Estimating Peer Effects Using Partial Network Data

Vincent Boucher and Elysée Aristide Houndetoungan[†]

June 2022

Abstract

We study the estimation of peer effects through social networks when researchers do not observe the entire network structure. Special cases include sampled networks, censored networks, misclassified links, and aggregated relational data. We assume that researchers can obtain a consistent estimator of the distribution of the network. We show that this assumption is sufficient for estimating peer effects using a linear-in-means model. We provide an empirical application to the study of peer effects on students academic achievement using the widely used Add Health database and show that network data errors have a first-order downward bias on estimated peer effects.

JEL Codes: C31, C36, C51

Keywords: Social networks, Peer effects, Missing variables, Measurement errors

[†]Boucher: Department of Economics, Université Laval, CRREP, CREATE, CIRANO; email: vincent.boucher@ecn.ulaval.ca. Houndetoungan: Cy Cergy Paris Université and Thema; email: aristide.houndetoungan@cyu.fr. We would like to thank Isaiah Andrews, Yann Bramoullé and Bernard Fortin for their helpful comments and insights, as always. We would also like to thank Eric Auerbach, Arnaud Dufays, Stephen Gordon, Chih-Sheng Hsieh, Arthur Lewbel, Tyler McCormick, Angelo Mele, Francesca Molinari, 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. An R package, including all replication codes is available at: https://github.com/ahoundetoungan/PartialNetwork.

1 Introduction

There is a large and growing literature on the impact of peer effects in social networks.[1] However, since eliciting network data is expensive (Breza et al., 2020), relatively few data sets contain comprehensive network information, and existing ones are prone to data errors. Despite some recent contributions, existing methodologies for the estimation of peer effects with incomplete or erroneous network data either focus on a specific kind of sampling or errors, or they are highly computationally demanding.

In this paper, we propose a unifying framework that allows for the estimation of peer effects under 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. Our methodology is computationally attractive and flexible enough to cover cases in which, for example, network data are sampled (Chandrasekhar and Lewis, 2011; Liu, 2013), censored (Griffith, 2019), missclassified (Hardy et al., 2019; Lewbel et al., 2019), or summarized by aggregated relational data (ARD; Breza et al. (2020); Alidaee et al. (2020)). Our central assumption is that the researcher is able to estimate a network formation model using some partial information about the network structure. Leveraging recent contributions on the estimation of network formation models, we show that this assumption is sufficient to identify and estimate peer effects.

We propose two estimators. First, we present a computationally attractive estimator based on a simulated generalized method of moment (SGMM). The moments are built using draws from the (estimated) network formation model. We study the finite sample properties of our SGMM estimator in the context of ARD using Monte Carlo simulations. We show that the estimator performs very well when the distribution of the true network is relatively informative about the realized network structure in the data. Second, we present a flexible Bayesian estimator allowing to exploit the

[1] For recent reviews, see Boucher and Fortin (2016), Bramoullé et al. (2020), Breza (2016), and De Paula (2017).

entire structure of the data generating process. Although the computational cost is greater, we exploit recent computational advances in the literature (e.g. Mele (2017); Hsieh et al. (2019)) and show that the estimator can be successfully implemented on common-sized data sets. In particular, we use our estimator to study peer effects on academic achievement using the widely used Add Health database. We find that data errors have first-order downward bias on the estimated endogenous effect.

Our SGMM estimator is inspired by the literature on error-in-variable models with repeated observations.[2] Using a network formation model, we obtain a consistent estimator of the distribution of the true network. We then use this estimated distribution to obtain different draws from the distribution of the network. Our moment condition can be seen as a biased-corrected version of the instrumental strategy proposed by Bramoullé et al. (2009), in which we substitute the true network with the draws from the estimated distribution. We show that our moment conditions are asymptotically valid for any fixed number of draws from the estimated distribution of the network. This property implies a significant reduction in the computational cost of the method, compared to methods based on the integration of the moment conditions (e.g. Chandrasekhar and Lewis (2011)).

Importantly, our SGMM strategy requires only the (partial) observation of a *single* cross-section, as opposed to, for example, the approach of Zhang (2020). This feature is due to two main properties of the model. First, we can consistently estimate the distribution of the mismeasured variable (i.e. the network) using a single (partial) observation of the variable. Second, in the absence of measurement error, valid instruments for the endogenous peer variable are available (Bramoullé et al., 2009).

We explore the finite sample properties of our instrumental variable estimator using Monte Carlo simulations. We consider the case in which the researcher only has access to aggregated relational data (ARD). In particular, we show that our

[2]See Bound et al. (2001) for a review and Chen et al. (2011) for a review focused on nonlinear models.

SGMM estimator can be successfully combined with the network formation estimators proposed by Breza et al. (2020) and Alidaee et al. (2020). As long as ARD are sufficiently informative in regard to the linking probabilities, our SGMM estimator successfully recovers the model's parameters.

Our Bayesian estimator is based on the likelihood and therefore uses more information about the structure of the model, leading to more precise estimates. In the context of this estimator, the estimated distribution for the network acts as a prior distribution, and the inferred network structure is updated through a Markov chain Monte Carlo (MCMC) algorithm. Our approach relies on data augmentation (Tanner and Wong, 1987), which treats the network as an additional set of parameters to be estimated. This approach saves us from integrating over the $2^{N(N-1)}$ potential networks compatible with the data, which would create an important computational issue. In particular, our MCMC builds on recent developments from the empirical literature on network formation (e.g. Mele (2017); Hsieh et al. (2019, 2020)). We show that the computational cost of our estimator is reasonable and that it can easily be applied to standard data sets.

We study the impact of errors in adolescents' friendship network data for the estimation of peer effects in education (Calvó-Armengol et al., 2009). We show that the widely used Add Health database features many missing links—around 30% of the within-school friendship nominations are coded with error—and that these data errors strongly bias the estimated peer effects. Specifically, we estimate a model of peer effects on students' academic achievement. Our Bayesian estimator probabilistically reconstructs the missing links, and we obtain a consistent estimator of peer effects. The bias due to data errors is qualitatively important. Our estimated endogenous peer effect coefficient is 1.6 times larger than the one obtained by assuming that the data contains no errors.

Related Literature

This paper contributes to the recent literature on the estimation of peer effects when the network is either not entirely observed or is observed with noise. In particular, our framework is valid when network data are either sampled, censored, misclassified, or consist of aggregate relational data.[3] We unify these strands in the literature and provide a flexible and computationally tractable framework for estimating peer effects with incomplete or erroneous network data.

Sampled networks and censoring: 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. 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)). Thirkettle (2019) also focuses on global network statistics, 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. Finally, Zhang (2020) studies program evaluation in a context in which networks are locally sampled and affected by a single type measurement error (either false positives or false negatives, but not both). Assuming that the researcher has access to two measurements of the network for each sampled unit, she presents a nonparametric estimator of the treatment and spillover effects.

Relatedly, Griffith (2019) explores the impact of imposing an upper bound to the number of links when eliciting network data, e.g. "Name your five best friends." He shows, analytically and through simulations, that these bounds may bias the

[3]For related literature that studies the estimation of peer effects when researchers have no network data, see Manresa (2016); De Paula et al. (2018a); Souza (2014); Lewbel et al. (2019).

estimates significantly. He presents a bias-correction method and explores the impact of censoring using two empirical applications. He finds that censoring underestimates peer effects.

We contribute to this literature by proposing two simple and flexible estimators for the estimation of peer effects based on the linear-in-means model (as opposed to network statistics as in Hsieh and Lee (2016) and Thirkettle (2019)). Our estimators do not require many observations of the sampled network (contrary to Zhang (2020)). Similar to Griffith (2019), we find—using the Add Health database—that sampling leads to an underestimation of peer effects, although we find that censoring has negligible impact in the context of peer effects on academic achievement.

Our Bayesian estimator is similar in spirit to the two-stage GMM estimator proposed by Chandrasekhar and Lewis (2011), but it is computationally advantageous. Indeed, their GMM estimator is based on an unconditional moment condition, which requires integrating over the entire set of networks that are compatible in the data. In Chandrasekhar and Lewis (2011), there are $2^{N(N-1)-M}$ such networks, where M is the number of sampled pairs of individuals. Even for small networks, the computational cost is substantial whenever sampling is nontrivial. Our SGMM estimator does not suffer from this computational cost and can produce precise estimates with as little as three network simulations. While our Bayesian estimator is more computationally demanding, we exploit recent developments from the empirical literature on network formation (e.g. Mele (2017); Hsieh et al. (2019)) and show that it is computationally tractable, even when no link is sampled (e.g. with ARD), which would otherwise require integrating over the $2^{N(N-1)}$ networks compatible with the data.

Misclassification: 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 estimator of the treatment effect. Lewbel et al. (2019) show that the estimation of the linear-in-means model is consistent and

that inference is valid if the expected number of misclassified links grows at a rate slower than $O(\sqrt{N})$.

Our model allows for misclassification of all links with positive probability. As in Hardy et al. (2019), we use a network formation model to estimate the probability of false positives and false negatives. However, our two-stage strategy—estimating the network formation model and then the peer effect model—allows for greater flexibility. In particular, our network formation model is allowed to flexibly depend on covariates. This is empirically important, as networks typically feature homophily on observed characteristics (e.g. Currarini et al. (2010); Bramoullé et al. (2012)).

Aggregate relational data: Breza et al. (2020) propose to estimate network effects using aggregate relation data (ARD). These are obtained from such survey questions as, "How many of your friends have trait X?" They present a network formation model that can be estimated using only ARD. They show the validity of their methodology using two empirical applications in which the outcome of interest depends on the summary statistics of the network. Alidaee et al. (2020) present an alternative estimator allowing to recover nonparametrically the linking probability through ARD. Using a low-rank assumption, they present a simple penalized regression.

We show that these recent methodologies can also be applied to the study of peer effects using the linear-in-means model, which significantly expands the scope of the potential applications of these approaches. In particular, we study the finite sample performance of our SGMM estimator using Monte Carlo simulations and show that we can successfully recover the simulated parameters when combined with the network formation estimators proposed by Breza et al. (2020) or by Alidaee et al. (2020).

A main contribution of this paper is that our estimators can be applied to each of the previously mentioned data issues or to their combination. Our two-step approach—first estimating the network formation and then the peer effects—is flexible and computationally attractive. To reduce the implementation costs, we also present an easy-to-use R package—named `PartialNetwork`—for applying our esti-

mators. The package allows replicating all simulations and empirical applications in the paper, including the estimator proposed by Breza et al. (2020). The package is available online at: `https://github.com/ahoundetoungan/PartialNetwork`. Additional implementation examples are provided in the vignette accompanying the package. For example, we show that the implementation of our Bayesian estimator, combined with the estimator for ARD proposed by Breza et al. (2020), is straightforward and computationally tractable.

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 our SGMM estimator and study its performance when combined with the estimators for ARD proposed by Breza et al. (2020) and Alidaee et al. (2020). In Section 4, we present our Bayesian estimation strategy. In Section 5, we present our application to peer effects on academic achievement. Section 6 concludes.

2 The Model

We now formally present our model. We first describe the *linear-in-means* model (Manski, 1993; Bramoullé et al., 2009), arguably the most widely used model for studying peer effects in networks (see Bramoullé et al. (2020) for a recent review). We then introduce our main assumption, characterizing what is known about the structure of the network.

2.1 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 $\mathbf{A}$.[4] Most of our results hold for any function f which preserves the independence among groups (see Assumption 2 below).

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.[5] The parameter α therefore captures the impact of the average outcome of one's peers on their behavior (the endogenous peer effect). The parameter $\boldsymbol{\beta}$ captures the impact of one's characteristics on their behavior (the individual effects). The parameter $\boldsymbol{\gamma}$ captures the impact of the average characteristics of one's peers on their behavior (the contextual peer effects).

The following set of assumptions summarizes our setup.

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

Assumption 2. *The population is partitioned into* $M > 1$ *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.

Assumption 3. *For each group, the outcome and individual characteristics are observed, i.e.* $(\mathbf{y}_r, \mathbf{X}_r)$, $r = 1, ..., M$, *are observed.*

Assumption 4. *Exogeneity:* $\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$.

Assumption 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. Assumption 2 is introduced for exposition purposes; for example, the data could

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

[5] Note that Boucher and Bramoullé (2020) recently showed that (1) is valid (coherent, complete, and microfounded), even when y_i is binary.

consist of a collection of small villages (Banerjee et al., 2013) or schools (Calvó-Armengol et al., 2009). Our methods extend to alternative assumptions such as those proposed by Lee (2004) and Lee et al. (2010).[6]

Assumption 3 implies that the data is composed of a subset of fully sampled groups.[7] A similar assumption is made by Breza et al. (2020). Note that Assumption 4 does not impose restrictions on the dependence between the errors and the network structure. We present the network formation process in details in the next section.

2.2 Partial Network Information

In this paper, we relax the costly assumption that the adjacency matrix $\mathbf{A}$ is observed. We assume instead that sufficient information about the network is observed so that a network formation model can be estimated.

More formally, we let $\mathcal{A}$ denote the observed information about the true network structure. That is, $\mathcal{A}$ is a function of the true network $\mathbf{A}$ and potentially of individuals' characteristics (see Example 4). We impose no particular structure on $\mathcal{A}$ but discuss important examples below (see Examples 1–4).

We assume that links are generated as follows:

$$P(a_{ij}) \equiv P(a_{ij}|\boldsymbol{\rho}_0) \propto \exp\{a_{ij} Q(\boldsymbol{\rho}_0, \mathbf{w}_{ij})\}, \tag{2}$$

where Q is some known continuous function, $\mathbf{w}_{ij} = \mathbf{w}_{ij}(\mathbf{X})$ is a vector of observed characteristics for the pair ij, and $\boldsymbol{\rho}_0$ is the true value of $\boldsymbol{\rho}$, a vector of parameters to be estimated. Note that the assumption that the set of observed characteristics of the pairs $\mathbf{w}_{ij}$ are function of $\mathbf{X}$ implies that the network is *exogenous*. We omit the dependence of $P(a_{ij})$ on $\mathbf{w}_{ij}$ to simplify the notation.

As will be made clear, our estimation strategy requires that the econometrician be able to draw samples from (a consistent estimator) $P(\mathbf{A})$. Thus, and for the sake

[6] The authors assume that the adjacency matrix $\mathbf{A}$ is bounded in row- and column-sums.
[7] Contrary to Liu et al. (2017) or Wang and Lee (2013), for example.

of simplicity, we focus on distributions that are conditionally independent across links (i.e. $P(a_{ij}|\mathbf{A}_{-ij}) = P(a_{ij})$), as in (2), although this is not formally required.[8]

We now present our main assumption.

