Estimating Peer Effects Using Partial Network Data

Vincent Boucher* and Aristide Houndetoungan†

April 2020

Abstract

We study the estimation of peer effects through social networks when researchers do not observe the network structure. Instead, we assume that researchers know (have a consistent estimate of) the *distribution* of the network. We show that this assumption is sufficient for the estimation of peer effects using a linear-in-means model. We present and discuss important examples where our methodology can be applied. In particular, we provide an empirical application to the study of peer effects on students' academic achievement.

JEL Codes: C31, C36, C51
Keywords: Social networks, Peer effects, Missing variables, Measurement errors

*Corresponding author. Department of Economics, Université Laval, CRREP and CREATE; email: vincent.boucher@ecn.ulaval.ca
†Department of Economics, Université Laval, CRREP; email: elysee-aristide.houndetoungan.1@ulaval.ca

We would like to thank Bernard Fortin for his helpful comments and insights, as always. We would also like to thank Eric Auerbach, Yann Bramoullé, Arnaud Dufays, Stephen Gordon, Chih-Sheng Hsieh, Arthur Lewbel, Tyler McCormick, Angelo Mele, Onur Özgür, Eleonora Patacchini, Xun Tang, and Yves Zenou for helpful comments and discussions. Thank you also to the participants of the Applied/CDES seminar at Monash University, the Economic seminar at the Melbourne Business School, the Econometric workshop at the Chinese University of Hong Kong, and the Centre of Research in the Economics of Development workshop at the Université de Namur.

This research uses data from Add Health, a program directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by Grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is given to Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from Grant P01-HD31921 for this research.

1 Introduction

There is a large and growing literature on the impact of peer effects in social networks.[1] However eliciting network data is expensive (Breza et al., 2017), and since networks must be sampled completely (Chandrasekhar and Lewis, 2011), there are few existing data sets that contain detailed network information.

In this paper, we explore the estimation of the widely used linear-in-means model (e.g. Manski (1993), Bramoullé et al. (2009)) when the researcher does not observe the entire network structure. Specifically, we assume that the researcher knows the *distribution* of the network but not necessarily the network itself. An important example is when a researcher is able to estimate a network formation model using some partial information about the network structure (e.g. Breza et al. (2017)). Other examples are when the researcher observes the network with noise (e.g. Hardy et al. (2019)) or only observes a subsample of the network (e.g. Chandrasekhar and Lewis (2011)).

We present an instrumental variable estimator and show that we can adapt the strategy proposed by Bramoullé et al. (2009), which uses instruments constructed using the powers of the interaction matrix. Specifically, we use two different draws from the distribution of the network. One draw is used to approximate the endogenous explanatory variable, while the other is used to construct the instruments.

We show that since the true networks and the two approximations are drawn from the same distribution, the instruments are uncorrelated with the approximation error and are therefore valid. We explore the properties of the estimator using Monte Carlo simulations. We show that the method performs well, even when the distribution of the network is diffuse and when we allow for group-level fixed effects.

We also present a Bayesian estimator. The estimator imposes more structure but allows cover cases for which the instrumental variable strategy fails.[2] Our estimator is general enough that it can be applied to many peer-effect models having misspecified networks (e.g. Chandrasekhar and Lewis (2011), Hardy et al. (2019), or Griffith (2019)). The approach relies on data augmentation (Tanner and Wong, 1987). The assumed distribution for the network acts as a prior distribution, and the inferred network structure is updated through the Markov chain Monte Carlo (MCMC) algorithm.

[1] For recent reviews, see Boucher and Fortin (2016), Bramoullé et al. (2019), Breza (2016), and De Paula (2017).

[2] We also provide a classical version of the estimator (using an expectation maximization algorithm) in Appendix 9.6, which is similar to the strategies used by Griffith (2018) and Hardy et al. (2019).

We present numerous examples of settings in which our estimators are implementable. In particular, we present an implementation of our instrumental variable estimator using the network formation model developed by Breza et al. (2017). We show that the method performs very well. We also show that the recent estimator proposed by Alidaee et al. (2020) works well but is less precise.

We also present an empirical application. We explore the impact of errors in the observed networks using data on adolescents' friendship networks. We show that the widely used Add Health database features many missing links: only 70% of the total number of links are observed. We estimate a model of peer effects on students' academic achievement. We show that our Bayesian estimator reconstructs these missing links and obtains a valid estimate of peer effects. In particular, we show that disregarding missing links underestimates the endogenous peer effect on academic achievement.

This paper contributes to the recent literature on the estimation of peer effects when the network is either not entirely observed or observed with noise. Chandrasekhar and Lewis (2011) show that models estimated using sampled networks are generally biased. They propose an analytical correction as well as a two-step general method of moment (GMM) estimator. Liu (2013) shows that when the interaction matrix is not row-normalized, instrumental variable estimators based on an out-degree distribution are valid, even with sampled networks. Relatedly, Hsieh et al. (2018) focus on a regression model that depends on global network statistics. They propose analytical corrections to account for non-random sampling of the network (see also Chen et al. (2013)).

Hardy et al. (2019) look at the estimation of (discrete) treatment effects when the network is observed noisily. Specifically, they assume that observed links are affected by iid errors and present an expectation maximization (EM) algorithm that allows for a consistent estimate of the treatment effect. Griffith (2018) also presents an EM algorithm to impute missing network data. Griffith (2019) explores the impact of imposing an upper bound to the number of links when eliciting network data. He shows, analytically and through simulations, that these bounds may bias the estimates significantly.

Relatedly, some papers derive conditions under which peer effects can be identified even without any network data. De Paula et al. (2018a) and Manresa (2016) use panel data and present models of peer effect having an unknown network structure. Both approaches require observing a large number of periods and some degree of *sparsity* for the interaction network. De Paula et al. (2018a) prove a global identification result and estimate their model using an

adaptive elastic net estimator, while Manresa (2016) uses a lasso estimator, while assuming no endogenous effect and deriving its explicit asymptotic properties.

Souza (2014) studies the estimation of a linear-in-means model when the network is not known. He presents a pseudo-likelihood model in which the true (unobserved) network is replaced by its expected value, given a parametric network formation model. He formally derives the identified set and applies his methodology to study the spillover effects of a randomized intervention.

Thirkettle (2019) focuses on the estimation of a given network statistic (e.g. some centrality measure), assuming that the researcher only observes a random sample of links. Using a structural network formation model, he derives bounds on the identified set for both the network formation model and the network statistic of interest. Lewbel et al. (2019) use a similar strategy but focus on the estimation of a linear-in-means model and assume a network formation model having conditionally independent linking probabilities. They show that their estimator is point-identified given some exclusion restrictions.

We contribute to the literature by proposing two estimators for the linear-in-means model, in a cross-sectional setting, when the econometrician does not know the true social network but rather knows the *distribution* of true network. Our estimators are both simple to implement and flexible. In particular, they can be used when network formation models can be estimated given only limited network information (e.g. Breza et al. (2017) or Graham (2017)) or when networks are observed imperfectly (e.g. Chandrasekhar and Lewis (2011), Griffith (2019), or Hardy et al. (2019)). We show that having partial information about network structure (as opposed to no information) allows the development of flexible and easily implementable estimators. Finally, we also present an easy-to-use R package—named `PartialNetwork`—for implementing our estimators and examples, including the estimator proposed by Breza et al. (2017). The package is available online at: `https://github.com/ahoundetoungan/PartialNetwork`.

The remainder of the paper is organized as follows. In Section 2, we present the econometric model as well as the main assumptions. In Section 3, we present an instrumental variable estimator. In Section 4, we present our Bayesian estimation strategy. In Section 5, we present important economic contexts in which our method is implementable. In Section 6, we present an empirical application in which the network is only partly observed. Section 7 concludes with a discussion of the main results, limits, and challenges for future research.

2 The Linear-in-Means Model

Let $\mathbf{A}$ represent the $N \times N$ *adjacency matrix* of the network. We assume a directed network: $a_{ij} \in \{0, 1\}$, where $a_{ij} = 1$ if i is linked to j. We normalize $a_{ii} = 0$ for all i and let $n_i = \sum_j a_{ij}$ denote the number of links of i. Let $\mathbf{G} = f(\mathbf{A})$, the $N \times N$ *interaction matrix* for some function f. Unless otherwise stated, we assume that $\mathbf{G}$ is a row-normalization of the adjacency matrix $\mathbf{A}$.[3] Our results extend to alternative specifications of f.

We focus on the following model:

$$\mathbf{y} = c\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \alpha\mathbf{G}\mathbf{y} + \mathbf{G}\mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y}$ is a vector of an outcome of interest (e.g. academic achievement), c is a constant, $\mathbf{X}$ is a matrix of observable characteristics (e.g. age, gender...), and $\boldsymbol{\varepsilon}$ is a vector of errors. The parameter α therefore captures the impact of the average outcome of one's peers on their behaviour (the endogenous effect). The parameter $\boldsymbol{\beta}$ captures the impact of one's characteristics on their behaviour (the individual effects). The parameter $\boldsymbol{\gamma}$ captures the impact of the average characteristics of one's peers on their behaviour (the contextual effects).

This *linear-in-means* model (Manski, 1993) is perhaps the most widely used model for studying peer effects in networks (see Bramoullé et al. (2019) for a recent review). In this paper, we contrast with the literature by assuming that the researcher does not know the interaction matrix $\mathbf{G}$. Specifically, we assume instead that the researcher knows the distribution of the interaction matrix.

The next assumption summarizes our set-up.

Assumption 1. *We maintain the following assumptions:*

(1.1) $|\alpha| < 1/\|\mathbf{G}\|$ *for some submultiplicative norm* $\|\cdot\|$.

(1.2) The distribution $P(\mathbf{A})$ *of the true network* $\mathbf{A}$ *(which potentially depends on* $\mathbf{X}$*) is known.*

(1.3) The population is partitioned in $M > 1$ *groups, where the size* N_r *of each group* $r = 1, ..., M$ *is bounded. The probability of a link between individuals of different groups is equal to* 0.

(1.4) For each group, the outcome and individual characteristics are observed, i.e. $(\mathbf{y}_r, \mathbf{X}_r)$, $r = 1, ..., M$, *are observed.*

(1.5) The network is exogenous in the sense that $\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}, \mathbf{G}] = \mathbf{0}$.

Assumption 1.1 ensures that the model is coherent and that there exists a unique vector $\mathbf{y}$ compatible with (1). When $\mathbf{G}$ is row-normalized, $|\alpha| < 1$ is sufficient.

[3] In such a case, $g_{ij} = a_{ij}/n_i$ whenever $n_i > 0$, while $g_{ij} = 0$ otherwise.

Assumption 1.2 states that the researcher knows the distribution of the true network $\mathbf{A}$. Of course, knowledge of $P(\mathbf{A})$ is sufficient for $P(\mathbf{G})$, since $\mathbf{G} = f(\mathbf{A})$ for some known function f. Assumption 1.2 is weaker than assuming that the econometrician observes the entire network structure. In Section 5, we discuss some important examples where Assumption 1.2 is reasonable for important economic contexts. In particular, we present examples from the literature on network formation models that allow for a consistent estimation of $P(\mathbf{A})$ using only partial network information.

As will be made clear, our estimation strategy requires that the econometrician be able to draw iid samples from $P(\mathbf{A})$. As such, and for the sake of simplicity, all of our examples will be based on network distributions that are conditionally independent across links (i.e. $P(a_{ij}|\mathbf{A}_{-ij}) = P(a_{ij})$), although this is not formally required.[4]

Assumption 1.3 is by no means necessary; however, it simplifies the exposition and ensures a law of large numbers (LLN) in our context. We refer the reader to Lee (2004) and Lee et al. (2010) for more general, alternative sufficient conditions.

Assumption 1.4 implies that the data is composed of a subset of fully sampled groups.[5] A similar assumption is made by Breza et al. (2017). Note that we assume that the network is exogenous (Assumption 1.5) mostly to clarify the presentation of the estimators. In Section 7, we discuss how recent advances for the estimation of peer effects in endogenous networks can be adapted to our context.

Finally, note that Assumption 1 *does not* imply that one can simply proxy $\mathbf{G}$ in (1) using a draw $\hat{\mathbf{G}}$ from $P(\mathbf{G})$. The reason is that for any vector $\mathbf{w}$, $\hat{\mathbf{G}}\mathbf{w}$ generally does not converge to $\mathbf{G}\mathbf{w}$ as N goes to infinity. In other words, knowledge of $P(\mathbf{G})$ and $\mathbf{w}$ is not sufficient to obtain a consistent estimate of $\mathbf{G}\mathbf{w}$. We discuss some exceptions in Section 7.

3 Estimation Using Instrumental Variables

As discussed in the introduction, we show that it is possible to estimate (1) given only partial information on network structure. To understand the intuition, note that it is not necessary to observe the complete network structure to observe $\mathbf{y}$, $\mathbf{X}$, $\mathbf{G}\mathbf{X}$, and $\mathbf{G}\mathbf{y}$. For example, one could simply obtain $\mathbf{G}\mathbf{y}$ from survey data: "What is the average value of your friends' y?"

However, the observation of $\mathbf{y}$, $\mathbf{X}$, $\mathbf{G}\mathbf{X}$, and $\mathbf{G}\mathbf{y}$ is not sufficient for the estimation of (1).

[4] A prime example of a network distribution that is not conditionally independent is the distribution for an exponential random graph model (ERGM), e.g. Mele (2017). See also our discussion in Section 7.

[5] Contrary to Liu et al. (2017) or Wang and Lee (2013), for example.

The reason is that $\mathbf{Gy}$ is endogenous; thus, a simple linear regression would produce biased estimates. (e.g. Manski (1993), Bramoullé et al. (2009)).

The typical instrumental approach to deal with this endogeneity is to use instruments based on the structural model, i.e. instruments constructed using second-degree peers (e.g. $\mathbf{G}^2\mathbf{X}$, see Bramoullé et al. (2009)). These are less likely to be found in survey data. Indeed, we could doubt the informativeness of questions such as: "What is the average value of your friends' average value of their friends' x?"

Under the assumption that the network is observed, the literature has focused mostly on efficiency: that is, how to construct the optimal set of instruments (e.g. Kelejian and Prucha (1998) or Lee et al. (2010)). Here, we are interested in a different question. We would like to understand how much information on the network structure is needed to construct relatively "good" instruments for $\mathbf{Gy}$? As we will discuss, it turns out that even very imprecise estimates of $\mathbf{G}$ allow for constructing valid instruments.

We present valid instruments in Proposition 1 and Proposition 2 below. We also study the properties of the implied estimators using Monte Carlo simulations. Unless otherwise stated, these simulations are performed as follows: we simulate 100 groups of 50 individuals each. Within each group, each link (i, j) is drawn from a Bernoulli distribution with probability:

$$p_{ij} = \frac{\exp\{c_{ij}/\lambda\}}{1 + \exp\{c_{ij}/\lambda\}}, \tag{2}$$

where $c_{ij} \sim N(0, 1)$, and $\lambda > 0$.

This approach is convenient since it allows for some heterogeneity among linking probabilities. Moreover, λ can easily control the spread of the distribution, and hence the quality of the approximation of the true network.[6] Indeed, when $\lambda \to 0$, $p_{ij} \to 1$ whenever $c_{ij} > 0$, while $p_{ij} \to 0$ whenever $c_{ij} < 0$. Similarly, as $\lambda \to \infty$, $p_{ij} \to 1/2$. Then, simulations are very precise for $\lambda \to 0$ and very imprecise (and homogeneous) for $\lambda \to \infty$.

We also let $\mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2]$, where $x_i^1 \sim N(0, 5^2)$ and $x_i^2 \sim Poisson(6)$. We set the true value of the parameters to: $\alpha = 0.4$, $\beta_0 = 2$, $\beta_1 = 1$, $\beta_2 = 1.5$, $\gamma_1 = 5$, and $\gamma_2 = -3$. Finally, we let $\varepsilon_i \sim N(0, 1)$.

We now present our formal results. To clearly expose the argument, we first start by discussing the special case where there are no contextual effects: $\boldsymbol{\gamma} = \mathbf{0}$. The model in (1) can

[6] The true network and the approximations are drawn from the same distribution.

therefore be rewritten as:

$$\mathbf{y} = c\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \alpha\mathbf{G}\mathbf{y} + \boldsymbol{\varepsilon}.$$

The following proposition holds.

Proposition 1. *Assume that* $\boldsymbol{\gamma} = \mathbf{0}$*. There are two cases:*

1. *Suppose that* $\mathbf{G}\mathbf{y}$ *is observed and let* $\boldsymbol{H}$ *be an interaction matrix, correlated with* $\mathbf{G}$*, and such that* $\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}, \mathbf{H}] = \mathbf{0}$*. Then,* $\mathbf{H}\mathbf{X}$*,* $\mathbf{H}^2\mathbf{X}$*,... are valid instruments.*

2. *Suppose that* $\mathbf{G}\mathbf{y}$ is not *observed and let* $\tilde{\mathbf{G}}$ *and* $\hat{\mathbf{G}}$ *be two draws from the distribution* $P(\mathbf{G})$*. Then,* $\hat{\mathbf{G}}\mathbf{X}$*,* $\hat{\mathbf{G}}^2\mathbf{X}$*,... are valid instruments when* $\tilde{\mathbf{G}}\mathbf{y}$ *is used as a proxy for* $\mathbf{G}\mathbf{y}$*.*

First, suppose that $\mathbf{G}\mathbf{y}$ is observed directly from the data; then, any instrument correlated with the usual instruments $\mathbf{G}\mathbf{X}, \mathbf{G}^2\mathbf{X}$,... while being exogenous are valid. Note that a special case of the first part of Proposition 1 is when $\mathbf{H}$ is drawn from $P(\mathbf{G})$. However, the instrument remains valid if the researcher uses the *wrong* distribution $P(\mathbf{G})$.[7] A similar strategy is used by Kelejian and Piras (2014) and Lee et al. (2020) in a different context. An example, presented in Section 5.2, is when $P(\mathbf{G})$ is estimated imprecisely in small samples.

Of course, the specification error on $P(\mathbf{G})$ must be independent of $\boldsymbol{\varepsilon}$. Note also that if the specification error is too large, the correlation between $\mathbf{G}\mathbf{y}$ and $\mathbf{H}\mathbf{X}$ will likely be weak. It is also worth noting that the first part of Proposition 1 does not depend on the assumption that groups are entirely sampled (i.e. Assumption 1.4).

When $\mathbf{G}\mathbf{y}$ is *not* observed directly, however, specification errors typically produce invalid instruments. Note also that the estimation requires two draws from $P(\mathbf{G})$ instead of just one. To see why, let us rewrite the model as:

$$\mathbf{y} = c\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \alpha\tilde{\mathbf{G}}\mathbf{y} + [\boldsymbol{\eta} + \boldsymbol{\varepsilon}],$$

where $\boldsymbol{\eta} = \alpha[\mathbf{G}\mathbf{y} - \tilde{\mathbf{G}}\mathbf{y}]$ is the approximation error for $\mathbf{G}\mathbf{y}$. Suppose also that $\hat{\mathbf{G}}\mathbf{X}$ is used as an instrument for $\mathbf{G}\mathbf{y}$.

