Estimating Social Network Models with Missing Links*

Arthur Lewbel, Xi Qu, and Xun Tang

July 28, 2023

Abstract

We propose an adjusted 2SLS estimator for social network models when some existing network links are missing from the sample (due, e.g., to recall errors by survey respondents, or lapses in data input). In the feasible structural form, missing links make all covariates endogenous and add a new source of correlation between the structural errors and endogenous peer outcomes (in addition to simultaneity), thus invalidating conventional estimators used in the literature. We resolve these issues by rescaling peer outcomes with estimates of missing rates and constructing instruments that exploit properties of the noisy network measures. We apply our method to study peer effects in household decisions to participate in a microfinance program in Indian villages. We find that ignoring missing links and applying conventional instruments would result in a sizeable upward bias in peer effect estimates.

*We are grateful to seminar and conference participates at CalTech, Chinese University of Hong Kong, Northwestern, Texas Camp Econometrics, U Chicago, U Penn, UT Austin, and Vanderbilt for useful comments and suggestions. Lewbel and Tang are grateful for the financial support from National Science Foundation (Grant SES-1919489). Any or all remaining errors are our own.

1 Introduction

In many social and economic environments, an individual's behavior or outcome (such as a consumption choice or a test score) depends not only on his or her own characteristics, but also on the behavior and characteristics of other individuals. Call such dependence between two individuals a *link*. A *social network* consists of a group of individuals, some of whom are linked to others. The econometrics literature on social networks has largely focused on disentangling various channels of social effects based on observed outcomes and characteristics of network members. These include identifying the effects on each individual's outcome of (i) the individual's own characteristics (*individual effects*), (ii) the characteristics of people linked to the individual (*contextual effects*), and (iii) the outcomes of people linked to the individual (*peer effects*). See Blume et al. (2011) and Graham (2020) for extensive surveys about identifying such effects in social network models.

A popular approach for estimating social network models is to use two-stage least squares (2SLS). This requires researchers to construct instruments for the endogenous peer outcomes, using *perfect knowledge* of the network structure, as given by the *adjacency* matrix (i.e., the matrix that lists all links in the network). See, for example, Bramoullé et al. (2009), Kelejian and Prucha (1998), Lee (2007), and Lin (2010). In practice, samples of network links are often collected from survey responses. Such samples may suffer from an issue of missing links, due, e.g., to recall errors or misunderstandings by survey respondents, or lapses in data input. These missing links can be viewed as *one-sided* misclassification errors. An existing link between two individuals may be misclassified as non-existent, but the sample does not erroneously impute links between those who are not linked.

Missing links in the sample pose major methodological challenges for estimators like 2SLS. To see this, consider a data-generating process (DGP) from which a large number of independent networks (i.e., groups) are drawn. Each group has n individual members. The issues we raise and the solutions we propose also apply to other contexts, such as a large number of independent networks with different sizes or a single growing network, but are easiest to illustrate in the context of many independent, identically sized groups.

Suppose that in each group, a vector of individual outcomes $y \in \mathbb{R}^n$ is determined by a structural model:

$$y = \lambda G y + X\beta + \varepsilon, \text{ where } E(\varepsilon|X, G) = 0.$$

In this model, the adjacency matrix, G, is an n-by-n matrix of dummy variables that describes the group's network: the element in row j and column k of G equals one if individual j is linked to member k, and zero otherwise.[1] Here X is an n-by-K matrix of exogenous covariates, and ε is an n-vector of structural errors. The random arrays y, G, X, and ε all vary across the groups in the sample, while the coefficients λ and β are the same across groups. We drop group subscripts for clarity.

For simplicity we have for now omitted contextual effects, i.e., a term defined as $GX\gamma$. We also omit group-level fixed effects for now. Extensions of our results that deal with these are provided later.

The regressors in the model are GY and X. While X is exogenous, the regressors GY are correlated with ε. The issue of simultaneity arises here, because any one individual's outcome depends on, and is determined simultaneously with, the outcomes of other group peers. A simple estimator of the peer effect λ and individual effects β that deals with this simultaneity problem is 2SLS, using GX or G^2X as instruments for GY, as in Bramoullé et al. (2009).[2]

But now suppose that, in each group, a researcher does not observe G perfectly, but instead observes a noisy measure H, which differs from G by randomly missing some actual existing links while correctly reporting others. The goal now is to estimate λ and β from the "*feasible*" structural form:

$$y = \lambda H y + X\beta + u, \tag{1}$$

where $u \equiv [\varepsilon + \lambda(G - H)y]$ is a vector of *composite* errors.

The missing links in H aggravate endogeneity issues in (1) in two important ways. First, they lead to correlation between X and u through $\lambda(G - H)y$, a component in u that is

[1]This is a "local-aggregate" network model, where the endogenous effect depends on the *aggregate* outcome of those linked to an individual. It differs from a "local-average" network model, where the endogenous effect is represented by the *average* outcome of those linked peers.