Assumption 5 (Partial Network Information)**.** *Given network information $\mathcal{A}$ and the parametric model (2), there exists an estimator $\hat{\boldsymbol{\rho}}_N$, such that $\hat{\boldsymbol{\rho}}_N \to_p \boldsymbol{\rho}_0$ as $N \to \infty$.*

Assumption 5 implies that, using (5), the researcher has access to "sufficient information" about the network structure to obtain a consistent estimator of the distribution of the true network $P(\mathbf{A})$. We denote such a consistent estimator by $\hat{P}(\mathbf{A}) \equiv P(\mathbf{A}|\hat{\boldsymbol{\rho}}, \mathcal{A})$. We omit the dependence on $\mathcal{A}$ to simplify the notation when this does not create confusion. Note that we can use $\hat{P}(\mathbf{A})$ to obtain the consistent estimator $\hat{P}(\mathbf{G})$, since $\mathbf{G} = f(\mathbf{A})$ for some known function f.[9]

Importantly, it should be noted that even if the econometrician has access to a consistent estimator of the distribution of the true network, it *does not* imply that they could simply proxy $\mathbf{G}$ in (1) using a draw $\tilde{\mathbf{G}}$ from the distribution $\hat{P}(\mathbf{G})$. The reason is that for any vector $\mathbf{z}$, $\tilde{\mathbf{G}}\mathbf{z}$ generally does not converge to $\mathbf{G}\mathbf{z}$ as the number of individuals N goes to infinity. In other words, knowledge of $\hat{P}(\mathbf{G})$ and $\mathbf{z}$ is not sufficient to obtain a consistent estimator of $\mathbf{G}\mathbf{z}$. To see why, note that $(\mathbf{G}\mathbf{z})_i = \sum_{j=1}^{N} g_{ij} z_j$. Under Assumption 2, the set of j such that $g_{ij} \neq 0$ is bounded, so a consistent estimator of $P(\mathbf{G})$ is not sufficient to obtain a consistent estimator of $\mathbf{G}\mathbf{X}$. This is also true under the alternative assumptions in Lee (2004) and De Paula et al. (2018b),[10] but it contrasts from models based on a large network asymptotic (e.g. Auerbach (2022)).

[8] 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).

[9] Since $\mathbf{A}$ takes a finite number of values, so does $\mathbf{G}$, and $P(\mathbf{G})$ is a multinomial distribution. As such it is a continuous function of the distribution $P(\mathbf{A})$, irrespective of the function f linking $\mathbf{A}$ and $\mathbf{G}$.

[10] Lee (2004) requires that $\mathbf{G}$ be bounded in row- and column-sums (in absolute value), whereas De Paula et al. (2018b) assume that $\mathbf{A}$ is bounded in row-sums.

Assumption 5 covers a large range of cases in which networks are partially observed. We specifically discuss four leading examples in which Assumption 5 holds: *sampled networks* (Example 1), *censored networks* (Example 2), *misclassified network links* (Example 3), and *aggregated relational data* (Example 4).

Example 1 (Sampled Networks)**.** *Suppose that we observe the realizations of a_{ij} for a random sample of m pairs (e.g. Chandrasekhar and Lewis (2011)). Here $\mathcal{A}$ can be represented by an $N \times N$ binary matrix $\mathbf{A}^{obs}$. Conley and Udry (2010) present such a sampling scheme and ask individuals about their relationship with a random sample of other individuals, for example, "Do you know person X?"*

Such a setup is sufficient to consistently estimate flexible network formation models such as the one in Graham (2017). A simpler example would be to assume that the network formation model (2) is equal to $P(a_{ij} = 1) \propto \exp\{\mathbf{w}_{ij}\boldsymbol{\rho}\}$. In this case, a simple logistic regression provides a consistent estimator of $\boldsymbol{\rho}$.

Given this consistent estimator for $\boldsymbol{\rho}$ and the assumed parametric model, we can compute an estimator for the distribution of the true network $\hat{P}(\mathbf{A}) = P(\mathbf{A}|\hat{\boldsymbol{\rho}}, \mathcal{A})$. Here, $\hat{P}(a_{ij}) = a_{ij}^{obs}$ for any sampled link ij, while $\hat{P}(a_{ij}) \propto \exp\{a_{ij}\mathbf{w}_{ij}\hat{\boldsymbol{\rho}}\}$ for the remaining unsampled links.

Example 2 (Censored Network Data)**.** *As discussed in Griffith (2019), network data is often censored. This typically arises when surveyed individuals are asked to name only $T > 1$ links (among the N possible links they may have). Here, $\mathcal{A}$ can be represented by an $N \times N$ binary matrix $\mathbf{A}^{obs}$. We can use censored network data to estimate a network formation model such as $P(a_{ij} = 1) \propto \exp\{\mathbf{w}_{ij}\boldsymbol{\rho}\}$, for example.*

For simplicity, let us assume as in Griffith (2019) that each link has the same probability of being censored. Then, the parameters in $\boldsymbol{\rho}$ (other than the constant) are identified from the observed ratios $\sum_{ij} a_{ij} w_{ij}^k / \sum_{ij} w_{ij}^k$ (for observable characteristic k), as these sufficient statistics are not biased by censoring. To identify the constant, note that we can easily compute the likelihood of the censored degree distribution (i.e.

$n_i = \sum_j a_{ij}$). *That is* $P(n_i = t|\mathbf{W};\boldsymbol{\rho})$ *for observed* $t < T$ *and* $P(n_i \geq T|\mathbf{W};\boldsymbol{\rho})$ *for observed* $t = T$*, which allows identifying the constant.*[11]

Once such estimator of $\boldsymbol{\rho}$ *is obtained, we can compute an estimator for the distribution of the true network* $\hat{P}(\mathbf{A}) = P(\mathbf{A}|\hat{\boldsymbol{\rho}}, \mathcal{A})$*. In particular,* $P(a_{ij}|\hat{\boldsymbol{\rho}}, a_{ij}^{obs} = 1) = 1$ *because observed links necessarily exist. Also note that for all individuals* i*, such that* $n_i < T$*, we have* $P(a_{ij}|\hat{\boldsymbol{\rho}}, a_{ij}^{obs}) = a_{ij}^{obs}$ *for all* j*, as their network data are not censored. Here, the structural model is only used to obtain the probability of links that are not observed for individuals whose links are potentially censored, i.e.* $P(a_{ij}|\hat{\boldsymbol{\rho}}, a_{ij}^{obs} = 0) \propto \exp\{a_{ij}\mathbf{w}_{ij}\hat{\boldsymbol{\rho}}\}$ *for all* ij*, such that* $n_i \geq T$*.*

Example 3 (Misclassification)**.** *Hardy et al. (2019) study cases in which networks are observed but may include misclassified links (i.e. false positives and false negatives). Here,* $\mathcal{A}$ *can be represented by an* $N \times N$ *binary matrix* $\mathbf{A}^{mis}$*. The (consistent) estimation of (2) in such a context follows directly from the existing literature on misclassification in binary outcome models, e.g. Hausman et al. (1998).*

Let q_1 *be the probability of false positives and* q_0 *be the probability of false negatives (both being elements of* $\boldsymbol{\rho}$*). The estimator for the distribution of the true network is given by* $P(a_{ij} = 1|\hat{\boldsymbol{\rho}}, a_{ij}^{mis}) = a_{ij}^{mis}(1 - \hat{q}_1) + (1 - a_{ij}^{mis})\hat{q}_0$*.*

Example 4 (Aggregated Relational Data)**.** *Aggregated relational data (ARD) are obtained from survey questions such as, "How many friends with trait 'X' do you have?" Here,* $\mathcal{A}$ *can be represented by an* $N \times K$ *matrix of integer values, where* K *is the number of traits that individuals were asked about.*

Building on McCormick and Zheng (2015), Breza et al. (2020) proposed a novel approach for the estimation of network formation models using only ARD. They assume:

$$P(a_{ij} = 1) \propto \exp\{\nu_i + \nu_j + \zeta \mathbf{z}_i' \mathbf{z}_j\}, \tag{3}$$

[11] A simple alternative, when the data of only a few individuals is censored, is to estimate the model only for individuals for whom $n_i < T$. We follow this strategy in Section 5.

where ν_i, ν_j, ζ, $\mathbf{z}_i$, and $\mathbf{z}_j$ are not observed by the econometrician but follow parametric distributions. Here, parameters ν_i and ν_j can be interpreted as i and j's propensities 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 (like attracts like) on an abstract latent space (e.g. Hoff et al. (2002)).

Breza et al. (2020) show that it is possible to use aggregate relational data (ARD) to recover the values of the variables in (3). In particular, letting $\boldsymbol{\rho} = [\{\nu_i\}_i, \{\mathbf{z}_i\}_i, \zeta]$, Breza et al. (2019) provide sufficient conditions for the consistent estimation of $\boldsymbol{\rho}$.

Contrary to Examples 1–3, ARD does not provide information on any specific links; [12] *therefore, the predicted distribution of the true network is $P(a_{ij}|\hat{\boldsymbol{\rho}}, \mathcal{A}) = P(a_{ij}|\hat{\boldsymbol{\rho}})$, which is given by Equation (3). Here, it is worth emphasizing that the observed network information (i.e. ARD) is not very informative about the particular network structure in the data. In this sense, it could be viewed as a worse case scenario.*

3 Simulated Generalized Method of Moment Estimators

In this section, we present estimator based on a Simulated Generalized Method of Moments. The key observation underlying our approach is that it is not necessary to observe the complete network structure to observe $\mathbf{y}$, $\mathbf{X}$, $\mathbf{GX}$, and $\mathbf{Gy}$. For example, one could simply obtain $\mathbf{Gy}$ from survey data, "What is the average value of your friends' y?"

However, even observing $\mathbf{y}$, $\mathbf{X}$, $\mathbf{GX}$, and $\mathbf{Gy}$, the model (1) cannot be simply estimated using simple linear regression. The reason is that $\mathbf{Gy}$ is endogenous; thus, a linear regression would produce biased estimates (e.g. Manski (1993), Bramoullé

[12] That is, unless ARD include the degree distribution with some individuals reporting having no links at all.

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?"

However, the above discussion is encouraging. When researchers observe $[\mathbf{y}, \mathbf{X}, \mathbf{GX}, \mathbf{Gy}]$, information about the network is only used as a mean to construct a valid instrument, e.g. following Bramoullé et al. (2009). Then, if $\mathbf{GX}$ and $\mathbf{Gy}$ are observed, the following simple estimator can be used.

Proposition 1. *[Conditions] Suppose that* $\mathbf{GX}$ *and* $\mathbf{Gy}$ *are observed. Let* $\boldsymbol{H}$ *be a matrix such that: (1) at least one column of* $\mathbf{H}^k\mathbf{X}$ *is (strongly) correlated with* $\mathbf{Gy}$*, conditional on* $[\mathbf{1}, \mathbf{X}, \mathbf{GX}]$ *for* $k \geq 2$*, and (2)* $\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}, \mathbf{H}] = \mathbf{0}$*. Finally, define the matrix* $\mathbf{Z} = [\mathbf{1}, \mathbf{X}, \mathbf{GX}, \mathbf{H}^2\mathbf{X}, \mathbf{H}^3\mathbf{X}...]$.

[Results] Then, under classical assumptions (e.g. Cameron and Trivedi (2005), Proposition 6.1), the (linear) GMM estimator based on the moment function $\frac{1}{N}\sum_i \mathbf{Z}_i'\boldsymbol{\varepsilon}_i$ *is consistent and asymptotically normally distributed with the usual asymptotic variance-covariance matrix.*

Condition (1) is the relevancy condition, while condition (2) is the exogeneity condition.[13] While Proposition 1 holds for any matrix $\mathbf{H}$ such that conditions (1) and (2), the most sensible example in our context is when $\mathbf{H}$ is constructed using a draw from $\hat{P}(\mathbf{G})$. A similar strategy is used by Kelejian and Piras (2014), König et al. (2019) and Lee et al. (2020) in a different context.

Importantly, the moment conditions remain valid even when the researcher uses a *mispecified* estimator of the distribution $P(\mathbf{G})$, as long as the specification error

[13] While (for simplicity) in Proposition 1, we use the entire matrix $\mathbf{X}$ in order to generate the instruments $\mathbf{HX}$, in practice, one should avoid including instruments (i.e. columns of $\mathbf{HX}$) that are weakly correlated with $\mathbf{Gy}$.

on $P(\mathbf{G})$ does not induce a correlation with $\boldsymbol{\varepsilon}$.[14] This could be of great practical importance, especially if the estimation of $\hat{P}(\mathbf{G})$ suffers from small sample bias.

3.1 The General Case

We now turn to the more general case where $\mathbf{Gy}$ and $\mathbf{GX}$ are not observed. To simplify the notation, for the remainder of this section, we define $\{\dot{\mathbf{G}}^{(r)}\}_{r=1}^{R}$, $\{\ddot{\mathbf{G}}^{(s)}\}_{s=1}^{S}$ and $\{\dddot{\mathbf{G}}^{(t)}\}_{t=1}^{T}$ as sequences of independent draws from $\hat{P}(\mathbf{G})$. We will also let $\dot{\mathbf{G}}$, $\ddot{\mathbf{G}}$, and $\dddot{\mathbf{G}}$ be single independent draws from $\hat{P}(\mathbf{G})$.

Before presenting our estimation strategy, we want to emphasize that we cannot simply proxy $\mathbf{Gy}$ and $\mathbf{GX}$ using draws form $\hat{P}(\mathbf{G})$. To see why, consider the simple case for which $\boldsymbol{\gamma} = \mathbf{0}$, so the observation of $\mathbf{GX}$ is inconsequential. We have, substituting $\mathbf{Gy}$ with $\ddot{\mathbf{G}}\mathbf{y}$:

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

where $\boldsymbol{\eta} = \alpha[\mathbf{G} - \ddot{\mathbf{G}}]\mathbf{y}$ is the error due to the approximation of $\mathbf{Gy}$ by $\ddot{\mathbf{G}}\mathbf{y}$. Importantly, the approximation error does *not* vanish as n grows, due to the fact that individuals belong to bounded groups (see Assumption 2).[15] Moreover, since $\mathbf{y}$ is a function of $\mathbf{G}$ (but not of $\ddot{\mathbf{G}}$), we typically have $\mathbb{E}(\mathbf{Gy})_i|\mathbf{z}_i \neq \mathbb{E}(\ddot{\mathbf{G}}\mathbf{y})_i|\mathbf{z}_i$ for some vector of instruments $\mathbf{z}_i$ and all i.

