network model where the network is completely unobserved. They require a panel data structure where researchers observe outcomes across multiple periods for a single *fixed* latent 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 to be constant over time.[4] With many time periods and aided by some notion of sparsity, they propose a consistent estimator for the social effects.

[3] Referring to potential omission of friends, Patacchini and Venanzoni (2014) say that, "in the large majority of cases (more than 94%), students tend to nominate best friends who are students in the same school and thus are systematically included in the network (and in the neighborhood patterns of social interactions)". Liu et al. (2014) report that "less than 1% of the students in our sample show a list of ten best friends, less than 3% a list of five males and roughly 4% a list of five females. On average, they declare that they have 4.35 friends with a small dispersion around this mean value (standard deviation equal to 1.41), and in the large majority of cases (more than 90%) the nominated best friends are in the same school." Lin (2015) says, "this nomination constraint only affects a small portion of our sample, as less than 10% of the sample have listed five male or female friends. Therefore, this restriction should not have a significant impact on the results." This last speculation is precisely what our first set of results establishes: that consistency of estimates will not be effected if the number of omitted (and hence misclassified) links is sufficiently small.

[4] In a general framework of a single large network, the setup in De Paula, Rasul and Souza (2018) is analogous to an unobserved block diagonal adjacency matrix, with each block defined by a time period *and each being identical to all the others*. Such a setup is motivated by a long panel of observations of the same fixed group of people, e.g., each block corresponds to the same classroom of friends being observed in a different time period.

The assumptions we need to deal with unobserved networks (in the second part of our paper) are motivated by a different data structure, and therefore differ fundamentally from those in De Paula, Rasul and Souza (2018). First, our method allows the unobserved network structure to vary across groups (e.g., classes or villages), and can be applied in a cross-sectional setting where the network varies across groups within a single period. We do not require a panel structure with the network fixed over time. Asymptotics in our case are defined in terms of the number of groups (each of which only needs to be observed once), not time periods.

Second, our identification argument differs qualitatively from De Paula, Rasul and Souza (2018) in that we capitalize on the relationship between the reduced-form impacts of multiple individual characteristics on outcomes. Our identification strategy also entails a mild exclusion restriction, such as the absence of contextual effects for certain characteristics. This is an assumption others have also used in the literature, e.g. Graham and Hahn (2005). We use these assumptions to disentangle the structural social effects from moments of the network structure in the reduced-form coefficients of individual characteristics.

Third, our identification strategy is constructive, and thus leads to a simple two-stage estimator that has a closed form. The estimator is easy to compute, and attains standard consistency and asymptotic normality.

# 3 The Model

Let $y = (y_1, ..., y_n)' \in \mathbb{R}^n$ be a vector of individual outcomes, let $\iota = (1, ..., 1)'$ and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)'$ be $n$-dimensional column vectors, and let $X = (x_1, ..., x_n)'$ be an $n$-by-$K$ matrix that consists of $n$ vectors of exogenous regressors $x_i \in \mathbb{R}^K$. Let $G^*$ be the $n$-by-$n$ adjacency matrix that summarizes the actual, unobserved link structure in the network, with $G^*_{ij} = 1$ if $i$ and $j$ are linked and $G^*_{ij} = 0$ otherwise. Let $G^*_{ii} = 0$ by convention in the literature. Define a row-normalized adjacency matrix $G$ by $G_{ij} = G^*_{ij} / \left(\sum_{j'} G^*_{ij'}\right)$, with $\sum_j$ summing over $j = 1, 2, ..., n$. By construction, each row in $G$ sums up to one. Throughout the paper, we maintain that $\min_i \sum_j G^*_{ij} > 0$ with probability one. This means there are no isolated individuals in the network, or equivalently no rows of zeros in $G^*$. This condition is standard in the literature and it ensures that the row-normalized adjacency matrix $G$ is well-defined.

We assume a linear social network model where outcomes are determined by:

$$y = \alpha\iota + \lambda G y + X\beta + GX\gamma + \varepsilon. \tag{1}$$

To reiterate, $\lambda \in \mathbb{R}$ is a non-zero endogenous peer effect, $\beta \in \mathbb{R}^K$ is a vector of exogenous individual effects, $\gamma \in \mathbb{R}^K$ is a vector of contextual effects, and $\alpha \in \mathbb{R}$ is the structural intercept. We assume that $|\lambda| < 1$.

Our goal is to identify and estimate the social effects $(\lambda, \beta', \gamma')'$ and $\alpha$, using observations of $(y, X)$ but allowing for misclassification or unobservability of $G^*$ and $G$. The data report $(y_i, x_i)_{i=1,2,...,n}$ over a single large network with $n$ individuals as in (1). As discussed in the intro-

duction, this general framework subsumes models where observations in the data are from multiple, unlinked networks. Such models corresponds to the special case where $G^*$ is block-diagonal.

# 4 Misclassified Links

A typical instrumental-variable method like Bramoullé, Djebbari and Fortin (2009) estimates $\lambda, \beta, \gamma$ and $\alpha$ via two-stage least squares (2SLS), using $G^2X$ as instruments for $Gy$ on the right-hand side of (1). This corresponds to using information about friends of friends as instruments. If some network links are misclassified in the sample, then the right hand side vectors $Gy$ and $GX$ will be misspecified, and the instruments $G^2X$ will be measured with errors. This then raises questions about the validity of the IV method when some links may be misclassified. We consider the impact of misclassified links on the consistency and the asymptotic distribution of the IV estimator. In particular, we show in this section that the IV estimators remain consistent for $\lambda, \beta, \gamma$ and $\alpha$, as long as the expected number of misclassified links increases at a rate slower than the sample size $n$. Furthermore, the estimators are root-n asymptotically normal and the conventional standard errors remain valid if this rate of misclassified links increased at a rate slower than $\sqrt{n}$.

Our results provide good news for applied researchers. As long as the number of misclassified links increases at a sufficiently slow rate, the presence of these misclassified links can be completely ignored; standard estimators will be consistent and have the same limiting distribution as if all the links were correctly measured.

## 4.1 Mismeasured instruments in 2SLS

Suppose the sample data does not report the actual adjacency matrix $G^*$, but provides instead a proxy measure $H^*$, whose off-diagonal components $H^*_{ij} \in \{0, 1\}$ are random misclassification of $G^*_{ij}$. Let $H$ be the row-normalization of $H^*$, i.e., $H_{ij} = H^*_{ij} / \left(\sum_{j'} H^*_{ij'}\right)$. Assume $\min_{i \leq n} \sum_j H^*_{ij} > 0$ with probability one so that the row normalization is well-defined.

A feasible IV estimator for $(\alpha, \beta, \gamma, \lambda)$ uses $H^2X$ as instruments for $Hy$. Let $\Delta \equiv H - G$ denote the misclassification error, and write the structural form in (1) as:

$$y = \alpha\iota + \lambda Hy + X\beta + HX\gamma + \widetilde{\varepsilon}, \tag{2}$$

where

$$\widetilde{\varepsilon} \equiv \varepsilon - \lambda\Delta y - \Delta X\gamma.$$

Let $R \equiv (\iota, Hy, X, HX)$ denote an $n$-by-$(2K + 2)$ matrix of explanatory variables in (2), let $V \equiv (\iota, H^2X, X, HX)$ denote an $n$-by-$(3K + 1)$ matrix of instruments. The IV estimator is:

$$(\hat{\alpha}, \hat{\lambda}, \hat{\beta}', \hat{\gamma}')' = \left[R'V(V'V)^{-1}V'R\right]^{-1} R'V(V'V)^{-1}V'y. \tag{3}$$

We first show this estimator is consistent when the order of misclassification error is small in the following sense.

**Assumption 1** $E\left(\sum_i \sum_j \left|H^*_{ij} - G^*_{ij}\right|\right) = O(n^s)$ *for some* $s < 1$.

Assumption 1 requires the expected number of misclassified links to increase at a rate slower than the sample size $n$. This condition holds, for example, if misclassification exists only for a subset of individuals of order $O(n^s)$ with $s < 1$ and the expected number of misclassified links is fixed for each individual in this subset. In contrast, this assumption will not hold if every component of $G^*$ is misclassified independently with some probability $\rho \in (0, 1)$. In this case, the expected number of misclassified links would be $\rho n(n-1)$, which is $O(n^2)$.

We show that the estimation error of the IV estimator in (3) has a stochastic order of $O_p(n^{-1/2} \vee n^{s-1})$. To see the intuition, notice that

$$(\hat{\alpha}, \hat{\lambda}, \hat{\beta}', \hat{\gamma}')' - (\alpha, \lambda, \beta', \gamma')' = \left[\frac{R'V}{n}\left(\frac{V'V}{n}\right)^{-1}\frac{V'R}{n}\right]^{-1}\frac{R'V}{n}\left(\frac{V'V}{n}\right)^{-1}\frac{V'\widetilde{\varepsilon}}{n}. \tag{4}$$

Under regularity conditions in Assumption 2, $(V'R)/n$ and $(V'V)/n$ both converge in probability to constant matrices with rank $(2K+1)$, and the last term on the right-hand side of (4) can be decomposed into

$$\frac{1}{n}V'\widetilde{\varepsilon} = \frac{1}{n}V'\varepsilon - \frac{1}{n}\lambda V'\Delta y - \frac{1}{n}V'\Delta X\gamma. \tag{5}$$

Under exogeneity of $G^*, H^*$, the first term in this decomposition is $O_p(n^{-1/2})$ by the Chebyshev's Inequality. The order of the second and third terms in (5) depends on the order of the misclassification errors and, as show in the appendix, is $O_p(n^{s-1})$. Combining these results, we conclude that the estimation errors in (4) is $O_p(n^{-1/2} \vee n^{s-1})$. Thus the 2SLS estimator using $H^2X$ as an instruments for $Hy$ is consistent.

The rest of this Section formally states this result and sufficient conditions required for it to hold. Define a sequence of random vectors $\{\xi_n\}_{n=1,2,\ldots,\infty}$ to be *bounded* (denoted as "$\xi_n < \infty$") if there exists a finite constant $\bar{\xi}$ such that $\Pr\{\|\xi_n\| \leq \bar{\xi}\} = 1$ for all $n$, where $\|\cdot\|$ denotes the Euclidean norm. Let $\sup_i$ be shorthand for $\sup_{i\in\{1,\ldots,n\}}$. In what follows we suppress the dependence on sample size $n$ in the notation for matrices $G^*$ and $H^*$.

**Assumption 2** *(i)* $\varepsilon$ *is independent from* $(X, G^*, H^*)$*;* $\varepsilon_i$ *is independent across* $i \leq n$*, with* $E(\varepsilon_i) = 0$ *for all* $i$ *and* $\sup_i E(\varepsilon_i^2) < \infty$*. (ii)* $\sup_i \sum_j G^*_{ij}$*,* $\sup_j \sum_i G^*_{ij}$*,* $\sup_i \sum_j H^*_{ij}$*,* $\sup_j \sum_i H^*_{ij}$ *are bounded. (iii)* $\frac{1}{n}V'V$ *and* $\frac{1}{n}V'R$ *converge in probability to constant matrices with rank* $(2K+1)$*, and* $\sup_i E\left(\ |x_i'x_i|\ \mid G, H\right) < \infty$*.*

Part (i) of Assumption 2 states that $X, G^*$ are exogenous, and that $\varepsilon$ is independent of the proxy $H^*$. Part (ii) requires the row and column sums of $G^*, H^*$ to be bounded, which implies the column sums of $G$ and $H$ are bounded. The row sums of $G$ and $H$ are equal to one by construction.

**Proposition 1** *Under Assumptions 1 and 2,*

$$(\hat{\alpha}, \hat{\lambda}, \hat{\beta}', \hat{\gamma}')' - (\alpha, \lambda, \beta', \gamma') = O_p(n^{-1/2} \vee n^{s-1}).$$

With $s < 1$, it then follows from this proposition that the IV estimator $(\hat{\alpha}, \hat{\lambda}, \hat{\beta}', \hat{\gamma}')$ using instruments $H^2X$ is consistent. Furthermore, if $s < 1/2$, the effect of misclassification vanishes fast enough so that it does not affect the root-n rate of convergence or the asymptotic distribution of these IV estimators. In the appendix, we provide similar results for alternative quasi-maximum likelihood estimators (QMLE) which treats $H^*$ as the true adjacency matrix in the likelihood.

### 4.2 Related applications

Our findings in Section 4.1 are applicable to a number of data scenarios that arise naturally in many contexts, including the following.

**1. Missing links across groups**

Often network observations can be partitioned into naturally defined groups of individuals, such as classes or cohorts at schools, or villages in developing countries, or more general neighborhoods within geographic boundaries. In such cases, link data may only be collected or recorded within these groups, but not across groups. For example data collected within schools may not record friendships between individuals who go to different schools. Alternatively, the data may report links across groups, but a researcher may choose to ignore these cross-group links and only consider the links within each group, with the goal of building a tractable econometric model that applies the law of large numbers across the groups. In either situations, one may be concerned about whether ignoring possible links between groups affects the inference of social effects. Our findings in Section 4.1 show that such links can often be safely ignored.

To fix ideas, consider a sample of $n$ individuals who are partitioned into $L$ groups. The actual $n$-by-$n$ adjacency matrix $G^*$ consists of links between individuals from different groups as well as individuals within each group. However, the data only report links within each group. That is, the adjacency matrix reported in the data takes a block-diagonal form $H^* \equiv diag\{H_1^*, ..., H_L^*\}$. Each block $H_l^*$ correctly reports the links within group $l$ (that is, all block-diagonal elements in $G^*$), but all links that exist outside these $L$ diagonal blocks are misclassified as non-existent.

Our results in Section 4.1 suggest that if the order of misclassification is small in the sense of Assumption 1, then one can ignore the misclassification issue and construct an IV estimator using instruments $H^2X$ (with $H$ being a row-normalization of $H^*$ that contains misclassification errors). The IV estimator is still consistent. If in addition the order of misclassification is restricted to be smaller than $\sqrt{n}$, then the estimator converges at root-n rate to a zero mean normal distribution and the usual formula for asymptotic variance remains valid.

**2. Panel data with time-varying links**

Consider a panel data setting similar to De Paula, Rasul and Souza (2018), where the sample contains many realizations of outcomes over a latent network. Suppose, in addition, the sample only reports the links in one of the time periods (e.g., the initial or the final period in the sampling process). It is likely that the true network evolves over time (e.g., individuals may stop being friends, or form new friendships over time). How do such unobserved changes in the network over time affect the inference of social effects when observations are pooled across time?

Our findings show that, if the expected changes in links over time are sufficiently small, then the unobserved changes in the network can be safely ignored. The 2SLS estimator can still be root-n consistent and the usual formula for standard errors can remain valid for inference.