The validity of the instrument therefore requires $\mathbb{E}[\boldsymbol{\eta} + \boldsymbol{\varepsilon}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}] = 0$, and in particular:

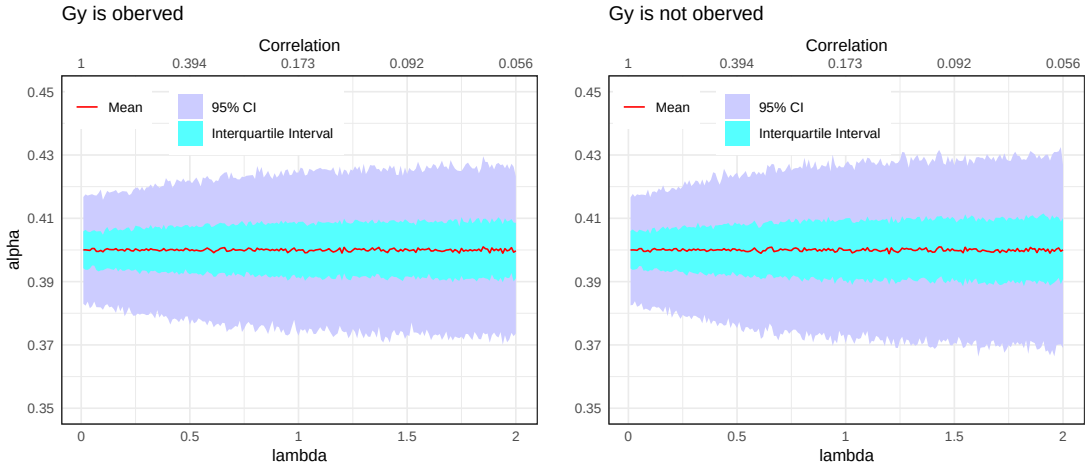
$$\mathbb{E}[\mathbf{G}\mathbf{y}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}] = \mathbb{E}[\tilde{\mathbf{G}}\mathbf{y}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}],$$

which is true since $\mathbf{G}$ and $\tilde{\mathbf{G}}$ are drawn from the same distribution.

[7] We would like to thank Chih-Sheng Hsieh and Arthur Lewbel for discussions on this important point.

Table 1 presents the results of the Monte Carlo simulations, which are in line with the above discussion. Figure 1 shows that the estimator is still centred and precise, even when the constructed networks are really imprecise estimates of the true network. Finally, note that this also implies a non-intuitive property: if $\boldsymbol{\gamma} = \mathbf{0}$, and if $\mathbf{GX}$ is observed, but not $\mathbf{Gy}$, then $\mathbf{GX}$ is not a valid instrument since it is correlated with the approximation error $\boldsymbol{\eta}$.

Figure 1: Peer effect without contextual effects



The graph shows estimates of α for 1000 replications of the model without contextual effects for various values of λ. The upper x-axis reports the average correlation between two independent network draws using the distribution given by equation (2).

Of course, Proposition 1 assumes that there are no contextual effects. We show that a similar result holds when $\boldsymbol{\gamma} \neq \mathbf{0}$. However, to estimate (1) using an instrumental variable approach, we must assume that $\mathbf{GX}$ is *observed*. The reason is that there are no natural instruments for $\mathbf{GX}$. In Section 4, we present an alternative estimation strategy that does not require the observation of $\mathbf{GX}$.

We have the following:

Proposition 2. *Assume that* $\mathbf{GX}$ *is observed. There are two cases:*

1. *Suppose that* $\mathbf{Gy}$ *is observed and let* $\boldsymbol{H}$ *be an interaction matrix, correlated with* $\mathbf{G}$*, and such that* $\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}, \mathbf{H}] = \mathbf{0}$*. Then,* $\mathbf{H}^2\mathbf{X}$*,* $\mathbf{H}^3\mathbf{X}$*,... are valid instruments.*

2. *Suppose that* $\mathbf{Gy}$ is not *observed and let* $\tilde{\mathbf{G}}$ *and* $\hat{\mathbf{G}}$ *be two draws from the distribution* $P(\mathbf{G})$*. Then,* $\hat{\mathbf{G}}^2\mathbf{X}$*,* $\hat{\mathbf{G}}^3\mathbf{X}$*,... are valid instruments when* $\tilde{\mathbf{G}}\mathbf{y}$ *is used as a proxy for* $\mathbf{Gy}$*, if* $\tilde{\mathbf{G}}\mathbf{X}$ *is added as additional explanatory variables.*

Table 1: Simulation results without contextual effects.

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - **Gy is Observed**					
Estimation results					
$Intercept = 2$	2.000	0.242	1.830	1.992	2.162
$\alpha = 0.4$	0.400	0.013	0.391	0.401	0.409
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	1,816.759	281.478	1,623.922	1,800.054	1,995.179
Hausman	1.198	1.607	0.120	0.566	1.669
Sargan	0.905	1.315	0.088	0.402	1.168
$N = 50$, $M = 100$ - **Gy is not observed - same draw**					
Estimation results					
$Intercept = 2$	4.348	0.287	4.156	4.332	4.535
$\alpha = 0.4$	0.271	0.015	0.261	0.272	0.282
$\beta_1 = 1$	1.002	0.003	0.999	1.001	1.004
$\beta_2 = 1.5$	1.503	0.006	1.498	1.503	1.507
Tests					
F-test	26,656.064	2,108.805	25,237.919	26,492.586	27,972.810
Hausman	245.060	36.134	220.376	242.230	267.029
Sargan	1.939	2.768	0.208	0.910	2.452
Validity					
$cor(\eta_i, \hat{x}_{i,1})$	−0.367	0.018	−0.380	−0.367	−0.355
$cor(\eta_i, \hat{x}_{i,2})$	−0.269	0.017	−0.280	−0.269	−0.257
$N = 50$, $M = 100$ - **Gy is not observed - different draws**					
Estimation results					
$Intercept = 2$	2.001	0.264	1.809	1.994	2.175
$\alpha = 0.4$	0.400	0.014	0.390	0.400	0.410
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	1,824.774	280.901	1,623.689	1812.479	2,014.936
Hausman	69.842	17.204	57.169	69.691	81.438
Sargan	0.891	1.245	0.082	0.431	1.143
Validity					
$cor(\eta_i, \hat{x}_{i,1})$	0.000	0.014	−0.010	0.000	0.010
$cor(\eta_i, \hat{x}_{i,2})$	0.000	0.014	−0.010	0.000	0.010

Note: Number of simulations: 1000, $\lambda = 1$. Instruments: $\boldsymbol{GX}$ if $\boldsymbol{Gy}$ is observed, $\boldsymbol{G}^c\boldsymbol{X}$ if $\boldsymbol{Gy}$ is not observed and approximated by $\boldsymbol{G}^c\boldsymbol{y}$. Additional results for alternative instruments and $\lambda = +\infty$ are available in Table 9, Table 10, and Table 11 of Appendix 9.1.

The first part of Proposition 2 is a simple extension of the first part of Proposition 1. The second part of Proposition 2 requires more discussion. Essentially, it states that $\hat{\mathbf{G}}^2\mathbf{X}$, $\hat{\mathbf{G}}^3\mathbf{X}$, ... are valid instruments when the following *expanded model* is estimated:

$$\mathbf{y} = c\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \alpha\tilde{\mathbf{G}}\mathbf{y} + \mathbf{G}\mathbf{X}\boldsymbol{\gamma} + \tilde{\mathbf{G}}\mathbf{X}\check{\boldsymbol{\gamma}} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{3}$$

where the true value of $\check{\boldsymbol{\gamma}}$ is $\mathbf{0}$.

To understand why the introduction of $\tilde{\mathbf{G}}\mathbf{X}\check{\boldsymbol{\gamma}}$ is needed, recall that the constructed instrument must be uncorrelated with the approximation error $\boldsymbol{\eta}$. This correlation is conditional on the explanatory variables, that contain $\mathbf{G}$. In particular, it implies that generically,

$$\mathbb{E}[\mathbf{G}\mathbf{y}|\mathbf{X}, \mathbf{G}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}] \neq \mathbb{E}[\tilde{\mathbf{G}}\mathbf{y}|\mathbf{X}, \mathbf{G}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}].$$

It turns out that adding the auxiliary variable $\tilde{\mathbf{G}}\mathbf{X}$ as a covariate is sufficient to restore the result, i.e.

$$\mathbb{E}[\mathbf{G}\mathbf{y}|\mathbf{X}, \mathbf{G}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}] = \mathbb{E}[\tilde{\mathbf{G}}\mathbf{y}|\mathbf{X}, \mathbf{G}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}].$$

Table 2 presents the simulations' results. We see that most of the estimated parameters are not biased. However, we also see that estimating the expanded model, instead of the true one, comes at a cost. Due to multicollinearity, the estimation of $\boldsymbol{\gamma}$ is contaminated by $\tilde{\mathbf{G}}\mathbf{X}$, and the parameters are biased. Figure 3 also shows that the estimation of α remains precise, even as the value of λ increases.

Proposition 1 and Proposition 2 therefore show that the estimation of (1) is possible, even with very limited information about the network structure. We conclude this section by discussing how one can adapt this estimation strategy while allowing for group-level unobservables.

3.1 Group-Level Unobservables

A common assumption is that each group in the population is affected by a common shock, unobserved by the econometrician (e.g. Bramoullé et al. (2009)). As such, for each group $r = 1, ..., M$, we have:

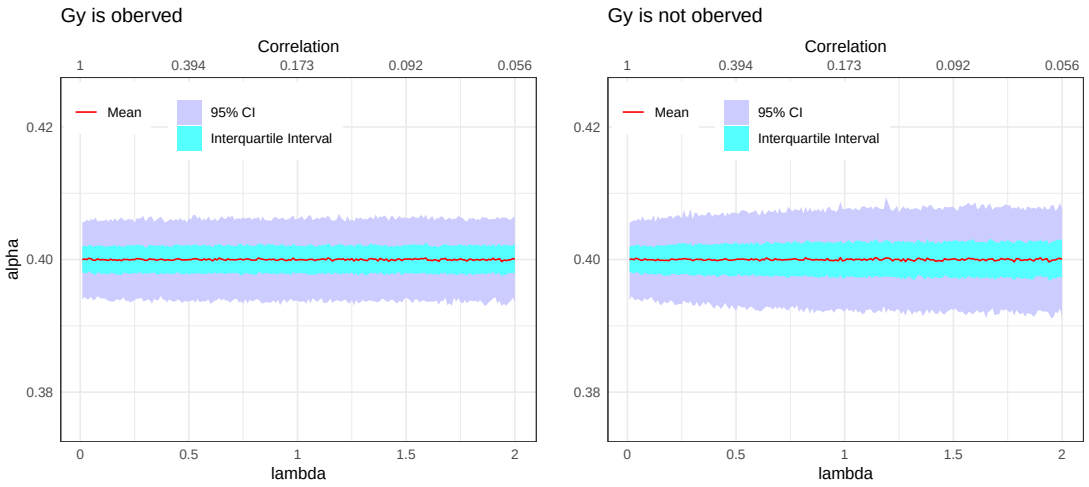
$$\mathbf{y}_r = c_r\mathbf{1}_r + \mathbf{X}_r\boldsymbol{\beta} + \alpha\mathbf{G}_r\mathbf{y}_r + \mathbf{G}_r\mathbf{X}_r\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_r,$$

Table 2: Simulation results with contextual effects.

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - Instrument: $(\tilde{\mathbf{G}})^2\mathbf{X}$ - $\boldsymbol{Gy}$ is observed					
Estimation results					
$Intercept = 2$	1.996	0.177	1.879	1.997	2.115
$\alpha = 0.4$	0.400	0.003	0.398	0.400	0.402
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
$\gamma_1 = 5$	5.000	0.021	4.985	5.000	5.015
$\gamma_2 = -3$	−2.999	0.029	−3.018	−2.999	−2.980
Tests					
F-test	18295.381	2049.380	16864.174	18258.774	19581.640
Hausman	1.202	1.624	0.127	0.568	1.593
Sargan	1.046	1.559	0.103	0.448	1.321
$N = 50$, $M = 100$ - Instrument: $(\hat{\mathbf{G}})^2\mathbf{X}$ - $\boldsymbol{Gy}$ is not observed					
Estimation results					
$Intercept = 2$	1.987	0.207	1.844	1.983	2.128
$\alpha = 0.4$	0.400	0.004	0.397	0.400	0.402
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
$\gamma_1 = 5$	5.357	0.021	5.342	5.356	5.370
$\gamma_2 = -3$	−2.381	0.038	−2.408	−2.378	−2.355
$\hat{\gamma}_1 = 0$	−0.356	0.024	−0.372	−0.356	−0.339
$\hat{\gamma}_2 = 0$	−0.617	0.038	−0.643	−0.618	−0.592
Tests					
F-test	13562.892	1402.029	12583.175	13547.357	14445.031
Hausman	17.051	8.277	11.093	15.779	22.061
Sargan	1.003	1.425	0.125	0.470	1.267

Note: Number of simulations: 1000, $\lambda = 1$. Additional results for $\lambda = +\infty$ are available in Table 12 of Appendix 9.1.

Figure 2: Peer effect with contextual effects



The graph shows estimates of α for 1000 replications of the model with contextual effects for various values of λ. The upper x-axis reports the average correlation between two independent network draws using the distribution given by equation (2).

where c_r is not observed, $\mathbf{1}_r$ is a N_r-dimensional vector of ones, N_r is the size of the group r, $\boldsymbol{G}_r$ is the sub-interaction matrix in the group r, and $\boldsymbol{\varepsilon}_r$ is the vector of error terms in the group r

Under Assumption 1.3, it is not possible to obtain a consistent estimate of $\{c_r\}_{r=1}^m$ since the number of observations used to estimate each c_r is bounded. This is known as the *incidental parameter problem*.[8] A common strategy is to use deviations from the group average and to estimate the model in deviations (e.g. Bramoullé et al. (2009)).

Let $\mathbf{J} = diag\{\mathbf{I}_{N_r} - \frac{1}{N_r}\mathbf{1}_r\mathbf{1}_r'\}$ be the group-differentiating matrix, where $\mathbf{I}_{N_r}$ is the identified matrix of dimension N_r. The operator *diag* generates a blog-diagonal matrix in which each group is a block.[9] We can rewrite:

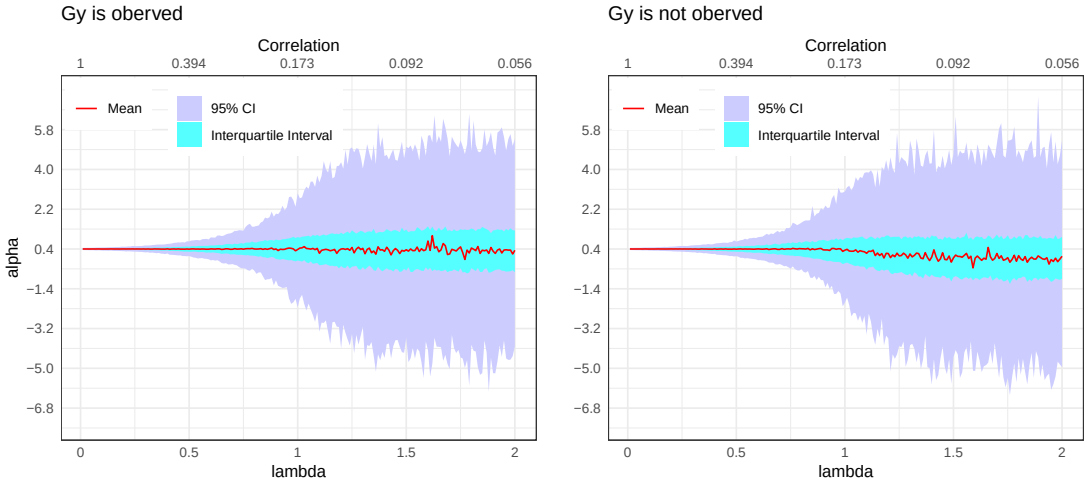
$$\mathbf{Jy} = \mathbf{JX}\boldsymbol{\beta} + \alpha\mathbf{JGy} + \mathbf{JGX}\boldsymbol{\gamma} + \boldsymbol{J\varepsilon}.$$

Note that the results of Propositions 1 and 2 extend directly. Figure 3 shows that the estimation performs well; however, the loss of information can be large. Indeed, as λ increases, not only does the correlation between the true network and the constructed network decrease, but the linking probabilities become homogeneous. Then, it becomes hard to distinguish between the

[8] See Lancaster (2000) for a review.
[9] Then, $\mathbf{Jw}$ gives $\mathbf{w}$ minus the group average of $\mathbf{w}$.

(almost uniform) network effects and the group effects. In practice, we therefore expect our approach to perform well when the distribution of the true network exhibits heterogeneous linking probabilities.

Figure 3: Peer effect with fixed effects



The graph shows α estimates for 1000 replications of the model with fixed effects for various values of λ. The above x-axis reports the average correlation between two independent draws of graphs from the distribution given by equation (2). Complete estimates for $\lambda \in \{1, +\infty\}$ are presented in Tables 13 and 14 of the Appendix 9.1.

For example, Table 3 presents the estimation results when we assume that the network formation process is a function of the observed characteristics. Specifically:

$$p(a_{ij} = 1|\mathbf{x}_i, \mathbf{x}_j) = \Phi(-4.5 + |x_{i,1} - x_{j,1}| - 2|x_{i,2} - x_{j,2}|),$$

where Φ is the cumulative distribution for the standardized normal distribution. As such, the network features *heterophily* with respect to the first variable and *homophily* with respect to the second variable.[10] As anticipated, the estimation performs well. We now present our likelihood-based estimator.

4 Likelihood-Based Estimators

The approach developed in the previous section assumes that $\mathbf{GX}$ is observed. When it is not, the instrumental variable estimators fail. We therefore present a likelihood-based estimator.

[10] That is, individuals with different values of x_1 and similar values of x_2 are more likely to be linked.

Table 3: Simulation results with subpopulation unobserved fixed effects (3).