Proposition 2 below however shows that the asymptotic bias induced by $\mathbb{E}\eta_i|\mathbf{z}_i \neq 0$ can be bounded, and be very small in practice for carefully constructed vectors of instruments.

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

[15] As opposed, for example, to the context studied by Auerbach (2019).

3.1.1 Asymptotically Biased Estimator

Before presenting our main result (Theorem 1), we present a simple, asymptotically biased, linear GMM estimator. The presentation of such this estimator is useful for two reasons. First, simulations show that the asymptotic bias turns out to be negligible in many cases, especially for moderate group sizes. Moreover, the estimator is computationally attractive since the estimator can be written in closed-form. Second, the estimator helps to understand the logic underlying the estimator defined in Theorem 1, which we can view as a bias-corrected version of this simple linear GMM estimator. Proposition 2 formalizes.

Proposition 2. *[Conditions] Assume that* $\mathbf{GX}$ *is observed. Let* $\ddot{\mathbf{S}} = [\mathbf{1}, \mathbf{X}, \mathbf{GX}, \ddot{\mathbf{G}}\mathbf{X}, \ddot{\mathbf{G}}\mathbf{y}]$ *and* $\dot{\mathbf{Z}} = [\mathbf{1}, \mathbf{X}, \mathbf{GX}, \ddot{\mathbf{G}}\mathbf{X}, \dot{\mathbf{G}}^2\mathbf{X}, \dot{\mathbf{G}}^3\mathbf{X}, ...]$. *Denote by* $\hat{\boldsymbol{\theta}}$ *the linear GMM estimator based on the (pseudo) moment function* $\frac{1}{N}\sum_i \dot{\mathbf{Z}}_i[\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i]$, *and define the sensitivity matrix*

$\mathbf{M}_N = [(\ddot{\mathbf{S}}'\dot{\mathbf{Z}}/N)\mathbf{W}(\dot{\mathbf{Z}}'\ddot{\mathbf{S}}/N)]^{-1}(\ddot{\mathbf{S}}'\dot{\mathbf{Z}}/N)\mathbf{W}$.

[Result] Then, under classical assumptions (see proof), the asymptotic bias of $\hat{\boldsymbol{\theta}}$ *is given by* $\alpha\mathbf{M}_0 \operatorname{plim}[\dot{\mathbf{Z}}'(\mathbf{G} - \ddot{\mathbf{G}})\mathbf{y}/N]$. *Moreover, letting* $\mathbf{W} = \mathbf{I}$ *minimizes the asymptotic bias in the sense of minimizing the Frobenius norm of* $\mathbf{M}$.

While there are no obvious way to get a consistent estimate of the asymptotic bias (because $\mathbf{y}$ is a function of $\mathbf{G}$) simulations show that the bias is very small in practice.[16]

The intuition behind Proposition 2 comes from the literature on error-in-variable models with repeated observations (e.g. Bound et al. (2001)). The instrumental variable uses two independent draws from the (estimated) distribution of the true network. One draw is used to proxy the unobserved variable (i.e. $\mathbf{Gy}$), while the other is used to proxy the instrument (i.e. $\mathbf{GX}$). This approach greatly reduces the bias as compared to a situation in which only one draw would be used.[17]

[16] See Section 3.2 below.
[17] Simulations available uppon request.

The argument in Proposition 2 is very similar to the one in Andrews et al. (2017), although here perturbation with respect to the true model is not *local*.[18] We also show that we expect $\mathbf{W} = \mathbf{I}$ to minimize the asymptotic bias. Our result therefore provides a theoretical justification for the simulations in Onishi and Otsu (2021) who show that the using the identity matrix to weight the moments greatly reduce the bias in the context studied by Andrews et al. (2017).

Conceptually, Proposition 2 states that the simple linear GMM estimator $\hat{\boldsymbol{\theta}}$ is such $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta_0} + F(\boldsymbol{\theta}_0)$ for some known function F. It is therefore possible to use this insight to define a consistent estimator of $\boldsymbol{\theta}_0$. This is what we do in Theorem 1 below.

3.1.2 Main Result

We show that we can include the bias directly in the moment function, which provides a consistent estimator of $\boldsymbol{\theta}$. Let $\tilde{\boldsymbol{\theta}} = [c, \boldsymbol{\beta}', \boldsymbol{\gamma}']'$.

Theorem 1. *[Conditions] Let* $\dot{\mathbf{Z}}^{(r)} = [\mathbf{1}, \mathbf{X}, \dot{\mathbf{G}}^{(r)}\mathbf{X}, (\dot{\mathbf{G}}^{(r)})^2\mathbf{X}, (\dot{\mathbf{G}}^{(r)})^3\mathbf{X}, ...]$ *and* $\ddot{\mathbf{V}}^{(s)} = [\mathbf{1}, \mathbf{X}, \ddot{\mathbf{G}}^{(s)}\mathbf{X}]$. *Consider also the following (simulated) moment function:*

$$\frac{1}{RST}\sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{t=1}^{T}\dot{\mathbf{Z}}_i^{(r)\prime}\left[(\mathbf{I} - \alpha\dddot{\mathbf{G}}^{(t)})_i\left(\mathbf{y} - (\mathbf{I} - \alpha\ddot{\mathbf{G}}^{(s)})^{-1}\ddot{\mathbf{V}}^{(s)}\tilde{\boldsymbol{\theta}}\right)\right] \tag{4}$$

[Result] For any positive integers R, S and T, the (simulated) GMM estimator based on (4) is consistent and asymptotically normal, i.e. $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$, *where $\Delta\boldsymbol{\theta}^*$ and $\mathbf{V}$ are defined in (10) and (12) in the Online Appendix and where $\Delta\boldsymbol{\theta}^* \to 0$ as $N \to \infty$.*

Here the asymptotic variance $\mathbf{V}$ includes the sampling uncertainty, the estimating uncertainty of $\hat{\boldsymbol{\rho}}$, as well as the simulating uncertainty (which decreases as R, S and T grow). The term $\Delta\boldsymbol{\theta}^*$ is introduced due to the fact the our moment condition is equal to zero only asymptotically, and it therefore allows to center the distribution of

[18] See page 1562 in Andrews et al. (2017).

$\hat{\boldsymbol{\theta}}$ for finite N. Similar corrections can be found for penalized likelihood models (e.g. Fan and Li (2001), Fan and Peng (2004), Dzemski (2019))

Importantly, and similar to standard methods of simulated moments (e.g. Gourieroux and Monfort (1993)), the estimation is consistent even for a finite number of simulations (i.e. R, S, and T are finite). Our estimator therefore does not suffer from the curse of dimensionality faced by Chandrasekhar and Lewis (2011).[19]

To understand the intuition behind the use of the moment function (4), let us replace $\mathbf{y} = (\mathbf{I} - \alpha\mathbf{G})^{-1}(\mathbf{V}\tilde{\boldsymbol{\theta}} + \boldsymbol{\varepsilon})$ in equation (4). In doing so, we can rewrite the term inside the triple summation in (4) as a sum of two terms. First, $\dot{\mathbf{Z}}_i^{(r)\prime}\boldsymbol{\varepsilon}_i$, which would be the standard moment function if $\mathbf{G}$ was observed (e.g. Bramoullé et al. (2009)), and second:

$$\dot{\mathbf{Z}}_i^{(r)\prime}\left[(\mathbf{I} - \alpha\dddot{\mathbf{G}}^{(t)})_i\left((\mathbf{I} - \alpha\mathbf{G})^{-1}\mathbf{V} - (\mathbf{I} - \alpha\ddot{\mathbf{G}}^{(s)})^{-1}\ddot{\mathbf{V}}^{(s)}\right)\tilde{\boldsymbol{\theta}}\right]. \tag{5}$$

That second term can be viewed as the bias-correction term of the moment function.

We show that the expectation both terms are equal to 0 as N grows to ∞. For the first term, this is true since the simulated network is exogenous. For the second term, this is true because $\mathbf{G}$, $\dot{\mathbf{G}}$, $\ddot{\mathbf{G}}$ and $\dddot{\mathbf{G}}$ are asymptotically drawn from the same distribution.

For the special case in which $\mathbf{G}$ is observed, one could substitute $\dot{\mathbf{G}}^{(r)} = \ddot{\mathbf{G}}^{(s)} = \dddot{\mathbf{G}}^{(t)} = \mathbf{G}$ for all r, s, t and the first term would simply be $\mathbf{Z}_i'\boldsymbol{\varepsilon}_i$, while the second term would be exactly equal to $\mathbf{0}$.

Note that Theorem 1 is valid whether or not $\mathbf{Gy}$ and $\mathbf{GX}$ are observed since the moment function (4) does not use information on $\mathbf{Gy}$ or $\mathbf{GX}$. It is of course possible to include this additional information, if either one is observed.

For example, assume that $\mathbf{GX}$ is observed, and replace $\ddot{\mathbf{V}}$ with $\mathbf{V}$ in (5). After

[19] The unconditional moment condition in Chandrasekhar and Lewis (2011) is based on the (Monte Carlo) integration of the moment condition $\mathbb{E}(\mathbf{Z}^{(s)\prime}\boldsymbol{\varepsilon}|\mathbf{G})$ over $\mathbf{G}$. Suppose that there are m pairs of individuals with unknown link status, there would be 2^m network structures to integrate from.

some algebra, we obtain:[20]

$$\dot{\mathbf{Z}}_i^{(r)\prime}\left[(\mathbf{I}-\alpha\dddot{\mathbf{G}}^{(t)})_i(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(s)})^{-1}[\alpha(\mathbf{G}-\ddot{\mathbf{G}})\mathbf{y}]\right], \tag{6}$$

plus a term that is equal to zero in expectation.

We can see that the last term in bracket corresponds to the source of the asymptotic bias in Proposition 2: the approximation error of $\mathbf{Gy}$ by $\ddot{\mathbf{G}}\mathbf{y}$. The instruments used are however different and, here, are such that (6) does not converge to 0 as N grows. As for Proposition 2, this is due to the fact that $\mathbf{y}$ is a function of $\mathbf{G}$. It is however possible to construct moment functions for the particular cases for which $\mathbf{GX}$ or $\mathbf{Gy}$ are observed. These can be found in Corollary 1 and 2 of the Online Appendix.

The results of this Section show that the estimation of (1) is possible, even with very limited information about the network structure. In Section 3.2, we study the finite sample properties of our approach using network formation models estimated on aggregated relational data (Breza et al. (2020); see Example 4).

3.2 Finite Sample Performance Using ARD

In this section, we study the small sample performance of the estimator presented in Section 3 when the researcher only has access to ARD (as in Example 4). First, we simulate network data using the model proposed by Breza et al. (2020) and simulate outcomes using the linear in means model (1) conditional on the simulated networks. Second, we estimate the network formation model using the Bayesian estimator proposed by Breza et al. (2020) (yielding $\hat{\boldsymbol{\rho}}_B$), as well as using the classical estimator proposed by Alidaee et al. (2020) (yielding $\hat{\boldsymbol{\rho}}_A$). Third, we estimate the linear-in-means model, using our the estimators presented in Proposition 2 and Theorem 1, based on $\hat{\boldsymbol{\rho}}_A$ and $\hat{\boldsymbol{\rho}}_B$.

[20] See the Online Appendix

Recall that

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

where ν_i, ν_j, ζ, $\mathbf{z}_i$, and $\mathbf{z}_j$ are not observed by the econometrician but follow parametric distributions. We refer the interested reader to McCormick and Zheng (2015), Breza et al. (2020), and Breza et al. (2019) for a formal discussion of the model, including its identification and consistent estimation.

To study the finite sample performance of our instrumental strategy in this context, we simulate 20 groups of 250 individuals each. Within each subpopulation, we simulate the ARD responses as well as a series of observable characteristics. The details of the Monte Carlo simulations can be found in the Online Appendix.

Importantly, the model in (7) is based on a single population framework. Thus, the network formation model must be estimated separately for each of the 20 groups. With only 250 individuals in each group, we therefore expect significant small-sample bias.

We contrast the estimator proposed by Breza et al. (2020) with the one proposed by Alidaee et al. (2020). Whereas Breza et al. (2020) present a parametric Bayesian estimator, Alidaee et al. (2020) propose a (nonparametric) 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.[21] Note, however, that their method only yields a consistent estimator of $\hat{P}(\mathbf{A})$ if the true network is effectively low rank.

Very intuitively, the low-rank assumption implies that linking probabilities were generated from a small number of parameters.Importantly, the model (7) is not necessarily low rank; for example, if the individuals' locations on Earth (i.e. the $\mathbf{z}_i$'s) are uniformly distributed and there are only very few large cities, then the model produced is not low rank and the method proposed by Alidaee et al. (2020) performs

[21] The authors developed user-friendly packages in R and Python. See their paper for links and details.

poorly. Intuitively, in such a case, the fact that individual i lives in city A poorly explains the linking probabilities.

We compare the performance of both estimators as we vary the concentration parameter (that is, κ, see the Online Appendix for details). This has the effect of changing the *effective rank* of the linking probabilities: increasing κ decreases the effective rank.[22] We therefore expect the estimator proposed by Alidaee et al. (2020) to perform better for larger values of κ.

A summary of the results is presented in Tables 1 and 2. The complete results can be found in Tables C.1 and C.2 of the Online Appendix. Table 1 presents the results for the special case where $\mathbf{GX}$ are observed in the data. The table displays the performance of the simple (biased) estimator in Proposition 2 and that of our simulated GMM (see Corollary 1) when the network formation model is estimated by Breza et al. (2020) and Alidaee et al. (2020).

When $\kappa = 0$, the network formation is not low-rank. This disproportionately affects the estimator of Alidaee et al. (2020). Although asymptotically biased, the performance of the IV in Proposition 2 is very good and even seems to outperform our SGMM when using the estimator proposed by Alidaee et al. (2020). Moreover, the overidentification test points (on average) to a violation of the exclusion restriction.

When $\kappa = 15$ the estimators proposed by Breza et al. (2020) and Alidaee et al. (2020) perform similarly and so do our estimators. Both the IV and SGMM estimators recover the true value of α.

We now turn to the more general case where $\mathbf{GX}$ are not observed. Note that in that case, Proposition 2 does not apply. Table 2 presents the performance of our SGMM estimator (Theorem 1) when the network formation process is estimated using the estimators proposed by Breza et al. (2020) and Alidaee et al. (2020), as well as when we assume that the researcher knows the true distribution of the network.