**3. Limits on reported links**

Many friendship surveys place an upper limit on how many friends one can list. As a result, for the subset of individuals who have more friends than the upper limit, some links with friends will be misclassified as non-existent. This source of misclassification satisfies Assumption 1 if the number of such individuals grows at a sufficiently slow rate. The 2SLS coefficient estimates then remain consistent despite this source of error in the measurement of the network. Moreover, the usual standard errors on these coefficients also remain valid as long as the number of such popular individuals grows at a rate slower than root-n.

**4. Undirected graphs with directed data**

In many applications, $G^*$ is assumed to be symmetric (corresponding to an undirected network graph), so if individual $i$ is linked to $j$ then $j$ is assumed to be linked to $i$. However, if $i$ reports being linked with $j$ and $j$ does not report being linked to $i$, then whatever entry is put in positions $i, j$ and $j, i$ risks being mismeasured. As long as the number of such cases does not grow too quickly with $n$, our results show this problem can be safely ignored.

# 5 Unreported Links

In this section we consider the more difficult problem of identifying and estimating the model coefficients $\alpha$, $\beta$, $\gamma$, and $\lambda$ when the network is unobserved, so no $H^*$ matrix is available. Since the vast majority of survey data does not include information on links across individuals, these results have potentially very many applications.

To make such identification possible, we assume the unobserved adjacency matrix $G^*$ is block diagonal (or, exploiting the results of the previous section, is close to block diagonal with only a small number of links between blocks). Suppose we know what block/group each individual belongs to. For example, if the blocks are villages, then we know the village each individual in our sample is from, but we do not observe who is linked with whom within each village.

We assume the unobserved network structure within each block is a draw from some underlying unknown distribution of possible networks. Then, based on reduced form coefficients estimated across groups, we show we can recover the model coefficients along with some features of the distributions of networks across groups. We provide a corresponding closed form consistent estimator of the model coefficients, along with standard root-n asymptotics.

## 5.1 Baseline model with many groups

To introduce the main idea in this section, first consider a baseline model where the sample consists of $L$ independent groups, or networks. Each group involves $n_l$ individuals and has an $n_l \times n_l$ adjacency matrix $G_l$. The adjacency matrices vary across these groups, and are *not* reported in

the data. For now, assume that no links exist between individuals from different groups. (We relax this assumption later in Section 5.8.) This baseline scenario fits in the framework introduced in Section 3, with $G$ consisting of $L$ diagonal blocks $\{G_l\}_{l \leq L}$ and zero entries outside these blocks.

We can rewrite (1) as a sequence of networks/groups indexed by $l = 1, 2, ..., L$:

$$y_l = \alpha\iota + \lambda G_l y_l + X_l\beta + G_l X_l\gamma + \varepsilon_l, \tag{6}$$

where $y_l$ and $\varepsilon_l$ are $n_l \times 1$ vectors, $\iota$ is an $n_l \times 1$ vector of ones, and $X_l$ an $n_l \times K$ matrix.

This model is interesting in its own right, because in practice data sets are often collected from multiple independent groups of individuals. One example is the Add Health data set used extensively in the literature on social networks. In that example, $L$ is the number of school-grades in the sample, and $n_l$ the number of students in each school-grade $l$. In the Add Health data, each $G_l$ is observed, while we will consider the more difficult problem in which each $G_l$ is not observed.

De Paula et al (2018) show joint identification and estimation of the coefficients $\alpha$, $\beta$, $\gamma$, and $\lambda$ in the model of equation (6) where $G_l$ is unobserved, by assuming $G_l = G_1$ for $l = 2, ..., L$, so that $G_l$ is identical for all groups $l$. They envision panel data with many time periods indexed by $l$, in which outcomes are realized repeatedly over the same unobserved network (e.g. scores from multiple tests taken by the same classroom of students over time). Their model is therefore finitely parameterized, because the unknown parameters are the constant coefficients and all elements in a fixed adjacency matrix $G_1$, while the number of available observations goes to infinity as $L \to \infty$.

Across groups $l = 1, ..., L$ in the sample, we instead assume that each unobserved $G_l$ is a random draw from some underlying unknown distribution of adjacency matrices, which are not reported in the data. For example, our data could consist of outcomes and covariates from the members of $L$ different classrooms or villages, each observed only once. We don't make specific assumptions about how the unobserved network in each classroom or village is formed, but instead assume each is an independent draw from some latent distribution of possible networks.

Our method implicitly assumes the definition of groups is observed in the sample. That is, we know to which group $l$ each individual in our sample belongs. This is justified because in practice groups are often defined by publicly observed information. Examples include geographic boundaries such as rural villages in India studied in Banerjee et al (2017) where each $l$ is a village, or registration/enrollment records such as class enrollment in the Add Health data set, where each $l$ is a school-grade.

### 5.2 Identifying Assumptions

We maintain the following formal assumptions regarding the data-generating process. To ease exposition and notation, suppose for now that $X_l$ does not include any group-level variables. This means none of the columns in $X_l$ consists of $n_l$ identical entries. Such group-level variables are easily accommodated by our method; details of how to do so are deferred to Section 5.6.1.

**Assumption 2.1** *(Independent groups)* $(G_l, X_l, \varepsilon_l)$ *are i.i.d. across groups* $l = 1, .., L$.

**Assumption 2.2** *(Exogenous networks)* $E(\varepsilon_l|G_l, X_l) = 0$ for all $l$.

To fix ideas, suppose all groups in data-generating process share the same size $n_l = n$ (we will later relax this by dividing the population into subgroups $s$, and allow both group size and some of the coefficients to vary by $s$). Let $X_{l,ck}$ denote the $k$-th column in $X_l$. That is, $X_{l,ck}$ is an $n \times 1$ vector of the $k$-th regressor for all members in group $l$. Let $w_l \equiv (1, X'_{l,c1}, X'_{l,c2}, ..., X'_{l,cK})'$ denote a $(nK + 1) \times 1$ vector that stacks regressors for all individuals in group $l$.

**Assumption 2.3** *(Independence)* $G_l$ *is independent of* $X_l$ *for each* $l$.[5]

**Assumption 2.4** *(Rank conditions) (i)* $E(w_l w'_l)$ *exists and is non-singular. (ii)* $I - \lambda G_l$ *is invertible with probability one. (iii)* $E(M_l) < \infty$ and $E(M_l G_l) < \infty$, *where* $M_l \equiv (I - \lambda G_l)^{-1}$.

In Assumption 2.4, $I$ is the $n_l \times n_l$ identity matrix. Invertibility of the matrix $M_l$ could require that $|\lambda| < 1$, which is a common assumption. Given Assumption 2.4, we can obtain the reduced form of (6) as

$$y_l = M_l \left(\alpha\iota + X_l\beta + G_l X_l\gamma + \varepsilon_l\right), \, l = 1, ..., L. \tag{7}$$

Our method for identification can be readily generalized to where Assumptions 2.2 and 2.3 hold conditional on additional exogenous regressors instead of unconditionally. We omit that extension in our derivations below to save on notation.

To obtain identification, we will require two additional assumptions. One, given by Assumption 2.5 in the next Section, rules out some pathological cases in which identification fails. The second is an exclusion restriction that will be discussed at length in Section 5.4.

### 5.3 Identification

**Lemma 1** *Under Assumptions 2.1-2.4, the following reduced-form parameters are identified:*

$$\begin{aligned} \mu_k &\equiv E\left(\beta_k M_l + \gamma_k M_l G_l\right) \text{ for } k = 1, ..., K; \\ \mu_0 &\equiv \alpha/(1 - \lambda). \end{aligned}$$

For each characteristic indexed by $k \leq K$, the $(i, j)$th-component in $\mu_k$ is the marginal effect of the $k$-th characteristic of individual $j$ on the mean outcome of individual $i$ in the reduced form (7) under Assumptions 2.1-2.4. We refer to $\mu_k, k \leq K$ as the *reduced-form coefficients* throughout the paper.

The intuition for identification of the reduced-form coefficients is as follows. Let $y_{l,i}$ denote the outcome for student $i$ in group $l$. By construction,

$$E(y_{l,i}|X_l) = \mu_0 + e_i E(M_l) X_l\beta + e_i E(M_l G_l) X_l\gamma, \tag{8}$$

where $e_i$ denotes an $1 \times n$ row unit vector whose $i$-th component is 1. This equation holds because $G_l$, and hence $M_l$, are independent from $X_l$ in Assumption 2.3 and $E(M_l\varepsilon_l|X_l) = E\left[M_l E(\varepsilon_l|X_l, G_l)|X_l\right]$

[5] This condition can be replaced by "$E(G^s|X)$ is mean independent of $X$ for $s = 1, 2, .., \infty$".

$= 0$ in Assumption 2.2. The equality in (8) also uses the fact that the block-diagonality and row-normalization of $G$ imply

$$\alpha M_l \iota = \alpha \left[\sum_{s=0}^{\infty} (\lambda G_l)^s\right] \iota = \mu_0 \iota,\ l = 1, ..., L.$$

The right-hand side of (8) is linear in all $nK$ components in $X_l$.

For each $i \leq n$, regressing $(y_{l,i})_{l=1,...,L}$ on $(X_l)_{l=1,...,L}$ leads to consistent estimators for the intercept $\mu_0$ and $nK$ slope coefficients in front of all components in $X_l$. (Consistency is defined as $L \to \infty$.) The rank condition in Assumption 2.4 guarantees the identification of these reduced-form coefficients. In such a regression of $y_{l,i}$ on $X_l$, the slope coefficient for the $k$-th regressor of individual $j$ in the group is $\beta_k \left[e_i E(M_l) e_j'\right] + \gamma_k \left[e_i E\left(M_l G_l\right) e_j'\right]$, where, for a generic $n \times n$ matrix $Q$, the product $e_i Q e_j'$ returns the $(i,j)$-th component in $Q$. Thus by regressing $y_{l,i}$ on $X_l$ for each $i = 1, ..., n$, we obtain consistent estimators for all $n^2 K$ slope coefficients. Rearranging and packing these estimators into matrices leads to consistent estimators for the $K$ matrices of reduced-form coefficients $\mu_k$, for $k = 1, ..., K$.

Next, we relate these reduced-form coefficients $\mu_0, \mu_k$ to structural parameters $\alpha, \lambda, \beta, \gamma$. To do so, we require some mild conditions that rules out pathological cases.

**Assumption 2.5** *(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 $c \in \mathbb{R}$.*

Part (i) of this assumption rules out the pathological case where two of $K$ regressors have proportional contextual and peer effects. Part (i) holds, for example, if $\gamma_K = 0$ (one of the regressors has no contextual effect) while $\beta_K$ and $\beta_k, \gamma_k$ are all nonzero. Part (ii) 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. This condition is testable in principle, using the sample data $(y_l, X_l)_{l=1,2,...,L}$.

The next lemma establishes a simple linear relation between the reduced-form coefficients $\mu_k$ and the structural parameters $(\lambda, \beta, \gamma)$. This relation provides the foundation for our constructive identification strategy and estimation method.

**Lemma 2** *Suppose Assumptions 2.1-2.5 hold. Then for each $k = 1, ..., K$ the equation*

$$a_k \mu_k + b_k \mu_K = I \tag{9}$$

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

$$\begin{pmatrix} a_k \\ b_k \end{pmatrix} = \begin{pmatrix} \beta_k & \beta_K \\ \gamma_k & \gamma_K \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ -\lambda \end{pmatrix}. \tag{10}$$

*Proof of Lemma 2.* For any $k = 1, ..., K$, the inverted matrix on the right-hand side of (10) has full rank under condition (i) in Assumption 2.5. 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,

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

Next, we need to show that for each $k$, $(a_k, b_k)$ as defined in (10) is the *unique* solution for (9). In other words, we need to show there exists no $(\tilde{a}_k, \tilde{b}_k) \neq (a_k, b_k)$ such that

$$(\tilde{a}_k - a_k)\mu_k + (\tilde{b}_k - b_k)\mu_K = \mathbf{0}. \tag{11}$$

Consider three mutually exclusive cases.

Case 1: $\tilde{a}_k = a_k$, $\tilde{b}_k \neq b_k$. Then (11) requires $\mu_K = \mathbf{0}$.

Case 2: $\tilde{a}_k \neq a_k$, $\tilde{b}_k = b_k$. Then (11) requires $\mu_k = \mathbf{0}$. This in turn implies $\mu_K$ must 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 $\mu_k = -\frac{\tilde{b}_k - b_k}{\tilde{a}_k - a_k}\mu_K$, which is a scalar multiple of $\mu_K$. Again, this implies that in order for (9) to hold for $(\tilde{a}_k, \tilde{b}_k)$, $\mu_K$ must be a scalar multiple of $I$. In each of these cases, the implication of (11) contradicts part (ii) of Assumption 2.5. □

The reduced-form coefficients $\mu_0$ and $\mu_k$ are identified by Lemma 1. Therefore, for each $k < K$, $(a_k, b_k)$ can be recovered as the unique solution of (9). Lemma 2 implies that these constants $(a_k, b_k)_{k<K}$ are related to $(\lambda, \beta, \gamma)$ in a system of linear equations:

$$\begin{pmatrix} \beta_k & \beta_K \\ \gamma_k & \gamma_K \end{pmatrix} \begin{pmatrix} a_k \\ b_k \end{pmatrix} = \begin{pmatrix} 1 \\ -\lambda \end{pmatrix} \text{ for } k = 1, ..., K-1. \tag{12}$$

Besides, by the row normalization of $G$,

$$m_k \equiv (\iota'\mu_k\iota)/n = \frac{\beta_k + \gamma_k}{1 - \lambda} \text{ for } k = 1, ..., K, \tag{13}$$

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

Combining (12) and (13), we have a linear system of $2(K-1) + K$ equations for the $2K + 1$ unknown parameters in $\theta \equiv (\lambda, \beta', \gamma')'$ with $\beta \equiv (\beta_1, \beta_2, ..., \beta_K)'$ and $\gamma \equiv (\gamma_1, \gamma_2, ..., \gamma_K)'$. The rank of the coefficient matrix in such a linear system is at most $2K - 1$ because $a_k m_k + b_k m_K = 1$ for all $k < K$ by construction. For example, consider the case with $K = 3$. Then the linear system is:

$$\begin{pmatrix} 0 & a_1 & 0 & b_1 & 0 & 0 & 0 \\ 0 & 0 & a_2 & b_2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & a_1 & 0 & b_1 \\ 1 & 0 & 0 & 0 & 0 & a_2 & b_2 \\ m_1 & 1 & 0 & 0 & 1 & 0 & 0 \\ m_2 & 0 & 1 & 0 & 0 & 1 & 0 \\ m_3 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ m_1 \\ m_2 \\ m_3 \end{pmatrix}. \tag{14}$$

The rank of this coefficient matrix in (14) is bounded above by 5.[6] For general cases with $K > 3$, we write (14) as

$$\underbrace{\begin{pmatrix} 0_{(K-1)\times 1} & A & 0_{(K-1)\times K} \\ \iota_{(K-1)\times 1} & 0_{(K-1)\times K} & A \\ m & I & I \end{pmatrix}}_{\boldsymbol{\pi}} \underbrace{\begin{pmatrix} \lambda \\ \beta \\ \gamma \end{pmatrix}}_{\theta} = \underbrace{\begin{pmatrix} \iota_{(K-1)\times 1} \\ 0_{(K-1)\times 1} \\ m \end{pmatrix}}_{d}, \tag{15}$$

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

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

The rank of the coefficient matrix on the left side of (15) is at most $2K-1$. It is strictly less than $2K-1$ only for pathological values of parameters in the data-generating process. Moreover, the coefficient matrix, and hence its rank, is identified and can be inferred from sample data.