[2]If the model included contextual effects $GX\gamma$ in its structural form, then G^2X could be used as instruments for Gy, otherwise use of GX as instruments suffices.

due to the measurement error in the adjacency matrix. As a result, unlike using GX or G^2X as instruments when G is perfectly reported in the sample, 2SLS estimates based on the feasible instruments HX or H^2X that are constructed from the noisy network measure would be inconsistent due to a failure of instrument exogeneity.

Second, these missing links cause an additional source of endogeneity in Hy. Like Gy, the feasible Hy is correlated with the model error ε due to simultaneity. But in addition, Hy is also correlated with u through the measurement errors in $\lambda(G-H)y$. For all these reasons, standard 2SLS estimators of this model become inconsistent in the presence of missing links.[3]

In this paper, we introduce an *adjusted-2SLS* estimator, which resolves these challenges and consistently estimates (λ, β) using alternative valid instruments constructed from H despite the missing links. We first introduce the main idea for a benchmark case, where actual links in G are missing randomly from H in the sample at an unknown rate $p \in (0,1)$. Later, we extend our method to allow the missing rates $p(X)$ to depend on covariates.

Our method is based on a series of new insights that have not been explored in the literature. *First*, we observe that by rescaling the noisy measure of peer outcomes Hy with the inverse probability of reporting correctly $1/(1-p)$, we restore the exogeneity of X in a *rescaled* structural form. Formally, this means if we reparametrize (1) as

$$y = \lambda^* Hy + X\beta + v \text{ with } \lambda^* \equiv \lambda/(1-p), \tag{2}$$

then the reparametrized composite errors $v \equiv \varepsilon + (\lambda G - \lambda^* H)y$ satisfy $E(v|X,G) = 0$. This holds regardless of how the actual network G is formed, as long as $E(\varepsilon|X,G) = 0$.

Second, despite the restored exogeneity of X in (2), conventional instruments such as HX or H^2X remain invalid, because the reparametrized errors v depend on H. To address this issue, we provide alternative functions of H and X that are valid instruments. To give an example, we show that if conditional on (G,X), the observed H is an unsymmetrized noisy measure with links missing independently, then $H'X$ is uncorrelated with v (where H'

[3]While we focus on the 2SLS estimator in this paper, the same arguments apply to show that conventional maximum likelihood, and the generalized least squares estimators based on (1) are also inconsistent.

denotes the transpose of H). This result holds regardless of whether the actual adjacency matrix G is known to be symmetric (i.e., with all links being undirected) or asymmetric (i.e., consisting of directed links). Therefore, we can use $H'X$ as valid instruments in an adjusted-2SLS where peer outcomes are rescaled by $1/(1-p)$. To the best of our knowledge, no other paper in the literature has proposed this use of $H'X$ as instruments.

For a different scenario where the noisy measures H's are symmetrized while the actual G is known to be symmetric with all links being undirected, we provide an alternative way to construct valid instruments, based on observing two different H matrices.[4] For example, in our empirical application, for an undirected link between two households A and B, we observe two proxy measures of the same link: whether A visited B, and whether B visited A. This yields two different observed H matrices corresponding to the same true undirected link in the G matrix. Observing these two H matrices allows us to construct valid instruments. This method of constructing instruments from multiple noisy measures can also be applied when H are unsymmetrized, regardless of whether the actual G is known to be symmetric or not.

Third, under either of these two scenarios above that permit construction of valid instruments (that is, the sample reports either a single unsymmetrized noisy measure H, or two independent measures that may or may not be symmetrized), we provide simple methods to identify and estimate the unknown missing rate p.[5]

Building on these insights, we construct an adjusted 2SLS estimator for (λ, β), and provide its limiting distribution as the number of groups grows to infinity. This estimator essentially applies 2SLS to the rescaled peer outcomes $Hy/(1-p)$ in (2), using our proposed new instruments and a sample analog estimator for the missing rate p. The estimator is easy to implement, and we demonstrate good finite-sample performance in monte carlo simulation.

We then generalize the model and our estimator in several directions. We show how

[4]We also show yet another way to construct valid instruments is to use nonlinear functions of X.

[5]The approach we take in this step differs from, and is simpler than, other papers that use multiple measures to deal with misclassification in discrete explanatory variables (e.g. Mahajan (2006), Lewbel (2007), and Hu (2008)). This is because, for implementing our adjusted-2SLS, it is only necessary to estimate the missing rate p, rather than the distribution of outcomes conditional on the actual G.

to include contextual effects (a term defined as $GX\gamma$) as well as group-level fixed effects into the structural form in (1). We also allow the missing rates p to be heterogeneous and depend on individual covariates in X.