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - Instrument: $\mathbf{J}(\tilde{\mathbf{G}})^2\mathbf{X}$ - $\hat{\mathbf{y}}$ is observed					
Estimation results					
$\alpha = 0.4$	0.400	0.006	0.396	0.400	0.404
$\beta_1 = 1$	1.000	0.007	0.995	1.000	1.005
$\beta_2 = 1.5$	1.500	0.020	1.486	1.499	1.514
$\gamma_1 = 5$	5.000	0.008	4.995	5.000	5.005
$\gamma_2 = -3$	−2.999	0.030	−3.021	−2.998	−2.979
Tests					
F-test	1123.431	178.101	999.270	1116.900	1242.319
Hausman	1.039	1.503	0.114	0.472	1.289
Sargan	1.037	1.370	0.111	0.509	1.458
$N = 50$, $M = 100$ - Instrument: $\mathbf{J}(\hat{\mathbf{G}})^2\mathbf{X}$ - $\hat{\mathbf{y}}$ is not observed					
Estimation results					
$\alpha = 0.4$	0.399	0.015	0.389	0.398	0.408
$\beta_1 = 1$	1.002	0.013	0.994	1.002	1.011
$\beta_2 = 1.5$	1.418	0.054	1.380	1.419	1.453
$\gamma_1 = 5$	4.743	0.046	4.713	4.743	4.775
$\gamma_2 = -3$	−3.655	0.252	−3.843	−3.669	−3.490
$\hat{\gamma}_1 = 0$	0.256	0.046	0.224	0.255	0.286
$\hat{\gamma}_2 = 0$	0.788	0.280	0.609	0.794	0.987
Tests					
F-test	1153.330	200.889	1003.857	1147.411	1277.991
Hausman	161.862	60.319	117.502	154.631	197.888
Sargan	9.257	13.256	0.825	4.052	12.405

Note: Number of simulations: 1000. In each group, the fixed effect is generated as $0.3x_{1,1} + 0.3x_{3,2} - 1.8x_{50,2}$. The network's true distribution follows the network formation model, such that $p_{ij} = \Phi\left(-4.5 + |x_{i,1} - x_{j,1}| - 2|x_{i,2} - x_{j,2}|\right)$, where Φ represents the cumulative distribution function of $\mathcal{N}(0, 1)$.

Accordingly, more structure must be imposed on the errors $\boldsymbol{\varepsilon}$.[11]

To clarify the exposition, we will focus on the network adjacency matrix $\mathbf{A}$ instead of the interaction matrix $\mathbf{G}$. Of course, this is without any loss of generality. Given parametric assumptions for $\boldsymbol{\varepsilon}$, one can write the log-likelihood of the outcome as:[12]

$$\ln \mathcal{P}(\mathbf{y}|\mathbf{A}, \boldsymbol{\theta}), \tag{4}$$

where $\boldsymbol{\theta} = [\alpha, \boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\sigma}']'$, $\boldsymbol{\sigma}$ are unknown parameters from the distribution of $\boldsymbol{\varepsilon}$. Note that $\mathbf{y} = (\mathbf{I}_N - \alpha\mathbf{G})^{-1}(c\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon})$ and $(\mathbf{I}_N - \alpha\mathbf{G})^{-1}$ exist under our Assumption 1.1.

If the adjacency matrix $\mathbf{A}$ was observed, then (4) could be estimated using a simple maximum likelihood estimator (as in Lee et al. (2010)) or using Bayesian inference (as in Goldsmith-Pinkham and Imbens (2013)).

Since $\mathbf{A}$ is not observed, an alternative would be to focus on the unconditional likelihood, i.e.

$$\ln \mathcal{P}(\mathbf{y}|\boldsymbol{\theta}) = \ln \sum_{\mathbf{A}} \mathcal{P}(\mathbf{y}|\mathbf{A}, \boldsymbol{\theta}) P(\mathbf{A}).$$

A similar strategy is proposed by Chandrasekhar and Lewis (2011) using a GMM estimator.

One particular issue with estimating $\ln \mathcal{P}(\mathbf{y}|\boldsymbol{\theta})$ is that the summation is not tractable. Indeed, the sum is over the set of possible adjacency matrices, which contain $2^{N(N-1)}$ elements. Then, simply simulating networks from $P(\mathbf{A})$ and taking the average is likely to lead to poor approximations.[13] A classical way to address this issue is to use an EM algorithm (Dempster et al., 1977). The interested reader can consult Appendix 9.6 for a presentation of such an estimator. Although valid, we found that the Bayesian estimator proposed in this section is less restrictive and numerically outperforms its classical counterpart.

For concreteness, we will assume that $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_N)$; however, it should be noted that our

[11] Lee (2004) presents a quasi maximum-likelihood estimator that does not require such a specific assumption for the distribution of the error term. His estimator could be used alternatively. As well, as will be made clear, our approach can be used for a large class of extremum estimators, following Chernozhukov and Hong (2003), and in particular for GMM estimators, as in Chandrasekhar and Lewis (2011).

[12] Note that under Assumption 1.3, the likelihood can be factorized across groups.

[13] That is: $\ln \mathcal{P}(\mathbf{y}|, \boldsymbol{\theta}) \approx \ln \frac{1}{S} \sum_{s=1}^{S} \mathcal{P}(\mathbf{y}|\mathbf{A}_s, \boldsymbol{\theta})$, where $\mathbf{A}_s$ is drawn from $P(\mathbf{A})$. This is the approximation suggested by Chandrasekhar and Lewis (2011) (see their Section 4.3). In their case, they only need to integrate over the $m < N(N-1)$ pairs that are not sampled. Still, the number of compatible adjacency matrices is 2^m. As such, the approach is likely to produce bad approximations.

approach is valid for a number of alternative assumptions. We have for $\mathbf{G} = f(\mathbf{A})$,

$$\begin{aligned}\ln \mathcal{P}(\mathbf{y}|\mathbf{A},\boldsymbol{\theta}) &= -N\ln(\sigma) + \ln|\mathbf{I}_N - \alpha\mathbf{G}| - \frac{N}{2}\ln(\pi) \\ &\quad -\frac{1}{2\sigma^2}[(\mathbf{I}_N - \alpha\mathbf{G})\mathbf{y} - c\mathbf{1}_N - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\mathbf{X}\boldsymbol{\gamma}]'[(\mathbf{I}_N - \alpha\mathbf{G})\mathbf{y} - c\mathbf{1}_N - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\mathbf{X}\boldsymbol{\gamma}].\end{aligned}$$

Since $\mathbf{A}$ is not observed, we follow Tanner and Wong (1987) and Albert and Chib (1993), and we use data augmentation to evaluate the posterior distribution of $\boldsymbol{\theta}$. That is, instead of focusing on the posterior $p(\boldsymbol{\theta}|\mathbf{y},\mathbf{A},\mathbf{X})$, we focus on the posterior $p(\boldsymbol{\theta},\mathbf{A}|\mathbf{y},\mathbf{X})$, treating $\mathbf{A}$ as another set of unknown parameters.

Indeed, it is possible to obtain draws from $p(\boldsymbol{\theta},\mathbf{A}|\mathbf{y},\mathbf{X})$ using the following MCMC:

Algorithm 1. *The MCMC goes as follows for $t = 1,...,T$, starting from any $\mathbf{A}_0, \boldsymbol{\theta}_0$.*

1. *Propose $\mathbf{A}^*$ from the proposal distribution $q_A(\mathbf{A}^*|\mathbf{A}_{t-1})$ and accept $\mathbf{A}^*$ with probability*

$$\min\left\{1, \frac{\mathcal{P}(\mathbf{y}|\boldsymbol{\theta}_{t-1},\mathbf{A}^*)q_A(\mathbf{A}_{t-1}|\mathbf{A}^*)P(\mathbf{A}^*)}{\mathcal{P}(\mathbf{y}|\boldsymbol{\theta}_{t-1},\mathbf{A}_{t-1})q_A(\mathbf{A}^*|\mathbf{A}_{t-1})P(\mathbf{A}_{t-1})}\right\}.$$

2. *Draw α^* from the proposal $q_\alpha(\cdot|\alpha_{t-1})$ and accept α^* with probability*

$$\min\left\{1, \frac{\mathcal{P}(\mathbf{y}|\mathbf{A}_t;\boldsymbol{\beta}_{t-1},\boldsymbol{\gamma}_{t-1},\alpha^*)q_\alpha(\alpha_{t-1}|\alpha^*)P(\alpha^*)}{\mathcal{P}(\mathbf{y}|\mathbf{A}_t;\boldsymbol{\theta}_{t-1})q_\alpha(\alpha^*|\alpha_{t-1})P(\alpha_{t-1})}\right\}.$$

3. *Draw $[\boldsymbol{\beta},\boldsymbol{\gamma},\sigma]$ from their conditional distributions.*

Detailed distributions for Steps 2 and 3 can be found Appendix 9.4. Step 1, however, involves some additional complexities. Indeed, the idea is the following: starting from a given network formation model (i.e. $P(\mathbf{A})$), one has to be able to draw samples from the posterior distribution of $\mathbf{A}$, given $\mathbf{y}$. This is not a trivial task. The strategy used here is to rely on a Metropolis–Hastings algorithm, a strategy that has also been used in the related literature on ERGMs (e.g. Snijders (2002), Mele (2017)).

The acceptance probability in Step 1 of Algorithm 1 clearly exposes the role of the assumed distribution for the true network $P(\mathbf{A})$, i.e. the prior distribution of $\mathbf{A}$. This highlights the importance of $P(\mathbf{A})$ for the identification of the model. Since $\boldsymbol{\theta}$ and $\mathbf{A}$ are unobserved, we have $N(N-1)+k$ parameters to estimate, where k is the number of dimensions of $\boldsymbol{\theta}$. In particular, if $P(a_{ij} = 1) = 1/2$ for all i, j, then the probability of acceptance in Step 1 of Algorithm 1 reduces

to:

$$\min\left\{1, \frac{\mathcal{P}(\mathbf{y}|\boldsymbol{\theta}_{t-1}, \mathbf{A}^*)q_A(\mathbf{A}_{t-1}|\mathbf{A}^*)}{\mathcal{P}(\mathbf{y}|\boldsymbol{\theta}_{t-1}, \mathbf{A}_{t-1})q_A(\mathbf{A}^*|\mathbf{A}_{t-1})}\right\},$$

which only depends on the likelihood of the model and $q_A(\cdot|\cdot)$.[14] We explore the impact of the information encoded in $P(\mathbf{A})$ on the identification of $\boldsymbol{\theta}$ using Monte Carlo simulations later in this section.

One issue, however, is that there is no general rule for selecting the network proposal distribution $q_A(\cdot|\cdot)$. A natural candidate is a Gibbs sampling algorithm for each link, i.e. change only one link ij at every step t and propose a_{ij} according to its marginal distribution:

$$a_{ij} \sim P(\cdot|\mathbf{A}_{-ij}, \mathbf{y}) = \frac{\mathcal{P}(\mathbf{y}|a_{ij}, \mathbf{A}_{-ij})P(a_{ij}|\mathbf{A}_{-ij})}{\mathcal{P}(\mathbf{y}|1, \mathbf{A}_{-ij})P(a_{ij}=1|\mathbf{A}_{-ij}) + \mathcal{P}(\mathbf{y}|0, \mathbf{A}_{-ij})P(a_{ij}=0|\mathbf{A}_{-ij})},$$

where $\mathbf{A}_{-ij} = \{a_{kl}; k \neq i, l \neq j\}$. In this case, the proposal is always accepted.

However, it has been argued that Gibbs sampling could lead to slow convergence (e.g. Snijders (2002), Chatterjee et al. (2013)), especially when the network is *sparse* or exhibits a high level of *clustering*. For example, Mele (2017) and Bhamidi et al. (2008) propose different blocking techniques that are meant to improve convergence.

Here, however, the realization of Step 1 involves an additional computational issue since evaluating the likelihood ratio in Step 1 requires comparing the determinants $|\mathbf{I} - \alpha f(\mathbf{A}^*)|$ for each proposed $\mathbf{A}^*$, which is computationally intensive. In particular, taking $\mathbf{G}^* = f(\mathbf{A}^*)$ to be a row-normalization of $\mathbf{A}^*$, changing a single element of $\mathbf{A}^*$ results in a change in the entire corresponding row of $\mathbf{G}^*$. Still, comparing the determinant of two matrices that differ only in a single row is relatively fast. Moreover, when $\mathbf{G} = \mathbf{A}$, Hsieh et al. (2019a) propose a blocking technique that facilitates the computation of the determinant.

Since the appropriate blocking technique depends strongly on $P(\mathbf{A})$ and the assumed distribution for $\boldsymbol{\varepsilon}$, we use the Gibbs sampling algorithm for each link of the simulations and estimations presented in this paper, adapting the strategy proposed by Hsieh et al. (2019a) to our setting (see Proposition 3 in Appendix 9.3). This can be viewed as a *worse-case* scenario. We encourage researchers to try other updating schemes if Gibbs sampling performs poorly in their specific contexts. In particular, we present a blocking technique in Appendix 9.3 that is also implemented in our R package `PartialNetwork`.[15]

Table 4 presents the Monte Carlo simulations using Algorithm 1. The simulated population

[14] In this case, the model would not be identified since there would be more parameters to estimate than there are observations.

[15] Available at: https://github.com/ahoundetoungan/PartialNetwork.

is the same as in Section 3; however, for computational reasons, we limit ourselves to $M = 50$ groups of $N = 30$ individuals each. As expected, the average of the means of the posterior distributions are centred relatively on the parameters' true values. Note, however, that due to the smaller number of groups and the fact that we performed only 200 simulations, the results in Table 4 may exhibit small sample as well as simulation biases.

Table 4: Simulation results with a Bayesian method.

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 30$, $M = 50$					
Estimation results					
$Intercept = 2$	1.873	0.893	1.312	1.815	2.486
$\alpha = 0.4$	0.398	0.025	0.383	0.398	0.414
$\beta_1 = 1$	1.003	0.027	0.982	1.002	1.019
$\beta_2 = 1.5$	1.500	0.019	1.489	1.501	1.512
$\gamma_1 = 5$	5.011	0.167	4.909	5.009	5.117
$\gamma_2 = -3$	−2.987	0.135	−3.084	−2.983	−2.887
$\sigma^2 = 1$	1.018	0.113	0.946	1.016	1.089

Note: Simulation results for 200 replications of the model with unobserved exogenous effects estimated by a Bayesian method where the graph precision parameter λ is set to 1.

5 Network Formation Models

Our main assumption (Assumption 1.2) is that the researcher has access to the true distribution of the observed network. An important special case is when the researcher has access to a consistent estimate of this distribution. For concreteness, in this section we assume that links are generated as follows:

$$\mathbb{P}(a_{ij} = 1) \propto \exp\{Q(\boldsymbol{\theta}, \mathbf{w}_{ij})\}, \tag{5}$$

where Q is some known function, $\mathbf{w}_{ij}$ is a vector of (not necessarily observed) characteristics for the pair ij, and $\boldsymbol{\theta}$ is a vector of parameters to be estimated.

An important feature of such models is that their estimation may not necessarily require the observation of the entire network structure. To understand the intuition, assume a simple

logistic regression framework:

$$\mathbb{P}(a_{ij} = 1) = \frac{\exp\{\mathbf{x}_{ij}\boldsymbol{\theta}\}}{1 + \exp\{\mathbf{x}_{ij}\boldsymbol{\theta}\}},$$

where $\mathbf{x}_{ij}$ is a vector of *observed* characteristics of the pair ij. Here, note that $\mathbf{s} = \sum_{ij} a_{ij}\mathbf{x}_{ij}$ is a vector of sufficient statistics. In practice, this therefore means that the estimation of $\boldsymbol{\theta}$ only requires the observation of such sufficient statistics.

To clarify this point, consider a simple example where individuals are only characterized by their gender and age. Specifically, assume that $\mathbf{x}_{ij} = [1, \mathbb{1}\{gender_i = gender_j\}, |age_i - age_j|]$. Then, the set of sufficient statistics is resumed by (1) the number of links, (2) the number of same-gender links, and (3) average age difference between linked individuals.

Note that these statistics are much easier (and cheaper) to obtain than the entire network structure; however, they nonetheless allow for estimating the distribution of the true network.

Of course, in general, the simple logistic regression above might be unrealistically simple as the probability of linking might depend on unobserved variables.

In this section, we discuss some examples of network formation models that can be estimated using only partial information about the network. We subdivide such models into two categories: models that can be estimated using sampled network data and latent surface models.

5.1 Sampled Network

As discussed in Chandrasekhar and Lewis (2011), sampled data can be used to estimate a network formation model under the assumptions that (1) the sampling is exogenous and (2) links are conditionally independent, i.e. $P(a_{ij}|\mathbf{A}_{-ij}) = P(a_{ij})$, as in (5).

Indeed, if the sampling was done, for example, as a function of the network structure, the estimation of the network formation model would likely be biased. Also, if the network formation model is such that links are *not* conditionally independent, then consistent estimation usually requires the observation of the entire network structure.[16]

An excellent illustration of a compatible sampling scheme is presented in Conley and Udry (2010). Rather than collecting the entire network structure, the authors asked the respondents about their relationship with a random sample of the other respondents: "Have you ever gone to ____ for advice about your farm?" If the answer is "Yes," then a link is assumed between

[16] Or at least requires additional network summary statistics, such as individual degree or clustering coefficients; see Boucher and Mourifié (2017) or Mele (2017).

the respondents.

Since the pairs of respondents for which the "Yes/No" question is asked are random, the estimation of a network formation model with conditionally independent links gives consistent estimates. If, in addition, the individual characteristics of the sampled pair of respondents cover the set of observable characteristics for the entire set of respondents, one can compute the predicted probability that any two respondents are linked.

For concreteness, consider the simple model presented above, such that:

$$\mathbb{P}(a_{ij} = 1) = \frac{\exp\{\mathbf{x}_{ij}\boldsymbol{\theta}\}}{1 + \exp\{\mathbf{x}_{ij}\boldsymbol{\theta}\}},$$

where $\mathbf{x}_{ij} = [1, \mathbb{1}\{gender_i = gender_j\}, |age_i - age_j|]$.

Then, as long as the random sample of pairs for which the "Yes/No" question is asked includes both men and women and includes individuals of different ages, then these sampled pairs allow for a consistent estimation of $\boldsymbol{\theta}$. As such, for any two respondents the (predicted) probability of a link is given by $\hat{p}_{ij} = \exp\{\mathbf{x}_{ij}\hat{\boldsymbol{\theta}}\}/(1 + \exp\{\mathbf{x}_{ij}\hat{\boldsymbol{\theta}}\})$, where $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$.

The argument can be extended to models featuring an unobserved degree of heterogeneity. Specifically, Graham (2017) studies the following *undirected* network formation model:

$$\mathbb{P}(a_{ij} = 1) = \frac{\exp\{\mathbf{x}_{ij}\boldsymbol{\theta} + \nu_i + \nu_j\}}{1 + \exp\{\mathbf{x}_{ij}\boldsymbol{\theta} + \nu_i + \nu_j\}},$$

where ν_i and ν_j are unobserved. He presents a *tetrad logit* estimator based on the assumption that only a random sample of links are observed (his Assumption 2), as in Conley and Udry (2010).

Graham (2017) shows that $\boldsymbol{\theta}$ can be recovered consistently given some regularity conditions on the asymptotic behaviour of the model (his Assumption 4, which is implied by our Assumption 1.3). Once the consistent estimator $\hat{\boldsymbol{\theta}}$ is recovered, the predicted probabilities are given by:

$$\hat{P}(\mathbf{A}|\mathbf{n}) = \frac{\exp\{\sum_{ij:j<i} a_{ij}\mathbf{x}_{ij}\hat{\boldsymbol{\theta}}\}}{\sum_{B\in\mathcal{A}} \exp\{\sum_{ij:j<i} b_{ij}\mathbf{x}_{ij}\hat{\boldsymbol{\theta}}\}}, \tag{6}$$

where $\mathbf{n} = [n_1, ..., n_n]$ is the *degree sequence*, and $\mathcal{A}$ is the set of adjacency matrices that have the same degree sequence as $\mathbf{A}$, i.e. $n_i = \sum_{j\neq i} a_{ij} = \sum_{j\neq i} b_{ij}$ for all i and all $\mathbf{B} \in \mathcal{A}$ (see Graham (2017), equation (3)).