It will be useful for our empirical application (and for the exclusion restrictions to be introduced in the next Section) to generalize our identification result, to the case where the population can be partitioned into $s = 1, 2, ..., S$ subsets based on observable information. Each subset pertains to a distinctive environment with a potentially different vector of structural parameters $\theta^{(s)} \equiv (\lambda^{(s)}, \beta^{(s)}, \gamma^{(s)}) \in \mathbb{R}^{2K+1}$. For example, an environment could be defined by the size of group. That is, all groups with size $n = n^{(s)}$ in the sample are considered as independent draws from a sub-population indexed by $s$. We can allow the structural parameters to potentially vary across groups with different sizes.

By repeating the argument above for each environment indexed by $s$, we can construct $S$ linear systems

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

with $\pi^{(s)}, d^{(s)}$ defined as in (15) for each environment $s$. We then stack these $S$ systems to get

$$\boldsymbol{\Pi\theta} = \mathbf{D},$$

where $\boldsymbol{\theta}$ and $\mathbf{D}$ are column vectors that stack $\theta^{(s)}$ and $d^{(s)}$ respectively for $s \leq S$; and $\boldsymbol{\Pi}$ is a block-diagonal matrix with diagonal blocks being $\pi^{(s)}, s \leq S$. In addition we append to this system the additional exclusion restrictions $\mathbf{R}\boldsymbol{\theta} = c$ where $\mathbf{R}, c$ are known a priori (see the next Section for examples).

**Theorem 1** *$\boldsymbol{\theta}$ is identified if $[\boldsymbol{\Pi}; \mathbf{R}]$ has full rank.*

This theorem follows immediately from Lemma 1 and 2. In the next Section, we discuss types of exclusion restrictions a researcher can use for constructing $R$.

[6] 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.

### 5.4 Exclusion restrictions

To obtain the full rank condition needed for identification, we require a (vector valued) exclusion restriction of the form $\mathbf{R}\boldsymbol{\theta} = c$. The dimension of $c$, corresponding to the number of required restrictions on the coefficients $\theta$, depends on both $K$, the number of regressors in $X$, and on the number of environments $S$. For example, with $S = 1$, two linear restrictions on $\theta$ will generally suffice for identifying $\theta$. We must be careful to ensure that the chosen restrictions do actually make the augmented coefficient matrix $[\Pi; R]$ attain full rank $2K + 1$ given the structure of $\Pi$, analogous to the difference between the order condition and the rank condition in standard linear regression identification. We provide examples below.

There are two types of exclusion restrictions one can consider. The *first type* of exclusion restriction specifies that a given regressor in $X_k$ has either no contextual effect or no individual effect, i.e., an element of $\beta$ or of $\gamma$ is assumed to equal zero. Graham and Hahn (2005) use such an exclusion restriction to identify a linear-in-means social interaction model, which is the special case of social networks where all group members are linked with equal weights.

In the above example where $K = 3$ and $S = 1$ it could suffice to assume that one regressor $X_k$ has no contextual effect ($\gamma_k^{(1)} = 0$) but non-zero individual effects ($\beta_k^{(1)} \neq 0$) while another regressor $X_{k'}$ has no individual effects ($\beta_{k'}^{(1)} = 0$) but non-zero contextual effects ($\gamma_{k'}^{(1)} \neq 0$). (We need $\beta_k, \gamma_{k'}$ to be nonzero so that $(a_k, b_k, a_{k'}, b_{k'})$ are well-defined.) For general cases with $K > 3$ and $S = 1$, the matrix $[\Pi; R]$ has full rank generically when $R$ is defined by the exclusion restrictions that there exist $k$, $k' < K$ with $\gamma_k = 0$, $\beta_{k'} = 0$ and $\beta_k, \gamma_{k'}$ being nonzero. However, restricting two regressors to both have nonzero individual effects but no contextual effects would not suffice to make $[\mathbf{\Pi}; \mathbf{R}]$ have full rank.

The *second type* of exclusion restriction exploits variation in environments $s$ as described in the previous Section, along with structural coefficient restrictions across environments. To fix ideas, suppose that the data contains just two different environments, so $S = 2$. For example, these two environments could correspond to two different group sizes. Analogous restrictions can immediately be constructed for $S > 2$. Let $n^{(s)}$ denote the size of group $s$, for $s = 1, 2$. Suppose further that peer effects $\lambda$ vary with the group size whereas individual effects $\beta$ and contextual effects $\gamma$ do not. In this case, a linear system like that in (15) can be constructed by including two unknown group-size-specific peer effects $\lambda^{(1)} \neq \lambda^{(2)}$, stacking the two linear systems (15) for different group sizes and appending it with any additional exclusion restrictions of the first type we may have

available.[7] That is,

$$\begin{pmatrix} 0 & 0 & A^{(1)} & 0 \\ \iota & 0 & 0 & A^{(1)} \\ m^{(1)} & 0 & I & I \\ 0 & 0 & A^{(2)} & 0 \\ 0 & \iota & 0 & A^{(2)} \\ 0 & m^{(2)} & I & I \\ & r & & \end{pmatrix} \begin{pmatrix} \lambda^{(1)} \\ \lambda^{(2)} \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} \iota \\ 0 \\ m^{(1)} \\ \iota \\ 0 \\ m^{(2)} \\ 0 \end{pmatrix} \tag{16}$$

where $\left(m^{(s)}, A^{(s)}\right)$ for $s = 1, 2$ are constructed as in (15), using reduced-form coefficients from regressions that only use groups with $n^{(s)}$ members respectively, and $r$ consists of row vectors that summarize additional exclusion restrictions of the first type. The coefficient matrix on the left of of (16) and the vector of constants on the right of this equation are both identified. Therefore, $\lambda^{(1)}, \lambda^{(2)}, \beta, \gamma$ are jointly identified, provided the coefficient matrix on the left side of (16) has full rank $2K + 2$. In this case, a single exclusion restriction of the first type, such as a zero contextual effect for a single regressors, would suffice generically.

This exclusion example assumed that $\lambda^{(1)} \neq \lambda^{(2)}$, which by Lemma 2 means that $(a_k^{(s)}, b_k^{(s)}, m_k^{(s)})$ will differ across $s = 1, 2$. If we instead had $\lambda^{(1)} = \lambda^{(2)}$, then the linear system in (16) (when $r$ only reflects a single exclusion restriction of the first type) would not have sufficient rank for identification, and an additional restriction of the first type would be needed for identification. The question of whether $\lambda^{(1)} \neq \lambda^{(2)}$ can be tested. In particular, given the above assumption that $\beta$ and $\gamma$ are the same across the group sizes, the reduced-form parameter $m_k^{(s)}$ will vary by group size $s$ if and only if $\lambda^{(1)} \neq \lambda^{(2)}$.

The assumption that $\beta$ and $\gamma$ do not vary by group size can be relaxed. For example, if the individual effects $\beta$ are the same across groups but contextual effects vary, so $\gamma^{(1)} \neq \gamma^{(2)}$, then the full rank condition require for identification can still hold by assuming that one of the regressors has no contextual effect regardless of group sizes.

For our empirical application in Section 7, we analyze students' test scores. There we divide classes into two sizes ($s = 1$ for small and $s = 2$ for large classes), and impose an exclusion of this second type that $\lambda$ varies by class size while $\beta$ and $\gamma$ do not. We then need one additional exclusion of the first type. 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 $\gamma$ corresponding to days of absence is zero.

[7] Stacking two linear systems alone by construction does not provide sufficient rank in the coefficient matrix in the linear system. At least one additional exclusion restriction of the first type is needed for point identification of the structural parameters.

### 5.5 Closed-form estimator

Here we describe a closed form estimator for $\theta$. The estimator is based on constructing 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 recover the structural coefficients from those reduced form estimates.

First consider the case of a single environment, so $S = 1$ and the only exclusion restrictions are of the first type, $\mathbf{R}\theta = c$. The extension to the second type of exclusion restrictions based on multiple environments is summarized at the end of this Section.

*Step One:* Linearly regress $y_l$ on $X_l$ to get estimated reduced form coefficients $\hat{\mu}_0 \in \mathbb{R}$ and $\hat{\mu}_k \in \mathbb{R}^{n\times n}$ for all $k$, using the simultaneous equations in (8). Let

$$\hat{m}_k \equiv (\iota'\hat{\mu}_k\iota)/n \text{ for } k = 1, 2, ..., K.$$

Note that at this stage one could test the condition of non-trivial marginal effects required by part (ii) of Assumption 2.5, using these estimates and standard errors.

*Step Two:* For each $k < K$, estimate the solution of (9), denoted $(\hat{a}_k, \hat{b}_k)$, using the extremum estimator

$$(\hat{a}_k, \hat{b}_k) \equiv \arg\min_{a_k, b_k \in \mathbb{R}} \textstyle\sum_{i,j} \left[e_i(a_k\hat{\mu}_k + b_k\hat{\mu}_K - I)e_j'\right]^2 \text{ for } k = 1, 2, ..., K.$$

This is not by itself closed-form and so may entail a numerical search. However, one can in closed-form use implications of (9) to construct a smaller linear system that can be solved by matrix inversion for $(a_k, b_k)$. These linear system equalities include that diagonal components in $a_k\hat{\mu}_k + b_k\hat{\mu}_K$ sum to $n$ while the off-diagonal ones need to add up to 0. These linear system based estimates could be used as consistent starting values for the above extremum estimator.[8]

*Step Three:* Given the estimates from Step Two, the closed-form estimator of the structural parameters $\hat{\theta} \equiv (\hat{\lambda},\, \hat{\beta}_1,\, ...,\, \hat{\beta}_K,\, \hat{\gamma}_1,\, ...,\, \hat{\gamma}_K)'$ is defined as:

$$\hat{\theta} \equiv \hat{\Psi}^{-1}\hat{v},$$

where $\hat{\Psi}$ is a coefficient matrix formed by stacking the linear systems in (12) and (13) with the additional equations derived from exclusion restrictions, and removing redundant rows to attain linear independence. For example, in the case with $K = 3$ above:

$$\hat{\Psi} \equiv \begin{pmatrix} 0 & \hat{a}_1 & 0 & \hat{b}_1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & \hat{a}_1 & 0 & \hat{b}_1 \\ 0 & 0 & \hat{a}_2 & \hat{b}_2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & \hat{a}_2 & \hat{b}_2 \\ \hat{m}_3 & 0 & 0 & 1 & 0 & 0 & 1 \\ & & & R & & & \end{pmatrix} \text{ and } \hat{v} \equiv \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ \hat{m}_3 \\ c \end{pmatrix}.$$

[8] One can also exploit the fact that $(a_k, b_k)$ is over-identified in (9) to construct closed form estimates $(\hat{a}_k, \hat{b}_k)$ that are a (possibly weighted) average of $\binom{n\times n}{2}$ closed-form estimates, each of which uses two of the total of $n \times n$ equalities in (9).

and $R\theta = c$ represents the additional equalities due to exclusion restrictions (such as zero contextual or direct effects by certain regressors).

These steps describe the estimator for a single environment, where $S = 1$. For multiple environments, steps 1 and 2 are first implemented separately for each environment. Then for step 3, stack the estimated $\pi^{(s)}$ and $R$ matrices and the $d^{(s)}$ and $c$ vectors, as described in the previous Sections (and again removing redundant rows) to obtain $\hat{\Psi}$ and $\hat{v}$, giving $\hat{\theta} \equiv \hat{\Psi}^{-1}\hat{v}$.

## 5.6 Details in implementation

### 5.6.1 Group-level variables

The method we described in Sections 5.3, 5.4 and 5.5 immediately extends to accommodates group-level variables. Suppose each group $l$ has a row vector of group-level characteristic $z_l \in R^P$ shared by all group members, such as attributes of the teacher when each group is an elementary school class. The structural form is

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

with $\delta \in R^P$ being a column vector of coefficients that reflect the impact of group characteristics on individual outcomes. One could interpret $\delta$ as a source of "correlated effects". Let Assumption 2.1, 2.2 and 2.3 hold with $X_l$ replaced by $(X_l, z_l)$, and let part (ii) of Assumption 2.4 hold with $w_l \equiv (1, z_l, X'_{l,c1}, X'_{l,c2}, ..., X'_{l,cK})$. The reduced-form is now

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

A regression therefore identifies the reduced-form coefficients for $z_l$, denoted $\nu_p \equiv \delta_p/(1-\lambda)$ for all $p = 1, ..., P$, in addition to $\mu_0$ and $(\mu_k)_{k\leq K}$ as defined above. This uses an implication of the row normalization $(I - \lambda G_l)^{-1}\iota = \frac{1}{1-\lambda}\iota$. Following the same argument as in Section 5.3 and 5.4, one can identify $\lambda, \beta, \gamma, \alpha$ from $\mu_0, (\mu_k)_{k\leq K}$ alone, using appropriate exclusion restrictions. It then follows that $\delta$ is identified from the reduced-form coefficients of $z_l$.[9] For estimation, use $\hat{\delta} = \hat{\nu}(1-\hat{\lambda})$, where $\hat{\lambda}$ is the peer effect estimates in Section 5.5, and $\hat{\nu}$ the OLS estimates for slope coefficients of $z_l$ in the reduced-form regression in (17).

### 5.6.2 Dimension reduction

In the first step regressions of $(y_{l,i})_{l\leq L}$ on $(X_l)_{l\leq L}$ for each $i \leq n$, we need the number of observations (groups) $L$ in the sample to be large relative to the dimension of regressors $nK$ (with $n$ being the number of individuals in a group and $K$ being the dimension of each individual's characteristics). In practice, it is possible that $L$ is not large relative to $nK$. In these cases it is still possible to proceed, by adding some assumptions that restrict the correlations among the characteristics of individuals within groups.