We see that the performance of our estimator, is strongly affected by the quality

[22] We refer the interested reader to Alidaee et al. (2020) for a formal discussion of the effective rank and of its importance for their estimator.

of the first-stage network formation estimator. When based on either the estimator proposed by Breza et al. (2020) or Alidaee et al. (2020), for $\kappa = 0$ or $\kappa = 15$, our SGMM estimator performs poorly.

This poor performance of our SGMM estimator in a context where both **Gy** and **GX** are unobserved was anticipated. This is due to two main reasons. *First*, the consistency of the network formation estimator in Breza et al. (2019) holds as the size of each sub-population goes to infinity, while the consistency of our estimator holds as the number of (bounded) sub-populations goes to infinity. This should affect the performance of our estimator, when based on *estimated* network formation models, but when based on the true distribution of the network.

Second, as discussed in Example 4, ARD provides very little information about the realized network structure in the data (as opposed, for example, as for censoring issues, see Example 2). Then, if the true distribution is vague in the sense that most predicted probabilities are away from 0 or 1, we expect the estimation to be imprecise. This is what happens when $\kappa = 15$, where our estimation based on the true distribution of the network is very imprecise, in a context where the network affects the outcome through both **Gy** and **GX**.

In the next section, we present a likelihood based estimator, which uses more information on the data generating process of the outcome, in order to improve the precision of the estimation.

4 Estimation Using Bayesian Inference

In the previous section, we showed that in a context in which **Gy** and **GX** are unobserved, and that the distribution of the network is imprecise, the SGMM estimator can be imprecise. In this section, we therefore present a likelihood-based estimator. Accordingly, greater structure must be imposed on the errors $\boldsymbol{\varepsilon}$.[23]

[23] One could also use our estimator to solve a classical estimator such as a GMM estimator, like that proposed by Chandrasekhar and Lewis (2011), using either the approach proposed by Chernozhukov

Table 1: Simulation results with ARD and observed $\boldsymbol{GX}$

Parameter	Breza et al.		Alidaee et al.	
	Mean	Std. Dev	Mean	Std. Dev
$\kappa = 0$				
IV (Proposition 2)				
$\alpha = 0.4$	0.392	(0.009)	0.372	(0.027)
Sargan test	1.217	(1.761)	7.779	(10.386)
F-test (Weak inst.)	4259.622	(1140.825)	597.055	(251.886)
SGMM (Corollary 1)				
$\alpha = 0.4$	0.392	(0.010)	0.492	(0.057)
$\kappa = 15$				
IV (Proposition 2)				
$\alpha = 0.4$	0.400	(0.007)	0.399	(0.007)
Sargan test	1.063	(1.496)	2.608	(1.496)
F-test (Weak inst.)	5306.551	(1519.394)	1767.274	(1519.394)
SGMM (Corollary 1)				
$\alpha = 0.4$	0.400	(0.009)	0.428	(0.009)

Note: For each case, we generated 20 independent subnetworks of 250 individuals each. In each subnetwork, the spherical coordinates of individuals are generated from a von Mises–Fisher distribution with a location parameter $(1, 0, 0)$ and intensity parameter κ. Details of the simulation exercise can be found in the Online Appendix B. Predicted probabilities are computed using the mean of the posterior distribution. Sargan test is the statistic of the overidentification test. F-test is the Fisher-statistic of the Weak instruments test. We chose the weight associated with the nuclear norm penalty to minimize the RMSE through cross-validation. This value of $\lambda = 600$ is smaller than the recommended value in Alidaee et al. (2020). See Table C.1 for the estimated values of the other parameters. Instruments are build using only second-degree peers, i.e. $\mathbf{G^2X}$.

Table 2: Simulation results with ARD and unobserved $\boldsymbol{GX}$ (SGMM, Theorem 1)

Parameter	Breza et al.		Alidaee et al.		True distribution	
	Mean	Std. Dev	Mean	Std. Dev	Mean	Std. Dev
			$\kappa = 0$			
$\alpha = 0.4$	0.717	(0.463)	0.700	(0.268)	0.400	(0.056)
			$\kappa = 15$			
$\alpha = 0.4$	0.603	(0.069)	0.870	(0.202)	0.434	(0.394)

Note: For each case, we generated 20 independent subnetworks of 250 individuals each. In each subnetwork, the spherical coordinates of individuals are generated from a von Mises–Fisher distribution with a location parameter $(1, 0, 0)$ and intensity parameter κ. Details of the simulation exercise can be found in the Online Appendix. Predicted probabilities are computed using the mean of the posterior distribution. Sargan test is the statistic of the overidentification test. F-test is the Fisher-statistic of the Weak instruments test. See Table C.2 for the estimated values of the other parameters. Instruments are build using only second-degree peers, i.e. $\mathbf{G^2X}$.

To clarify the exposition, we will focus on the network adjacency matrix $\mathbf{A}$ rather than 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:[24]

$$\ln \mathcal{P}(\mathbf{y}|\mathbf{A}, \boldsymbol{\theta}), \tag{8}$$

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

If the adjacency matrix $\mathbf{A}$ was observed, then $\boldsymbol{\theta}$ 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.

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

and Hong (2003), or the Bayesian empirical likelihood approach by Chib et al. (2018).

[24] Note that under Assumption 2, the likelihood can be factorized across groups.

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 $P(\mathbf{A})$ (or rather from $\hat{P}(\mathbf{A})$) as proposed by Chandrasekhar and Lewis (2011), and taking the average likely lead to poor approximations. A classical way to address this issue is to use an EM algorithm (Dempster et al., 1977). 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 approach is valid for a number of alternative assumptions, as long as it yeilds a computationally tractable likelihood. We have, for $\mathbf{G} = f(\mathbf{A})$,

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

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 $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{A})$, we focus on the posterior $p(\boldsymbol{\theta}, \mathbf{A}|\mathbf{y}, \mathcal{A})$, treating $\mathbf{A}$ as another set of unknown parameters. Note that we now make the dependence on $\mathcal{A}$ explicit for clarity.

Indeed, the identification of the model rests on the a priori information of $\mathbf{A}$. A sensible prior for $\mathbf{A}$ is the consistent estimator of its distribution, i.e. $\hat{P}(\mathbf{A}|\mathcal{A}) \equiv P(\mathbf{A}|\hat{\boldsymbol{\rho}}, \mathcal{A})$. One may wish, however, to also use the information regarding the sampling uncertainty around $\hat{P}(\mathbf{A}|\mathcal{A})$. This is very similar to inference for two-step estimators in a classical setting: estimation uncertainty in the first stage must be

accounted for to provide valid inference in the second stage (see Cameron and Trivedi (2005), Section 6.6, for a discussion).

Let $\pi(\boldsymbol{\rho}|\mathcal{A})$ be the prior density on $\boldsymbol{\rho}$. How to obtain $\pi(\boldsymbol{\rho}|\mathcal{A})$, depending on whether $\hat{\boldsymbol{\rho}}$ is obtained using a Bayesian or classical setting, is discussed in Examples 5 and 6 below. Given $\pi(\boldsymbol{\rho}|\mathcal{A})$, it is possible to obtain draws from the posterior distribution $p(\boldsymbol{\rho}, \mathbf{A}|\mathbf{y}, \mathcal{A})$ using the following MCMC:

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

1. *Draw $\boldsymbol{\rho}^*$ from the proposal distribution $q_\rho(\boldsymbol{\rho}^*|\boldsymbol{\rho}_{t-1})$ and accept $\boldsymbol{\rho}^*$ with probability*

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

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

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

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

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

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

As discussed, Step 1 accounts for the sampling uncertainty around the true value of $\boldsymbol{\rho}$. If the true value of $\boldsymbol{\rho}$ was known (instead of being estimated) Step 1 would not be required. Step 1 shows that the flexibility of the network formation model comes at a cost. For example, Graham (2017) and Breza et al. (2020) propose network formation models for which the number of parameters is $O(N_r)$. In turn, this large number of parameters increases the computational cost of Step 1.

Example 5 (Priors from the Asymptotic Distribution of $\boldsymbol{\rho}$)**.** *In a classical setting, and under the usual assumptions, the estimation of (2) produces a estimator* $\hat{\boldsymbol{\rho}}$ *of* $\boldsymbol{\rho}_0$ *as well as an estimator of the asymptotic variance of* $\hat{\boldsymbol{\rho}}$*, i.e.* $\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}})$*. In this case, we define the prior density* $\pi(\boldsymbol{\rho})$ *as the density of a multivariate normal distribution with mean* $\hat{\boldsymbol{\rho}}$ *and variance-covariance matrix* $\hat{\mathbf{V}}(\hat{\boldsymbol{\rho}})$*.*

Example 6 (Priors from the Posterior Distribution of $\boldsymbol{\rho}$)**.** *In a Bayesian setting, the estimation of* $\boldsymbol{\rho}$ *from the network formation model (2) results in draws from the posterior distribution of* $\boldsymbol{\rho}$*. It is therefore natural to use such a posterior distribution as the prior distribution of* $\mathbf{A}$ *for the estimation based on (8). Performing such a sequential Bayesian updating approach comes with a well-known numerical issue.*[25]

Indeed, the evaluation of the acceptance ratio in Step 1 of Algorithm 1 below requires the evaluation of the density of $\boldsymbol{\rho}$ *at different values. Ideally, one would use the draws from the posterior distribution of* $\boldsymbol{\rho}$ *from the first step (network formation model) and perform a nonparametric kernel density estimation of the posterior distribution. However, when the dimension of* $\boldsymbol{\rho}$ *is large, the kernel density estimation may be infeasible in practice.*

This is especially true for very flexible network formation models, such as the one proposed by Breza et al. (2020), for which the number of parameters to estimate is $O(N_r)$*. In such a case, it might be more reasonable to use a more parametric approach or to impose additional restrictions on the dependence structure of* $\boldsymbol{\rho}$ *across dimensions.*[26]

Detailed distributions for Steps 3 and 4 can be found in the Online Appendix. Step 2, however, involves requires some discussion. Indeed, the idea is the following: given the prior information $P(\mathbf{A}|\boldsymbol{\rho}_{t-1}, \mathcal{A})$, one must be able to draw samples from the posterior distribution of $\mathbf{A}$, given $\mathbf{y}$. This is not a trivial task.

[25] See Thijssen and Wessels (2020) for a recent discussion.

[26] For example, if we assume that the posterior distribution of $\boldsymbol{\rho}$ is jointly normal, the estimation of the mean and variance-covariance matrix is straightforward, even in a high-dimensional setting. Simulations (not reported) suggest that this approach performs well in practice.

In particular, 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, i.e. $a_{ij} \sim P(\cdot|\mathbf{A}_{-ij}, \mathbf{y}, \mathcal{A})$, 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, achieving Step 2 involves an additional computational issue because 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. (2019) propose a blocking technique that facilitates the computation of the determinant.

In any case, note that the computational complexity of Step 2 depends strongly on $P(\mathbf{A}|\boldsymbol{\rho}_{t-1}, \mathcal{A})$, which is a function of the assumed network formation model (2) and of the observed information about the network structure $\mathcal{A}$. For censored network data, for example, most of the network structure is observed (see Example 2). This implies that $P(a_{ij}|\boldsymbol{\rho}_{t-1}, \mathcal{A}) \in \{0, 1\}$ for most pairs ij. As such, few entries of $\mathbf{A}$ must be updated in Step 2. The opposite is true for ARD (see Example 4) for which all entries of $\mathbf{A}$ must be updated.

Then, the appropriate blocking technique depends strongly on $P(\mathbf{A}|\boldsymbol{\rho}_{t-1}, \mathcal{A})$ as well as on the assumed distribution for $\boldsymbol{\varepsilon}$. For the simulations and estimations presented in this paper, we use the Gibbs sampling algorithm for each link, adapting the strategy proposed by Hsieh et al. (2019) to our setting (see Proposition 3 in the

Online Appendix. This can be viewed as a *worse-case* scenario. Nonetheless, the Gibbs sampler performs reasonably well in practice, even for ARD;[27] however, 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 the Online Appendix that is also implemented in our R package `PartialNetwork`.[28]

It is important to note that the complexity of Step 2 is not limited to our Bayesian approach. Classical estimators, such as GMM estimators, face a similar challenge in requiring the integration over the entire set of networks. The strategy used here is to rely on a Metropolis–Hastings algorithm, a strategy that has also been successfully implemented in the related literature on ERGMs (e.g. Snijders (2002); Mele (2017, 2020); Badev (2018); Hsieh et al. (2019, 2020)).

Finally, note that for simple network formation models, it is possible to jointly estimate $\boldsymbol{\rho}$ and $\boldsymbol{\theta}$ within the same MCMC instead of using the two-step procedure described above. In that case, Step 1 can simply be replaced by:

1'. Draw $\boldsymbol{\rho}^*$ from the proposal distribution $q_\rho(\boldsymbol{\rho}^*|\boldsymbol{\rho}_{t-1})$ and accept $\boldsymbol{\rho}^*$ with probability

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

Here, $P(\mathcal{A}|\boldsymbol{\rho}^*)$ is the likelihood of the network information $\mathcal{A}$ assuming the network formation model in (2). Note that $\pi(\boldsymbol{\rho})$, the prior density on $\boldsymbol{\rho}$, no longer depends on $\mathcal{A}$ and can be chosen arbitrarily (e.g. uniform).

This approach would work well for simple models, such as the ones discussed in Examples 1 and 3. It is impractical, however, for more involved models, such as the one proposed by Breza et al. (2020).

[27] Simulations available upon request.

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

5 Imperfectly Measured Networks

In this section, we assume that the econometrician has access to network data but that the data may contain errors. 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 use the widely used AddHealth database and show that network data errors have a first-order impact on the estimated peer effects. Specifically, we focus on a subset of schools from the "In School" sample that each have less than 200 students. Table 3 displays the summary statistics.

Table 3: 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's 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's 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	2.912	0.794	2.333	3.5

Note: We only keep the 33 schools having less than 200 students from the In-School sample. The variable GPA is computed by taking the average grade for English, Mathematics, History, and Science, letting $A = 4$, $B = 3$, $C = 2$, and $D = 1$. Thus, higher scores indicate better academic achievement.

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 censoring: students can only report up to five male and five female friends. We also allow for censoring but show that censoring is not the most important issue with the Add Health data. To understand why, we discuss the organization of the data.