Note that computing $\hat{P}(\mathbf{A})$ therefore requires knowledge of the degree sequence, but this

information can easily be incorporated as a survey question: "How many people have you gone to for advice about your farm?" Also, as noted by Graham (2017), the computation of the normalizing term in (6) is not tractable for networks of moderate size. As such, the predicted probabilities cannot be computed directly and must be simulated, for example using the sequential importance sampling algorithm proposed by Blitzstein and Diaconis (2011).

5.2 Latent Surface Models

Recently, McCormick and Zheng (2015) and Breza et al. (2017) have proposed a novel approach for the estimation of network formation models represented by:

$$\mathbb{P}(a_{ij} = 1) \propto \exp\{\nu_i + \nu_j + \zeta \mathbf{z}_i' \mathbf{z}_j\}, \tag{7}$$

where ν_i, ν_j, ζ, $\mathbf{z}_i$, and $\mathbf{z}_j$ are not observed by the econometrician but follow parametric distributions. As in Graham (2017), ν_i and ν_j can be interpreted as i and j's propensity to create links, irrespective of the identity of the other individual involved. The other component, $\zeta \mathbf{z}_i' \mathbf{z}_j$, is meant to capture homophily on an abstract latent space (e.g. Hoff et al. (2002)).

Breza et al. (2017) show that it is possible to use aggregate relational data (ARD) to recover the values of the variables in (7) and therefore obtain an estimate of $\mathbb{P}(a_{ij} = 1)$. ARD are obtained from survey questions such as: "How many friends with trait 'X' do you have?" We refer the interested reader to McCormick and Zheng (2015) and Breza et al. (2017) for a formal discussion of the model. Here, we discuss the intuition using a simple analogy.

Suppose that individuals are located according to their geographical position on Earth. Suppose also that there are a fixed number of cities on Earth in which individuals can live. The econometrician does not know the individuals' location on Earth nor do they know the location of the cities. In model (7), $\mathbf{z}_i$ represent i's position on Earth.

Suppose that the researcher has data on ARD for a subset of the population. In the context of our example, ARD data are count variables of the type: "How many of your friends live in city A?"[17] Given (7) and parametric assumptions for the distribution of ν' and $\mathbf{z}$'s, the goal is to use ARD responses to infer the positions and sizes of the cities on Earth, as well as the values for ν_i and $\mathbf{z}_i$.[18]

To understand the intuition behind the identification of the model, consider the following

[17] The general approach works for any discrete characteristic.

[18] One also needs the ARD traits of the entire population, which is similar to our Assumption 1.4. See Section C.I, Step II in Breza et al. (2017) for details.

example: suppose that individual i has many friends living in city A. Then, city A is likely located close to i's location. Similarly, if many individuals have many friends living in city A, then city A is likely a large city. Finally, if i has many friends from many cities, i likely has a large ν_i.

As mentioned above, we refer the interested reader to McCormick and Zheng (2015) and Breza et al. (2017) for a formal description of the method as well as formal identification conditions. Here, we provide Monte Carlo simulations for the estimators developed in Section 3, assuming that the true network follows (7). The details of the Monte Carlo simulations can be found in Appendix 9.2.

We simulate 20 groups of 250 individuals each. Within each subpopulation, we simulate the ARD responses as well as a series of observable characteristics (e.g. cities). We then estimate the model in (7) and compute the implied probabilities, $\mathbb{P}(a_{ij} = 1)$, which we used as the distribution of our true network.[19] We estimate peer effects using the instrumental variable strategy presented in Section 3. Results are presented in Tables 5 and 6.

Results show that the method performs relatively well when $\mathbf{Gy}$ is observed but slightly less well when $\mathbf{Gy}$ is not observed and when the model allows for group-level unobservables. Note, however, that one potential issue with this specific network formation model is that it is based on a single population setting (i.e. there is only one Earth). The researcher should keep in mind that the method should only be used on medium- to large-sized groups.

If the method proposed by Breza et al. (2017) performs well, note that our instrumental variable estimator does not require the identification of the structural parameters in (7). Indeed, the procedure only requires a consistent estimate of the linking probabilities.

As such, we could alternatively use the approach recently proposed by Alidaee et al. (2020). They present an alternative estimation procedure for models with ARD that does not rely on the parametric assumption in equation (7). They propose a penalized regression based on a low-rank assumption. One main advantage of their estimator is that it allows for a wider class of model and ensures that the estimation is fast and easily implementable.[20]

As for most penalized regressions, the estimation requires the user to select a tuning parameter, which effectively controls the weight of the penalty. We found that the value recommended by the authors is too large in the context of (7), using our simulated values. Since the choice of this tuning parameter is obviously context dependent, we recommend choosing it using a cross-validation procedure.

[19] We fix $\zeta = 1.5$ (i.e. ζ is not estimated) to mitigate part of the small sample bias. See our discussion below.
[20] The authors developed user-friendly packages in R and Python. See their paper for links and details.

To explore the properties of their estimator in our context, we do the following. First, we simulate data using (7), using the same specification as for Tables 5 and 6. Second, we estimate the linking probabilities using their penalized regression under different tuning parameters, including the optimal (obtained through cross-validation) and the recommended parameter (taken from Alidaee et al. (2020)). Third, we estimate the peer-effect model using our instrumental variable estimator.

Table 16 of Appendix 9.1 presents the results under alternative tuning parameters when $\mathbf{Gy}$ is observed and $\boldsymbol{\gamma} = 0$. We see that the procedure performs well but is less precise than when using the parametric estimation procedure. This is intuitive since the estimation procedure in Breza et al. (2017) imposes more structure (and is specified correctly in our context). The procedure proposed by Alidaee et al. (2020) is valid for a large class of models but is less precise.

Tables 17 and 18 also exemplify the results of Propositions 1 and 2. When $\mathbf{Gy}$ is observed, the estimation is precise, even if the network formation model is not estimated precisely. However, when $\mathbf{Gy}$ is not observed, small sample bias strongly affects the performance of the estimator.

Results from this section imply that, when $\mathbf{Gy}$ is observed, the estimator proposed by Alidaee et al. (2020) is the most attractive since it is less likely to be misspecified (and correlated with $\boldsymbol{\varepsilon}$). However, when $\mathbf{Gy}$ is not observed, the estimator from Breza et al. (2017) should be privileged. Of course, in the latter case, the validity of the results are based on the assumption that (7) is correctly specified.

6 Imperfectly Measured Networks

In this section, we assume that the econometrician has access to network data but that the data may contain errors. For example, Hardy et al. (2019) assume that some links are missing with some probability, while others are included falsely with some other probability.

To show how our method can be used to address these issues, we consider a simple example where we are interested in estimating peer effects on adolescents' academic achievements. We assume that we observe the network but that some links are missing.

To estimate this model, we use the AddHealth database. Specifically, we focus on a subset of schools from the "In School" sample that each have less than 200 students. Table 7 displays the summary statistics.

Most of the papers estimating peer effects that use this particular database have taken the network structure as given. One notable exception is Griffith (2019), looking at the issue of

Table 5: Simulation results using ARD with contextual effects (1000 replications).

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 250$, $M = 20$ - Instrument: $(\tilde{\mathbf{G}})^2\mathbf{X}$ - $\boldsymbol{Gy}$ is observed					
Estimation results					
$Intercept = 2$	1.991	0.222	1.845	1.992	2.141
$\alpha = 0.4$	0.400	0.006	0.396	0.400	0.404
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.005	1.496	1.500	1.504
$\gamma_1 = 5$	5.000	0.020	4.986	4.999	5.013
$\gamma_2 = -3$	−2.999	0.032	−3.020	−2.998	−2.977
Tests					
F-test	5473.171	1735.035	4232.103	5325.774	6528.537
Hausman	0.986	1.346	0.106	0.475	1.291
Sargan	1.045	1.461	0.108	0.465	1.353
$N = 250$, $M = 20$ - Instrument: $(\hat{\mathbf{G}})^2\mathbf{X}$ - $\boldsymbol{Gy}$ is not observed					
$Intercept = 2$	2.065	0.327	1.852	2.051	2.275
$\alpha = 0.4$	0.399	0.008	0.394	0.400	0.405
$\beta_1 = 1$	1.002	0.003	1.000	1.002	1.004
$\beta_2 = 1.5$	1.499	0.006	1.495	1.499	1.503
$\gamma_1 = 5$	5.411	0.020	5.397	5.411	5.425
$\gamma_2 = -3$	−2.403	0.040	−2.429	−2.402	−2.375
$\hat{\gamma}_1 = 0$	−0.383	0.023	−0.399	−0.384	−0.367
$\hat{\gamma}_2 = 0$	−0.608	0.038	−0.635	−0.609	−0.583
Tests					
F-test	4790.020	1407.596	3760.700	4682.825	5686.841
Hausman	70.940	19.503	57.143	70.430	82.175
Sargan	1.167	1.615	0.103	0.523	1.534
F-test	3,867.077	1,093.165	3,037.776	3,855.692	4,588.458
Hausman	228.290	49.002	194.110	227.981	261.617
Sargan	26.953	13.515	17.184	25.380	34.583

Note: Results without contextual effects are presented in Table 15 of Appendix 9.1.

Table 6: Simulation results with ARD and subpopulation unobserved fixed effects (1000 replications).

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 250$, $M = 20$ - Instrument: $\mathbf{J}(\tilde{\mathbf{G}})^2\mathbf{X}$ - $\boldsymbol{Gy}$ is observed					
Estimation results					
$\alpha = 0.4$	0.401	0.054	0.366	0.400	0.436
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
$\gamma_1 = 5$	4.999	0.067	4.956	4.999	5.043
$\gamma_2 = -3$	−3.001	0.082	−3.051	−2.999	−2.949
Tests					
F-test	169.973	55.243	130.603	165.188	205.152
Hausman	0.978	1.306	0.096	0.438	1.360
Sargan	0.919	1.393	0.084	0.402	1.228
$N = 250$, $M = 20$ - Instrument: $\mathbf{J}(\hat{\mathbf{G}})^2\mathbf{X}$ - $\boldsymbol{Gy}$ is not observed					
Estimation results					
$\alpha = 0.4$	0.477	0.077	0.426	0.474	0.528
$\beta_1 = 1$	1.001	0.003	0.999	1.001	1.003
$\beta_2 = 1.5$	1.499	0.007	1.495	1.499	1.504
$\gamma_1 = 5$	5.401	0.024	5.386	5.400	5.416
$\gamma_2 = -3$	−2.399	0.040	−2.425	−2.401	−2.374
$\hat{\gamma}_1 = 0$	−0.467	0.087	−0.524	−0.465	−0.408
$\hat{\gamma}_2 = 0$	−0.721	0.119	−0.795	−0.721	−0.641
Tests					
F-test	125.898	43.107	94.650	122.580	150.436
Hausman	23.096	11.431	15.100	21.238	29.881
Sargan	1.096	1.611	0.131	0.479	1.369

Table 7: Summary statistics.

Statistic	Mean	Std. Dev.	Pctl(25)	Pctl(75)
Female	0.540	0.498	0	1
Hispanic	0.157	0.364	0	0
Race				
White	0.612	0.487	0	1
Black	0.246	0.431	0	0
Asian	0.022	0.147	0	0
Other	0.088	0.283	0	0
Mother education				
High	0.310	0.462	0	1
$<$ High	0.193	0.395	0	0
$>$ High	0.358	0.480	0	1
Missing	0.139	0.346	0	0
Mother job				
Stay-at-home	0.225	0.417	0	0
Professional	0.175	0.380	0	0
Other	0.401	0.490	0	1
Missing	0.199	0.399	0	0
Age	13.620	1.526	13	14
GPA	2.088	0.794	1.5	2.667

top coding.[21] In practice, even if the schools are meant to be entirely sampled, we are still potentially missing many links. To understand why, we discuss the organization of the data.

Each adolescent is assigned to a unique identifier. The data includes ten variables for the ten potential friendships (maximum of 5 male and 5 female friends). These variables can contain missing values (no friendship was reported), an error code (the named friend could not be found in the database), or an identifier for the reported friends. This data is then used to generate the network's adjacency matrix $\mathbf{A}$.

Of course, error codes cannot be matched to any particular adolescent; as well, even in the case where the friendship variable refers to a valid identifier, however, the referred adolescent may still be absent from the database. A prime example is when the referred adolescent has been removed from the database by the researcher, perhaps due to other missing variables for these particular individuals. These missing links are quantitatively important as they account for roughly 30% of the total number of links (7,830 missing for 17,993 observed links). Figure 4 displays the distribution of the number of "unmatched named friends".[22]

[21] Although we are not exploring this issue here, our method can also be applied to analyse the impact of top coding.

[22] We focus on within-school friendships; thus, nominations outside of school are discarded.

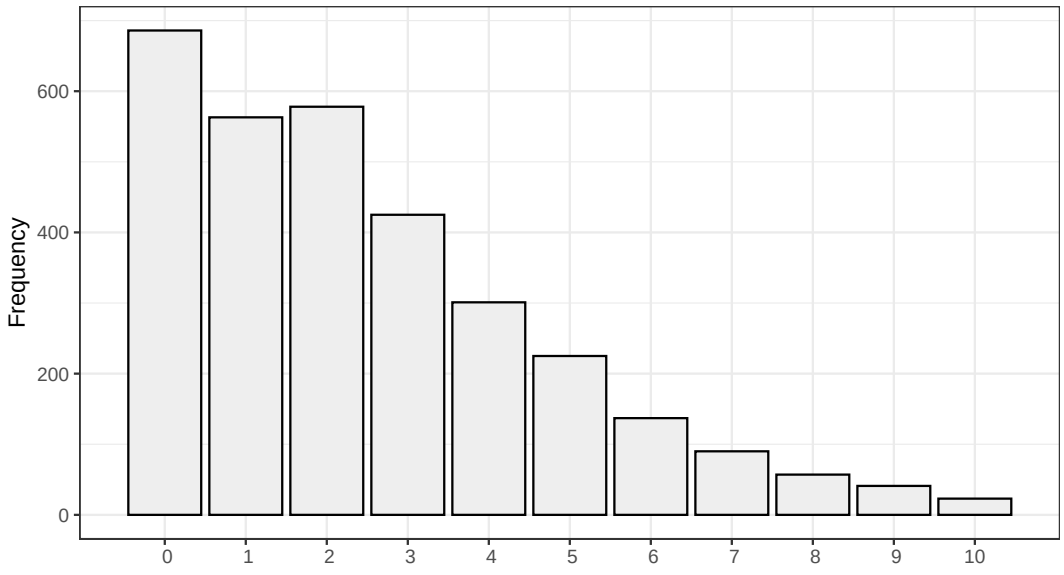


Figure 4: Frequencies of the number of missing links per adolescent.

Given that we observe the number of missing links for each individual, we can use the general estimator proposed in Section 4 to estimate the model. For this, however, we need one additional assumption.

Let N_s be the size of school s, n_i be the number of observed (matched) friends of i, and $\check{n}_i$ be the number of missing (unmatched) friends of i. We make the following assumption for the network formation process:

$$\begin{aligned} \mathbb{P}(a_{ij}=1) &= 1 \text{ if the link is observed in the data,} \\ \mathbb{P}(a_{ij}=1) &= \frac{\check{n}_i}{N_s-1-n_i} \text{ otherwise.} \end{aligned}$$

In large schools, this is equivalent to assuming that missing friendships are drawn at random among the remaining adolescents. In small schools, however, this approach disregards the dependence between links.

To understand why we cannot directly assume that the $\check{n}_i$ missing links are assigned randomly to the adolescents with whom i is not a friend, consider the following simple example. Suppose that there are only three adolescents, i, j, and k, and that i has no matched (observed) friends, but we know that they have one friend (i.e. $\check{n}_i = 1$). This friend can either be j or k. Selecting friendship relations at random therefore implies that a_{ij} and a_{ik} are perfectly (negatively)

correlated. This will lead many Metropolis–Hastings algorithms to fail. In particular, this is the case of the Gibbs sampling procedure.[23]

Assuming $\mathbb{P}(a_{ij} = 1) = \mathbb{P}(a_{ik} = 1) = \dfrac{\check{n}_i}{N_s - 1 - n_i} = 1/2$ for unobserved links circumvents this issue. Moreover, since this assumption only affects the prior distribution of our Bayesian inference procedure, it need not have important consequences on the posterior distribution. Table 8 presents the estimation results.[24] Importantly, we see that the estimated value for α is significantly larger when the network is reconstructed. Also notable is the fact that the reconstructed network captures an additional contextual peer effect: having a larger fraction of Hispanic friends significantly reduces academic achievement.[25] The remainder of the estimated parameters are roughly the same for both specifications.

7 Discussions

In this paper, we proposed two types of estimators that can estimate peer effects given that the researcher only has access to the distribution of the true network. In doing so, we abstracted from many important considerations. In this section, we discuss some limits, some areas for future research, and some general implications of our results.

7.1 Endogenous Networks

In this paper, we assumed away any endogeneity of the network structure (Assumption 1.5). As discussed in Section 2, this is done for the purposes of presentation. Indeed, there are mulitple ways to introduce, and correct for, endogenous networks, which lead to many possible models. In this section, we discuss how existing endogenous network corrections can be adapted to our setting. Specifically, we assume that there exists some unobserved variable correlated with both the network $\mathbf{A}$ and the outcome $\mathbf{y}$. This violates Assumption 1.5.

A first remark is that our ability to accommodate such an unobserved variable depends on the flexibility of the used network formation model. Indeed, some models can obtain estimates of the unobserved heterogeneity (e.g. Breza et al. (2017) or Graham (2017)). In such cases, one could simply include the estimated unobserved variables as additional explanatory variables. Johnsson and Moon (2015) discuss this approach as well as other control-function approaches

[23] This issue could be solved by updating an entire line of $\mathbf{A}$ for each step of the algorithm. However, this proves to be computationally intensive for networks of moderate size.

[24] Trace plots and posterior distributions are presented in Figures 5, 6, and 7 of Appendix 9.5.

[25] Note that one has to be careful in discussing the structural interpretation of the contextual effects. See Boucher and Fortin (2016) for a discussion.

Table 8: Posterior distribution.