[9] If $\delta$ does not vary across environments indexed by $s = 1, 2$, then the reduced-form coefficients for $z_l$ would vary across environments if and only if $\lambda^{(1)} \neq \lambda^{(2)}$. This provides us with yet another way to test the null $\lambda^{(1)} = \lambda^{(2)}$.

Suppose for each individual $i$ the vector of characteristics $x_{l,i} \in \mathbb{R}^K$ is uncorrelated with those of other group members $(x_{l,j})_{j \neq i}$. Then we may 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 $\bar{y}_i \equiv \frac{1}{L} \sum_{l' \leq L} y_{l',i}$, $\bar{x}_i \equiv \frac{1}{L} \sum_{l' \leq L} x_{l',i}$, and run $n$ lower-dimension regressions with $K$ regressors each. Specifically, regress $(\Delta y_{l,i})_{l \leq L}$ on $(\Delta x_{l,j})_{l \leq L}$ separately for each $j = 1, ..., n$. Next, repeat these steps with dependent variables being $(\Delta y_{l,j})$ for all other $j \neq i$. This leads to $n \times n \times K$ consistent estimators for the reduced-form coefficients in $\mu_k, k \leq K$. With $\mu_k$ recovered for all $k \leq K$, one can proceed and estimate the social effects as above.

If individual characteristics are correlated across group members, then we could instead use a partitioned regression approach to estimate $\mu_k, k \leq K$ via sequential, lower dimension regressions, imposing restrictions on the correlations of subsets of characteristics. Specifically, we could partition the vector of characteristics into $K_1 \cup K_2 = \{1, ..., K\}$, and let $(X_{l,1}, X_{l,2}) = X_l$ denote the corresponding partition of the matrix of regressors. For each $i$, regress $\{y_{l,i}\}_{l \leq L}$ and $\{X_{l,2}\}_{l \leq L}$ respectively on $\{X_{l,1}\}_{l \leq L}$ and get the residuals. Next, regress the residuals from $\{y_{l,i}\}_{l \leq L}$ on those from $\{X_{l,2}\}_{l \leq L}$ to consistently estimate the reduced-form slope coefficients for characteristics in $K_2$. Then plug these estimates into the original regression equation, and run a lower-dimension regression on $\{X_{l,1}\}_{l \leq L}$ to estimate the remaining reduced-form slope coefficients in $K_1$. If $K$ is too large, we could partition the vector of characteristics into more subvectors and apply the partitioned regression algorithm iteratively to estimate $\mu_k, k \leq K$.

### 5.6.3 Individual labels

Following convention in the literature, our method requires that the individual labels $i = 1, ..., n$ in each group be such that the random vectors/matrices $(X_l, G_l, \varepsilon_l)$ across the groups are drawn from the same joint distribution. This construct is also implicit in earlier papers which identify parameters in social network models from reduced-form coefficients $\mu_k$ (e.g., Bramoullé et al 2009 and Blume et al 2015).[10]

If the actual joint distribution of $(X_l, G_l, \varepsilon_l)$ in the data-generating process is exchangeable in the individual indices within each group, then the labeling of individuals has no impact on the asymptotic properties of the estimator, and individuals within each group can just be ordered randomly from 1 to $n$. If this exchangeability further holds across individuals in groups of different sizes, then the joint distribution of $(X_l, G_l, \varepsilon_l)$ will also be the same across groups of different sizes. This would then provide a source of over-identification in our method which we can use to improve the efficiency in estimation.

In practice, we may label (i.e. order) individuals based on some observed individual characteristic (e.g., date of birth, and hence exact age, in classroom data), in which case the assumed exchangeability would only be desirable conditional on that characteristic. In our empirical work,

[10] De Paula et al (2018) consider a more restricted environment where both the individual labels and the network links themselves are the same for every group $l$. In their case the assumption is that $l$ indexes time periods, and it is the same group that is observed very many times.

we compare estimates based on random labeling versus those based on date of birth, to assess the sensitivity of the estimates to the assumed labeling.

### 5.6.4 Variation in group sizes

As described in previous sections, our estimator can handle variation in group size $n_l$ by making each group size correspond to a different environment $s$ (recall that by definition, the reduced-form coefficients $\mu_k$ depend on the group size). However, in some samples there may not be enough observations of groups of each size to implement this estimator. We therefore propose two approaches for resolving such data deficiency, by pooling observations of groups with different sizes. One requires some uncorrelated assumptions, while the other imposes restrictions on the coefficient estimates across groups of different sizes.

The first approach exploits the dimension-reduction method introduced in Section 5.6.2. To fix ideas, first suppose the individual characteristics $x_{l,i} \in \mathbb{R}^K$ are *uncorrelated* across group members. Then, as explained in Section 5.6.2, one can estimate the reduced-form coefficients for each $i \leq n$ via a sequence of $n$ lower-dimension regressions, each involving $K$ explanatory variables only. In this case, one can account for variation in group sizes in each of these $n$ lower-dimension regressions by including dummy variables for group sizes and interacting them with the slope coefficients. This dimension reduction method can still be used if individual characteristics are correlated across group members, but in this case one would apply the partitioned regressions described in Section 5.6.2 to estimate reduced-form coefficients, again including group size dummies (and their interactions with slope coefficients) to control for variation in group sizes in each of the lower-dimension regressions.

It is possible that the number of observations of each group size is too small even for this first approach, or that the required uncorrelatedness assumptions are implausibly strong in a given application. We therefore also propose a second approach that can work even when the sample contains very few observations of some group sizes. However, a limitation of this second approach is that it requires the structural parameters $\lambda, \beta, \gamma, \alpha$ to be the same for all the group sizes being pooled. This second approach takes smaller groups, and augments them with additional simulated "pseudo-individuals" to artificially increase their size, and thereby make all groups being pooled the same size. The resulting pooled regressions then consistently estimate a weighted average of reduced-form coefficient matrices for groups of different sizes. Details are explained in Appendix A4.

With either of these approaches, one can define different environments $s$ corresponding to different ranges of group sizes. This then only requires pooling groups of relatively similar sizes. For example, in our empirical application where groups are student's classes, we define two environments defined as "small classes" and "large classes," and within each of these environments we combine (using the second approach above) a range of small class sizes and large class sizes, respectively.

### 5.7 Extension: endogenous networks

In practice, the formation of links between individuals in a network may depend on some individual demographic characteristics reported in the data. In this Section we discuss how to generalize our estimators to deal with this dependence.

For simplicity in exposition, let groups have identical sizes $n_l = n$. Suppose individual characteristics can be partitioned into $X_l = (X_l^a, X_l^e)$, with $X_l^e$ being an $n \times K_e$ matrix of excluded individual characteristics, i.e., covariates that may affect outcomes but do not affect link formation. Let $X_l^a$ be the remaining an $n$-by-$K_a$ matrix of individual characteristics that may affect individuals' outcomes, link formation decisions, or both. By construction, $K_a + K_e = K$. To illustrate, in our empirical application below, we let $X_l^e$ be students' days of absence from school and test scores from previous years, assuming that friendships are independent of test scores conditional on demographics such as proximity of age.

Our method above can then be applied after conditioning on $X_l^a$. Suppose the unknown network formation is given by $G_l = \zeta(X_l^a, u_l)$, which does not depend on excluded regressors in $X_l^e$. 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) $\varepsilon_l$ 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 $X_l^a$. Note these conditions allow the unobserved errors $\varepsilon_l$ and $u_l$ to be correlated conditional on $X_l^a$. Under these assumptions, $E(M_l|X_l)$ and $E(M_l G_l|X_l)$ only depend on $X_l^a$, 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 a slight abuse of notation, let $K_a$ and $K_e$ also denote the set of indices for characteristics in $X_l^a$, $X_l^e$ respectively, so, $K_a$ and $K_e$ partitions $\{1, 2, ..., K\}$. 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 now identify and estimate 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 $X_l^a$, using a reduced-form regression of $y_l$ on $X_l^e$ conditional on $X_l^a$. Then, for all $k \in K_e$, identify $\lambda, \beta_k, \gamma_k$ from $\mu_k(X_l^a)$, using the methods in Section 5.3 and 5.4. Finally, as before, we can 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 $\beta_k, \gamma_k$, $k \in K_e$ as identified in the previous step.[11].

[11] If we wanted to also recover the remaining model elements $\phi(\cdot)$ and $\mu_k(\cdot)$ for $k \in K_a$ from $\psi(\cdot)$, then doing so is possible, but would require additional functional form assumptions, e.g., index sufficiency in $\phi(\cdot)$ and $\mu_k(\cdot)$ for $k \in K_a$.

### 5.8 Extension: links across groups

In this section we bring together the two parts of this paper. Our model and estimator with an unobserved adjacency matrix assumes that there are no links between groups, i.e., no links between the blocks in the $(\sum_l n_l)$-by-$(\sum_l n_l)$ overall adjacency matrix $G^*$. It therefore treats $G^*$ as block-diagonal. However, our earlier result on estimation with mismeasured links in Section 4.1 showed that, when only links within blocks are observed, if the number of unobserved links between blocks is sufficiently small, then the IV estimator of linear coefficients remains consistent. This was the example in Section 4.2 of unobserved links between groups (blocks) in a near-block-diagonal $G^*$.

We here show that a similar analysis can be applied if there exist links between groups in our estimator with an *unobserved* adjacency matrix, provided that, as before, the number of such links is sufficiently small. In our empirical application, this corresponds to a relatively small number of students having links to students in other classes. By assuming that $G^*$ is block diagonal and hence that such links don't exist, we are introducing measurement error in the adjacency matrix.

This scenario is more challenging than in Section 4.2, because now all elements in $G^*$, including those inside the diagonal blocks, are not reported in the data. Our solution takes two steps. In the first step, we show that when links outside the diagonal blocks are sparse (in the sense of Assumption 1), then the reduced-form coefficients remain consistently estimable. Given this result, the second step is then to apply our identification method from Section 5.1 to recover the model coefficients from the reduced-form coefficients.

For simplicity, consider the case where the data contains $L$ groups with equal size $n$, so that the total number of individuals on the network is $nL$. Let $G_l^*$ denote an $n \times n$ adjacency matrix *within* the group $l$; and let $H^* \equiv diag\{G_1^*, ..., G_L^*\}$ denote an $nL \times nL$ block-diagonal matrix. Clearly, $H^*$ differs from $G^*$ whenever there are non-zero elements outside the diagonal blocks in $G^*$. Let $H$ be the row normalization of $H^*$. That is, $H \equiv diag\{H_1, ..., H_L\}$, with each $H_l$ being a row normalization of $G_l^*$. It is worth emphasizing that we are here considering a scenario where neither $G^*$ nor $H^*$ is observed.

The expected individual outcome is

$$E(y|X) = E[(I - \lambda H)^{-1}H]X\gamma + E[(I - \lambda H)^{-1}](X\beta + \alpha) + \widetilde{\eta}, \tag{19}$$

where the $nL \times 1$ vector $\widetilde{\eta}$ absorbs the errors resulted from replacing the actual adjacency $G$ with its block-diagonal approximation $H$. Notice that $(I - \lambda H)^{-1}$ and $(I - \lambda H)^{-1}H$ are both geometric series of a block-diagonal matrix $H$, and are also block-diagonal themselves. Consequently, we can write (19) as a system of group-specific equations:

$$E(y_l|X) = \mu_0\iota + \textstyle\sum_{k=1}^{K} \widetilde{\mu}_k X_{l,ck} + \widetilde{\eta}_l \text{ for } l = 1, 2, ..., L, \tag{20}$$

with $\widetilde{\mu}_k \equiv E[(I - \lambda H_l)^{-1}(\beta_k I + \gamma_k H_l)]$ being an $n \times n$ matrix of coefficients,[12] and $\widetilde{\eta}_l$ a group-specific

[12] We assume $G_l^*$ are drawn from the same marginal distribution in the data-generating process. Therefore $\tilde{\mu}_k$ is identical across groups and not indexed by $l = 1, 2, ..., L$.

$n \times 1$ subvector in $\widetilde{\eta} \equiv (\widetilde{\eta}_1', ..., \widetilde{\eta}_L')'$.[13] Let $\widetilde{\mu}_{k,ri}$ denote the $i$-th row of $\widetilde{\mu}_k$ for $i = 1, ..., n$. Let $y_{l,i}$ be the outcome for individual $i$ in group $l$. Then

$$E(y_{l,i}|X) = w_l' \widetilde{\Phi}_i + \widetilde{\eta}_{l,i},$$

where $\widetilde{\Phi}_i \equiv (\mu_0, \widetilde{\mu}_{1,ri}, ..., \widetilde{\mu}_{K,ri})'$. We estimate $\widetilde{\Phi}_i$, $i = 1, ..., n$ by regressing individual outcomes on all individual characteristics in each group:

$$\widehat{\Phi}_i \equiv \left(\sum_l w_l w_l'\right)^{-1} \left(\sum_l w_l y_{l,i}\right).$$

In the appendix we show that if the links outside the diagonal blocks in $G^*$ are sparse in the sense of Assumption 1, then the impact of misclassifying components outside the diagonal blocks, as in $\widetilde{\eta}_{l,i}$, vanishes as the network size increases ($L \to \infty$). With $\widetilde{\mu}_k$ consistently estimated, we can then apply the method from Section 5.1 to identify and infer social effects.

# 6 Simulations

## 6.1 IV and QMLE with misclassified links in a fixed sample

In this section we study the impact of misclassification rates on the performance of standard IV and QMLE estimators in a fixed sample. We use an existing data set from an empirical application in Lin and Lee (2010). We treat the adjacency matrix reported in that sample as the correct measure of the network. We then artificially introduce misclassification errors into this matrix, and evaluate how close the resulting estimates are to those based on the original adjacency matrix.

Lin and Lee (2010) model teenage pregnancy rates, using the model

$$Teen_i = \lambda \sum_{j=1}^{760} g_{ij} Teen_j + \beta_1 + Edu_i \beta_2 + Inco_i \beta_3 + FHH_i \beta_4 + Black_i \beta_5 + Phy_i \beta_6 + \varepsilon_i,$$

where $Teen_i$ is the teenage pregnancy rate in county $i$, which is the percentage of pregnancies occurring to females of 12-17 years old, $g_{ij}$ is the entry in the row-normalized network $G_n$. The original link matrix $G^*$ is constructed by $g_{ij}^* = 1$ if two counties are neighboring counties. $Edu_i$ is the education service expenditure (in units of \$100), $Inco_i$ is median household income (divided by 1000), $FHH_i$ is the percentage of female-headed households, $Black_i$ is the proportion of black population and $Phy_i$ is the number of physicians per 1000 population, all in county $i$. We focus on counties in the 10 Upper Great Plains States, including Colorado, Iowa, Kansas, Minnesota, Missouri, Montana, Nebraska, North Dakota, South Dakota, and Wyoming, which consist of 761 counties. More details of the data can be found in Lin and Lee (2010).