Each adolescent is assigned a unique identifier. The data includes ten variables for the ten potential friendships (maximum of five male and five 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, 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 1 displays the distribution of the number of "unmatched named friends."[29]

To use the methodology developed in Section 4, we first need to estimate a network formation model using the observed network data. In this section, we assume that links are generated using a simple logistic framework, i.e.

$$P(a_{ij} = 1) \propto \exp\{\mathbf{w}_{ij}\boldsymbol{\rho}\},$$

where $\mathbf{w}_{ij}$ is built to capture homophily on the observed characteristics of i and j (see Table 4). We estimate the network formation model on the set of individuals who nominate strictly less than five male and five female friends and for which we

[29] We focus on within-school friendships; thus, nominations outside of school are not treated as "unmatched friends."

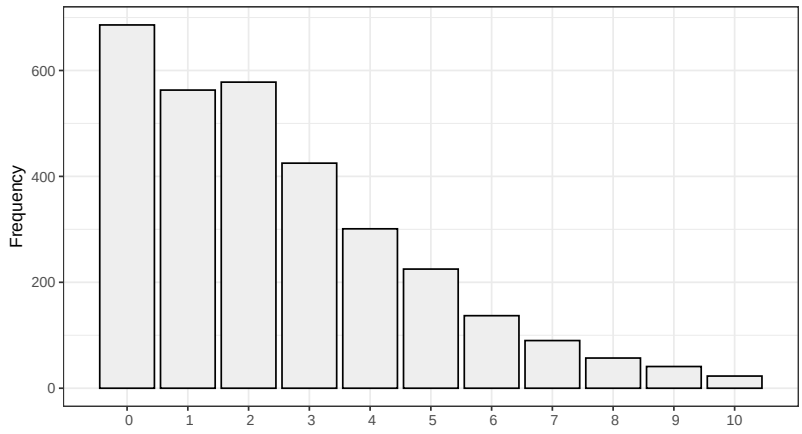

Figure 1: Frequencies of the number of missing links per adolescent.

observe no "unmatched friends." For these students, we know for sure that their friendship data is complete. Under a missing at random assumption, the estimation is consistent.

Table 4 presents the estimation results for the network formation model (panel *Network Formation*), as well as for the peer effect model (panel *Peer Effect Model*, column *Reconstructed Network*). We also present the results for the (inconsistent) estimation of the peer effect model under the assumption that the observed network is the true one (panel *Peer Effect Model*, column *Observed Network*).

We find evidence of homophily on all observed covariates, which is coherent with the literature (e.g. Currarini et al. (2010); Boucher and Mourifié (2017); Mele (2020)). The bias on the peer effects' coefficients (endogenous and contextual) is qualitatively important (and statistically significant). The estimated endogenous peer effect using the reconstructed network is 1.6 times larger than the one estimated assuming that the observed network is the true network. The reconstructed network also allows capturing a positive effect of having a high proportion of Hispanic friends on academic achievement.

The contribution of censoring to these biases is insignificant. In Table C.3 of the Online Appendix, we report estimates for two alternative cases: one in which we

disregard censoring and another in which censoring is the only data issue. The impact of controlling for censoring is almost null, coherent with the fact that only about 5% of the observations are potentially censored.

Our exercise shows that data errors are a first-order concern when using the Add Health database. We do not argue, however, that our estimated coefficients in Table 4 are necessarily causal because the friendship network is likely endogenous (e.g. Goldsmith-Pinkham and Imbens (2013); Hsieh and Van Kippersluis (2018); Hsieh et al. (2020)). The estimation of peer effects with partial endogenous network data is left for future research but is discussed in the next section.

5.1 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. Nonetheless, many important challenges remain, in particular with respect to the study of compatible models of network formation, which can be estimated under nonrandom sampling, for example.

Table 4: Posterior distribution.

	Observed network			Reconstructed network		
Statistic	Mean	Std. Dev.	t-stat	Mean	Std. Dev.	t-stat
Peer effect model						
Peer effects	0.350***	(0.024)	14.809	0.564***	(0.044)	12.842
Own effects						
Female	0.144***	(0.029)	5.018	0.120***	(0.031)	3.822
Hispanic	−0.083**	(0.042)	−1.980	−0.145**	(0.047)	3−.090
Race (White)						
Black	−0.230***	(0.045)	−5.070	−0.210***	(0.055)	−3.787
Asian	−0.091	(0.089)	−1.024	−0.115	(0.089)	−1.286
Other	0.055	(0.051)	1.064	0.036	(0.052)	0.686
Mother's education (High)						
<high	−0.122**	(0.039)	−3.147	−0.126**	(0.039)	−3.201
>high	0.140***	(0.034)	4.131	0.109**	(0.034)	3.208
missing	−0.060	(0.050)	−1.182	−0.065	(0.051)	−1.277
Mother's job (Stay-at-home)						
Professional	0.080*	(0.045)	1.793	0.078*	(0.044)	1.759
Other	0.003	(0.035)	0.078	−0.006	(0.035)	−0.168
Missing	−0.066	(0.047)	−1.400	−0.068	(0.048)	−1.427
Age	−0.073***	(0.010)	−7.606	−0.078***	(0.011)	−7.168
Contextual effects						
Female	0.011	(0.049)	0.221	0.072	(0.092)	0.779
Hispanic	0.060	(0.069)	0.868	0.353**	(0.115)	3.081
Race (White)						
Black	0.050	(0.058)	0.865	0.058	(0.075)	0.769
Asian	0.209	(0.186)	1.126	0.518	(0.579)	0.895
Other	−0.137	(0.089)	−1.537	−0.230	(0.185)	−1.241
Mother's education (High)						
<High	−0.269***	(0.070)	−3.841	−0.363**	(0.158)	−2.290
>High	0.072	(0.059)	1.213	0.062	(0.114)	0.540
Missing	−0.077	(0.093)	−0.835	−0.025	(0.197)	−0.128
Mother's job (Stay-at-home)						
Professional	−0.110	(0.080)	−1.373	0.032	(0.157)	0.204
Other	−0.101*	(0.060)	−1.686	−0.051	(0.114)	−0.451
Missing	−0.093	(0.085)	−1.086	−0.006	(0.188)	−0.029
Age	0.066***	(0.006)	11.323	0.091***	(0.010)	9.228
SE^2	0.523			0.515		
Network formation model						
Same sex				0.456***	(0.016)	28.208
Both Hispanic				0.343***	(0.023)	14.739
Both White				0.320***	(0.022)	14.741
Both Black				1.167***	(0.032)	36.121
Both Asian				0.278***	(0.043)	6.502
Mothers education <High				0.178***	(0.018)	10.063
Mothers education >High				0.075***	(0.016)	4.756
Mothers job Professional				−0.130***	(0.017)	−7.544
Age absolute difference				−0.648***	(0.010)	−64.105

Note: N = 3,126. Observed links = 17,993. Proportion of inferred network data = 77.7%. Significance levels: *** = 1%, ** = 5%, * = 10%. The explained variable is computed by taking the average grade for English, Mathematics, History, and Science, letting $A = 4$, $B = 3$, $C = 2$, and $D = 1$. Thus, higher scores indicate better academic achievement.

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*.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the sensitivity of parameter estimates to estimation moments," *The Quarterly Journal of Economics*, 132, 1553–1592.

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

Auerbach, E. (2019): "Measuring Differences in Stochastic Network Structure," *arXiv preprint arXiv:1903.11117*.

——— (2022): "Identification and estimation of a partially linear regression model using network data," *Econometrica*, 90, 347–365.

Badev, A. (2018): "Nash equilibria on (un) stable networks," *arXiv preprint arXiv:1901.00373*.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013): "The diffusion of microfinance," *Science*, 341, 1236498.

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.

Boucher, V. and Y. Bramoullé (2020): "Binary Outcomes and Linear Interactions," Tech. rep., Aix-Marseille School of Economics, France.

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.

Bound, J., C. Brown, and N. Mathiowetz (2001): "Measurement error in survey data," in *Handbook of econometrics*, Elsevier, vol. 5, 3705–3843.

Bramoullé, Y., S. Currarini, M. O. Jackson, P. Pin, and B. W. Rogers (2012): "Homophily and long-run integration in social networks," *Journal of Economic Theory*, 147, 1754–1786.

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

——— (2020): "Peer effects in networks: A survey," *Annual Review of Economics*, 12, 603–629.

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 (2019): "Consistently estimating graph statistics using Aggregated Relational Data," *arXiv preprint arXiv:1908.09881*.

——— (2020): "Using aggregated relational data to feasibly identify network structure without network data," *American Economic Review*, 110, 2454–84.

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

Cameron, A. C. and P. K. Trivedi (2005): *Microeconometrics: methods and applications*, Cambridge university press.

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.

Chen, X., H. Hong, and D. Nekipelov (2011): "Nonlinear models of measurement errors," *Journal of Economic Literature*, 49, 901–37.

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.

Chib, S., M. Shin, and A. Simoni (2018): "Bayesian estimation and comparison of moment condition models," *Journal of the American Statistical Association*, 113, 1656–1668.

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

Currarini, S., M. O. Jackson, and P. Pin (2010): "Identifying the roles of race-based choice and chance in high school friendship network formation," *Proceedings of the National Academy of Sciences*, 107, 4857–4861.

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.

Dzemski, A. (2019): "An empirical model of dyadic link formation in a network with unobserved heterogeneity," *Review of Economics and Statistics*, 101, 763–776.

Fan, J. and R. Li (2001): "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, 96, 1348–1360.

Fan, J. and H. Peng (2004): "Nonconcave penalized likelihood with a diverging number of parameters," *The annals of statistics*, 32, 928–961.

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

Gourieroux, C. and A. Monfort (1993): "Simulation-based inference: A survey with special reference to panel data models," *Journal of Econometrics*, 59, 5–33.

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

Griffith, A. (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*.

Hausman, J. A., J. Abrevaya, and F. M. Scott-Morton (1998): "Misclassification of the dependent variable in a discrete-response setting," *Journal of econometrics*, 87, 239–269.

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. Kováříík, 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 (2019): "A structural model for the coevolution of networks and behavior," *Review of Economics and Statistics*, 1–41.

Hsieh, C.-S. and L. F. Lee (2016): "A social interactions model with endogenous friendship formation and selectivity," *Journal of Applied Econometrics*, 31, 301–319.

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

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.

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.

König, M. D., X. Liu, and Y. Zenou (2019): "R&D networks: Theory, empirics, and policy implications," *Review of Economics and Statistics*, 101, 476–491.

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," *Journal 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.

——— (2020): "Does school desegregation promote diverse interactions? An equilibrium model of segregation within schools," *American Economic Journal: Economic Policy*, 12, 228–57.

Onishi, R. and T. Otsu (2021): "Sample sensitivity for two-step and continuous updating GMM estimators," *Economics Letters*, 198, 109685.

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.

Thijssen, B. and L. F. Wessels (2020): "Approximating multivariate posterior distribution functions from Monte Carlo samples for sequential Bayesian inference," *PloS one*, 15, e0230101.

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.

Withers, C. S. (1981): "Central limit theorems for dependent variables. I," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57, 509–534.

Zhang, L. (2020): "Spillovers of Program Benefits with Mismeasured Networks," *arXiv preprint arXiv:2009.09614*.

Online Appendix

A Appendix – GMM estimator: assumptions and proofs

Assumption 6. *For all* $s = 1, ..., S$ *and* $t = 1, ..., T$, $(\mathbf{I} - \alpha\dot{\mathbf{G}}^s)$ *and* $(\mathbf{I} - \alpha\ddot{\mathbf{G}}^t)$ *are non-singular*

In particular, when $\mathbf{G}$ is row-normalized (so $\dot{\mathbf{G}}^s$ and $\ddot{\mathbf{G}}^s$ are also row-normalized), Assumption 1 implies Assumption 6.

We consider an objective function for the GMM estimator of the usual form:

$$\left(\tfrac{1}{N}\textstyle\sum_i \mathbf{m}_i(\boldsymbol{\theta})\right)' \mathbf{W}_N \left(\tfrac{1}{N}\textstyle\sum_i \mathbf{m}_i(\boldsymbol{\theta})\right),$$

where $\mathbf{m}_i(\boldsymbol{\theta})$ is the moment function.

Assumption 7. $\mathbf{W}_N$ *is symmetric and positive definite with* $\mathbf{W}_N \to_p \mathbf{W}_0$ *for* $\mathbf{W}_0$ *also symmetric and positive definite.*

A.1 Proof of Proposition 2

Part 1: Asymptotic normality Define $\boldsymbol{\Delta} = \mathbf{G} - \ddot{\mathbf{G}}$ and rewrite $\boldsymbol{\eta} = \alpha[\mathbf{G} - \ddot{\mathbf{G}}]\mathbf{y} = \alpha\boldsymbol{\Delta}\mathbf{y}$. Let $\boldsymbol{\theta}_0^+ = (\boldsymbol{\theta}_0', 0, 0, \ldots, 0)'$ be the true value the parameter when regressors are defined as: $\ddot{\mathbf{S}} = [\mathbf{1}, \mathbf{X}, \mathbf{G}\mathbf{X}, \ddot{\mathbf{G}}\mathbf{X}, \ddot{\mathbf{G}}\mathbf{y}]$. The Linear GMM estimator can be written as:

$$\hat{\boldsymbol{\theta}} = \left[\left(\tfrac{\ddot{\mathbf{S}}'\dot{\mathbf{Z}}}{N}\right)\mathbf{W}_N\left(\tfrac{\dot{\mathbf{Z}}'\ddot{\mathbf{S}}}{N}\right)\right]^{-1}\left(\tfrac{\ddot{\mathbf{S}}'\dot{\mathbf{Z}}}{N}\right)\mathbf{W}_N\left(\tfrac{\dot{\mathbf{Z}}'\ddot{\mathbf{S}}}{N}\boldsymbol{\theta}_0^+ + \tfrac{\dot{\mathbf{Z}}'[\boldsymbol{\eta}+\boldsymbol{\varepsilon}]}{N}\right),$$
$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0^+ + \mathbf{M}_N\left(\tfrac{\dot{\mathbf{Z}}'\boldsymbol{\eta}+\dot{\mathbf{Z}}'\boldsymbol{\varepsilon}}{N}\right),$$

where $\mathbf{M}_N = \left[\left(\frac{\ddot{\mathbf{S}}'\dot{\mathbf{Z}}}{N}\right)\mathbf{W}_N\left(\frac{\dot{\mathbf{Z}}'\ddot{\mathbf{S}}}{N}\right)\right]^{-1}\left(\frac{\ddot{\mathbf{S}}'\dot{\mathbf{Z}}}{N}\right)\mathbf{W}_N$.
Let $\boldsymbol{\theta}_N^* = \mathbf{M}_N\,\mathbb{E}\left(\frac{\dot{\mathbf{Z}}'\boldsymbol{\eta}}{N}\right)$. Then,