	Observed network			Reconstructed network		
Statistic	Mean	Std. Dev.	t-stat	Mean	Std. Dev.	t-stat
Peer effect	0.350^{***}	0.022	15.519	0.526^{***}	0.036	14.468
Intercept	1.196^{***}	0.132	9.086	1.153^{***}	0.139	8.293
Own effects						
Female	-0.144^{***}	0.029	-5.013	-0.131^{***}	0.028	-4.690
Hispanic	0.084^{**}	0.042	1.999	0.122^{***}	0.044	2.745
Race						
Black	0.231^{***}	0.045	5.084	0.233^{***}	0.047	4.913
Asian	0.090	0.090	1.008	0.106	0.087	1.212
Other	-0.056	0.051	-1.085	-0.044	0.052	-0.862
Mother education						
$<$ High	0.122^{***}	0.039	3.125	0.120^{***}	0.038	3.136
$>$ High	-0.139^{***}	0.033	-4.165	-0.111^{***}	0.034	-3.317
Missing	0.060	0.051	1.179	0.058	0.052	1.131
Mother job						
Professional	-0.081^{*}	0.044	-1.833	-0.068	0.044	-1.551
Other	-0.003	0.035	-0.078	0.009	0.034	0.253
Missing	0.066	0.047	1.411	0.067	0.047	1.426
Age	0.073^{***}	0.009	7.749	0.076^{***}	0.010	7.824
Contextual effects						
Female	-0.012	0.049	-0.234	-0.048	0.071	-0.679
Hispanic	-0.061	0.069	-0.876	-0.275^{***}	0.091	-3.042
Race						
Black	-0.051	0.058	-0.868	-0.097	0.065	-1.478
Asian	-0.212	0.184	-1.150	-0.131	0.324	-0.404
Other	0.138	0.090	1.539	0.154	0.140	1.099
Mother education						
$<$ High	0.269^{***}	0.072	3.733	0.337^{***}	0.109	3.093
$>$ High	-0.071	0.060	-1.196	-0.118	0.086	-1.377
Missing	0.078	0.094	0.828	0.013	0.146	0.089
Mother job						
Professional	0.109	0.081	1.352	0.060	0.115	0.519
Other	0.101^{*}	0.060	1.684	0.057	0.088	0.647
Missing	0.093	0.087	1.080	0.055	0.136	0.400
Age	-0.066^{***}	0.006	-11.583	-0.087^{***}	0.008	-11.178
SE^2	0.523			0.493		

Note: N = 3,126. Observed links = 17,993. Missing links = 7,830. Significance levels: *** = 1%, ** = 5%, * = 10%.

in detail. Since their instrumental variable estimator is based on the higher-order link relations (i.e. $\mathbf{G}^2\mathbf{X}, \mathbf{G}^3\mathbf{X}, ...$), it is also valid for the instrumental variable estimator proposed in Section 3.[26]

However, this is not entirely satisfying since it assumes that the network formation is estimated consistently, independent of the peer-effect model.[27] The Bayesian estimator presented in Section 4 allows for more flexibility. Indeed, one could expand Algorithm 1 and perform the estimation of the network formation model jointly with the peer-effect model, as has been done, for instance, by Goldsmith-Pinkham and Imbens (2013), Hsieh and Van Kippersluis (2018), and Hsieh et al. (2019b)) in contexts where the network is observed. Essentially, instead of relying on the known distribution $P(\mathbf{A})$, one would need to rely on $P(\mathbf{A}|\mathbf{S}, \boldsymbol{\kappa}, \mathbf{z})$, where $\mathbf{S}$ is a matrix (possibly a vector) of observed statistics about $\mathbf{A}$ (e.g. a sample, a vector of summary statistics, or ARD), $\mathbf{z}$ is an unobserved latent variable (correlated with $\boldsymbol{\varepsilon}$), and $\boldsymbol{\kappa}$ is a vector of parameters to be estimated.

This contrasts, for example, with the network formation model presented in Section 5.2 where $\mathbf{S}$ (i.e. ARD) is sufficient for the estimation of all of the models' parameters. Here, the estimation of $\mathbf{z}$ and $\boldsymbol{\kappa}$ also requires knowledge about the likelihood of $\mathbf{y}$. The exploration of such models (especially their identification) goes far beyond the scope of the current paper and is left for future (exciting) research.

7.2 Large Populations and Partial Sampling

In this section, we discuss the estimation of (1) when the population cannot be partitioned into groups of bounded sizes (i.e. when Assumption 1.3 does not hold). Note that doing so will also most likely violate Assumption 1.4. Indeed, in large populations (e.g. cities, countries...), assuming that the individuals' characteristics $(\mathbf{y}, \mathbf{X})$ are observed for the entire population is unrealistic (that is, except for census data). This implies that strategies such as the one presented in Section 4 would also be unfeasible, irrespective of the network formation model.[28]

One would therefore have to rely on an instrumental strategy, such as the one presented in Section 3. In this section, we discuss the properties of the estimator in Section 3 in the context of a single, large sample.

To fix the discussion, assume for now that $\mathbf{x}_i = x_i$ has only one dimension and only takes

[26] Note that estimated unobserved variables can also be added as explanatory variables in the context of the estimator presented in Section 4.

[27] See, for example, Assumption 1 and Assumptions 6–11 in Johnsson and Moon (2015).

[28] Also, in terms of computational cost, the estimator presented in Section 4 is likely to be very costly for very large populations.

on a finite number of distinct values. Assume also that for any two i, j, $P(a_{ij} = 1) = \phi(x_i, x_j)$ for some known function ϕ. We have:

$$(\mathbf{Gx})_i = \sum_{j=1}^{N} g_{ij} x_j,$$

which we can rewrite as:

$$\sum_{j=1}^{N} g_{ij} x_j = \frac{\sum_x x (n_x/N) \frac{1}{n_x} \sum_{j:x_j=x} a_{ij}}{\sum_x (n_x/N) \frac{1}{n_x} \sum_{j:x_j=x} a_{ij}},$$

where n_x is the number of individuals having the trait x. Note that since we assumed that the support of x only takes a finite number of values, n_x goes to infinity with N. We therefore have (by a strong law of large numbers):

$$\frac{1}{n_x} \sum_{j:x_j=x} a_{ij} \to \phi(x_i, x), \tag{8}$$

and $n_x/N \to p(x)$, where $p(x)$ is the fraction of individuals with trait x in the population. We therefore have:

$$(\mathbf{Gx})_i \to \frac{\sum_x x p(x) \phi(x_i, x)}{\sum_x p(x) \phi(x_i, x)} \equiv \hat{x}(x_i).$$

For example, for an Erdös–Rényi network (i.e. $\phi(x, x') = \bar{\phi}$ for all x, x'), we have $\hat{x}(x_i) = \mathbb{E}x$.

This means that the knowledge of $\phi(\cdot, \cdot)$ is sufficient to construct a *consistent* estimate of $\mathbf{Gx}$. Note also that a similar argument allows constructing a consistent estimate for $\mathbf{Gy}$ or $\mathbf{G}^2\mathbf{X}$. As such, the instrumental variable strategy proposed in Section 3 can be applied even if $\mathbf{GX}$ is not observed.

Unfortunately, this approach relies on the (perhaps unrealistic) assumption that $P(a_{ij} = 1) = \phi(x_i, x_j)$, which here implies that individuals form an (asymptotically) infinite number of links.[29] When the number of links is bounded (e.g. De Paula et al. (2018b)), the average in (8) does not converge for each i. Note, however, that the first part of Proposition 2 still applies: if $\mathbf{Gy}$ and $\mathbf{GX}$ are observed, then the constructed (biased) instrument $\mathbf{H}^2\mathbf{X}$ drawn from a network formation model with bounded degrees is valid.

Finally, note that the argument presented here can be generalized. In particular, Parise and Ozdaglar (2019) recently proposed a means of approximating games on large networks using *graphon games*, i.e. games played directly on the network formation model. If the approach

[29] A special case, when the network is complete, is presented in Brock and Durlauf (2001).

is promising, its implications for the estimation of peer effects go far beyond the scope of this paper and are left for future research.

7.3 Survey Design

As discussed in Section 3, instrumental variable estimators are only valid if the researcher observes $\mathbf{GX}$. Also, Breza et al. (2017) and Alidaee et al. (2020) propose using ARD to estimate network formation models. Importantly, although ARD responses and $\mathbf{GX}$ are similar, they are not equivalent. For example, consider a binary variable (e.g. gender). One can obtain $\mathbf{GX}$ by asking questions such as "What fraction of your friends are female?" For ARD, the question would be "How many of your friends are female?" This suggests asking two questions. One related to the number of female friends and one related to the number of friends.[30]

For continuous variables (e.g. age), this creates additional issues. One can obtain $\mathbf{GX}$ by asking about the average age of one's friends, but ARD questions must be discrete: "How many of your friends are in the same age group as you?" Then, in practice, an approach could be to ask individuals about the number of friends they have, as well as the number of friends they have from multiple age groups: "How many of your friends are between X and Y years old?" Using this strategy allows construction of both the ARD and $\mathbf{GX}$.

Finally, an implication of Propositions 1 and 2 is that asking directly for $\mathbf{Gy}$ in the survey leads to a more robust estimation strategy. Indeed, the constructed instruments are valid even if the network formation model is misspecified.

7.4 Next Steps

In this paper, we proposed two estimators where peer effects can be estimated without having knowledge of the entire network structure. We found that, perhaps surprisingly, even very partial information on network structure is sufficient. However, there remains many important challenges, in particular with respect to the study of compatible models of network formation.

[30] Breza et al. (2017) do not require information on the number of friends, although this significantly helps the estimation.

References

Albert, J. H. and S. Chib (1993): "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, 88, 669–679.

Alidaee, H., E. Auerbach, and M. P. Leung (2020): "Recovering Network Structure from Aggregated Relational Data using Penalized Regression," *arXiv preprint arXiv:2001.06052*.

Atchadé, Y. F. and J. S. Rosenthal (2005): "On adaptive markov chain monte carlo algorithms," *Bernoulli*, 11, 815–828.

Bhamidi, S., G. Bresler, and A. Sly (2008): "Mixing time of exponential random graphs," in *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, IEEE, 803–812.

Blitzstein, J. and P. Diaconis (2011): "A sequential importance sampling algorithm for generating random graphs with prescribed degrees," *Internet mathematics*, 6, 489–522.

Boucher, V. and B. Fortin (2016): "Some challenges in the empirics of the effects of networks," *The Oxford Handbook on the Economics of Networks*, 277–302.

Boucher, V. and I. Mourifié (2017): "My friend far, far away: a random field approach to exponential random graph models," *The econometrics journal*, 20, S14–S46.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009): "Identification of peer effects through social networks," *Journal of econometrics*, 150, 41–55.

——— (2019): "Peer Effects in Networks: A Survey," *Annual Review of Economics, forthcoming*.

Breza, E. (2016): "Field Experiments, Social Networks, and Development," *The Oxford Handbook on the Economics of Networks*, 412–439.

Breza, E., A. G. Chandrasekhar, T. H. McCormick, and M. Pan (2017): "Using Aggregated Relational Data to feasibly identify network structure without network data," Tech. rep., National Bureau of Economic Research.

Brock, W. A. and S. N. Durlauf (2001): "Discrete choice with social interactions," *The Review of Economic Studies*, 68, 235–260.

Chandrasekhar, A. and R. Lewis (2011): "Econometrics of sampled networks," *Unpublished manuscript, MIT.[422]*.

Chatterjee, S., P. Diaconis, et al. (2013): "Estimating and understanding exponential random graph models," *The Annals of Statistics*, 41, 2428–2461.

Chen, X., Y. Chen, and P. Xiao (2013): "The impact of sampling and network topology on the estimation of social intercorrelations," *Journal of Marketing Research*, 50, 95–110.

Chernozhukov, V. and H. Hong (2003): "An MCMC approach to classical estimation," *Journal of Econometrics*, 115, 293–346.

Chib, S. and S. Ramamurthy (2010): "Tailored randomized block MCMC methods with application to DSGE models," *Journal of Econometrics*, 155, 19–38.

Conley, T. G. and C. R. Udry (2010): "Learning about a new technology: Pineapple in Ghana," *American economic review*, 100, 35–69.

De Paula, A. (2017): "Econometrics of network models," in *Advances in Economics and Econometrics: Theory and Applications: Eleventh World Congress (Econometric Society Monographs*, ed. by M. P. B. Honore, A. Pakes and L. Samuelson, Cambridge: Cambridge University Press, 268–323.

De Paula, Á., I. Rasul, and P. Souza (2018a): "Recovering social networks from panel data: identification, simulations and an application," *LACEA Working Paper Series*.

De Paula, Á., S. Richards-Shubik, and E. Tamer (2018b): "Identifying preferences in networks with bounded degree," *Econometrica*, 86, 263–288.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, 39, 1–22.

Goldsmith-Pinkham, P. and G. W. Imbens (2013): "Social networks and the identification of peer effects," *Journal of Business & Economic Statistics*, 31, 253–264.

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

Griffith, A. (2018): "Random assignment with non-random peers: A structural approach to counterfactual treatment assessment," *Working Paper*.

——— (2019): "Name Your Friends, but Only Five? The Importance of Censoring in Peer Effects Estimates using Social Network Data," *Working Paper*.

Hardy, M., R. M. Heath, W. Lee, and T. H. McCormick (2019): "Estimating spillovers using imprecisely measured networks," *arXiv preprint arXiv:1904.00136*.

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002): "Latent space approaches to social network analysis," *Journal of the American Statistical Association*, 97, 1090–1098.

Hsieh, C.-S., S. I. Ko, J. Kovářík, and T. Logan (2018): "Non-Randomly Sampled Networks: Biases and Corrections," Tech. rep., National Bureau of Economic Research.

Hsieh, C.-S., M. D. König, and X. Liu (2019a): "A Structural Model for the Coevolution of Networks and Behavior," *CEPR Discussion Paper No. DP13911*.

Hsieh, C.-S., L.-F. Lee, and V. Boucher (2019b): "Specification and estimation of network formation and network interaction models with the exponential probability distribution," .

Hsieh, C.-S. and H. Van Kippersluis (2018): "Smoking initiation: Peers and personality," *Quantitative Economics*, 9, 825–863.

Johnson, C. R. and R. A. Horn (1985): *Matrix analysis*, Cambridge University Press Cambridge.

Johnsson, I. and H. R. Moon (2015): "Estimation of peer effects in endogenous social networks: control function approach," *Review of Economics and Statistics*, 1–51.

Kelejian, H. H. and G. Piras (2014): "Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes," *Regional Science and Urban Economics*, 46, 140–149.

Kelejian, H. H. and I. R. Prucha (1998): "A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances," *The Journal of Real Estate Finance and Economics*, 17, 99–121.

Lancaster, T. (2000): "The incidental parameter problem since 1948," *Journal of econometrics*, 95, 391–413.

Lee, L.-F. (2004): "Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models," *Econometrica*, 72, 1899–1925.

Lee, L.-f., X. Liu, and X. Lin (2010): "Specification and estimation of social interaction models with network structures," *The Econometrics Journal*, 13, 145–176.

Lee, L.-f., X. Liu, E. Patacchini, and Y. Zenou (2020): "Who is the key player? A network analysis of juvenile delinquency," *ournal of Business and Economic Statistics, forthcoming.*

Lewbel, A., X. Qu, and X. Tang (2019): "Social networks with misclassified or unobserved links," *Working Paper.*

Liu, X. (2013): "Estimation of a local-aggregate network model with sampled networks," *Economics Letters*, 118, 243–246.

Liu, X., E. Patacchini, and E. Rainone (2017): "Peer effects in bedtime decisions among adolescents: a social network model with sampled data," *The econometrics journal*, 20, S103–S125.

Manresa, E. (2016): "Estimating the Structure of Social Interactions Using Panel Data," *Working paper.*

Manski, C. F. (1993): "Identification of endogenous social effects: The reflection problem," *Review of Economic Studies*, 60, 531–542.

McCormick, T. H. and T. Zheng (2015): "Latent surface models for networks using Aggregated Relational Data," *Journal of the American Statistical Association*, 110, 1684–1695.

Mele, A. (2017): "A structural model of Dense Network Formation," *Econometrica*, 85, 825–850.

Parise, F. and A. E. Ozdaglar (2019): "Graphon Games: A Statistical Framework for Network Games and Interventions," *Available at SSRN 3437293.*

Snijders, T. A. (2002): "Markov chain Monte Carlo estimation of exponential random graph models," *Journal of Social Structure*, 3, 1–40.

Souza, P. (2014): "Estimating network effects without network data," *PUC-Rio Working Paper.*

Tanner, M. A. and W. H. Wong (1987): "The calculation of posterior distributions by data augmentation," *Journal of the American statistical Association*, 82, 528–540.

Thirkettle, M. (2019): "Identification and Estimation of Network Statistics with Missing Link Data," *Working Paper.*

Wang, W. and L.-F. Lee (2013): "Estimation of spatial autoregressive models with randomly missing data in the dependent variable," *The Econometrics Journal*, 16, 73–102.

8 Appendix – Proofs

8.1 Proof of Proposition 1

The model can be written as:

$$\mathbf{y} = [\mathbf{I} - \alpha\mathbf{G}]^{-1}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}].$$

Or, using the geometric expansion:

$$\mathbf{y} = \sum_{k=0}^{\infty} \alpha^k \mathbf{G}^k \mathbf{X}\boldsymbol{\beta} + \sum_{k=0}^{\infty} \alpha^k \mathbf{G}^k \boldsymbol{\varepsilon}$$

or

$$\mathbf{G}\mathbf{y} = \sum_{k=0}^{\infty} \alpha^k \mathbf{G}^{k+1} \mathbf{X}\boldsymbol{\beta} + \sum_{k=0}^{\infty} \alpha^k \mathbf{G}^{k+1} \boldsymbol{\varepsilon}.$$

As such, any variable correlated with $\mathbf{GX}$, $\mathbf{G}^2\mathbf{X}$,... is also correlated with $\mathbf{Gy}$, conditional on $\mathbf{X}$. It remains to show that such variables are valid. For the first part of Proposition 1, we need:

$$\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}, \mathbf{HX}, \mathbf{H}^2\mathbf{X}, ...] = 0,$$

which is true by assumption. For the second part of Proposition 1, we need:

$$\mathbb{E}[\boldsymbol{\varepsilon} + \boldsymbol{\eta}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ...] = 0.$$

By Assumption 1.5, this is equivalent to:

$$\mathbb{E}[\boldsymbol{\eta}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ...] = 0,$$

which is equivalent to:

$$\mathbb{E}[\mathbf{Gy}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ...] = \mathbb{E}[\tilde{\mathbf{G}}\mathbf{y}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ...].$$

Using iterated expectations:

$$\mathbb{E}_{\mathbf{y}}\mathbb{E}[\mathbf{Gy}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ..., \mathbf{y}]|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ... = \mathbb{E}[\mathbf{G}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ...]\mathbb{E}[\mathbf{y}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ...],$$

and so $\mathbb{E}[\mathbf{Gy}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ...] = \mathbb{E}[\tilde{\mathbf{G}}\mathbf{y}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ...]$ follows since $\mathbf{G}$ $\hat{\mathbf{G}}$ and $\tilde{\mathbf{G}}$ are iid, which implies that $\mathbb{E}[\mathbf{G}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ...] = \mathbb{E}[\tilde{\mathbf{G}}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ...]$.