In our simulations, we take their empirical estimates as true values and randomly generate misclassified links using $h_{ij}^* = g_{ij}^* \cdot e_1 + (1 - g_{ij}^*) \cdot e_2$ for $i \neq j$, where $e_1$ and $e_2$ are independent Bernoulli random variables with probabilities $1 - \tau_1$ and $\tau_2$ of equalling one, respectively. Therefore,

[13] While $H_l$ depends only on the $l$-th block in $G^*$ (the adjacency matrix within group $l$), the subvector $\widetilde{\eta}_l$ depends on all links outside the diagonal blocks in $G^*$.

$\tau_1$ measures the misclassification probability that $h^*_{ij} = 0$ when the true $g^*_{ij} = 1$, and $\tau_2$ measures the misclassification probability that $h^*_{ij} = 1$ when the true $g^*_{ij} = 0$. We set $\tau_1 = n^{s-1}$ and $\tau_2 = 100n^{s-2}$ with different values of $s$ to see how the misclassification rate affects the IV estimator and QMLE. The sample size is $n = 761$. The original network matrix $G^*_n$ has $761 \times 760 = 578,360$ entries (diagonal entries are zero). The total number of existing links on the network is $4,606$.

We first report, for different values of the rate $s$, the corresponding misclassification probability $\tau$, and the expected total number of misclassified links.

**Table 6.1. Expected Number of Misclassified Links**

| $s$ | $\tau_1$ | $\tau_2$ | Mis. (1 to 0) | Mis. (0 to 1) | Total No. Mis. |
|---|---|---|---|---|---|
| 0.1 | 0.0026 | 0.0034 | 11.751 | 192.35 | 204.10 |
| 0.3 | 0.0096 | 0.0013 | 44.230 | 725.06 | 769.35 |
| 0.5 | 0.0363 | 0.0048 | 166.97 | 2733.1 | 2900.0 |
| 0.7 | 0.1366 | 0.0180 | 629.37 | 10302 | 10931 |
| 0.9 | 0.5151 | 0.0677 | 2372.4 | 38833 | 41206 |

The expected total number of misclassified links increases with the rate $s$. When $s = 0.9$, the misclassification from missing an existing link $\tau_1$ exceeds 0.5 and the expected total number of misclassified links exceeds 7% of the total links in the network. Notice $\tau_2$ is mostly much smaller than $\tau_1$, because we want our misreported matrix to have roughly similar sparsity to the real social network.

Next, we report IV and QML estimates using 200 Monte Carlo replications under $\varepsilon_i \sim N(0, \sigma^2)$ with $\sigma = 4.5$, which is the QML estimate of the error standard deviation from original data.

From these estimation results, we observe several patterns. First, when the misclassification rate is low ($s \leq 0.3$), both IV estimates and QMLE work fine. Estimates and the standard errors are close to those from the true network, consistent with our theoretical results. Secondly, when $s$ increases, the bias and inaccuracy of both estimators increases, as expected. When $s = 0.9$, estimates are severely biased for all the parameters, with estimates of $\lambda$ suffering the most. Lastly, QMLE has smaller standard errors compared to the IV estimates in all cases, which is consistent with the asymptotic efficiency of the QMLE.

**Table 6.2. Simulation Results with Low Misclassification Rates**

| True | $\lambda = 0.4$ | $\beta_1 = 7$ | $\beta_2 = -0.01$ | $\beta_3 = -0.2$ | $\beta_4 = 0.75$ | $\beta_5 = 0.14$ | $\beta_6 = -0.5$ |
|---|---|---|---|---|---|---|---|
| No mis. | 0.4096 | 6.8879 | −0.0097 | −0.2007 | 0.7498 | 0.1347 | −0.4924 |
| IV | (0.087) | (1.576) | (0.0064) | (0.042) | (0.064) | (0.056) | (0.185) |
| | [0.079] | [1.553] | [0.0074] | [0.044] | [0.063] | [0.052] | [0.195] |
| QMLE | 0.3815 | 7.2858 | −0.0097 | −0.2065 | 0.7582 | 0.1383 | −0.4970 |
| | (0.039) | (1.123) | (0.0064) | (0.039) | (0.061) | (0.055) | (0.184) |
| | [0.045] | [1.213] | [0.0074] | [0.042] | [0.061] | [0.053] | [0.197] |
| $s = 0.1$ | 0.4121 | 6.9081 | −0.0097 | −0.2037 | 0.7530 | 0.1369 | −0.4963 |
| IV | (0.089) | (1.594) | (0.0064) | (0.042) | (0.064) | (0.056) | (0.186) |
| | [0.080] | [1.570] | [0.0074] | [0.044] | [0.063] | [0.053] | [0.195] |
| QMLE | 0.3851 | 7.2856 | −0.0098 | −0.2089 | 0.7609 | 0.1401 | −0.5004 |
| | (0.040) | (1.131) | (0.0064) | (0.039) | (0.061) | (0.055) | (0.185) |
| | [0.045] | [1.223] | [0.0074] | [0.042] | [0.061] | [0.054] | [0.198] |
| $s = 0.3$ | 0.4232 | 6.8757 | −0.0099 | −0.2096 | 0.7600 | 0.1433 | −0.5027 |
| IV | (0.096) | (1.654) | (0.0065) | (0.042) | (0.064) | (0.056) | (0.187) |
| | [0.090] | [1.652] | [0.0075] | [0.044] | [0.064] | [0.054] | [0.198] |
| QMLE | 0.3930 | 7.2909 | −0.0099 | −0.2149 | 0.7678 | 0.1462 | −0.5061 |
| | (0.041) | (1.145) | (0.0064) | (0.039) | (0.061) | (0.055) | (0.186) |
| | [0.050] | [1.255] | [0.0075] | [0.042] | [0.062] | [0.055] | [0.200] |

Note: The table reports average estimates and average standard errors (in parentheses) from from 200 simulated samples. The sample standard deviation of 200 estimates are reported in square brackets.

**Table 6.3. Simulation Results with High Misclassification Rates**

| True | $\lambda = 0.4$ | $\beta_1 = 7$ | $\beta_2 = -0.01$ | $\beta_3 = -0.2$ | $\beta_4 = 0.75$ | $\beta_5 = 0.14$ | $\beta_6 = -0.5$ |
|---|---|---|---|---|---|---|---|
| $s = 0.5$ | 0.4578 | 6.7437 | −0.0101 | −0.2265 | 0.7801 | 0.1560 | −0.5173 |
| IV | (0.122) | (1.883) | (0.0065) | (0.042) | (0.064) | (0.057) | (0.189) |
| | [0.115] | [1.867] | [0.0077] | [0.045] | [0.065] | [0.056] | [0.205] |
| QMLE | 0.4135 | 7.3237 | −0.0101 | −0.2321 | 0.7881 | 0.1584 | −0.5196 |
| | (0.047) | (1.192) | (0.0065) | (0.039) | (0.062) | (0.056) | (0.188) |
| | [0.059] | [1.346] | [0.0077] | [0.044] | [0.064] | [0.057] | [0.208] |
| $s = 0.7$ | 0.4985 | 6.7212 | −0.0105 | −0.2563 | 0.8163 | 0.1719 | −0.5340 |
| IV | (0.206) | (2.679) | (0.0067) | (0.041) | (0.064) | (0.058) | (0.194) |
| | [0.200] | [2.756] | [0.0080] | [0.049] | [0.066] | [0.061] | [0.216] |
| QMLE | 0.4108 | 7.7762 | −0.0106 | −0.2607 | 0.8229 | 0.1734 | −0.5348 |
| | (0.065) | (1.333) | (0.0067) | (0.040) | (0.063) | (0.057) | (0.193) |
| | [0.084] | [1.595] | [0.0081] | [0.048] | [0.066] | [0.061] | [0.218] |
| $s = 0.9$ | 0.2829 | 9.4891 | −0.0110 | −0.2794 | 0.8477 | 0.1788 | −0.5402 |
| IV | (0.407) | (4.773) | (0.0068) | (0.041) | (0.064) | (0.059) | (0.197) |
| | [0.450] | [3.359] | [0.0083] | [0.051] | [0.068] | [0.063] | [0.224] |
| QMLE | 0.1871 | 10.581 | −0.0110 | −0.2802 | 0.8488 | 0.1792 | −0.5383 |
| | (0.069) | (1.471) | (0.0068) | (0.041) | (0.064) | (0.058) | (0.196) |
| | [0.135] | [1.958] | [0.0083] | [0.051] | [0.068] | [0.063] | [0.226] |

Note: The table reports average estimates and average standard errors (in parentheses) from from 200 simulated samples. The sample standard deviation of 200 estimates are reported in square brackets.

## 6.2 IV estimation using misclassified links

In the previous section we added simulated measurement error to a single empirically observed adjacency matrix. In this section we investigate the performance of the two-stage least-squares estimator with misclassified links using simulated data with a range of sample sizes.

The structural equation in our data-generating process (DGP) is $y = \alpha + \lambda Gy + X\beta + GX\gamma + \varepsilon$, where $X$ is an $n \times 2$ matrix that consists of two characteristics. The parameter values are: $\alpha = 1$; $\lambda = 0.8$; $\beta = (1.5, 2)'$ and $\gamma = (0.9, 0.6)'$. For each observation $i = 1, .., n$, the error terms $\varepsilon_i$ 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 $\{-1, 1, 2\}$, and the second from a standard normal. The links in the latent adjacency matrix $G^*$ (of which $G$ is a row normalization) are formed independently with probability $p_n = \mu/n$ for some constant $\mu < \infty$, so that the expected number of neighbors for each individual is $n \times (\mu/n) = \mu$. In

the data-generating process, non-existent links ($G^*_{ij} = 0$) are never misclassified, while existing links ($G^*_{ij} = 1$) are misclassified with probability $\tau_n = \rho n^{s-1}$ for some constant $\rho < \infty$. The expected number of misclassified links is therefore $p_n \times n \times n \times \tau_n = \mu \times \rho \times n^s = O(n^s)$.

We set $\mu = 20$, $\rho = 3$ and $s = 2/5$ in simulation, and experiment with network sizes $n = 250, 500, 1000, 2000$. For each network size, we simulate $T = 200$ samples from the DGP above, and calculate a two-stage least-squares estimator using $H^2X$ as (mismeasured) instruments, where $H$ is the row normalization of the network with misclassified links reported in the sample.

**Table 6.4. IV Estimator with Misclassified Links**

(# of simulated samples: 200)

| | $n = 500$ | | | | $n = 1000$ | | | | $n = 2000$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | m.s.e. | bias | std | a.s.e. | m.s.e. | bias | std | a.s.e. | m.s.e. | bias | std | a.s.e. |
| $\alpha$ | 10.85 | 0.937 | 3.166 | 3.257 | 5.092 | 0.384 | 2.229 | 2.219 | 2.095 | 0.373 | 1.402 | 1.534 |
| $\lambda$ | 0.074 | -0.069 | 0.264 | 0.272 | 0.034 | -0.025 | 0.184 | 0.184 | 0.014 | -0.027 | 0.116 | 0.127 |
| $\beta_1$ | 0.001 | -0.003 | 0.036 | 0.037 | 0.0006 | -0.0004 | 0.025 | 0.026 | 0.0003 | 0.0004 | 0.017 | 0.018 |
| $\beta_2$ | 0.002 | -0.010 | 0.048 | 0.046 | 0.0011 | 0.0001 | 0.034 | 0.032 | 0.0005 | 0.0024 | 0.023 | 0.023 |
| $\gamma_1$ | 0.182 | -0.047 | 0.425 | 0.436 | 0.090 | -0.084 | 0.289 | 0.298 | 0.040 | -0.026 | 0.199 | 0.206 |
| $\gamma_2$ | 0.312 | -0.038 | 0.558 | 0.573 | 0.144 | -0.051 | 0.377 | 0.394 | 0.067 | -0.025 | 0.258 | 0.272 |

Note: m.s.e (mean squared error), bias, std (standard deviation) are calculated using the empirical empirical distribution of 200 estimates. "a.s.e." is the average of standard errors in 200 samples.

Table 6.4 summarizes the performance of the two-stage least-squares estimator using the mismeasured instruments based on $H$. It appears that the estimators for $\alpha, \lambda, \beta, \gamma$ all converge at a root-n rate. The mean-squared errors for the coefficient of the discrete regressor is remarkably lower than that of continuous regressor.

## 6.3 Estimation with unreported links

In this Section we provide another simulation study of the asymptotic properties of our closed-form estimator for social effects in Section 5 where the links are completely unreported in the data. We simulate $S = 200$ samples, each of which consists of $L$ independent groups. Each group involves $n$ individuals, where $n$ is a fixed small integer, and no links exist across groups. Note there is slight abuse of notation in that we now use $n$ to denote the size of each group. This contrasts with Section 6.1 where $n$ denotes the number of all individuals on a single large network. As mentioned earlier, the DGP considered in the current section can be interpreted as a special case of a single large network with block-diagonal adjacency matrix.

The structural equation in our data-generating process (DGP) is $y = \alpha + \lambda Gy + X\beta + GX\gamma + \varepsilon$, 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 $\varepsilon_i$ 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 $N(1, \sigma)$ with $\sigma = 2$. 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 with probability 0.5.

**Table 6.5. Closed-form Estimates with Unreported 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 |
| $\lambda$ | 0.0197 | -0.0305 | 0.1374 | 0.0044 | -0.0162 | 0.0648 | 0.0017 | -0.0061 | 0.0409 | 0.0010 | -0.0069 | 0.0314 |
| $\beta_1$ | 0.7232 | 0.0288 | 0.8521 | 0.0143 | 0.0133 | 0.1190 | 0.0047 | 0.0123 | 0.0677 | 0.0024 | 0.0086 | 0.0487 |
| $\beta_2$ | 0.6762 | 0.0590 | 0.8223 | 0.0078 | 0.0130 | 0.0876 | 0.0031 | 0.0072 | 0.0553 | 0.0018 | 0.0074 | 0.0416 |
| $\gamma_1$ | 1.3511 | 0.2260 | 1.1430 | 0.2911 | 0.0808 | 0.5347 | 0.1009 | 0.0399 | 0.3159 | 0.0760 | 0.0357 | 0.2740 |
| $\gamma_3$ | 0.1192 | 0.0370 | 0.3441 | 0.0484 | 0.0151 | 0.2200 | 0.0225 | -0.0016 | 0.1505 | 0.0125 | 0.0061 | 0.1119 |
| $\alpha$ | 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 deviation are calculated from empirical distribution of coefficient estimates in 200 simulated samples.