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0^+ - \boldsymbol{\theta}_N^*) = \mathbf{M}_N N^{-1/2}\left(\dot{\mathbf{Z}}'\boldsymbol{\eta} + \dot{\mathbf{Z}}'\boldsymbol{\varepsilon} - \mathbb{E}(\dot{\mathbf{Z}}'\boldsymbol{\eta})\right).$$

We can apply the Central Theorem Limit (CTL) to the term $N^{-1/2}\left(\dot{\mathbf{Z}}'\boldsymbol{\eta} + \dot{\mathbf{Z}}'\boldsymbol{\varepsilon} - \mathbb{E}(\dot{\mathbf{Z}}'\boldsymbol{\eta})\right)$ which can be written as $N^{-1/2}\sum_m \boldsymbol{\zeta}_{N,m}^{\hat{P}}$, where $\boldsymbol{\zeta}_{N,m}^{\hat{P}}$ is the sum $\dot{\mathbf{Z}}_i'\boldsymbol{\eta}_i + \dot{\mathbf{Z}}_i'\boldsymbol{\varepsilon}_i - \mathbb{E}(\dot{\mathbf{Z}}_i'\boldsymbol{\eta}_i)$ in the m-th sub-group. If the network distribution were observed, the CLT would follow directly when M grows to ∞, because the sub-groups are independents by Assumption 2 and then $(\boldsymbol{\zeta}_{N,m}^{P})_m$ would be a sequence of independent variables. In the current setup, $\boldsymbol{\zeta}_{N,m}^{\hat{P}}$'s are not independent because they all depend on $\hat{\boldsymbol{\rho}}$ which is random. However since $\boldsymbol{\zeta}_{N,m}^{\hat{P}} \xrightarrow{p} \boldsymbol{\zeta}_{N,m}^{P}$, then $\boldsymbol{\zeta}_{N,m}^{\hat{P}}$'s are asymptotically independent. We use the CLT for *ℓ-mixing* array (see Withers, 1981). Under regular conditions, in particular, $\mathbb{V}(N^{-1/2}\sum_m \boldsymbol{\zeta}_{N,m}^{\hat{P}}) < \infty$ as M grows to ∞, the CLT implies that $N^{-1/2}\sum_m \boldsymbol{\zeta}_{N,m}^{\hat{P}}$ converges in distribution to a $N(\mathbf{0}, \boldsymbol{\Sigma}_0)$, where $\boldsymbol{\Sigma}_0 = \lim_{N\to\infty}\boldsymbol{\Sigma}_N$ and $\boldsymbol{\Sigma}_N = \mathbb{V}(N^{-1/2}\sum_m \boldsymbol{\zeta}_{N,m}^{\hat{P}})$.
$\boldsymbol{\Sigma}_0$ can be estimated by $\boldsymbol{\Sigma}_N$, where

$$\boldsymbol{\Sigma}_N = \mathbb{E}(\mathbb{V}(N^{-1/2}\textstyle\sum_m \boldsymbol{\zeta}_{N,m}^{\hat{P}}|\hat{P})) + \mathbb{V}(\mathbb{E}(N^{-1/2}\textstyle\sum_m \boldsymbol{\zeta}_{N,m}^{\hat{P}}|\hat{P})),$$
$$\boldsymbol{\Sigma}_N = N^{-1}\,\mathbb{E}(\textstyle\sum_m \mathbb{V}(\boldsymbol{\zeta}_{N,m}^{\hat{P}}|\hat{P})) + N^{-1}\,\mathbb{V}(\textstyle\sum_m \mathbb{E}(\boldsymbol{\zeta}_{N,m}^{\hat{P}}|\hat{P})).$$

The term $N^{-1}\,\mathbb{E}(\sum_m \mathbb{V}(\boldsymbol{\zeta}_{N,m}^{\hat{P}}|\hat{P}))$ is due to the error term of the model $\boldsymbol{\varepsilon}$ and the fact that the true network is not observed whereas $N^{-1}\,\mathbb{V}(\sum_m \mathbb{E}(\boldsymbol{\zeta}_{N,m}^{\hat{P}}|\hat{P}))$ is due to uncertainty associated with the estimation of P.
Besides, by the law of large numbers (LLN), $\mathbf{M}_N \xrightarrow{p} \mathbf{M}_0$. As a result,

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0^+ + \boldsymbol{\theta}_N^*) \xrightarrow{d} N(0, \mathbf{M}_0\boldsymbol{\Sigma}_0\mathbf{M}_0'). \tag{9}$$

The asymptotic bias of $\hat{\boldsymbol{\theta}}$ is then given by $\operatorname{plim}\boldsymbol{\theta}_N^* = \alpha\mathbf{M}_0\operatorname{plim}[\dot{\mathbf{Z}}'(\mathbf{G} - \ddot{\mathbf{G}})\mathbf{y}/N]$.

Part 2: Choice of W Let $\mathbf{K} = \dot{\mathbf{Z}}'\boldsymbol{\Delta}\mathbf{G}^2/N$ if $\boldsymbol{\gamma} = \mathbf{0}$ and $\mathbf{K} = \dot{\mathbf{Z}}'\boldsymbol{\Delta}/N$ otherwise. Consider $\|\mathbf{MK}\|_F = \sqrt{trace(\mathbf{K}'\mathbf{M}'\mathbf{MK})} = \sqrt{trace(\mathbf{KK}'\mathbf{M}'\mathbf{M})}$. We have:

$$(1/N^2)\mathbf{MM}' = [\ddot{\mathbf{S}}'\dot{\mathbf{Z}}\mathbf{W}\dot{\mathbf{Z}}'\ddot{\mathbf{S}}]^{-1}\ddot{\mathbf{S}}'\dot{\mathbf{Z}}\mathbf{W}\mathbf{W}\dot{\mathbf{Z}}'\ddot{\mathbf{S}}[\ddot{\mathbf{S}}'\dot{\mathbf{Z}}\mathbf{W}\dot{\mathbf{Z}}'\ddot{\mathbf{S}}]^{-1}$$

Let $\mathbf{W} = \mathbf{C}'\mathbf{C}$ and let $\mathbf{B} = \ddot{\mathbf{S}}'\dot{\mathbf{Z}}\mathbf{C}'$. We have:

$$(1/N^2)\mathbf{MM}' = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{C}\mathbf{C}'\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}$$

Now, define $\mathbf{J}' = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{C} - (\mathbf{B}'(\mathbf{C}')^{-1}\mathbf{C}^{-1}\mathbf{B})^{-1}\mathbf{B}'(\mathbf{C}')^{-1}$. We have:

$$(1/N^2)\mathbf{MM}' = \mathbf{J}'\mathbf{J} + (\mathbf{B}'(\mathbf{C}')^{-1}\mathbf{C}^{-1}\mathbf{B})^{-1} = \mathbf{J}'\mathbf{J} + (\ddot{\mathbf{S}}'\dot{\mathbf{Z}}\dot{\mathbf{Z}}'\ddot{\mathbf{S}})^{-1}$$

Therefore, we have:

$$(1/N^2)\|\mathbf{M}\|_F = \sqrt{trace(\mathbf{J}'\mathbf{J} + (\ddot{\mathbf{S}}'\dot{\mathbf{Z}}\dot{\mathbf{Z}}'\ddot{\mathbf{S}})^{-1})} = \sqrt{trace(\mathbf{J}'\mathbf{J}) + trace((\ddot{\mathbf{S}}'\dot{\mathbf{Z}}\dot{\mathbf{Z}}'\ddot{\mathbf{S}})^{-1})}$$

When $\mathbf{W} = \mathbf{I}$, we have that $\mathbf{J} = \mathbf{0}$ and the Frobenius norm of $\mathbf{M}$ is given by $N^2\sqrt{trace((\ddot{\mathbf{S}}'\dot{\mathbf{Z}}\dot{\mathbf{Z}}'\ddot{\mathbf{S}})^{-1})}$.

A.2 Proof of Theorem 1

A.2.1 Validity of the moment condition and consistency

Let

$$\mathbf{m}_i(\boldsymbol{\theta}) = \tfrac{1}{ST}\textstyle\sum_s\sum_t \dot{\mathbf{Z}}_i^{(s)\prime}(\mathbf{I} - \alpha\dot{\mathbf{G}}^{(s)})_i\left(\mathbf{y} - (\mathbf{I} - \alpha\ddot{\mathbf{G}}^{(t)})^{-1}\ddot{\mathbf{V}}^{(t)}\tilde{\boldsymbol{\theta}}\right)$$

By construction, we have: $\operatorname{plim}\mathbb{E}\left(\mathbf{m}_i(\boldsymbol{\theta})\right) = \mathbf{0}$ for any finite S and T.

To see why, lets substitute $\mathbf{y} = (\mathbf{I} - \alpha\mathbf{G})^{-1}(\mathbf{V}\tilde{\boldsymbol{\theta}} + \boldsymbol{\varepsilon})$ in the moment function. We

have:

$$\begin{aligned}\mathbf{m}_i(\boldsymbol{\theta}) &= \tfrac{1}{ST}\sum_s\sum_t \dot{\mathbf{Z}}_i^{(s)\prime}(\mathbf{I}-\alpha\dot{\mathbf{G}}^{(s)})_i\left[(\mathbf{I}-\alpha\mathbf{G})^{-1}\mathbf{V}-(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(t)})^{-1}\ddot{\mathbf{V}}^{(t)}\right]\tilde{\boldsymbol{\theta}}\\ &+ \tfrac{1}{S}\sum_s \dot{\mathbf{Z}}_i^{(s)\prime}(\mathbf{I}-\alpha\dot{\mathbf{G}}^{(s)})_i(\mathbf{I}-\alpha\mathbf{G})^{-1}\boldsymbol{\varepsilon}\end{aligned}$$

It is therefore sufficient that for any s,t:

$$\mathbb{E}\left(\dot{\mathbf{Z}}_i^{(s)\prime}(\mathbf{I}-\alpha\dot{\mathbf{G}}^{(s)})_i(\mathbf{I}-\alpha\mathbf{G})^{-1}\boldsymbol{\varepsilon}\right)=\mathbf{0},$$

which holds directly from Assumption 1 and the specification of the network formation process (see Assumption 5), and that

$$\operatorname{plim}\mathbb{E}\left[\dot{\mathbf{Z}}_i^{(s)\prime}(\mathbf{I}-\alpha\dot{\mathbf{G}}^{(s)})_i\left[(\mathbf{I}-\alpha\mathbf{G})^{-1}\mathbf{V}-(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(t)})^{-1}\ddot{\mathbf{V}}^{(t)}\right]\tilde{\boldsymbol{\theta}}\right]=\mathbf{0},$$

where plim is the probability limit.

Note that, for N fixed, $\mathbf{G}$, $\dot{\mathbf{G}}^{(s)}$, and $\ddot{\mathbf{G}}^{(t)}$ have a finite number of possible realizations. Thus, $\mathbb{E}\left[\dot{\mathbf{Z}}_i^{(s)\prime}(\mathbf{I}-\alpha\dot{\mathbf{G}}^{(s)})_i(\mathbf{I}-\alpha\mathbf{G})^{-1}\mathbf{V}\tilde{\boldsymbol{\theta}}\right]$ can be written as a continuous function of P, the distribution of the adjacency matrix, and $\hat{P}$, a consistent estimator of that distribution.[30] Analogously, $\mathbb{E}\left[\dot{\mathbf{Z}}_i^{(s)\prime}(\mathbf{I}-\alpha\dot{\mathbf{G}}^{(s)})_i(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(t)})^{-1}\ddot{\mathbf{V}}^{(t)}\tilde{\boldsymbol{\theta}}\right]$ also can be written as a same function with the difference that P is replaced by $\hat{P}$. As both functions are continuous and since $\hat{P}$ converges in probability to P (see Assumption 5), their limits are obtained by replacing $\hat{P}$ by P and then equal. As such, we have:

$$\operatorname{plim}\mathbb{E}\left[\dot{\mathbf{Z}}_i^{(s)\prime}(\mathbf{I}-\alpha\dot{\mathbf{G}}^{(s)})_i(\mathbf{I}-\alpha\mathbf{G})^{-1}\mathbf{V}\tilde{\boldsymbol{\theta}}\right]=\operatorname{plim}\mathbb{E}\left[\dot{\mathbf{Z}}_i^{(s)\prime}(\mathbf{I}-\alpha\dot{\mathbf{G}}^{(s)})_i(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(t)})^{-1}\ddot{\mathbf{V}}^{(t)}\tilde{\boldsymbol{\theta}}\right].$$

The GMM estimator

$$\hat{\boldsymbol{\theta}}=\arg\max\left(\tfrac{1}{N}\textstyle\sum_i\mathbf{m}_i(\boldsymbol{\theta})\right)'\mathbf{W}_N\left(\tfrac{1}{N}\textstyle\sum_i\mathbf{m}_i(\boldsymbol{\theta})\right)$$

[30] The function would essentially composed of summations and products of terms depending on P and $\hat{P}$.

is therefore consistent under Assumptions 2-7 and an identification condition, which we now discuss.

A.2.2 Concentration and Identification

While the GMM estimator could be solved numerically, it is helpful to concentrate the objective function around α. Let $\mathbf{R}(\alpha) = \frac{1}{ST}\sum_s\sum_t \dot{\mathbf{Z}}^{(s)\prime}(\mathbf{I} - \alpha\dot{\mathbf{G}}^{(s)})(\mathbf{I} - \alpha\ddot{\mathbf{G}}^{(t)})^{-1}\ddot{\mathbf{V}}^{(t)}$ and $\mathbf{D}(\alpha) = \frac{1}{ST}\sum_s\sum_t \dot{\mathbf{Z}}^{(s)\prime}(\mathbf{I} - \alpha\dot{\mathbf{G}}^{(s)})$ so that $\frac{1}{N}\sum_i \mathbf{m}_i(\boldsymbol{\theta}) = \mathbf{D}(\alpha)\mathbf{y} - \mathbf{R}(\alpha)\tilde{\boldsymbol{\theta}}$.

The gradient of the objective function with respect to $\tilde{\boldsymbol{\theta}}$ is:

$$-2\left(\tfrac{1}{N}\mathbf{D}(\alpha)\mathbf{y} - \mathbf{R}(\alpha)\tilde{\boldsymbol{\theta}}\right)'\mathbf{W}_N\mathbf{R}(\alpha).$$

The hessian is therefore: $\mathbf{R}'(\alpha)\mathbf{W}_N\mathbf{R}(\alpha)$ which is semi-positive definite. Hence, the following assumption ensures identification of $\tilde{\boldsymbol{\theta}}$, conditional on α.