8.2 Proof of Proposition 2

The model can be written as:

$$\mathbf{y} = [\mathbf{I} - \alpha\mathbf{G}]^{-1}[\mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}].$$

Or, using the geometric expansion:

$$\mathbf{y} = \sum_{k=0}^{\infty} \alpha^k \mathbf{G}^k \mathbf{X}\boldsymbol{\beta} + \sum_{k=0}^{\infty} \alpha^k \mathbf{G}^k + \mathbf{G}\mathbf{X}\boldsymbol{\gamma} \sum_{k=0}^{\infty} \alpha^k \mathbf{G}^k \boldsymbol{\varepsilon}$$

or

$$\mathbf{G}\mathbf{y} = \sum_{k=0}^{\infty} \alpha^k \mathbf{G}^{k+1} \mathbf{X}\boldsymbol{\beta} + \sum_{k=0}^{\infty} \alpha^k \mathbf{G}^{k+2} \mathbf{X}\boldsymbol{\gamma} + \sum_{k=0}^{\infty} \alpha^k \mathbf{G}^{k+1} \boldsymbol{\varepsilon}.$$

As such, any variable correlated with $\mathbf{G}^2\mathbf{X}$, $\mathbf{G}^3\mathbf{X}$,... is also correlated with $\mathbf{G}\mathbf{y}$, conditional on $\mathbf{X}$ and $\mathbf{G}\mathbf{X}$. It remains to show that such variables are valid. For the first part of Proposition 2, we need:

$$\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}, \mathbf{G}\mathbf{X}, \mathbf{H}^2\mathbf{X}, \mathbf{H}^3\mathbf{X}, ...] = 0,$$

which is true by assumption. For the second part of Proposition 1, we need:

$$\mathbb{E}[\boldsymbol{\varepsilon} + \boldsymbol{\eta}|\mathbf{X}, \mathbf{G}\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, \hat{\mathbf{G}}^3\mathbf{X}, ...] = 0.$$

By Assumption 1.5, this is equivalent to:

$$\mathbb{E}[\boldsymbol{\eta}|\mathbf{X}, \mathbf{G}\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, \hat{\mathbf{G}}^3\mathbf{X}, ...] = 0,$$

which is equivalent to:

$$\mathbb{E}[\mathbf{G}\mathbf{y}|\mathbf{X}, \mathbf{G}\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, \hat{\mathbf{G}}^3\mathbf{X}, ...] = \mathbb{E}[\tilde{\mathbf{G}}\mathbf{y}|\mathbf{X}, \mathbf{G}\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, \hat{\mathbf{G}}^3\mathbf{X}, ...].$$

Using iterated expectations:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{y}}\mathbb{E}[\mathbf{G}\mathbf{y}|\mathbf{X}, \mathbf{G}\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, \hat{\mathbf{G}}^3\mathbf{X}, ..., \mathbf{y}]|\mathbf{X}, \mathbf{G}\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, \hat{\mathbf{G}}^3\mathbf{X}, ... \\
&= \mathbb{E}[\mathbf{G}|\mathbf{X}, \mathbf{G}\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, \hat{\mathbf{G}}^3\mathbf{X}, ...]\mathbb{E}[\mathbf{y}|\mathbf{X}, \mathbf{G}\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, \hat{\mathbf{G}}^3\mathbf{X}, ...],
\end{aligned}$$

and so $\mathbb{E}[\mathbf{G}\mathbf{y}|\mathbf{X}, \mathbf{G}\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, \hat{\mathbf{G}}^3\mathbf{X}, ...] = \mathbb{E}[\tilde{\mathbf{G}}\mathbf{y}|\mathbf{X}, \mathbf{G}\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, \hat{\mathbf{G}}^3\mathbf{X}, ...]$ follows since $\mathbf{G}$ $\hat{\mathbf{G}}$ and $\tilde{\mathbf{G}}$ are iid, which implies that $\mathbb{E}[\mathbf{G}|\mathbf{X}, \mathbf{G}\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, \hat{\mathbf{G}}^3\mathbf{X}, ...] = \mathbb{E}[\tilde{\mathbf{G}}|\mathbf{X}, \mathbf{G}\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, \hat{\mathbf{G}}^3\mathbf{X}, ...]$.

9 Appendix – Supplementary Material

9.1 Additional Monte-Carlo Results

Table 9: Simulation results without contextual effects (2).

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - **Gy is Observed**					
Estimation results					
$Intercept = 2$	2.011	0.263	1.844	2.007	2.182
$\alpha = 0.4$	0.399	0.014	0.390	0.400	0.409
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	1302.788	232.211	1145.790	1289.808	1445.453
Hausman	1.210	1.713	0.099	0.509	1.642
Sargan	0.970	1.382	0.095	0.420	1.300
$N = 50$, $M = 100$ - **Gy is not observed - same draw**					
Estimation results					
$Intercept = 2$	4.843	0.324	4.630	4.828	5.044
$\alpha = 0.4$	0.244	0.017	0.234	0.245	0.256
$\beta_1 = 1$	1.002	0.003	1.000	1.002	1.004
$\beta_2 = 1.5$	1.503	0.007	1.498	1.503	1.507
Tests					
F-test	26588.232	2113.817	25076.054	26534.073	27959.250
Hausman	339.782	44.092	310.264	338.783	368.245
Sargan	2.261	3.229	0.237	1.067	3.085
Validity					
$cor(\eta_i, \hat{x}_{i,1})$	−0.403	0.019	−0.416	−0.403	−0.390
$cor(\eta_i, \hat{x}_{i,2})$	−0.296	0.017	−0.307	−0.296	−0.284
$N = 50$, $M = 100$ - **Gy is not observed - different draws**					
Estimation results					
$Intercept = 2$	2.018	0.302	1.828	2.020	2.213
$\alpha = 0.4$	0.399	0.016	0.389	0.399	0.409
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.007	1.495	1.500	1.505
Tests					
F-test	1311.739	231.331	1150.204	1305.262	1457.470
Hausman	71.466	18.214	58.611	70.960	82.160
Sargan	1.000	1.365	0.095	0.456	1.389
Validity					
$cor(\eta_i, \hat{x}_{i,1})$	−0.001	0.015	−0.011	0.000	0.009
$cor(\eta_i, \hat{x}_{i,2})$	−0.001	0.014	−0.010	−0.001	0.009

Note: Number of simulations: 1000, $\lambda = +\infty$. Instruments: $\boldsymbol{GX}$ if $\boldsymbol{Gy}$ is observed, $\hat{\mathbf{G}}\boldsymbol{X}$ if $\boldsymbol{Gy}$ is not observed and approximated by $\tilde{\mathbf{G}}\boldsymbol{y}$.

Table 10: Simulation results without contextual effects (3).

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - **Gy is Observed**					
Estimation results					
$Intercept = 2$	2.001	0.188	1.867	1.990	2.125
$\alpha = 0.4$	0.400	0.010	0.393	0.401	0.407
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	2776.682	416.894	2489.415	2755.846	3036.762
Hausman	1.673	2.285	0.157	0.756	2.307
Sargan	2.889	2.356	1.156	2.274	3.959
$N = 50$, $M = 100$ - **Gy is not observed - same draw**					
Estimation results					
$Intercept = 2$	3.719	0.274	3.520	3.702	3.905
$\alpha = 0.4$	0.306	0.015	0.296	0.307	0.316
$\beta_1 = 1$	1.001	0.003	0.999	1.001	1.003
$\beta_2 = 1.5$	1.502	0.006	1.498	1.502	1.506
Tests					
F-test	38566.204	5520.495	34806.901	38162.018	42221.999
Hausman	21.208	11.433	13.058	19.173	28.217
Sargan	247.860	32.667	225.385	246.456	268.697
Validity					
$cor(\eta_i, \hat{\hat{x}}_{i,1})$	0.000	0.014	−0.009	0.000	0.009
$cor(\eta_i, \hat{\hat{x}}_{i,2})$	0.000	0.014	−0.010	0.000	0.010
$N = 50$, $M = 100$ - **Gy is not observed - different draws**					
Estimation results					
$Intercept = 2$	2.002	0.202	1.857	1.999	2.137
$\alpha = 0.4$	0.400	0.011	0.392	0.400	0.408
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	2798.114	418.630	2508.203	2795.093	3071.257
Hausman	218.584	34.913	192.694	217.089	240.687
Sargan	2.828	2.232	1.112	2.293	3.888
Validity					
$cor(\eta_i, \hat{\hat{x}}_{i,1})$	0.000	0.014	−0.009	0.000	0.009
$cor(\eta_i, \hat{\hat{x}}_{i,2})$	0.000	0.014	−0.010	0.000	0.010

Note: Number of simulations: 1000, $\lambda = 1$. Instruments: $\left\{(\tilde{\mathbf{G}})^k \boldsymbol{X}, k = 1, 2\right\}$.

Table 11: Simulation results without contextual effects.

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - **Gy is Observed**					
Estimation results					
$Intercept = 2$	2.003	0.190	1.877	1.997	2.126
$\alpha = 0.4$	0.400	0.010	0.393	0.400	0.407
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	2582.805	396.149	2310.803	2566.537	2834.415
Hausman	1.648	2.135	0.175	0.803	2.350
Sargan	2.962	2.547	1.153	2.313	4.007
$N = 50$, $M = 100$ - **Gy is not observed - same draw**					
Estimation results					
$Intercept = 2$	4.088	0.312	3.885	4.068	4.283
$\alpha = 0.4$	0.285	0.017	0.275	0.287	0.296
$\beta_1 = 1$	1.001	0.003	0.999	1.001	1.004
$\beta_2 = 1.5$	1.502	0.007	1.498	1.502	1.507
Tests					
F-test	37084.837	5430.758	33178.701	36580.801	40451.244
Hausman	24.804	12.850	15.008	23.435	32.737
Sargan	351.747	38.676	327.470	349.879	376.039
Validity					
$cor(\eta_i, \hat{\hat{x}}_{i,1})$	0.000	0.015	−0.011	−0.001	0.010
$cor(\eta_i, \hat{\hat{x}}_{i,2})$	0.000	0.014	−0.010	0.000	0.010
$N = 50$, $M = 100$ - **Gy is not observed - different draws**					
Estimation results					
$Intercept = 2$	2.006	0.216	1.866	2.010	2.143
$\alpha = 0.4$	0.400	0.012	0.392	0.400	0.407
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.007	1.495	1.500	1.505
Tests					
F-test	2605.493	395.108	2339.618	2608.021	2860.739
Hausman	298.171	43.582	267.427	296.419	325.031
Sargan	3.001	2.464	1.170	2.417	4.076
Validity					
$cor(\eta_i, \hat{\hat{x}}_{i,1})$	0.000	0.015	−0.011	−0.001	0.010
$cor(\eta_i, \hat{\hat{x}}_{i,2})$	0.000	0.014	−0.010	0.000	0.010

Note: Number of simulations: 1000, $\lambda = +\infty$. Instruments: $\left\{(\tilde{\mathbf{G}})^k \boldsymbol{X}, k = 1, 2\right\}$.

Table 12: Simulation results with contextual effects (2).

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - Instrument: $(\tilde{\mathbf{G}})^2\mathbf{X}$ - $\boldsymbol{Gy}$ is observed					
Estimation results					
$Intercept = 2$	2.004	0.181	1.880	2.007	2.126
$\alpha = 0.4$	0.400	0.003	0.398	0.400	0.402
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
$\gamma_1 = 5$	5.001	0.021	4.987	5.001	5.015
$\gamma_2 = -3$	−3.001	0.029	−3.020	−3.001	−2.981
Tests					
F-test	16233.492	1898.917	14917.312	16163.443	17491.015
Hausman	1.239	1.768	0.102	0.504	1.627
Sargan	1.005	1.364	0.104	0.487	1.289
$N = 50$, $M = 100$ - Instrument: $(\hat{\mathbf{G}})^2\mathbf{X}$ - $\boldsymbol{Gy}$ is not observed					
Estimation results					
$Intercept = 2$	2.002	0.217	1.855	2.005	2.150
$\alpha = 0.4$	0.400	0.004	0.397	0.400	0.403
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
$\gamma_1 = 5$	5.357	0.021	5.343	5.357	5.371
$\gamma_2 = -3$	−2.380	0.037	−2.405	−2.381	−2.357
$\hat{\gamma}_1 = 0$	−0.356	0.024	−0.372	−0.356	−0.340
$\hat{\gamma}_2 = 0$	−0.620	0.035	−0.642	−0.621	−0.597
Tests					
F-test	10741.676	1124.978	9978.928	10687.760	11475.206
Hausman	22.119	10.011	14.423	21.247	28.586
Sargan	0.956	1.304	0.107	0.464	1.263

Note: Number of simulations: 1000, $\lambda = +\infty$.

Table 13: Simulation results with subpopulation unobserved fixed effects.

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - Instrument: $\mathbf{J}(\tilde{\mathbf{G}})^2\mathbf{X}$ - $\hat{\mathbf{y}}$ is observed					
Estimation results					
$\alpha = 0.4$	0.379	1.703	−0.122	0.398	0.938
$\beta_1 = 1$	1.000	0.005	0.997	1.000	1.003
$\beta_2 = 1.5$	1.500	0.009	1.496	1.500	1.505
$\gamma_1 = 5$	5.018	1.318	4.584	5.001	5.415
	4.970	1.372	4.565	4.992	5.421
$\gamma_2 = -3$	−2.967	2.710	−3.861	−2.998	−2.156
Tests					
F-test	2.860	2.270	1.132	2.364	4.045
Hausman	1.029	1.398	0.105	0.484	1.392
Sargan	0.820	1.184	0.066	0.335	1.078
$N = 50$, $M = 100$ - Instrument: $\mathbf{J}(\hat{\mathbf{G}})^2\mathbf{X}$ - $\hat{\mathbf{y}}$ is not observed					
Estimation results					
$\alpha = 0.4$	0.345	1.609	−0.246	0.265	0.869
$\beta_1 = 1$	1.000	0.005	0.997	1.000	1.002
$\beta_2 = 1.5$	1.500	0.008	1.495	1.500	1.505
$\gamma_1 = 5$	5.351	0.170	5.284	5.343	5.408
$\gamma_2 = -3$	−2.378	0.123	−2.416	−2.373	−2.331
$\hat{\gamma}_1 = 0$	−0.307	1.442	−0.769	−0.236	0.237
$\hat{\gamma}_2 = 0$	−0.534	2.490	−1.346	−0.407	0.398
Tests					
F-test	2.773	2.189	1.072	2.304	3.773
Hausman	1.114	1.563	0.105	0.525	1.483
Sargan	0.889	1.329	0.083	0.381	1.206

Note: Number of simulations: 1000, $\lambda = 1$. In each group, the fixed effect is generated as $0.3x_{1,1} + 0.3x_{3,2} - 1.8x_{50,2}$.

Table 14: Simulation results with subpopulation unobserved fixed effects (2).

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - Instrument: $\mathbf{J}(\tilde{\mathbf{G}})^2\mathbf{X}$ - $\hat{\mathbf{y}}$ is observed					
Estimation results					
$\alpha = 0.4$	0.623	4.670	−0.475	0.341	1.345
$\beta_1 = 1$	1.001	0.015	0.997	1.000	1.003
$\beta_2 = 1.5$	1.499	0.019	1.494	1.500	1.504
$\gamma_1 = 5$	4.826	3.644	4.263	5.042	5.688
$\gamma_2 = -3$	−3.357	7.485	−4.491	−2.913	−1.572
Tests					
F-test	1.025	1.023	0.293	0.710	1.451
Hausman	0.979	1.527	0.090	0.396	1.264
Sargan	0.665	1.184	0.042	0.203	0.801
$N = 50$, $M = 100$ - Instrument: $\mathbf{J}(\hat{\mathbf{G}})^2\mathbf{X}$ - $\hat{\mathbf{y}}$ is not observed					
Estimation results					
$\alpha = 0.4$	−0.071	3.561	−0.949	−0.047	0.876
$\beta_1 = 1$	0.999	0.010	0.996	0.999	1.002
$\beta_2 = 1.5$	1.500	0.018	1.495	1.501	1.506
$\gamma_1 = 5$	5.305	0.377	5.208	5.306	5.406
$\gamma_2 = -3$	−2.364	0.192	−2.411	−2.355	−2.306
$\hat{\gamma}_1 = 0$	0.062	3.183	−0.777	0.051	0.846
$\hat{\gamma}_2 = 0$	0.117	5.512	−1.360	0.062	1.459
Tests					
F-test	1.063	1.046	0.329	0.741	1.486
Hausman	0.993	1.499	0.095	0.409	1.255
Sargan	0.670	1.117	0.041	0.207	0.817

Note: Number of simulations: 1000, $\lambda = +\infty$. In each group, the fixed effect is generated as $0.3x_{1,1} + 0.3x_{3,2} - 1.8x_{50,2}$.

Table 15: Simulation results with ARD: without contextual effects (1000 replications).

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 250$, $M = 20$ - Instrument: $\tilde{\mathbf{G}}\mathbf{X}$ - $\boldsymbol{Gy}$ is observed					
Estimation results					
$Intercept = 2$	2.031	0.941	1.471	2.048	2.608
$\alpha = 0.4$	0.398	0.051	0.368	0.397	0.429
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	152.722	63.626	104.971	145.232	189.435
Hausman	1.037	1.467	0.098	0.435	1.352
Sargan	1.003	1.407	0.096	0.422	1.335
$N = 250$, $M = 20$ - Instrument: $\left\{\left(\tilde{\mathbf{G}}\right)^k \mathbf{X}, k = 1, 2\right\}$ -$\boldsymbol{Gy}$ is observed					
Estimation results					
$Intercept = 2$	1.999	0.429	1.721	1.985	2.254
$\alpha = 0.4$	0.400	0.023	0.386	0.400	0.415
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	427.225	146.605	319.567	417.901	517.949
Hausman	1.026	1.467	0.116	0.491	1.339
Sargan	3.000	2.442	1.218	2.342	4.089
$N = 250$, $M = 20$ - Instrument: $\hat{\mathbf{G}}\mathbf{X}$ - $\boldsymbol{Gy}$ is not observed					
Estimation results					
$Intercept = 2$	1.854	1.117	1.182	1.879	2.569
$\alpha = 0.4$	0.408	0.061	0.368	0.407	0.445
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.007	1.496	1.500	1.505
Tests					
F-test	144.813	60.270	100.244	138.238	178.694
Hausman	30.339	14.849	19.275	27.985	39.404
Sargan	1.105	1.576	0.109	0.506	1.450
$N = 250$, $M = 20$ - Instrument: $\left\{\left(\hat{\mathbf{G}}\right)^k \mathbf{X}, k = 1, 2\right\}$ - $\boldsymbol{Gy}$ is not observed					
Estimation results					
$Intercept = 2$	1.969	0.531	1.599	1.962	2.318
$\alpha = 0.4$	0.402	0.029	0.383	0.402	0.422
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.007	1.496	1.500	1.505
Tests					
F-test	419.464	142.058	314.376	407.311	505.246
Hausman	171.162	53.370	133.847	169.651	205.461
Sargan	3.274	2.575	1.346	2.613	4.443

Table 16: Simulation results with nuclear ARD: without contextual effects, and $\hat{\boldsymbol{y}}$ is observed (1000 replications).