**Table 6.6. Closed-form Estimates with Unreported 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 |
| $\lambda$ | 0.0181 | -0.0340 | 0.1305 | 0.0037 | -0.0086 | 0.0603 | 0.0017 | -0.0037 | 0.0417 | 0.0007 | -0.0059 | 0.0258 |
| $\beta_1$ | 0.0151 | 0.0199 | 0.1216 | 0.0031 | 0.0024 | 0.0556 | 0.0015 | 0.0051 | 0.0389 | 0.0006 | -0.0020 | 0.0238 |
| $\beta_2$ | 0.0118 | 0.0184 | 0.1071 | 0.0022 | 0.0044 | 0.0463 | 0.0008 | 0.0028 | 0.0283 | 0.0004 | -0.0017 | 0.0207 |
| $\gamma_1$ | 1.4307 | 0.2101 | 1.1805 | 0.2747 | 0.0443 | 0.5236 | 0.1233 | 0.0255 | 0.3510 | 0.0546 | 0.0279 | 0.2326 |
| $\gamma_3$ | 0.1422 | 0.0448 | 0.3753 | 0.0373 | 0.0006 | 0.1937 | 0.0209 | 0.0016 | 0.1448 | 0.0105 | 0.0184 | 0.1010 |
| $\alpha$ | 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 deviation are calculated from empirical distribution of coefficient estimates in 200 simulated samples.

We estimate the model using the two-step method in Section 5.5. 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 5.6.2. Table 6.5 and 6.6 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 calculated for $S = 200$ simulate samples. We increase the sample size $L$, i.e. the number of groups in each sample, from $L = 60$ to $L = 240$.

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 increases in sample size. In fact the reduction in m.s.e. between $L = 60$ and $L = 120$ is even more dramatic 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 rather small between the DGP with small group sizes $n = 10$ and $n = 20$. This illustrates a desirable feature of our two-step 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 is deterministic and 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 5.6.2. 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 is key for the highly encouraging performance reported in Tables 6.5 and 6.6.

# 7 Peer Effects in Tennessee Elementary Schools

We next apply our method for dealing with unobserved adjacency matrices from Section 5 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. The STAR sample data does not report any measure of links among the students, and so is a natural environment for applying our method of estimation in Section 5.

The typical method of evaluating potential peer effects in a model without link data is to assume a linear-in-means specification. That is, each group member (in this case, each student in a class) has outcomes that are assumed to depend on classmates by including the class average (or leave-one-out class average) outcome as a regressor, or by including class average covariates as regressors. This is equivalent to assuming an adjacency matrix where each student in a class is linked to all the others with equal weights. Examples of papers that use this method to estimate contextual effects of pre-assignment covariates include effects of student-teacher races in Dee (2004), gender ratios in Whitmore (2005), or a composite of peer characteristics in Graham (2008) and Sojourner (2013). Boozer and Cacciola (2001) use experimental variation in class quality (fraction of students exposed in the previous year to small classes) as an instrument to identify peer effects in our STAR data, assuming a linear-in-means specification.

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 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. In particular, we find we can reject the linear-in-means specification and random Poisson link formation in the STAR data. We further use our structural model estimates to perform counterfactual calculations.

### 7.1 Data description

The sample consists of a cohort of students who were in kindergarten between 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 (13-25 students). Students who joined the cohort at STAR schools after 1985-1986 were also included in the experiment throughout latter years.

Our sample consists of 258 classes that had at least 15 but no more than 25 students. The total number of students in the sample is 5,189. For ease of comparison, we partition the classes in the sample into smaller (with 15-20 students) and larger (with 21-25 students) classes. Table 7.1 reports summary statistics of the students' mathematics test scores in the second and third grade ($s2$ and $s3$) 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 *s2* using the overall mean and standard deviation of raw scores of all classes in the sample.

**Table 7.1. Summary Statistics**

| | Small class (122 obs) | | | | Large class (136 obs) | | | |
|---|---|---|---|---|---|---|---|---|
| | mean | median | std dev | range | mean | median | std dev | range |
| *s3* | 620.7 | 618.0 | 40.88 | [487.0, 774.0] | 616.6 | 616.0 | 40.15 | [510.0, 774.0] |
| *s2* | 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: *s3*: raw scores for 3rd grade math; *s2*: 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 |
|---|---|---|---|
| *s3* | 0.001 | *abs* | 0.401 |
| *s2* | <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 is a linear specification that incorporates direct, contextual and peer effects:

$$\begin{aligned} s3_{l,i} = \alpha \quad & +\lambda \textstyle\sum_j G_{ij} s3_{l,j} + \beta_1 abs_{l,i} + \beta_2 mot_{l,i} + \beta_3 s2_{l,i} + \delta tec_l \\ & +\gamma_1 \textstyle\sum_j G_{ij} mot_{jl} + \gamma_2 \sum_j G_{ij} s2_{l,j} + \varepsilon_{l,i}, \end{aligned}$$

where $l$ is an index for classes, and $i$ and $j$ are indexes for individual members in a class. For each pair $i$ and $j$, $G_{ij}$ is the row-normalized unobserved zero or nonzero link between $i$ and $j$. As noted before, the coefficient $\lambda$ reflects endogenous peer effects, $(\beta_1, \beta_2, \beta_3)$ are direct individual effects, and $(\gamma_1, \gamma_2)$ are exogenous contextual effects. The coefficient $\delta$ is the marginal impact of teacher experience, which is a possible source of correlated effects.

Note 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 the peer effects. In contrast, a student's self-reported motivation score *mot* and his Grade 2 math score $s2$ are allowed to have contextual effects ($\gamma_1, \gamma_2 \neq 0$) in addition to peer effects. These assumptions are less restrictive than those required for identification in a linear-in-means specification.

We also assume the individual effects $\beta$ and contextual effects $\gamma$ are the same in small and large classes, while the structural intercept $\alpha$, the peer effect $\lambda$ and the correlated effect $\delta$ are potentially different in small and large classes. This constraint on $\beta$ and $\gamma$ is an example of exclusion restrictions across environments.

An advantage of our method is that it does not require explicit modeling or parametrization of the link formation process. In general, network formation may depend on student demographics.

As a first-order control for such dependence, we partition the classes in the sample into those with higher or lower dispersion in birthdays.[14] Consistent with previous literature, we maintain that the model coefficients do not vary with factors related to link formation such as birthday dispersion. The estimates for social effects reported below are sample-size-weighted averages of estimates obtained conditional on birthday dispersion.

### 7.3 Estimation results

Table 7.3 reports our estimates for social effects as well as structural intercepts $\alpha$. Standard errors are calculated using $B = 200$ 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* | $\lambda$ | 0.8478*** | (0.0159) | 0.9208*** | (0.0280) |
| *Group* | $\delta$ | 0.0709 | (0.2885) | 0.2032 | (0.2609) |
| *Constant* | $\alpha$ | 94.543*** | (26.221) | 48.126*** | (14.450) |
| | | est. | | (s.e.) | |
| *Direct* | $\beta_1$ | -0.3639** | | (0.1604) | |
| | $\beta_2$ | 0.0384 | | (0.0602) | |
| | $\beta_3$ | 23.356*** | | (5.0284) | |
| *Context* | $\gamma_1$ | -0.0118 | | (0.0728) | |
| | $\gamma_2$ | 13.129** | | (6.0902) | |

Notes: Standard errors are computed using $B = 200$ 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 $\lambda$ 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 $\lambda$ estimates are largely consistent with earlier findings that used 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. Another linear-in-means estimate is Graham (2008), who using GMM reports a social interaction effect of 0.86 for normalized math scores ($\gamma - 1$ in his notation). Although the estimate of peer effects is similar across specifications, we later test and

[14] 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.

reject the linear-in-means specification, and also obtain estimates of both direct and contextual effects.

Unlike these previous papers, we obtain different estimates of peer effects in large versus small classes. The bigger $\lambda$ in larger classes could be due to students having more options to form links/friendships in larger classes. This could on average lead to better matches of friends, and hence be conducive to more productive relationships.

Our estimates also show that the number of days absent from school has a statistically significantly but small impact on a student's test performance. Self-reported motivation scores have no significant direct impact on one's own test score, or contextual influence on classmates' performance. On the other hand, students' performance in the second grade ($s2$) has statistically significant positive impacts on their scores in the third grade ($s3$) both through the direct and contextual effects. A unit (one standard deviation) increase in a student's score in the second grade would improve his raw score in the third grade by 23.36 This impact is significant at the 1% level. In addition, we find that a unit increase in friends' Grade 2 scores increases a student's own Grade 3 score by 13.13. Such a contextual effect is smaller than the direct effect of one's own Grade 2 score.

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 $\alpha$, which can be interpreted as a proxy benchmark, 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 several tests related to model specification, using the estimates and bootstrap standard errors calculated from the preceding section.

First, we perform a test of our model specification. The test exploits the fact that our model is over-identified given our exclusion restrictions. Specifically, the last step of our estimator leads, in our specification, to a system of fifteen linear equations for seven parameters. Our estimator chooses parameter values to minimize the distance between the left-and right-hand side of the linear system, with the distance measured as the Euclidean norm, or dot-product, of the difference.

To test the linear specification of social network and the exclusion restrictions, we use the minimized objective function in the last step as our test statistic.[15] Under the null of correct linear specification and exclusion restrictions, this quantity is asymptotically zero. To test if the minimized objective function is zero, we use $B = 200$ bootstrap samples to estimate the sampling

[15] Note that our estimator does not lend itself to the use of classical J-tests for over-identification 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. Once these reduced-form coefficient estimates are calculated, the linear system used in the last step is deterministic.

distribution of this statistic and calculate p-values under the null. The results show no statistically significant evidence against the null: The p-values are 0.580 for classes where students' birthdays are less dispersed, and 0.375 for classes with higher birthday dispersion.

Our model only imposes regularity on the adjacency matrix data generating process. We can therefore use our model to test commonly proposed models of the adjacency matrix. We next test two different null hypotheses: the linear-in-means specification of the adjacency matrix, and a Poisson random network formation process, where links are drawn independently from a heterogenous Bernoulli distribution.

In the linear-in-means specification of social interaction, the adjacency matrix $G$ is a constant $n \times n$ matrix with all components being $1/n$. This means that $G^s = G$ for all any integer $s$, and that for all individual characteristics $k$,

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

This implies that the off-diagonal components in $\mu_k$ must be identical. We calculated Wald test statistics using a $6 \times 6$ leading principal minor of the reduced-form coefficient for $s2$ (standardized Grade 2 score) in each of the subpopulations defined by the sample size and the birthday dispersion of students. The test statistics are reported in the following table:

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

| | small class (p-val) | large class (p-val) |
|---|---|---|
| low disp. | 98.258 (<.001) | 72.948 (<.001) |
| high disp. | 47.398 (.017) | 63.117 (<.001) |

The number of restrictions, which equals the degrees of freedom, of each test is $d.f. = 6 \times 6 - 6 - 1 = 29$, which makes 42.557 be the critical value for each test at the 5% level. We reject the linear-in-means social interaction specification at the 5% level in all four subpopulation defined by the class size and birthday dispersion.

Next, we construct classical minimum distance (CMD) tests for the null hypothesis of Poisson random network formation, controlling for class sizes and birthday dispersion. Specifically, the null hypothesis posits a random link formation process where each element of $G^*$ equals one with some success probability, and equals zero with one minus that probability, independent of all the other elements of $G^*$. $G^*$ is then row normalized to yield $G$. The success probability takes one of three possible values $p \in (0,1)^3$ depending on the difference between the two students' birthdays.

The CMD objective function for estimating link formation probabilities is constructed as follows. For a generic vector $p \in (0,1)^3$, simulate a large number $S$ of $n \times n$ networks by drawing independently from a Bernoulli distribution with corresponding success probability. Then define the objective function $\hat{Q}_S(p)$ as the weighted sum of the distance between model-implied marginal effects $S^{-1}\sum_s (I - \hat{\lambda} G_s)^{-1}(\hat{\beta}_k I + \hat{\gamma}_k G_s)$ and the reduced-form coefficients $\hat{\mu}_k$ in first-step regressions controlling for class sizes and birthday dispersion. In particular, we define the distance between matrices as two differences in average diagonal and off-diagonal components respectively.

Our statistic for testing the null of Poisson random network is the minimized objective function under the optimal choice of weight matrix in CMD, which is constructed using bootstrap standard errors. For each test within a subpopulation (defined by class size and birthday dispersion), the degree of freedom of the limit distribution under the null is 3.[16] The wald statistics are reported in the following table:

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

| | small class (p-val) | large class (p-val) |
|---|---|---|
| low disp. | 61.276 (<.001) | 159.09 (<.001) |
| high disp. | 39.348 (<.001) | 115.752 (<.001) |

Thus we reject the null of Poisson random network formation in all subpopulations.

We conclude that the link formation process is more complicated than either everyone linking with everyone (i.e., linear-in-means), or independent random links.

## 7.5 Counterfactuals: complete network and alternative peer effects

Given the popularity of the linear-in-means specifications, our first counterfactual exercise is to use the structural estimates from Table 7.3 to predict counterfactual outcomes of Grade 3 math scores if the network were to be replaced by a linear-in-means model. For each class, we calculate the within-class average change in Grade 3 scores under this counterfactual change (post-change minus before-change). The expected outcome in each class under this change is calculated by replacing the unknown random adjacency matrix $G_l$, which enters $M_l = (I - \lambda G_l)^{-1}$ in the reduced form, with one where every entry is $1/n$. Table 7.6 reports the average changes in group means across the classes in each sub-population defined by class size and birthday dispersion.

[16] This is because the restrictions (# 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 $\dim(p) = 3$.

**Table 7.6: Changes in Outcome under the Linear-in-Means Network**

| | Est. mean $\Delta$ | 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 $\Delta$: average changes in class means of G3 math scores in a network with equal weights on all neighbors.

Table 7.6 shows that the overall changes induced by a complete network are relatively small, compared with the observed standard deviation of 40 for Grade 3 raw math scores in the data (see Table 7.1). We also report p-values of t-tests for equal sample means with unequal variance in Table 7.6. The counterfactual changes in the classes with higher birthday dispersions are statistically insignificant. Among classes with less dispersed birthdays, those with fewer students benefit from a complete network while those with more students see a lower class average. But both effects are insignificant at 12% level.

These results could explain why the previous literature that assumed a linear-in-means specification obtained peer effect estimates similar in magnitude to ours, despite the fact that our tests in Table 7.4 reject the linear-in-means specification.

While not always statistically significant, the difference in the signs of changes in small versus large classes reported in Table 7.6 is suggestive, and might be explained as follows: Replacing the actual adjacency matrix with a complete network essentially amounts to redistributing weights onto classmates who were previously not friends. This could impact a student's score in both directions, depending on whether the counterfactual "new friends" would have a positive or negative impact on a student's test performance. This average effect of potential new friends appears to differ between small and large classes.