Assumption 8. $\mathbf{R}'(\alpha)\mathbf{W}_N\mathbf{R}(\alpha)$ *is non-singular for all* α.

A sufficient condition is that $\mathbf{R}'(\alpha)$ has full rank, for all α.
Under Assumption 8, the estimator of $\tilde{\boldsymbol{\theta}}$ conditionally on α is
$\hat{\tilde{\boldsymbol{\theta}}}(\alpha) = (\mathbf{R}'(\alpha)\mathbf{W}_N\mathbf{R}(\alpha))^{-1}\mathbf{R}'(\alpha)\mathbf{W}_N\mathbf{D}(\alpha)\mathbf{y}$.
The objective function can then be concentrated around α, as

$$\left(\tfrac{1}{N}\mathbf{D}(\alpha)\mathbf{y} - \mathbf{R}(\alpha)\hat{\tilde{\boldsymbol{\theta}}}(\alpha)\right)'\mathbf{W}_N\left(\tfrac{1}{N}\mathbf{D}(\alpha)\mathbf{y} - \mathbf{R}(\alpha)\hat{\tilde{\boldsymbol{\theta}}}(\alpha)\right).$$

While we cannot easily ensure that this function is globally convex in α, the behavior of the function is easy to study since α is uni-dimensional and has a bounded support (see Assumption 1). Note that numerically minimizing the concentrated objective function implicitly checks Assumption 8. Our final assumption ensures identification of α.

Assumption 9. *The concentrated objective function admits an interior global minimum.*

A.2.3 Asymptotic Normality

The moment function depends on $\hat{P}$, the estimated distribution of the network. For the sake of clarification in this section, we write the moment function as $\mathbf{m}_i^{\hat{P}}(\boldsymbol{\theta})$ instead of $\mathbf{m}_i(\boldsymbol{\theta})$. This allows to distinguish $\mathbf{m}_i^{\hat{P}}(\boldsymbol{\theta})$, the biased (but consistent) moment function we use, from $\mathbf{m}_i^{P}(\boldsymbol{\theta})$ which is the unbiased moment function we would use if the true distribution of $\mathbf{G}$ were known.

Taking the first derivative of the objective function with respect to $\boldsymbol{\theta}$, we have:

$$\tfrac{1}{N}\textstyle\sum_i \frac{\partial \mathbf{m}_i^{\hat{P}}(\boldsymbol{\theta})'}{\partial \boldsymbol{\theta}} \mathbf{W}_N \tfrac{1}{N}\textstyle\sum_i \mathbf{m}_i^{\hat{P}}(\boldsymbol{\theta}) = \mathbf{0}.$$

By applying the mean value theorem to $\dfrac{1}{N}\displaystyle\sum_i \mathbf{m}_i^{\hat{P}}(\boldsymbol{\theta})$, we have:

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) = -\left[\tfrac{1}{N}\textstyle\sum_i \frac{\partial \mathbf{m}_i^{\hat{P}}(\hat{\boldsymbol{\theta}})'}{\partial \boldsymbol{\theta}} \mathbf{W}_N \tfrac{1}{N}\textstyle\sum_i \frac{\partial \mathbf{m}_i^{\hat{P}}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}'}\right]^{-1} \tfrac{1}{N}\textstyle\sum_i \frac{\partial \mathbf{m}_i^{\hat{P}}(\hat{\boldsymbol{\theta}})'}{\partial \boldsymbol{\theta}} \mathbf{W}_N N^{-1/2} \textstyle\sum_i \mathbf{m}_i^{\hat{P}}(\boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}^*$ is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$.

Let

$$\mathbf{H}_N = \left[\tfrac{1}{N}\textstyle\sum_i \frac{\partial \mathbf{m}_i^{\hat{P}}(\hat{\boldsymbol{\theta}})'}{\partial \boldsymbol{\theta}} \mathbf{W}_N \tfrac{1}{N}\textstyle\sum_i \frac{\partial \mathbf{m}_i^{\hat{P}}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}'}\right]^{-1} \tfrac{1}{N}\textstyle\sum_i \frac{\partial \mathbf{m}_i^{\hat{P}}(\hat{\boldsymbol{\theta}})'}{\partial \boldsymbol{\theta}} \mathbf{W}_N$$

and

$$\boldsymbol{\theta}_N^* = \tfrac{1}{N}\mathbf{H}_N \textstyle\sum_i \mathbb{E}(\mathbf{m}_i^{\hat{P}}(\boldsymbol{\theta}_0)). \tag{10}$$

We have

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 + \boldsymbol{\theta}_N^*) = -\mathbf{H}_N N^{-1/2} \textstyle\sum_i \tilde{\mathbf{m}}_i^{\hat{P}}(\boldsymbol{\theta}_0),$$

where $\tilde{\mathbf{m}}_i^{\hat{P}}(\boldsymbol{\theta}_0) = \mathbf{m}_i^{\hat{P}}(\boldsymbol{\theta}_0) - \mathbb{E}(\mathbf{m}_i^{\hat{P}}(\boldsymbol{\theta}_0))$. Note that $\boldsymbol{\theta}_N^* \xrightarrow{p} \mathbf{0}$ since $\hat{P}$ is consistent.

Let $\boldsymbol{\zeta}_{N,m}^{\hat{P}}$ be the sum of $\tilde{\mathbf{m}}_i^{\hat{P}}(\boldsymbol{\theta}_0)$ over the m-th sub-network and let $\boldsymbol{\zeta}_{N,m}^{P}$ be the same statistic where $\hat{P}$ is replaced by P. We have $\sum_i \tilde{\mathbf{m}}_i^{\hat{P}}(\boldsymbol{\theta}_0) = \sum_m \boldsymbol{\zeta}_{N,m}^{\hat{P}}$. We need to apply the Central Limit Theorem (CLT) to $N^{-1/2}\displaystyle\sum_m \boldsymbol{\zeta}_{N,m}^{\hat{P}}$. If the network distribution were observed, the CLT would follow directly when M grows to ∞,

because the sub-groups are independents by Assumption 2 and then $(\boldsymbol{\zeta}^{P}_{N,m})_m$ would be a sequence of independent variables. In the current setup, $\boldsymbol{\zeta}^{\hat{P}}_{N,m}$'s are not independent because they all depend on $\hat{\boldsymbol{\rho}}$ which is random. However since $\boldsymbol{\zeta}^{\hat{P}}_{N,m} \xrightarrow{p} \boldsymbol{\zeta}^{P}_{N,m}$, then $\boldsymbol{\zeta}^{\hat{P}}_{N,m}$'s are asymptotically independent. We use the CLT for *ℓ-mixing* array (see Withers, 1981). Under regular conditions, in particular, $\mathbb{V}(N^{-1/2}\sum_m \boldsymbol{\zeta}^{\hat{P}}_{N,m}) < \infty$ as M grows to ∞, the CLT implies that $N^{-1/2}\sum_m \boldsymbol{\zeta}^{\hat{P}}_{N,m}$ converges in distribution to a $N(\mathbf{0}, \boldsymbol{\Sigma}_0)$, where $\boldsymbol{\Sigma}_0 = \lim_{N\to\infty} \boldsymbol{\Sigma}_N$ and $\boldsymbol{\Sigma}_N = \mathbb{V}(N^{-1/2}\sum_m \boldsymbol{\zeta}^{\hat{P}}_{N,m})$.

In addition to the uncertainty due to $\boldsymbol{\varepsilon}$, the variance $\boldsymbol{\Sigma}_0$ includes uncertainty due to $\hat{\boldsymbol{\rho}}$ which is prior estimated using a network formation model.

Besides, by the law of large numbers (LLN), $\frac{1}{N}\sum_i \frac{\partial \mathbf{m}^{\hat{P}}_i(\hat{\boldsymbol{\theta}})'}{\partial \boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\Delta} m^{P}_0 = \mathbb{E}(\frac{\partial \mathbf{m}^{\hat{P}}_i(\hat{\boldsymbol{\theta}})'}{\partial \boldsymbol{\theta}})$. Thus $\mathbf{H}_N \xrightarrow{p} \mathbf{H}_0 = (\boldsymbol{\Delta}\mathbf{m}^{P}_0 \mathbf{W}_0 \boldsymbol{\Delta}\mathbf{m}^{P\prime}_0)^{-1}\boldsymbol{\Delta}\mathbf{m}^{P}_0\mathbf{W}_0$. As a result,

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 + \boldsymbol{\theta}^*_N) \xrightarrow{d} N(0, \mathbf{H}_0\boldsymbol{\Sigma}_0\mathbf{H}_0'). \tag{11}$$

This result implies that $\hat{\boldsymbol{\theta}}$ is not $\sqrt{N}$-consistent to $\boldsymbol{\theta}_0$. However, as $\lim_{N\to\infty} \boldsymbol{\theta}^*_N = 0$, $\hat{\boldsymbol{\theta}}$ is consistent.[31] This result is sufficient to perform statistical tests on $\boldsymbol{\theta}_0$ when M is large since the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is $N(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*_N, \mathbb{V}(\hat{\boldsymbol{\theta}}))$, where $\boldsymbol{\theta}^*_N$ is approximately equal to zero and asymptotic variance is

$$\mathbb{V}(\hat{\boldsymbol{\theta}}) = \frac{\mathbf{H}_0\boldsymbol{\Sigma}_0\mathbf{H}_0'}{N} \tag{12}$$

On the one hand, $\mathbf{H}_0$ can be estimated consistently by,

$$\hat{\mathbf{H}}_0 = N\left[\sum_i \frac{\partial \mathbf{m}^{\hat{P}}_i(\hat{\boldsymbol{\theta}})'}{\partial \boldsymbol{\theta}}\mathbf{W}_N \sum_i \frac{\partial \mathbf{m}^{\hat{P}}_i(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'}\right]^{-1} \sum_i \frac{\partial \mathbf{m}^{\hat{P}}_i(\hat{\boldsymbol{\theta}})'}{\partial \boldsymbol{\theta}}\mathbf{W}_N.$$

[31] In addition, if $\lim_{N\to\infty} \sqrt{N}\boldsymbol{\theta}^*_N = o_p(1)$, then $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \mathbf{H}_0\boldsymbol{\Sigma}_0\mathbf{H}_0')$.

On the other hand, $\mathbf{\Sigma}_0$ can be estimated by $\mathbf{\Sigma}_N$, where

$$\mathbf{\Sigma}_N = \mathbb{E}(\mathbb{V}(N^{-1/2}\sum_m \boldsymbol{\zeta}_{N,m}^{\hat{P}}|\hat{P})) + \mathbb{V}(\mathbb{E}(N^{-1/2}\sum_m \boldsymbol{\zeta}_{N,m}^{\hat{P}}|\hat{P})),$$
$$\mathbf{\Sigma}_N = N^{-1}\,\mathbb{E}(\sum_m \mathbb{V}(\boldsymbol{\zeta}_{N,m}^{\hat{P}}|\hat{P})) + N^{-1}\,\mathbb{V}(\sum_m \mathbb{E}(\boldsymbol{\zeta}_{N,m}^{\hat{P}}|\hat{P})).$$

The term $N^{-1}\,\mathbb{E}(\sum_m \mathbb{V}(\boldsymbol{\zeta}_{N,m}^{\hat{P}}|\hat{P}))$ is due to the error term of the model $\boldsymbol{\varepsilon}$ and the fact that the true network is not observed whereas $N^{-1}\,\mathbb{V}(\sum_m \mathbb{E}(\boldsymbol{\zeta}_{N,m}^{\hat{P}}|\hat{P}))$ is due to uncertainty associated with the estimation of P. In practice we estimate, we can compute the estimator of $\boldsymbol{\zeta}_{N,m}^{\hat{P}}$, $m = 1, \rightarrow, M$. Thus, we can also estimate $\sum_m \mathbb{V}(\boldsymbol{\zeta}_{N,m}^{\hat{P}}|\hat{P})$ and $\sum_m \mathbb{E}(\boldsymbol{\zeta}_{N,m}^{\hat{P}}|\hat{P})$. With several simulations from the distribution of $\hat{P}$, we finally compute $\mathbf{\Sigma}_N$.[32]

A.3 Additional Results

A.3.1 Bias correction for GX observed

Let us replace $\ddot{\mathbf{V}}$ with $\mathbf{V}$ in (5). We have:

$$\dot{\mathbf{Z}}_i^{(r)\prime}\left[(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(t)})_i\left((\mathbf{I}-\alpha\mathbf{G})^{-1}-(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(s)})^{-1}\right)\mathbf{V}\tilde{\boldsymbol{\theta}}\right]$$

which is equal to:

$$\dot{\mathbf{Z}}_i^{(r)\prime}\left[(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(t)})_i\left((\mathbf{I}-\alpha\mathbf{G})^{-1}-(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(s)})^{-1}\right)((\mathbf{I}-\alpha\mathbf{G})\mathbf{y}+\boldsymbol{\varepsilon})\right]$$

which simplifies to:

$$\dot{\mathbf{Z}}_i^{(r)\prime}(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(t)})_i(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(s)})^{-1}[\alpha(\mathbf{G}-\ddot{\mathbf{G}}^{(s)})]\mathbf{y}+\dot{\mathbf{Z}}_i^{(r)\prime}\left[(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(t)})_i\left((\mathbf{I}-\alpha\mathbf{G})^{-1}-(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(s)})^{-1}\right)\boldsymbol{\varepsilon}\right]$$

Note that the expectation of the term $\dot{\mathbf{Z}}_i^{(r)\prime}\left[(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(t)})_i\left((\mathbf{I}-\alpha\mathbf{G})^{-1}-(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(s)})^{-1}\right)\boldsymbol{\varepsilon}\right]$ is null.

[32] Our R package offers tools to compute this variance. See also our online appendix.