Weight[†]		Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 250$, $M = 20$ - Instrument: $\tilde{\mathbf{G}}\mathbf{X}$ - $\boldsymbol{Gy}$ is observed						
$\tau = 200$	$\alpha = 0.4$	0.403	0.088	0.350	0.406	0.456
$\tau = 600$	$\alpha = 0.4$	0.396	0.139	0.322	0.396	0.468
$\tau = 1374$	$\alpha = 0.4$	0.399	0.242	0.279	0.396	0.516
$N = 250$, $M = 20$ - Instrument: $\left\{\left(\hat{\mathbf{G}}\right)^k \mathbf{X}, k = 1, 2\right\}$ - $\boldsymbol{Gy}$ is observed						
$\tau = 200$	$\alpha = 0.4$	0.400	0.036	0.378	0.401	0.422
$\tau = 600$	$\alpha = 0.4$	0.396	0.050	0.368	0.398	0.427
$\tau = 1374$	$\alpha = 0.4$	0.403	0.111	0.346	0.402	0.461

Note: The parameter τ corresponds to the parameter λ in Alidaee et al. (2020) and represents the weight associated to the nuclear norm. In our context, the recommended value of Alidaee et al. (2020) is given by $\tau = 1374$. The optimal value (in terms of RMSE) found through cross-validation is $\tau = 600$, while $\tau = 200$ gives a RMSE similar to the recommended value.

Table 17: Simulation results with nuclear ARD: $\hat{\boldsymbol{y}}$ is observed, $\tau = 600$ (1000 replications).

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 250$, $M = 20$ - Instrument: $\tilde{\mathbf{G}}\mathbf{X}^2$ - $\boldsymbol{Gy}$ is observed					
Estimation results					
$Intercept = 2$	2.001	0.267	1.825	1.999	2.181
$\alpha = 0.4$	0.400	0.011	0.393	0.400	0.406
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
$\gamma_1 = 5$	5.001	0.027	4.982	5.001	5.020
$\gamma_2 = -3$	−3.000	0.033	−3.022	−2.999	−2.977
Tests					
F-test	810.356	338.648	570.099	774.973	1005.605
Hausman	0.956	1.281	0.113	0.467	1.310
Sargan	1.033	1.417	0.107	0.515	1.395

Table 18: Simulation results with nuclear ARD: $\hat{\boldsymbol{y}}$ is not observed, $\tau = 600$ (1000 replications).

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 250$, $M = 20$ - Instrument: $\tilde{\mathbf{G}}\mathbf{X}$ - $\boldsymbol{Gy}$ is not observed					
Estimation results					
$Intercept = 2$	5.215	2.934	4.008	5.423	6.748
$\alpha = 0.4$	0.225	0.161	0.141	0.214	0.291
$\beta_1 = 1$	1.001	0.004	0.998	1.001	1.003
$\beta_2 = 1.5$	1.501	0.007	1.497	1.501	1.505
Tests					
F-test	21.033	13.187	11.490	18.861	28.535
Hausman	7.626	6.980	2.377	5.937	10.976
Sargan	2.488	3.432	0.258	1.191	3.313
$cor(\eta_i, \hat{x}_{i,1})$	−0.006	0.022	−0.020	−0.006	0.008
$cor(\eta_i, \hat{x}_{i,2})$	−0.036	0.029	−0.055	−0.036	−0.016
$N = 250$, $M = 20$ - Instrument: $\left\{\left(\hat{\mathbf{G}}\right)^k \mathbf{X}, k = 1, 2\right\}$ - $\boldsymbol{Gy}$ is not observed					
Estimation results					
$Intercept = 2$	4.481	1.503	3.469	4.493	5.474
$\alpha = 0.4$	0.266	0.082	0.211	0.264	0.321
$\beta_1 = 1$	1.001	0.004	0.998	1.001	1.003
$\beta_2 = 1.5$	1.501	0.007	1.496	1.501	1.506
Tests					
F-test	50.332	24.549	31.909	46.716	64.013
Hausman	57.141	40.279	27.257	48.421	79.545
Sargan	11.075	9.786	4.209	7.877	15.298
$cor(\eta_i, \hat{\hat{x}}_{i,1})$	−0.011	0.023	−0.025	−0.011	0.005
$cor(\eta_i, \hat{\hat{x}}_{i,2})$	−0.064	0.040	−0.092	−0.064	−0.036

9.2 ARD Simulations Setting

This section provides details about ARD simulation and model estimation using a MCMC method. We simulate the network for a population of 5000 individuals divided into $m = 20$ groups of $n = 250$ individuals. Within each group, the probability of a link is:

$$\mathbb{P}(a_{ij} = 1) \propto \exp\{\nu_i + \nu_j + \zeta \mathbf{z}_i' \mathbf{z}_j\}. \tag{9}$$

Since there is no connection between the groups, the networks are simulated and estimated independently. We first present how we simulate the data following the model (7).

9.2.1 ARD Simulation

The parameters are defined as follows: $\zeta = 1.5$, $\nu_i \sim \mathcal{N}(-1.25, 0.37)$, and $\mathbf{z}_i$ are distributed uniformly according to a von Mises–Fisher distribution. We use a hypersphere of dimension 3. We set the same values for the parameter for the 20 groups. We generate the probabilities of links in each network following Breza et al. (2017).

$$\mathbb{P}(a_{ij} = 1 | \nu_i, \nu_j, \zeta, \mathbf{z}_i, \mathbf{z}_j) = \frac{\exp\{\nu_i + \nu_j + \zeta \mathbf{z}_i' \mathbf{z}_j\} \sum_{i=1}^{N} d_i}{\sum_{ij} \exp\{\nu_i + \nu_j + \zeta \mathbf{z}_i' \mathbf{z}_j\}}, \tag{10}$$

where d_i is the degree defined by $d_i \approx \frac{C_p(0)}{C_p(\zeta)} \exp(\nu_i) \sum_{i=1}^{N} \exp(\nu_i)$, the function $C_p(.)$ is the normalization constant in the von Mises–Fisher distribution density function. After computing the probability of a link for any pair in the population, we sample the entries of the adjacency matrix using a Bernoulli distribution with probability (10).

To generate the ARD, we require the "traits" (e.g. cities) for each individual. We set $K = 12$ traits on the hypersphere. Their location $\mathbf{v}_k$ is distributed uniformly according to the von Mises–Fisher distribution. The individuals having the trait k are assumed to be generated by a von Mises–Fisher distribution with the location parameter $\mathbf{v}_k$ and the intensity parameter $\eta_k \sim |\mathcal{N}(4, 1)|$, $k = 1, \ldots, 12$.

We attribute traits to individuals given their spherical coordinates. We first define N_k, the number of individuals having the trait k:

$$N_k = \left\lfloor r_k \frac{\sum_{i=1}^{N} f_{\mathcal{M}}(\mathbf{z}_i | \mathbf{v}_k, \eta_k)}{\max_i f_{\mathcal{M}}(\mathbf{z}_i | \mathbf{v}_k, \eta_k)} \right\rfloor,$$

where $\lfloor x \rfloor$ stands for the greatest integer less than or equal to x, r_k is a random number uniformly distributed over $(0.8; 0.95)$, and $f_{\mathcal{M}}(\mathbf{z}_i|\mathbf{v}_k, \eta_k)$ is the von Mises–Fisher distribution density function evaluated at $\mathbf{z}_i$ with the location parameter $\mathbf{v}_k$ and the intensity parameter η_k.

The intuition behind this definition for N_k is that when many $\mathbf{z}_i$ are close to $\mathbf{v}_k$, many individuals should have the trait k.

We can finally attribute trait k to individual i by sampling a Bernoulli distribution with the probability f_{ik} given by:

$$f_{ik} = N_k \frac{f_{\mathcal{M}}(\mathbf{z}_i|\mathbf{v}_k, \eta_k)}{\sum_{i=1}^{N} f_{\mathcal{M}}(\mathbf{z}_i|\mathbf{v}_k, \eta_k)}.$$

The probability of having a trait depends on the proximity of the individuals to the trait's location on the hypersphere.

Once the traits for each individual and the network are generated, we can build the ARD.

9.2.2 Model Estimation

In practice, we only have the ARD and the traits for each individual. McCormick and Zheng (2015) propose a MCMC approach to infer the parameters in model (9).

However, the spherical coordinates and the degrees in this model are not identified. The authors solve this issue by fixing some $\mathbf{v}_k$ and use the fixed positions to rotate the latent surface back to a common orientation at each iteration of the MCMC using a Procrustes transformation. In addition, the total size of a subset b_k is constrained in the MCMC.

As discussed by McCormick and Zheng (2015), the numbers of $\mathbf{v}_k$ and b_k to be set as fixed depends on the dimension of hypersphere. In our simulations, $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_5$ are set as fixed to rotate back the latent space. When simulating the data, we let $\mathbf{v}_1 = (1, 0, 0)$, $\mathbf{v}_2 = (0, 1, 0)$, and $\mathbf{v}_3 = (0, 0, 1)$. This ensures that the fixed positions on the hypersphere are spaced, as suggested by the authors, to use as much of the space as possible, maximizing the distance between the estimated positions. We also constrain b_3 to its true value. The results do not change when we constrain a larger set of b_k

Following Breza et al. (2017), we estimate the link probabilities using the parameters' posterior distributions. The gregariousness parameters are computed from the degrees d_i and the parameter ζ using the following equation:

$$\nu_i = \log(d_i) - \log\Big(\sum_{i=1}^{N} d_i\Big) + \frac{1}{2}\log\Big(\frac{C_p(\zeta)}{C_p(0)}\Big).$$

9.3 Network Sampling

This section explains how we sample the network in Algorithm 1 using Gibbs sampling. As discussed above, a natural solution is to update only one entry of the adjacency matrix at every step t of the MCMC. The entry (i, j) is updated according to its conditional posterior distribution:

$$a_{ij} \sim P(\cdot|\mathbf{A}_{-ij}, \mathbf{y}) = \frac{\mathcal{P}(\mathbf{y}|a_{ij}, \mathbf{A}_{-ij})P(a_{ij}|\mathbf{A}_{-ij})}{\mathcal{P}(\mathbf{y}|1, \mathbf{A}_{-ij})P(a_{ij} = 1|\mathbf{A}_{-ij}) + \mathcal{P}(\mathbf{y}|0, \mathbf{A}_{-ij})P(a_{ij} = 0|\mathbf{A}_{-ij})}.$$

However, for each entry, we need to compute $\mathcal{P}(\mathbf{y}|0, \mathbf{A}_{-ij})$ and $\mathcal{P}(\mathbf{y}|1, \mathbf{A}_{-ij})$, which are the respective likelihoods of replacing a_{ij} by 0 or by 1. The likelihood computation requires the determinant of $(\mathbf{I} - \alpha\mathbf{G})$, which has a complexity $O(N^3)$, where N is the dimension of $\mathbf{G}$. This implies that we must compute $2N(N-1)$ times $\det(\mathbf{I} - \alpha\mathbf{G})$ to update the adjacency matrix at each step of the MCMC. As $\boldsymbol{G}$ is row-normalized, alternating any off-diagonal entry (i, j) in $\boldsymbol{A}$ between 0 and 1 perturbs all off-diagonal entries of the row i in $(\mathbf{I} - \alpha\mathbf{G})$. We show that $\boldsymbol{A}_{ij}$ and $\det(\mathbf{I} - \alpha\mathbf{G})$ can be updated by computing a determinant of an auxiliary matrix that requires only updating two entries.

Assume that we want to update the entry (i, j). Let h be the function defined in $\mathbb{N}$ such that $\forall\, x \in \mathbb{N}^*$, $h(x) = x$, and $h(0) = 1$. Let $\mathbf{L}$ be an N $\times$ N diagonal matrix, where $\mathbf{L}_{ii} = h(n_i)$, and n_i stands for the degree of i, while $\mathbf{L}_{kk} = 1$ for all $k \neq i$, and $\mathbf{W}$ is the matrix $\boldsymbol{G}$ where the row i of $\mathbf{W}$ is replaced by the row i of $\boldsymbol{A}$. Then, since the determinant is linear in each row, we can obtain $\mathbf{I} - \alpha\mathbf{G}$ by dividing the row i of $\mathbf{L} - \alpha\mathbf{W}$ by $h(n_i)$. We get:

$$\det(\mathbf{I} - \alpha\mathbf{G}) = \frac{1}{h(n_i)}\det(\mathbf{L} - \alpha\mathbf{W}).$$

When a_{ij} changes (from 0 to 1, or 1 to 0), note that only the entries (i, i) and (i, j) change in $\mathbf{L} - \alpha\mathbf{W}$. Two cases can be distinguished.

- If $a_{ij} = 0$ before the update, then the new degree of i will be $n_i + 1$. Thus, the entry (i, i) in $\mathbf{L} - \alpha\mathbf{W}$ will change from $h(n_i)$ to $h(n_i + 1)$ (since the diagonal of $\mathbf{W}$ equals 0) and the entry (i, j) from 0 to $-\alpha$. The new determinant is therefore given by:

$$\det(\mathbf{I} - \alpha\mathbf{G}^*) = \frac{1}{h(n_i + 1)}\det(\mathbf{L}^* - \alpha\mathbf{W}^*),$$

  where $\mathbf{G}^*$, $\mathbf{L}^*$, and $\alpha\mathbf{W}^*$ are the new matrices, once a_{ij} has been updated.

- If $a_{ij} = 1$ before the update, then the new degree of k will be $n_i - 1$. Thus the entry (i, i) in $\mathbf{L} - \alpha\mathbf{W}$ will change from $h(n_i)$ to $h(n_i - 1)$ and the entry (i, j) from $-\alpha$ to 0. The new determinant is therefore given by:

$$\det(\mathbf{I} - \alpha\mathbf{G}^*) = \frac{1}{h(n_i - 1)}\det(\mathbf{L}^* - \alpha\mathbf{W}^*).$$

Then, to update $\det(\mathbf{L}-\alpha\mathbf{W})$ when only the entries (i, i) and (i, j) change, we adapt the Lemma 1 in Hsieh et al. (2019a) as follows:

Proposition 3. *Let $\mathbf{e}_i$ be the i'th unit basis vector in $\mathbb{R}^N$. Let $\mathbf{M}$ denote an $N \times N$ matrix and $\mathbf{B}_{ij}(\mathbf{Q}, \epsilon)$ an $N \times N$ matrix as function of an $N \times N$ matrix $\mathbf{Q}$ and a real value ϵ, such that:*

$$\mathbf{B}_{ij}(\mathbf{Q}, \epsilon) = \frac{\mathbf{Q}\mathbf{e}_i\mathbf{e}_j'\mathbf{Q}}{1 + \epsilon\mathbf{e}_j'\mathbf{Q}\mathbf{e}_i}. \tag{11}$$

Adding a perturbation ϵ_1 in the (i, i)th position and a perturbation ϵ_2 in the (i, j)th position to the matrix $\mathbf{M}$ can be written as $\tilde{\mathbf{M}} = \mathbf{M} + \epsilon_1\mathbf{e}_i\mathbf{e}_i' + \epsilon_2\mathbf{e}_i\mathbf{e}_j'$.

(a) The inverse of the perturbed matrix can be written as:

$$\tilde{\mathbf{M}}^{-1} = \mathbf{M}^{-1} - \epsilon_1\mathbf{B}_{ii}(\mathbf{M}^{-1}, \epsilon_1) - \epsilon_2\mathbf{B}_{ij}\left(\mathbf{M}^{-1} - \epsilon_1\mathbf{B}_{ii}(\mathbf{M}^{-1}, \epsilon_1), \epsilon_2\right).$$

(b) The determinant of the perturbed matrix can be written as:

$$det\left(\tilde{\mathbf{M}}\right) = \left(1 + \epsilon_2\mathbf{e}_j'\left(\mathbf{M}^{-1} - \epsilon_1\mathbf{B}_{ii}(\mathbf{M}^{-1}, \epsilon_1)\mathbf{e}_i\right)\right)(1 + \epsilon_1\mathbf{e}_i'\mathbf{M}^{-1}\mathbf{e}_i)det(\mathbf{M}).$$

Proof. (a) By the Sherman–Morrison formula (Mele, 2017), we have:

$$\left(\mathbf{M} + \epsilon\mathbf{e}_i\mathbf{e}_j'\right)^{-1} = \mathbf{M}^{-1} - \epsilon\frac{\mathbf{M}^{-1}\mathbf{e}_i\mathbf{e}_j'\mathbf{M}^{-1}}{1 + \epsilon\mathbf{e}_j'\mathbf{M}^{-1}\mathbf{e}_i} = \mathbf{M}^{-1} - \epsilon\mathbf{B}_{ij}(\mathbf{M}, \epsilon).$$

Thus,

$$\begin{aligned}
\tilde{\mathbf{M}}^{-1} &= \left((\mathbf{M} + \epsilon_1\mathbf{e}_i\mathbf{e}_i') + \epsilon_2\mathbf{e}_i\mathbf{e}_j'\right)^{-1}, \\
\tilde{\mathbf{M}}^{-1} &= (\mathbf{M} + \epsilon_1\mathbf{e}_i\mathbf{e}_i')^{-1} - \epsilon_2\mathbf{B}_{ij}((\mathbf{M} + \epsilon_1\mathbf{e}_i\mathbf{e}_i')^{-1}, \epsilon_2), \\
\tilde{\mathbf{M}}^{-1} &= \mathbf{M}^{-1} - \epsilon_1\mathbf{B}_{ii}(\mathbf{M}^{-1}, \epsilon_1) - \epsilon_2\mathbf{B}_{ij}\left(\mathbf{M}^{-1} - \epsilon_1\mathbf{B}_{ii}(\mathbf{M}^{-1}, \epsilon_1), \epsilon_2\right).
\end{aligned}$$

(b) By the matrix determinant lemma (Johnson and Horn, 1985), we have:

$$\det\left(\mathbf{M}+\epsilon\mathbf{e}_i\mathbf{e}_j'\right)=(1+\epsilon\mathbf{e}_j'\mathbf{M}^{-1}\mathbf{e}_i)\det\left(\mathbf{M}\right).$$

It follows that:

$$\begin{aligned}
\det\left(\tilde{\mathbf{M}}\right)&=\det\left((\mathbf{M}+\epsilon_1\mathbf{e}_i\mathbf{e}_i')+\epsilon_2\mathbf{e}_i\mathbf{e}_j'\right),\\
\det\left(\tilde{\mathbf{M}}\right)&=(1+\epsilon_2\mathbf{e}_j'(\mathbf{M}+\epsilon_1\mathbf{e}_i\mathbf{e}_i')^{-1}\mathbf{e}_i)\det\left(\mathbf{M}+\epsilon_1\mathbf{e}_i\mathbf{e}_i'\right),\\
\det\left(\tilde{\mathbf{M}}\right)&=\left(1+\epsilon_2\mathbf{e}_j'\left(\mathbf{M}^{-1}-\epsilon_1\mathbf{B}_{ii}(\mathbf{M}^{-1},\epsilon_1)\mathbf{e}_i\right)\right)(1+\epsilon_1\mathbf{e}_i'\mathbf{M}^{-1}\mathbf{e}_i)\det\left(\mathbf{M}\right).
\end{aligned}$$

□

The method proposed above becomes computationally intensive when many entries must be updated simultaneously. We also propose an alternative method that allows updating the block for entries in $\boldsymbol{A}$. Let $\mathbf{D}=(\mathbf{I}-\alpha\mathbf{G})$; we can write:

$$\det(\mathbf{D})=\sum_{j=1}^{N}(-1)^{i+j}\mathbf{D}_{\mathrm{ij}}\delta_{\mathrm{ij}}, \tag{12}$$

where i denotes any row of $\mathbf{D}$ and δ_{ij} the minor[31] associated with the entry (i,j). The minors of row i do not depend on the values of entries in row i. To update any block in row i, we therefore compute the N minors associated to i and use this minor within the row. We can then update many entries simultaneously without increasing the number of times that we compute $\det(\mathbf{D})$. One possibility is to update multiple links simultaneously by randomly choosing the number of entries to consider and their position in the row. As suggested by Chib and Ramamurthy (2010), this method would help the Gibbs to converge more quickly. We can summarize how we update the row i as follows:

(a) Compute the N minors $\delta_{i1},\ldots,\delta_{in}$.