Would it be worthwhile to institute policies that encourage students to form additional links or friendships? The results in Table 7.6 suggests the impacts of such policies would be small, and could even have negative consequences based on class size.

In the next counterfactual exercise, we combine the complete adjacency matrix with alternative, hypothetical peer effects. Specifically, we swap the estimated peer effects between small and large classes (i.e., increase $\lambda$ to 0.9208 in small classes and decrease $\lambda$ to 0.8478 in large classes). The goal of this exercise is to assess how these differences in peer effect magnitudes interact with the contextual effects and other differences between small and large classes.

Table 7.7 reports the average changes in class means within each subpopulation defined by class sizes and birthday dispersion. It also reports p-values of t-tests for the significance of mean changes.

**Table 7.7: Impact of Counterfactual Peer Effects**

| | Est. mean $\Delta$ | 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 $\Delta$: average changes in class mean of G3 math scores when peer effects in small and large classes are swapped in a complete network.

The table shows that increasing peer effects in small classes would lead to significantly better Grade 3 performance, and reducing peer effects in large classes would yield worse performance. Again, there is evidence that the impact is statistically more significant in classes with less dispersed birthdays. In classes with low birthday dispersion, swapping the peer effects increases the magnitudes of the effects reported in Table 7.6. In particular the average changes in class means becomes highly statistically significant **in these classes**.

# 8 Conclusions

We provide two sets of results related to the estimation of social network models when the data does not report perfect measure of links. First, we characterize conditions under which IV and QML estimators of social network models remain consistent (and standard inference on these models remain valid) despite the presence of misclassified links in the observed network.

Second, we provide an original method for identifying and estimating social effects when the random latent network links are not reported in the data at all. In this case, we propose a simple two-step estimator for social effects. We apply our method to estimate the 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.

# Appendix A. Proofs

## A1. Proof of Proposition 1

For a generic matrix $A$, let $C_A$ denote the number of non-zero elements in $A$; let $A_{(i)}$, $A_{[k]}$ denote its $i$-th row and $k$-th column respectively; and let $A_{ij}$ denote its $(i,j)$-th component. Let $\Delta^* \equiv H^* - G^*$ be the difference between $H^*$ and $G^*$, where $H^*_{ii} = 0$. The difference between $H$ and $G$ is:

$$\Delta \equiv H - G = diag\left\{\left(\frac{1}{C_{G^*_{(1)}}}, ..., \frac{1}{C_{G^*_{(n)}}}\right)\right\}\Delta^* + diag\left\{\left(\frac{1}{C_{H^*_{(1)}}} - \frac{1}{C_{G^*_{(1)}}}, ..., \frac{1}{C_{H^*_{(n)}}} - \frac{1}{C_{G^*_{(n)}}}\right)\right\}H^*.$$

The right-hand side consists of a term that directly depends on $\Delta^*$ and a term due to potential wrong normalization. The second term is zero if the total number of links for each individual is reported correctly in the data despite misclassification. We first establish two lemmas that are used for proving Proposition 1.

**Lemma A1**. *Let $a$, $b$ be two random vectors in $\mathbb{R}^n$ such that $\sup_i E(|a_i| \,|G,H)$ and $\sup_j E(|b_j| \,|G,H)$ are bounded. Then $\frac{1}{n}a'\Delta b = O_p(n^{s-1})$ under Assumption 1.*

*Proof of Lemma A1.* Let $\sum_i$ and $\sum_j$ be shorthand for $\sum_{i=1}^n$ and $\sum_{j=1}^n$ respectively. By the triangular inequality,

$$\begin{aligned}
& \textstyle\sum_i \sum_j |\Delta_{ij}| = \sum_i \sum_j \left| \frac{C_{G^*_{(i)}} - C_{H^*_{(i)}}}{C_{G^*_{(i)}} C_{H^*_{(i)}}} H^*_{ij} + \frac{1}{C_{G^*_{(i)}}} \Delta^*_{ij} \right| \\
\leq\ & \textstyle\sum_i \sum_j \left( \frac{1}{C_{G^*_{(i)}} C_{H^*_{(i)}}} \left| C_{G^*_{(i)}} - C_{H^*_{(i)}} \right| H^*_{ij} + \frac{1}{C_{G^*_{(i)}}} \left| \Delta^*_{ij} \right| \right) \\
=\ & \textstyle\sum_i \left[ \frac{\left| C_{H^*_{(i)}} - C_{G^*_{(i)}} \right|}{C_{G^*_{(i)}} C_{H^*_{(i)}}} \left( \sum_j H^*_{ij} \right) + \frac{1}{C_{G^*_{(i)}}} \left( \sum_j \left| \Delta^*_{ij} \right| \right) \right] = \sum_i \left( \frac{\left| C_{H^*_{(i)}} - C_{G^*_{(i)}} \right|}{C_{G^*_{(i)}}} + \frac{1}{C_{G^*_{(i)}}} C_{\Delta^*_i} \right) \\
\leq\ & \textstyle\sum_i \left( \frac{1}{C_{G^*_{(i)}}} C_{\Delta^*_i} + \frac{1}{C_{G^*_{(i)}}} C_{\Delta^*_i} \right) \leq 2 \left( \sup_i \frac{1}{C_{G^*_{(i)}}} \right) C_{\Delta^*} = O_p(n^s).
\end{aligned}$$

where the second inequality holds because by definition $\left| C_{H^*_{(i)}} - C_{G^*_{(i)}} \right| \leq C_{\Delta^*_i}$ with probability one; and the last equality holds because $C_{\Delta^*}$ is $O_p(n^s)$ under Assumption 1. Furthermore,

$$\begin{aligned}
E\left( \left| \tfrac{1}{n} a'\Delta b \right| \middle| G, H \right) &\leq \tfrac{1}{n} E\left[ \textstyle\sum_i \sum_j |H_{ij} - G_{ij}| E\left( |a_i b_j| \mid G, H \right) \right] \\
&\leq \tfrac{1}{n} \left( \sup_{i,j} E\left( |a_i b_j| \mid G, H \right) \right) E\left( \textstyle\sum_i \sum_j |\Delta_{ij}| \right).
\end{aligned}$$

Because $\sup_{i,j} E(|a_i b_j||G,H)$ is bounded, we have $\frac{1}{n}a'\Delta b = O_p(n^{s-1})$. □

Let $V \equiv (1_n, H^2X, X, HX)$ as defined in the text.

**Lemma A2**. *Under Assumption 2, $\sup_i E(|V_{iq}||G,H) < \infty$, $\sup_i E(V_{iq}^2|G,H) < \infty$, and $\sup_i E(|y_i| \mid G,H) < \infty$.*

*Proof of Lemma A2.* Let $X_{[q]}$ denote the $q$-th column of $X$, and note

$$\begin{aligned}
\sup_i E[\left(H_{(i)}X_{[q]}\right)^2|G,H] &= \sup_i E\left[\left(\textstyle\sum_{j=1}^n H_{ij}x_{jq}\right)^2\middle| G,H\right] \\
&\leq \left(\sup_i \textstyle\sum_j |H_{ij}|\right)^2 \sup_j E(x_{jq}^2|H,G) < \infty; \\
\sup_i E\left[(H_{(i)}^2X_{[q]})^2|G,H\right] &= \sup_i E\left[\left(\textstyle\sum_k\sum_j H_{ik}H_{kj}x_{jq}\right)^2\middle| G,H\right] \\
&\leq \left(\sup_i \textstyle\sum_k |H_{ik}|\right)^2\left(\sup_k \textstyle\sum_j |H_{kj}|\right)^2 \sup_j E(x_{jq}^2|H,G) < \infty,
\end{aligned}$$

By the norm inequality, $\sup_i E(V_{iq}^2|G,H) < \infty$ implies $\sup_i E\left(|V_{iq}||G,H\right) < \infty$.

Note that the reduced form for $y$ is

$$y = M\left[\alpha_0 1_n + \textstyle\sum_k \left(\beta_{0k}I_n + \gamma_{0k}G\right)X_{[k]} + \varepsilon\right], \text{ where } M \equiv (I_n - \lambda_0 G)^{-1}.$$

It then follows that

$$\begin{aligned}
&\sup_i E\left(|y_i| \mid G,H\right) = \sup_i E\left[\left|\textstyle\sum_{j=1}^n M_{ij}\left(\alpha_0 + x_j'\beta_0 + \sum_{s=1}^n G_{js}x_s'\gamma_0 + \varepsilon_j\right)\right|\middle| G,H\right] \\
\leq\ &\sup_i \left|\textstyle\sum_j M_{ij}\right| \sup_j \left[|\alpha_0| + E\left(\left|x_j'\beta_0\right|\middle| G,H\right) + \textstyle\sum_s |G_{js}|E\left(\left|x_s'\gamma_0\right|\middle| G,H\right) + E\left(|\varepsilon_j|\right)\right].
\end{aligned}$$

Under Assumption 2, $E\left(\left|x_j'\beta_0\right|\middle| G,H\right) < \infty$ and $E\left(\left|x_s'\gamma_0\right|| G,H\right) < \infty$. Besides,

$$\sup_i \left|\textstyle\sum_j M_{ij}\right| = \sup_i \left|e_i(I-\lambda_0 G)^{-1}1_n\right| = \sup_i \left|e_i(\textstyle\sum_{l=0}^\infty \lambda_0^l G^l 1_n)\right| = \left|\frac{1}{1-\lambda_0}\right|.$$

It then follows that $\sup_i E\left(|y_i| \mid G,H\right) < \infty$. $\square$

Recall from the text that the estimation error of the IV estimator using instruments $H^2X$ is

$$(\widehat{\alpha},\widehat{\lambda},\widehat{\gamma}',\widehat{\beta}')' - (\alpha_0,\lambda_0,\gamma_0',\beta_0')' = \left[\frac{R'V}{n}(\frac{V'V}{n})^{-1}\frac{V'R}{n}\right]^{-1}\frac{R'V}{n}(\frac{V'V}{n})^{-1}\frac{V'\widetilde{\varepsilon}}{n}. \tag{21}$$

where

$$\frac{1}{n}V'\widetilde{\varepsilon} = \frac{1}{n}V'\varepsilon + \frac{1}{n}V'\Delta X\gamma_0 + \frac{1}{n}\lambda_0 V'\Delta y. \tag{22}$$

Thanks to Assumption 2 and Lemma A2, $\sup_i E(V_iV_i'|G,H) < \infty$. By the Chebyshev's inequality $\frac{1}{n}V'\varepsilon = O_p(n^{-1/2})$. Lemma A2 also suggests that $V$, $X\gamma_0$, $y$ all satisfy the dominance conditions on the vectors $a, b$ in Lemma A1. Thus the second and third terms on the right-hand side of (22) are $O_p(n^{s-1})$. Combining the results above, we have $\frac{1}{n}V'\widetilde{\varepsilon} = O_p(n^{-1/2}\vee n^{s-1})$. Under Assumptions 1 and 2, $\frac{1}{n}R'V$ converge to a matrix with rank $(2K+1)$ and is $O_p(1)$. It then follows from (21) that the stochastic order of this estimation error is $O_p(n^{-1/2}\vee n^{s-1})$.

## A2. QMLE

Assume $\varepsilon \sim (0, \sigma^2)$. The quasi log likelihood function for $y$ in eq (1) is

$$\ln L(\theta) = -\frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{n}\ln|I - \lambda H| - \frac{1}{2n\sigma^2}\varepsilon(\theta)'\varepsilon(\theta),$$

where $\theta = (\alpha, \lambda, \beta', \gamma', \sigma^2)'$, $\varepsilon(\theta) = y - \alpha\iota - \lambda Hy - X\beta - HX\gamma$. First order derivatives are:

$$\begin{aligned}
\frac{\partial \ln L(\theta)}{\partial \alpha} &= \frac{1}{n}\frac{1}{\sigma^2}\iota'\varepsilon(\theta);\ \frac{\partial \ln L(\theta)}{\partial \gamma} = \frac{1}{n}\frac{1}{\sigma^2}(HX)'\varepsilon(\theta);\ \frac{\partial \ln L(\theta)}{\partial \beta} = \frac{1}{n}\frac{1}{\sigma^2}X'\varepsilon(\theta); \\
\frac{\partial \ln L(\theta)}{\partial \lambda} &= \frac{1}{n}\frac{1}{\sigma^2}\varepsilon(\theta)'Hy - \frac{1}{n}tr[H(I-\lambda H)^{-1}];\ \frac{\partial \ln L(\theta)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{n}\frac{1}{2\sigma^4}\varepsilon(\theta)'\varepsilon(\theta),
\end{aligned}$$

where plugging in $y = \alpha_0\iota + \lambda_0 Gy + X\beta_0 + GX\gamma_0 + \varepsilon$, we can express

$$\begin{aligned}
\varepsilon(\theta) &= (\alpha_0 - \alpha)\iota + (\lambda_0 - \lambda)Hy + X(\beta_0 - \beta) + HX(\gamma_0 - \gamma) + \widetilde{\varepsilon} \text{ with} \\
\widetilde{\varepsilon} &= \varepsilon - \lambda_0\Delta y - \Delta X\gamma_0 \text{ the same as before.}
\end{aligned}$$

Then, terms in $\frac{\partial \ln L(\theta_0)}{\partial \theta}$ involve following three terms regarding to $\widetilde{\varepsilon}$, a linear term, a quadratic term, and a complicated term:

$$\frac{1}{n}a'\widetilde{\varepsilon},\ \frac{1}{n}(\widetilde{\varepsilon}'\widetilde{\varepsilon} - \sigma_0^2),\ \frac{1}{n}\left(\frac{1}{\sigma_0^2}\widetilde{\varepsilon}'Hy - tr[H(I-\lambda_0 H)^{-1}]\right).$$

We want to show that each one is $o_p(1)$. From the IV estimation proof, we know for any constant $a$, the linear term

$$\frac{1}{n}a'\widetilde{\varepsilon} = O_p(\frac{1}{\sqrt{n}} \vee n^{s-1}).$$

The quadratic term $\frac{1}{n}(\widetilde{\varepsilon}'\widetilde{\varepsilon} - \sigma_0^2)$ can be expressed as a linear-quadratic form of $\varepsilon$ as

$$\frac{1}{n}(\widetilde{\varepsilon}'\widetilde{\varepsilon} - \sigma_0^2) = \frac{1}{n}(\varepsilon'\varepsilon - \sigma_0^2) + \frac{1}{n}a'\varepsilon + \frac{1}{n}\varepsilon'B\varepsilon,$$

where $B$ satisfies $\sum_i \sum_j E(|b_{ij}||G, H) = O_p(n^s)$. Then, $\frac{1}{n}(\widetilde{\varepsilon}'\widetilde{\varepsilon} - \sigma_0^2) = O_p(\frac{1}{\sqrt{n}} \vee n^{s-1})$ because

$$E|\frac{1}{n}\varepsilon'B\varepsilon| \leq \frac{1}{n}\sup_{i,j} E|\varepsilon_i\varepsilon_j| \sum_{i=1}^n\sum_{j=1}^n E(|b_{ij}||G, H) \leq \frac{\sigma_0^2}{n}\sum_{i=1}^n\sum_{j=1}^n E(|b_{ij}||G, H) = O(n^{s-1}).$$