A.3.2 Corollaries

Corollary 1. *Assume that* $\mathbf{GX}$ *is observed, but that* $\mathbf{Gy}$ *is not observed. Let* $\dot{\mathbf{Z}}^{(s)} = [\mathbf{1}, \mathbf{X}, \mathbf{GX}, \dot{\mathbf{G}}^s\mathbf{GX}, (\dot{\mathbf{G}}^{(s)})^2\mathbf{GX}, ...]$, $\ddot{\mathbf{Z}}^{(s,t)} = [\mathbf{1}, \mathbf{X}, \ddot{\mathbf{G}}^{(t)}\mathbf{X}, \dot{\mathbf{G}}^s\ddot{\mathbf{G}}^{(t)}\mathbf{X}, (\dot{\mathbf{G}}^{(s)})^2\ddot{\mathbf{G}}^{(t)}\mathbf{X}, ...]$, $\mathbf{V} = [\mathbf{1}, \mathbf{X}, \mathbf{GX}]$, *and* $\ddot{\mathbf{V}}^{(t)} = [\mathbf{1}, \mathbf{X}, \ddot{\mathbf{G}}^{(t)}\mathbf{X}]$. *Then, the results from Theorem 1 hold for the following (simulated) moment function:*

$$\frac{1}{S}\sum_{s=1}^{S}\dot{\mathbf{Z}}_i^{(s)\prime}(\mathbf{I}-\alpha\dot{\mathbf{G}}^{(s)})_i\mathbf{y} - \frac{1}{ST}\sum_{t=1}^{T}\sum_{s=1}^{S}(\dot{\mathbf{Z}}^{(s)\prime}\mathbf{V} - \ddot{\mathbf{Z}}^{(s,t)\prime}\ddot{\mathbf{V}}^{(t)})_i\tilde{\boldsymbol{\theta}} - \frac{1}{ST}\sum_{t=1}^{T}\sum_{s=1}^{S}\ddot{\mathbf{Z}}_i^{(s,t)\prime}(\mathbf{I}-\alpha\dot{\mathbf{G}}^{(s)})_i(\mathbf{I}-\alpha\ddot{\mathbf{G}}^{(t)})^{-1}\ddot{\mathbf{V}}^{(t)}\tilde{\boldsymbol{\theta}} \tag{13}$$

Corollary 2. *Assume that* $\mathbf{Gy}$ *is observed, but that* $\mathbf{GX}$ *is not observed. Let* $\dot{\mathbf{Z}}^{(s)} = [\mathbf{1}, \mathbf{X}, \dot{\mathbf{G}}^{(s)}\mathbf{X}, (\dot{\mathbf{G}}^{(s)})^2\mathbf{X}, ...]$, *and* $\ddot{\mathbf{V}}^{(t)} = [\mathbf{1}, \mathbf{X}, \ddot{\mathbf{G}}^{(t)}\mathbf{X}]$. *Then, the results from Theorem 1 hold for the following (simulated) moment function:*

$$\frac{1}{S}\sum_{s=1}^{S}\dot{\mathbf{Z}}_i^{(s)\prime}(\mathbf{I}-\alpha\mathbf{G})_i\mathbf{y} - \frac{1}{ST}\sum_{t=1}^{T}\sum_{s=1}^{S}\dot{\mathbf{Z}}_i^{(s)\prime}\ddot{\mathbf{V}}_i^{(t)}\tilde{\boldsymbol{\theta}} \tag{14}$$

B Appendix – ARD Details

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:

$$P(a_{ij} = 1) \propto \exp\{\nu_i + \nu_j + \zeta\mathbf{z}_i'\mathbf{z}_j\}. \tag{15}$$

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).

B.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. (2020).

$$P(a_{ij} = 1|\nu_i, \nu_j, \zeta, \mathbf{z}_i, \mathbf{z}_j) = \frac{\exp\{\nu_i + \nu_j + \zeta \mathbf{z}_i' \mathbf{z}_j\} \sum_{i=1}^{N} d_i}{\sum_{ij} \exp\{\nu_i + \nu_j + \zeta \mathbf{z}_i' \mathbf{z}_j\}}, \tag{16}$$

where d_i is the degree defined by $d_i \approx \frac{C_p(0)}{C_p(\zeta)} \exp(\nu_i) \sum_{i=1}^{N} \exp(\nu_i)$, and 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 (16).

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 $\lfloor x \rfloor$ represents 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.

B.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 (15).

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 depend 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. (2020), 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:

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

C Appendix – Bayesian Inference

C.1 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 $\mathbf{\Lambda}_t = [\boldsymbol{\beta}_t,\ \boldsymbol{\gamma}_t]$ and σ_t^2 are drawn from their posterior conditional distributions, given as follows:

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

where,

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

C.2 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. For each entry, however, 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, as 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 $h(n_i + 1)$ (as the diagonal of $\mathbf{W}$ equals 0) and the entry (i, j) will change 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) will change 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. (2019) 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:*

$$\mathbf{B}_{ij}(\mathbf{Q}, \epsilon) = \frac{\mathbf{Q}\mathbf{e}_i\mathbf{e}_j'\mathbf{Q}}{1 + \epsilon\mathbf{e}_j'\mathbf{Q}\mathbf{e}_i}. \tag{17}$$

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* $\tilde{\mathbf{M}} = \mathbf{M} + \epsilon_1\mathbf{e}_i\mathbf{e}_i' + \epsilon_2\mathbf{e}_i\mathbf{e}_j'$.

1. *The inverse of the perturbed matrix can be written as:*

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

2. *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. 1. 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,

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

2. 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:

$$\det(\mathbf{D}) = \sum_{j=1}^{N}(-1)^{i+j}\mathbf{D}_{\mathrm{ij}}\delta_{\mathrm{ij}}, \tag{18}$$

where i denotes any row of $\mathbf{D}$ and δ_{ij} the minor[33] 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 with 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:

1. Compute the N minors $\delta_{i1}, \ldots, \delta_{in}$.
2. 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 $\Omega_{\boldsymbol{G}}$.

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

(a) 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 $r \leq \min(5, n_{\boldsymbol{G}})$ because the number of possibilities of links to consider grows exponentially with r.

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

(c) 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 1 and the formula (18) to quickly compute $det(\mathbf{D})$.

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

Table C.1: Simulation results with ARD and observed $\boldsymbol{GX}$

Parameter	Breza et al.		Alidaee et al.	
	Mean	Std. Dev	Mean	Std. Dev
	IV, $\kappa = 0$, $N = 250$, $M = 20$			
$\alpha = 0.4$	0.392	(0.009)	0.372	(0.027)
$\beta_1 = 1$	1.001	(0.003)	1.007	(0.004)
$\beta_2 = 1.5$	1.500	(0.006)	1.497	(0.008)
$\gamma_1 = 5$	5.398	(0.021)	5.504	(0.037)
$\gamma_2 = -3$	−2.400	(0.037)	−2.478	(0.059)
$\check{\gamma}_1 = 0$	−0.387	(0.023)	−0.364	(0.033)
$\check{\gamma}_2 = 0$	−0.589	(0.039)	−0.307	(0.068)
Sargan test	1.217	(1.761)	7.779	(10.386)
F-test (Weak inst.)	4259.622	(1140.825)	597.055	(251.886)
	IV, $\kappa = 15$, $N = 250$, $M = 20$			
$\alpha = 0.4$	0.400	(0.007)	0.399	(0.007)
$\beta_1 = 1$	1.000	(0.003)	1.000	(0.003)
$\beta_2 = 1.5$	1.500	(0.006)	1.500	(0.006)
$\gamma_1 = 5$	5.392	(0.019)	5.391	(0.019)
$\gamma_2 = -3$	−2.396	(0.039)	−2.402	(0.039)
$\check{\gamma}_1 = 0$	−0.390	(0.023)	−0.390	(0.023)
$\check{\gamma}_2 = 0$	−0.602	(0.039)	−0.544	(0.039)
Sargan test	1.063	(1.496)	2.608	(1.496)
F-test (Weak inst.)	5306.551	(1519.394)	1767.274	(1519.394)
	SMM, $\kappa = 0$, $N = 250$, $M = 20$			
$\alpha = 0.4$	0.392	(0.01)	0.492	(0.057)
$\beta_1 = 1$	1.001	(0.004)	1.002	(0.009)
$\beta_2 = 1.5$	1.500	(0.007)	1.496	(0.016)
$\gamma_1 = 5$	5.013	(0.034)	3.884	(0.295)
$\gamma_2 = -3$	−2.993	(0.052)	−4.048	(0.354)
	SMM, $\kappa = 15$, $N = 250$, $M = 20$			
$\alpha = 0.4$	0.400	(0.009)	0.428	(0.009)
$\beta_1 = 1$	1.000	(0.004)	0.999	(0.004)
$\beta_2 = 1.5$	1.500	(0.008)	1.499	(0.008)
$\gamma_1 = 5$	4.996	(0.034)	4.677	(0.034)
$\gamma_2 = -3$	−3.005	(0.055)	−3.387	(0.055)

Note: In each subnetwork, the spherical coordinates of individuals are generated from a von Mises–Fisher distribution with a location parameter $(1, 0, 0)$ and intensity parameter κ. Predicted probabilities are computed using the mean of the posterior distribution. Sargan test is the statistic of the overidentification test. F-test is the Fisher-statistic of the Weak instruments test. We chose the weight associated with the nuclear norm penalty to minimize the RMSE through cross-validation. This value of $\lambda = 600$ is smaller than the recommended value in Alidaee et al. (2020). Instruments are build using only second-degree peers, i.e. $\mathbf{G^2X}$.

Table C.2: Simulation results with ARD and unobserved $\boldsymbol{GX}$

Parameter	Breza et al.		Alidaee et al.		True distribution	
	Mean	Std. Dev	Mean	Std. Dev	Mean	Std. Dev
			SMM, $\kappa = 0$, $N = 250$, $M = 20$			
$\alpha = 0.4$	0.717	(0.463)	0.700	(0.268)	0.400	(0.056)
$\beta_1 = 1$	0.988	(0.022)	0.995	(0.017)	1.000	(0.015)
$\beta_2 = 1.5$	1.505	(0.03)	1.503	(0.029)	1.501	(0.021)
$\gamma_1 = 5$	1.778	(4.473)	1.512	(2.37)	4.991	(0.455)
$\gamma_2 = -3$	-2.205	(1.24)	-0.405	(0.955)	-3.005	(0.287)
			SMM, $\kappa = 15$, $N = 250$, $M = 20$			
$\alpha = 0.4$	0.603	(0.069)	0.870	(0.202)	0.434	(0.394)
$\beta_1 = 1$	0.989	(0.014)	0.984	(0.015)	0.998	(0.021)
$\beta_2 = 1.5$	1.504	(0.029)	1.509	(0.029)	1.501	(0.023)
$\gamma_1 = 5$	2.866	(0.566)	0.246	(1.973)	4.638	(3.887)
$\gamma_2 = -3$	-2.458	(0.379)	-1.539	(0.602)	-2.913	(1.037)

Note: In each subnetwork, the spherical coordinates of individuals are generated from a von Mises–Fisher distribution with a location parameter $(1, 0, 0)$ and intensity parameter κ. Predicted probabilities are computed using the mean of the posterior distribution. Sargan test is the statistic of the overidentification test. F-test is the Fisher-statistic of the Weak instruments test. We chose the weight associated with the nuclear norm penalty to minimize the RMSE through cross-validation. This value of $\lambda = 600$ is smaller than the recommended value in Alidaee et al. (2020). Instruments are build using only second-degree peers, i.e. $\mathbf{G^2X}$.

Table C.3: Posterior distribution by controlling for either censoring or missing links

	Censoring			Missing links		
Statistic	Mean	Std. Dev.	t-stat	Mean	Std. Dev.	t-stat
Peer effect model						
Peer effects	0.351***	(0.024)	14.718	0.565***	(0.041)	13.844
Own effects						
Female	0.144***	(0.029)	5.008	0.125***	(0.031)	4.050
Hispanic	−0.083**	(0.042)	−1.967	−0.142**	(0.047)	−3.046
Race (White)						
Black	−0.232***	(0.046)	−5.035	−0.210***	(0.056)	−3.778
Asian	−0.090	(0.090)	−0.999	−0.119	(0.091)	−1.308
Other	0.053	(0.051)	1.042	0.038	(0.052)	0.721
Mother's education. (High)						
<High	−0.123**	(0.039)	−3.182	−0.123**	(0.040)	−3.102
>High	0.140***	(0.034)	4.103	0.110**	(0.034)	3.216
Missing	−0.061	(0.052)	−1.172	−0.064	(0.051)	−1.235
Mother's job (Stay-at-home)						
Professional	0.083*	(0.044)	1.872	0.074*	(0.044)	1.686
Other	0.004	(0.035)	0.112	−0.007	(0.035)	−0.199
Missing	−0.065	(0.047)	−1.365	−0.069	(0.047)	−1.450
age	−0.074***	(0.010)	−7.745	−0.079***	(0.011)	−7.317
Contextual effects						
Female	0.009	(0.049)	0.177	0.053	(0.094)	0.569
Hispanic	0.062	(0.068)	0.905	0.356**	(0.114)	3.117
Race (White)						
Black	0.053	(0.059)	0.898	0.057	(0.075)	0.754
Asian	0.233	(0.184)	1.264	0.505	(0.552)	0.916
Other	−0.138	(0.089)	−1.538	−0.240	(0.182)	−1.317
Mother's education (High)						
<High	−0.271***	(0.071)	−3.833	−0.373**	(0.157)	−2.383
>High	0.069	(0.060)	1.154	0.067	(0.111)	0.606
Missing	−0.078	(0.094)	−0.833	−0.024	(0.191)	−0.127
Mother's job (Stay-at-home)						
Professional	−0.117	(0.080)	−1.474	0.024	(0.150)	0.157
Other	−0.107*	(0.060)	−1.767	−0.038	(0.112)	−0.345
Missing	−0.103	(0.087)	−1.190	0.004	(0.184)	0.024
Age	0.067***	(0.006)	11.441	0.091***	(0.009)	9.611
SE^2	0.523			0.513		
Network formation model						
Same sex	0.396***	(0.015)	27.043	0.383***	(0.015)	25.850
Both Hispanic	0.355***	(0.023)	15.245	0.301***	(0.024)	12.736
Both White	0.468***	(0.022)	21.537	0.289***	(0.024)	12.116
Both Black	0.859***	(0.032)	26.718	1.117***	(0.031)	36.066
Both Asian	0.164***	(0.045)	3.666	0.290***	(0.048)	6.039
Mothers education <High	0.212***	(0.016)	12.987	0.178***	(0.018)	9.936
Mothers education >High	0.037**	(0.015)	2.407	0.075***	(0.016)	4.668
Mothers job Professional	−0.077***	(0.017)	−4.592	−0.087***	(0.019)	−4.658
Age absolute diff	−0.616***	(0.010)	−64.156	−0.652***	(0.009)	−72.445

Note: N = 3,126. Observed links = 17,993.
Proportion of inferred network data: censoring = 4.8%, missing links = 76.2%. Significance levels: *** = 1%, ** = 5%, * = 10%. The explained variable is computed by taking the average grade for English, Mathematics, History, and Science, letting $A = 4$, $B = 3$, $C = 2$, and $D = 1$. Thus, lower scores indicate better academic achievement.