(b) Let $\Omega_{\boldsymbol{G}}$ be the entries to update in the row i, and $n_{\boldsymbol{G}}=|\Omega_{\boldsymbol{G}}|$ the number of entries in $\Omega_{\boldsymbol{G}}$.

(b.1) Choose r, the size of the block to update, as a random integer number such that $1\leq r\leq n_{\boldsymbol{G}}$. In practice, we choose $r\leq\min(5,n_{\boldsymbol{G}})$ since the number of possibilities of links to consider grows exponentially with r.

[31] The determinant of the submatrix of $\mathbf{M}$ by removing row i and column j.

(b.2) Choose the r random entries from $\Omega_{\boldsymbol{G}}$. These entries define the block to update.

(b.3) Compute the posterior probabilities of all possibilities of links inside the block and update the block (there are 2^r possibilities). Use the minors calculated at (a) and the formula (12) to quickly compute $det(\mathbf{D})$.

(b.4) Remove the r drawn positions from $\Omega_{\boldsymbol{G}}$ and let $n_{\boldsymbol{G}} = n_{\boldsymbol{G}} - r$. Replicate (b.1), (b.2), and (b.3) until $n_{\boldsymbol{G}} = 0$.

9.4 Posterior Distributions for Algorithm 1.

To compute the posterior distributions, we set prior distributions on $\tilde{\alpha}$, $\mathbf{\Lambda}$, and σ^2, where $\tilde{\alpha} = \log(\frac{\alpha}{1-\alpha})$ and $\mathbf{\Lambda} = [\boldsymbol{\beta}, \boldsymbol{\gamma}]$. In Algorithm 1, we therefore sample $\tilde{\alpha}$ and compute α, such that $\alpha = \frac{\exp(\tilde{\alpha})}{1+\exp(\tilde{\alpha})}$. Using this functional form for computing α ensures that $\alpha \in (0, 1)$. The prior distributions are set as follows:

$$\begin{aligned}
\tilde{\alpha} &\sim \mathcal{N}(\mu_{\tilde{\alpha}}, \sigma^2_{\tilde{\alpha}}), \\
\mathbf{\Lambda}|\sigma^2 &\sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{\Lambda}}, \sigma^2 \mathbf{\Sigma}_{\mathbf{\Lambda}}), \\
\sigma^2 &\sim IG(\frac{a}{2}, \frac{b}{2}).
\end{aligned}$$

For the simulations and estimations in this paper, we set: $\mu_{\tilde{\alpha}} = -1$, $\sigma_{\tilde{\alpha}}^{-2} = 2$, $\boldsymbol{\mu}_{\mathbf{\Lambda}} = \mathbf{0}$, ${\mathbf{\Sigma}_{\mathbf{\Lambda}}}^{-1} = \frac{1}{100}\boldsymbol{I}_K$, $a = 4$ and $b = 4$, where $\boldsymbol{I}_K$ is the identity matrix of dimension K and $K = \dim(\mathbf{\Lambda})$.

Following Algorithm 1, α is updated at each iteration t of the MCMC by drawing $\tilde{\alpha}^*$ from the proposal $\mathcal{N}(\tilde{\alpha}_{t-1}, \xi_t)$, where the jumping scale ξ_t is also updated at each t following Atchadé and Rosenthal (2005) for an acceptance rate of a^* targeted at 0.44. As the proposal is symmetrical, $\alpha^* = \frac{\exp(\tilde{\alpha}^*)}{1+\exp(\tilde{\alpha}^*)}$ is accepted with the probability:

$$\min\left\{1, \frac{\mathcal{P}(\mathbf{y}|\mathbf{A}_t, \mathbf{\Lambda}_{t-1}, \alpha^*)P(\tilde{\alpha}^*)}{\mathcal{P}(\mathbf{y}|\mathbf{A}_t, \boldsymbol{\theta}_{t-1})P(\tilde{\alpha}_t)}\right\}.$$

The parameters $\boldsymbol{\Lambda}_t = [\boldsymbol{\beta}_t,\, \boldsymbol{\gamma}_t]$ and σ_t^2 are drawn from their posterior conditional distributions, given as follows:

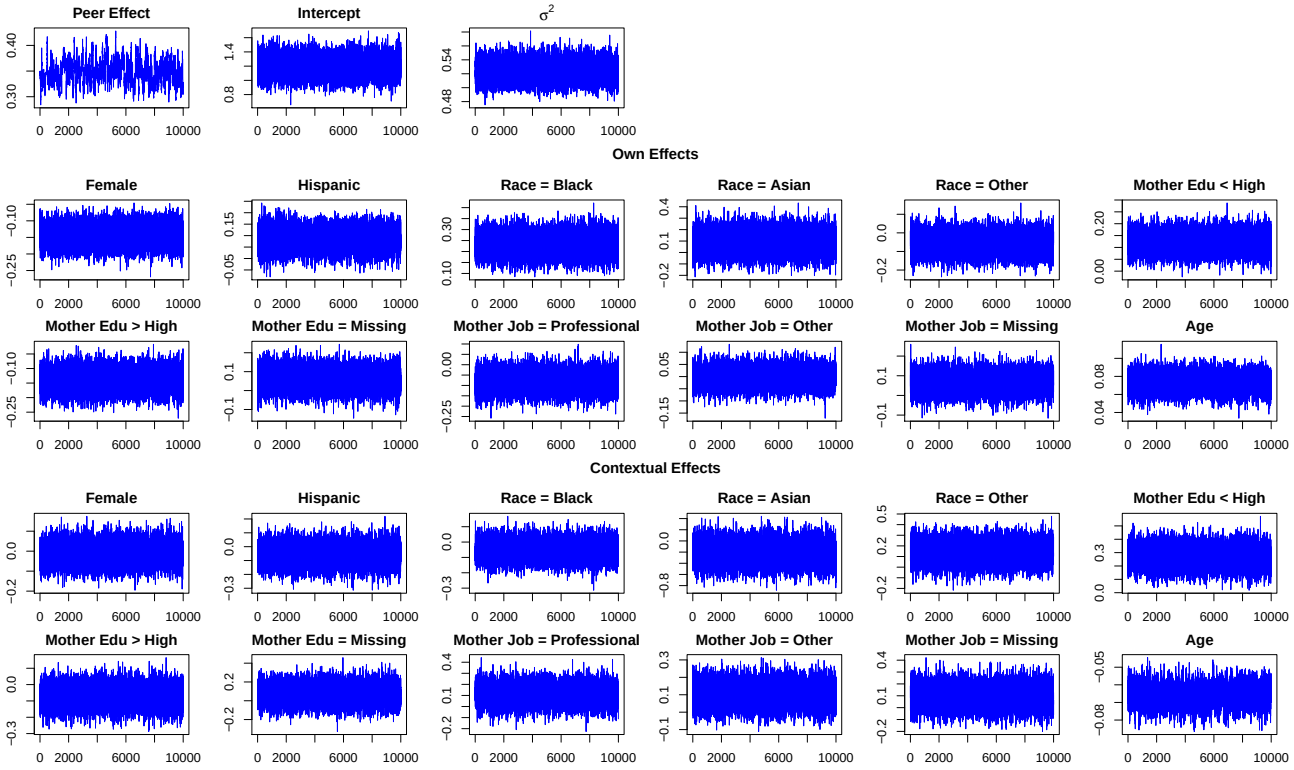
$$\boldsymbol{\Lambda}_t|\boldsymbol{y}, \boldsymbol{A}_t, \alpha_t, \sigma_{t-1}^2 \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{\boldsymbol{\Lambda}_t}, \sigma_{t-1}^2 \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\Lambda}_t}),$$
$$\sigma_t^2|\boldsymbol{y}, \boldsymbol{A}_t, \boldsymbol{\theta}_t \sim IG\left(\frac{\hat{a}_t}{2}, \frac{\hat{b}_t}{2}\right),$$

where,

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\Lambda}_t}^{-1} &= \mathbf{V}_t'\mathbf{V}_t + \boldsymbol{\Sigma}_{\boldsymbol{\Lambda}}^{-1}, \\
\hat{\boldsymbol{\mu}}_{\boldsymbol{\Lambda}_t} &= \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\Lambda}_t}\left(\mathbf{V}_t'(\boldsymbol{y} - \alpha_t \boldsymbol{G}_t\boldsymbol{y}) + \boldsymbol{\Sigma}_{\boldsymbol{\Lambda}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\Lambda}}\right), \\
\hat{a}_t &= a + N, \\
\hat{b}_t &= b + (\boldsymbol{\Lambda}_t - \boldsymbol{\mu}_{\boldsymbol{\Lambda}})'\boldsymbol{\Sigma}_{\boldsymbol{\Lambda}}^{-1}(\boldsymbol{\Lambda}_t - \boldsymbol{\mu}_{\boldsymbol{\Lambda}}) + (\boldsymbol{y} - \alpha_t\boldsymbol{G}_t\boldsymbol{y} - \mathbf{V}_t\boldsymbol{\Lambda}_t)'(\boldsymbol{y} - \alpha_t\boldsymbol{G}_t\boldsymbol{y} - \mathbf{V}_t\boldsymbol{\Lambda}_t), \\
\mathbf{V}_t &= [\mathbf{1},\ \boldsymbol{X},\ \boldsymbol{G}_t\boldsymbol{X}].
\end{aligned}$$

9.5 Empirical Application

Figure 5: Simulations using the observed network.



9.6 Expectation Maximization Algorithm

If $\mathbf{A}$ was observed, the objective would be to maximize the log-likelihood of the model with respect to $\boldsymbol{\theta}$:

$$\mathcal{P}(\mathbf{y}|\mathbf{A}, \mathbf{X}; \boldsymbol{\theta}).$$

Since $\mathbf{A}$ is unobserved, we propose to treat it as a latent variable. We therefore look for the value of $\boldsymbol{\theta}$ that maximizes:

$$\mathcal{P}(\mathbf{y}|\mathbf{X}; \boldsymbol{\theta}) = \sum_{\mathbf{A}} \mathcal{P}(\mathbf{y}|\mathbf{A}, \mathbf{X}; \boldsymbol{\theta}) P(\mathbf{A}).$$

Since the number potential network structures is huge, evaluating this expectation is unfeasible.[32] We therefore propose to maximize $\mathcal{P}(\mathbf{y}|\mathbf{X}; \boldsymbol{\theta})$ using an expectation maximization algorithm:

Algorithm 2. *Initialize* $\boldsymbol{\theta}_0$*, and for* $t = 0, ..., T$*, do the following:*

[32] For a population of only 5 individuals, the number of network structures is 2^{20}.

Figure 6: Simulations using the reconstructed network.

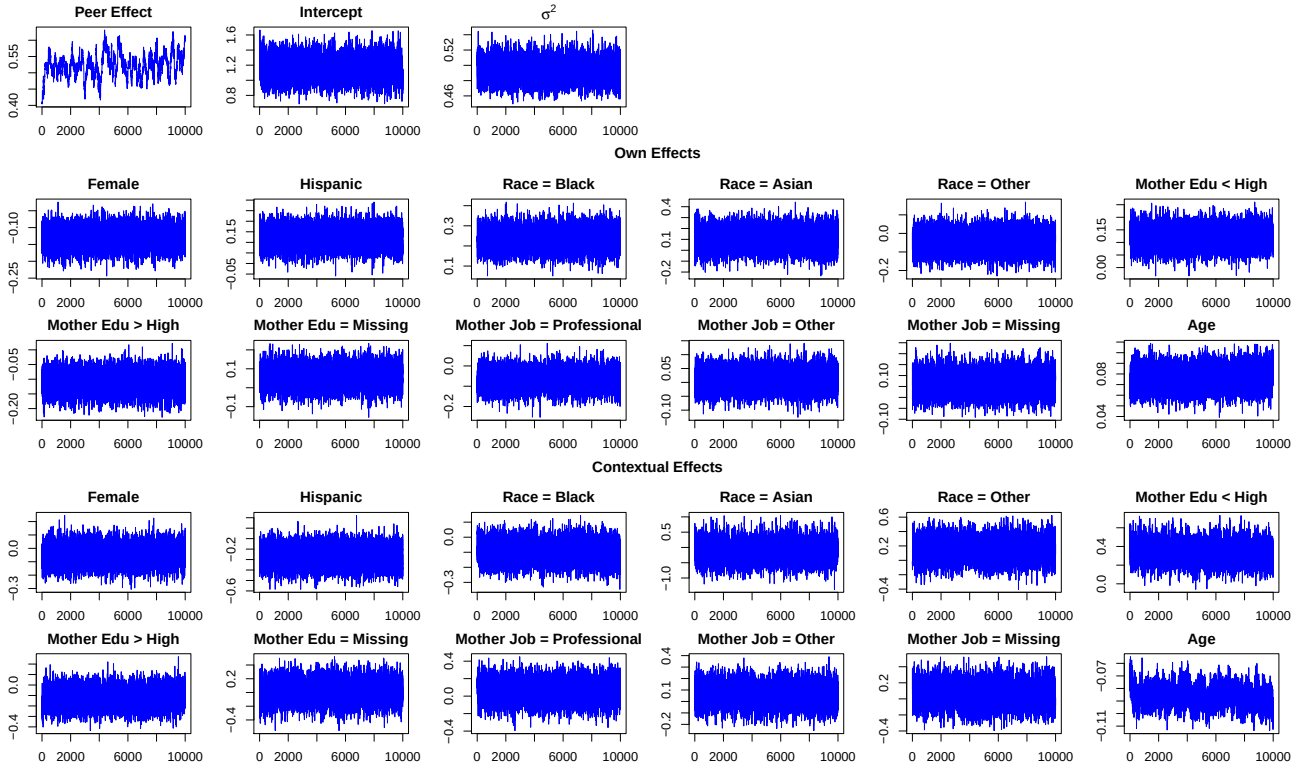
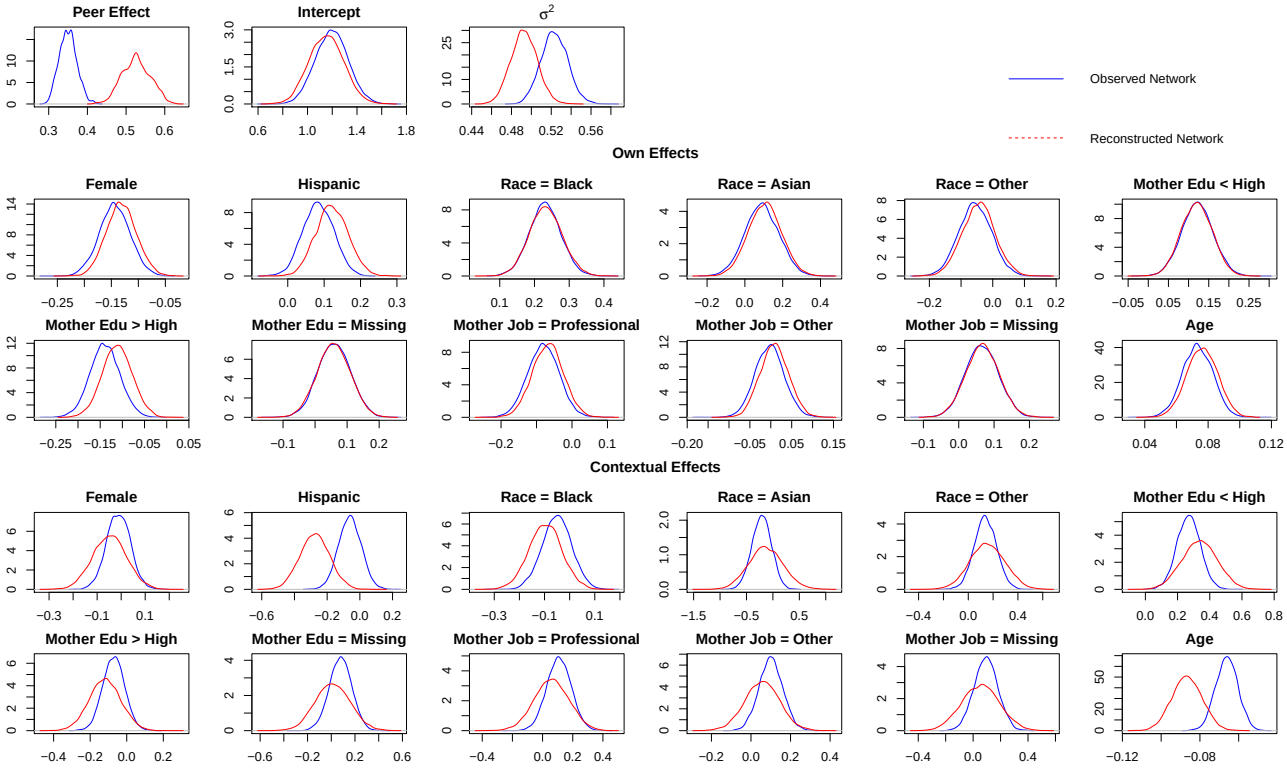


Figure 7: Posterior density.



1. *Use a Metropolis–Hastings algorithm (see Algorithm 1) to obtain the draws* $(\mathbf{A}_1, ..., \mathbf{A}_R)$ *from* $P(\mathbf{A}|\mathbf{y}, \boldsymbol{\theta}_t)$.

2. *Evaluate*
$$Q_t(\boldsymbol{\theta}) \approx \sum_{r=1}^{R} \mathcal{P}(\mathbf{y}|\mathbf{A}_r, \mathbf{X}; \boldsymbol{\theta}_t).$$

3. *Set* $\boldsymbol{\theta}_{t+1} = \arg\max Q_t(\boldsymbol{\theta})$.