The third term $\frac{1}{n}\left(\frac{1}{\sigma_0^2}\widetilde{\varepsilon}'Hy - tr[H(I-\lambda_0 H)^{-1}]\right)$ can be expressed as a linear term $\frac{1}{n}a'\widetilde{\varepsilon}$ plus $\frac{1}{n}\left(\frac{1}{\sigma_0^2}\varepsilon'H(I-\lambda_0 G)^{-1}\varepsilon - tr[H(I-\lambda_0 H)^{-1}]\right)$. With

$$\frac{1}{n\sigma_0^2}\varepsilon'H(I-\lambda_0 G)^{-1}\varepsilon = \frac{1}{n\sigma_0^2}E[\varepsilon'H(I-\lambda_0 G)^{-1}\varepsilon|G, H] + O_p(\frac{1}{\sqrt{n}}) = \frac{1}{n}tr[H(I-\lambda_0 G)^{-1}] + O_p(\frac{1}{\sqrt{n}})$$

from the LLNs of quadratic terms of $\varepsilon$ conditional on $G, H$, and applying $D^{-1} - E^{-1} = -E^{-1}(D - E)D^{-1}$ and

$$tr(ABC) = \sum_{k=1}^n\sum_{i=1}^n\sum_{j=1}^n A_{ki}B_{ij}C_{jk} \leq \sup_i \sum_{k=1}^n |A_{ki}| \cdot \sup_{j,k}|C_{jk}| \cdot \sum_{i=1}^n\sum_{j=1}^n |B_{ij}|$$

for any matrices $A$, $B$, $C$, $D$, and $E$, to our case with $A = H(I - \lambda_0 H)^{-1}$, $B = G - H$, $C = (I - \lambda_0 G)^{-1}$, $D = I - \lambda_0 G$ and $E = I - \lambda_0 H$, we have

$$\frac{1}{n} tr[H(I - \lambda_0 G)^{-1} - H(I - \lambda_0 H)^{-1}] = \frac{\lambda_0}{n} tr[H(I - \lambda_0 H)^{-1}(G - H)(I - \lambda_0 G)^{-1}] = O_p(n^{s-1}).$$

and hence,

$$\frac{\partial \ln L_n(\theta_0)}{\partial \theta} = O_p(\frac{1}{\sqrt{n}} \vee n^{s-1}).$$

Together with the boundedness of $\frac{\partial^2 \ln L_n(\widetilde{\theta})}{\partial \theta \partial \theta'}$ for any $\widetilde{\theta} \in \Theta$, our QMLE has the same order as the IV estimator:

$$\widehat{\theta}_{QMLE} - \theta_0 = \left( -\frac{\partial^2 \ln L_n(\widetilde{\theta})}{\partial \theta \partial \theta'} \right)^{-1} \frac{\partial \ln L_n(\theta_0)}{\partial \theta} = O_p(\frac{1}{\sqrt{n}} \vee n^{s-1}).$$

## A3. GMM Estimator

As in Section 5.1, let $w_l \equiv (1, X'_{l,[1]}, ..., X'_{l,[K]})'$ for each group $l$, and $X_{l,[k]}$ denote the $k$-th column of $X_l$. For each $i \leq \bar{n}$,

$$E(y_{l,i}|X_l) = \mu_0 + \textstyle\sum_{k=1}^{K} \mu_{k,(i)} X_{l,[k]} = w'_l \left( \mu_0, \mu_{1,(i)} ..., \mu_{K,(i)} \right)',$$

where $\mu_{k,(i)}$ denotes the $i$-th row of $\mu_k$. Let $\bar{y}_l$ denote the average outcome in group $l$, i.e., $\bar{y}_l \equiv \frac{1}{\bar{n}} \sum_{i=1}^{\bar{n}} y_{l,i}$. Let $\theta_0 \equiv (\lambda_0, \beta'_0, \gamma'_0)'$. Section 5.1 shows that the following moment conditions hold:

$$\left( \begin{array}{c} E[y_{l,1} w'_l (w_l w'_l)^{-1}] \\ \vdots \\ E[y_{l,\bar{n}} w'_l (w_l w'_l)^{-1}] \end{array} \right) \left( \begin{array}{c} \mathbf{0}_{1 \times \bar{n}} \\ D_k(\theta_0) \end{array} \right) = I_{\bar{n}} \text{ for } k < K \quad ,$$

$$(0, \iota_k) E[(w_l w'_l)^{-1} w_l \bar{y}_l] = \frac{\beta_{0k} + \gamma_{0k}}{1 - \lambda_0} \text{ for } k \leq K \quad , \tag{23}$$

and

$$(1, \mathbf{0}_{1 \times K\bar{n}}) E[(w_l w'_l)^{-1} w_l \bar{y}_l] = \mu_0,$$

where $\mathbf{0}_{1 \times m}$ is a 1-by-$m$ vector of zeros; $\iota_k$ is a 1-by-$K\bar{n}$ vector with the first $(k-1)\bar{n}$ and the last $(k+1)\bar{n}$ components being zeros and all other components being ones; and $D_k(\theta_0)$ is a $K\bar{n}$-by-$\bar{n}$ matrix defined as

$$D_k(\theta_0) \equiv \left\{ (e'_k, e'_K) \left[ \left( \begin{array}{cc} \beta_{0k} & \beta_{0K} \\ \gamma_{0k} & \gamma_{0K} \end{array} \right)^{-1} \left( \begin{array}{c} 1 \\ -\lambda_0 \end{array} \right) \right] \right\} \otimes I_{\bar{n}},$$

with $e'_k$ being a $K$-by-1 unit vector whose $k$-th component is 1.

A GMM estimator minimizes $g(w;\theta)' \Omega g(w;\theta)$ over $\theta$ subject to known linear constraints under Assumption 2.6, where $\Omega$ is a weight matrix and $g(w;\theta)$ is a vector of sample analogs to the moment conditions specified in (23). Compared with the two-step closed-form estimator proposed in the text, this GMM estimator is computationally more demanding, because it involves solving a high-dimensional nonlinear minimization problem.

## A4. Pooling groups with different sizes

In this section we explain how to impute smaller classes with simulated "pseudo-individuals" so that the class sizes are balanced in a pooled regression, which consistently estimates a weighted average of reduced-form coefficient matrices with varying class sizes.

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 5.1 hold conditional on either group size.For each group $l$ with $n_l = \underline{n}$, define an $\bar{n} \times K$ matrix $\tilde{X}_l$ 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, $\tilde{X}_l$ 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:

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

with $X_{l,ck}$ denoting the $k$-th column in $X_l$ as before. By construction, $E(\tilde{w}_l \tilde{w}'_l)$ does not vary across groups with different sizes.

For any large group $l$ with $n_l = \bar{n}$ and all $i \leq n_l$, we have $E(\tilde{w}_l y_{l,i} | n_l = \bar{n}) = E(\tilde{w}_l \tilde{w}'_l)\Phi_i(\bar{n})$, 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 $\mu_k$ 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 $E(\tilde{w}_l y_{l,i} | n_l = \underline{n}) = E(\tilde{w}_l \tilde{w}'_l)\Phi_i(\underline{n})$, 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 $i = 1, ..., \underline{n}$,

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

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 5.3 to estimate the structural parameters of social effects.

## A5. Proofs in Section 5

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

$$y_{l,i} = w_l'\delta_{l,i} + \tilde{\varepsilon}_{l,i},$$

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

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

with $\beta_{0k}, \gamma_{0k}$ being the $k$-th components in $\beta, \gamma$. Regressing $(y_{l,i})_{l\leq L}$ on $(w_l)_{l\leq L}$ gives:

$$\left(\textstyle\sum_l w_l w_l'\right)^{-1}\left(\textstyle\sum_l w_l y_{l,i}\right) = \underbrace{\left(\frac{1}{L}\textstyle\sum_l w_l w_l'\right)^{-1}}_{A_L}\underbrace{\left(\frac{1}{L}\textstyle\sum_l w_l w_l'\delta_{l,i}\right)}_{B_L} + \left(\frac{1}{L}\textstyle\sum_l w_l w_l'\right)^{-1}\underbrace{\left(\frac{1}{L}\textstyle\sum_l w_l\tilde{\varepsilon}_{l,i}\right)}_{C_L}.$$

As $L\to\infty$, $A_L \xrightarrow{p} E\left(w_l w_l'\right)^{-1}$ and $C_L \xrightarrow{p} E\left(w_l\tilde{\varepsilon}_{l,i}\right) = 0$ because of the weak law of large numbers and Assumptions 2.1, 2.2 and 2.3. Furthermore,

$$B_L \xrightarrow{p} E\left(w_l w_l'\delta_{l,i}\right) = E\left(w_l w_l'\right)E\left(\delta_{l,i}\right),$$

where the equality follows from Assumption 2.4. This implies

$$\left(\textstyle\sum_l w_l w_l'\right)^{-1}\left(\textstyle\sum_l w_l y_{l,i}\right) \xrightarrow{p} E(\delta_{l,i}).$$

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

**Consistency of $\widehat{\Phi}_i$ in Section 5.8.** The weak dependence in $y$ is guaranteed by the boundedness of column sum and row sum in $G^*$, so for any $nL$-dimensional constant $a$, $\frac{1}{L}(a'y - E(a'y|X)) = O_p(L^{-1/2})$. By construction, the composite error in (19), which absorbs the misclassification error, is:

$$\widetilde{\eta} = E\left(MG - \tilde{M}H\right)X\gamma_0 + E\left(M - \tilde{M}\right)(X\beta_0 + \alpha_0), \tag{24}$$

where $\tilde{M} \equiv (I - \lambda H)^{-1}$. As in the text, let $\widetilde{\eta}_l$ denote an $n$-by-1 subvector in $\widetilde{\eta}$ that correspond to group $l$ in the sample, and let $\widetilde{\eta}_{l,i}$ denote its $i$-th component. Under assumptions maintained in Section 5, $E(y_{l,i}|X) = w_l'\widetilde{\Phi}_i + \widetilde{\eta}_{l,i}$, where $X$ is $nL$-by-$K$ matrix of regressors for all individuals. By construction,

$$\begin{aligned}\widehat{\Phi}_i - \widetilde{\Phi}_i &= \left(\textstyle\sum_l w_l w_l'\right)^{-1}\left(\textstyle\sum_l w_l y_{l,i}\right) - \widetilde{\Phi}_i = \left(\frac{1}{L}\textstyle\sum_l w_l w_l'\right)^{-1}\left[\frac{1}{L}\textstyle\sum_l w_l E(y_{l,i}|X)\right] + O_p(L^{-1/2}) - \widetilde{\Phi}_i \\ &= O_p(L^{-1/2}) + \left(\frac{1}{L}\textstyle\sum_l w_l w_l'\right)^{-1}\left(\frac{1}{L}\textstyle\sum_l w_l\widetilde{\eta}_{l,i}\right).\end{aligned}$$

Under usual regularity conditions, $\left(\frac{1}{L}\sum_l w_l w_l'\right)^{-1} = O(1)$.

Then, it remains to show that $L^{-1}\sum_l w_l\widetilde{\eta}_{l,i} = O_p(L^{s-1})$. We decompose $\sum_l w_l\widetilde{\eta}_{l,i}$ into two parts: $\sum_l w_l[E\left(MG - \tilde{M}H\right)X\gamma]_{l,i}$ and $\sum_l w_l[E\left(M - \tilde{M}\right)(X\beta + \alpha)]_{l,i}$. Each part can be expressed as $\sum_{l=1}^{L}\sum_{q=l,(i)} w_l B_{lq} c_q$ satisfying that $\sum_{l=1}^{L}\sum_{q=l,(i)}|B_{lq}| = O(n^s)$, and

$\sup_{l,i} E|\tau_l c_{q=l,(i)}| = O(1)$. For the first part, $B = E[MG - \tilde{M}H]$ and $c = X\gamma$. For the second part, $B = E\left(M - \tilde{M}\right)$ and $c = X\beta + a$. From the proof in Appendix A2, we show that for both $B$, $\sum_{p=1}^{n}\sum_{j=1}^{n}|B_{pj}| = O(n^s) = O(L^s)$. Then, for any $i = 1, ..., n$,

$$E\left|\sum_{l=1}^{L} w_l \widetilde{\eta}_{l,i}\right| \leq \sum_{l=1}^{L}\sum_{q=l,(i)} E|w_l B_{lq} c_q| \leq \sup_{l,i} E|w_l c_{q=l,(i)}| \cdot \sum_{l=1}^{L}\sum_{q=l,(i)} |B_{lq}| = O(L^s).$$

Therefore,

$$\widehat{\Phi}_i - \widetilde{\Phi}_i = O_p(L^{-1/2}) + (\frac{1}{L}\sum_l w_l w_l')^{-1}(\frac{1}{L}\sum_l w_l \widetilde{\eta}_{l,i}) = O_p(L^{-1/2} \vee L^{s-1}).$$

When $s < 1/2$, the conventional asymptotic distribution holds as the effects of $\widetilde{\eta}$ will vanish asymptotically. The estimation procedure discussed in Section 4 can be directly applied to this case.

# References


Blume, Lawrence, William Brock, Steven Durlauf, and Rajshrj Tayaraman. 2015. Linear social interactions models. Journal of Political Economy 123 (2): 444-496.

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

Boozer, Michael A., and Stephen E. Cacciola. 2001, Inside the 'Black Box' of Project Star: Estimation of Peer Effects Using Experimental Data. Yale University Economic Growth Center Discussion Paper 2001.

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

Bramoullé, Y, Rachel Kranton, and Martin D'Amours. 2014. Strategic Interaction and Networks. American Economic Review, 104 (3): 898-930.

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

de Paula Áureo, Imran Rasul, and Pedro CL Souza, 2018. Recovering social networks from panel data: identification, simulations and an application, CeMMAP working papers CWP58/18.

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

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

Hsieh, Chis-Sheng and Xu Lin, 2017. Gender and racial peer effects with endogenous network formation. Regional Science and Urban Economics, 67: 135-147

Hanushek, Eric, 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.

Hsieh, Chis-Sheng and Lungfei Lee 2016. A social interaction model with endogenous friendship formation and selectivity. Journal of Applied Econoemtrics 31: 301–319.

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

Krueger, Alan B., and Diane 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, Lungfei. 2007. Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. Journal of Econometrics 140: 333–374.

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

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

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

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