Furthermore, we extend our method to the case of a single large network. In this case, the asymptotic experiment is to increase the number of individuals on a single network, rather than increasing the number of small groups with fixed sizes. For this extension we propose two possible settings where some form of weak dependence exists between the outcomes of individuals who are "sufficiently far" from each other, either in the sense of not being in the same group (Section 6.1) or in terms of a latent distance metric (Section 6.2). In either case, we show that under such weak dependence our adjusted 2SLS estimator, when pooled over individuals in the sample, still converges to the intended estimand.

Finally, we apply our method to estimate peer effects in household decisions to participate in a microfinance program in Indian villages, using data from Banerjee et al. (2013). We match the individual survey to the household survey there, yielding a sample of 4134 households from 43 villages in South India. The parameter of interest is the peer (endorsement) effect, which reflects how a household's decision is influenced by the microfinance program participation of other households to which it is linked. Survey information about visits between the households provides two symmetrized noisy measures of undirected links (i.e., two symmetric H matrices). We estimate missing rates in each of these two measures using our methodology, and then we apply these rates in our adjusted-2SLS procedure to estimate the endorsement peer effects.

We find that participation by another linked household increases a household's own participation rate by around 4.6%. This effect is economically significant, compared to the average participation rate of 18.2% in the sample. We also find that ignoring the missing links in the noisy measures and applying conventional 2SLS estimation results in a sizeable upward bias in the estimates of these peer effects.

Roadmap. Section 2 reviews the related literature, and explains our contribution in its context. Section 3 specifies the model, and illustrates the main ideas in a benchmark model with independent and identical missing rates. Section 4 defines an analog estimator for missing rates, and provides our adjusted-2SLS estimator for social effects. Section 5 extends the method to more general settings with contextual effects, heterogeneous missing rates, or group fixed effects. Section 6 shows how our estimator works when the sample consists of a single, large network. Section 7 presents monte carlo simulation results. Section 8 applies our method to analyze peer effects in microfinance participation in India. Proofs are collected in the Appendix.

2 Related Literature

Models with misclassified binary or discrete variables have been studied extensively in the econometrics literature. Aigner et al. (1973), Klepper (1988), Bollinger (1996), and Molinari (2008) point-identify or set-identify such models using various restrictions on the misclassification rates; Mahajan (2006), Lewbel (2007), and Hu (2008) exploit exogenous instruments to deal with misclassified explanatory variables.

Estimation of peer effects in social networks with measurement errors in the links is an increasingly important topic. Shalizi and Rinaldo (2013) note the challenge of dealing with missing network links in Random Graph Models. Advani and Malde (2018) show that even a relatively low misreporting rate can lead to large bias in causal effect estimates. Butts (2003) proposes a hierarchical Bayesian model to infer social structure in the presence of measurement errors. Chandrasekhar and Lewis (2011) show how egocentrically sampled network data can be used to predict the full network in a graphical reconstruction process. Liu (2013a) shows that when the adjacency matrix is not row-normalized, instrumental variable estimators based on an out-degree distribution can be valid.

Goldsmith-Pinkham and Imbens (2013) examine network endogeneity and investigate simultaneously alternative definitions of links and the possibility of peer effects arising through multiple networks. They explicitly model network formation, with estimation based on maximum likelihood, using a Bayesian approach for computational convenience and feasibility. Hardy et al. (2019) estimate treatment effects on a social network when the reported links are a noisy representation of true spillover pathways. They use a mixture model that accounts for missing links as unobserved network heterogeneity, and estimate it using an Expectation-Maximization algorithm. This approach requires a parametric model of how links are determined and treatment is assigned, and requires enumerating the likelihood conditional on all possible treatment exposures (which in turn depends on the latent unobserved network). Auerbach (2022) studies a network model where links are correctly measured but both peer and contextual effects interact with unobserved individual heterogeneity that affects link formation.

In contrast with these papers, we focus on social effect parameters in a linear social network model, and exploit implications of randomly missing links for identification. Our method does not require modeling the formation of actual links. Our estimator is essentially a rescaled 2SLS, which has closed form and is easy to compute.

Boucher and Houndetoungan (2020) estimate peer effects when the social networks in the sample are subject to measurement issues, such as missing or misclassified links. Their method can be applied when the researcher only has access to aggregated relational data, but assumes the researcher knows, or has a consistent estimator of, the distribution of the actual network. They construct instruments by drawing from this distribution, and use 2SLS to estimate the peer effects. In comparison, the method we propose does not require such prior knowledge or estimates of network distribution.

Griffith (2021) studies the case where links are censored in the sample (e.g., when each individual is restricted to naming 5 or fewer links with other people, even if the actual number of people the individual is linked with is larger). Griffith (2021) analytically characterizes the bias in a reduced-form regression (i.e., when the outcome vector y is regressed on the exogenous variables X and GX). Moreover, for a model with *no* endogenous peer effect ($\lambda = 0$), Griffith (2021) shows that the bias can be consistently estimated under an order invariance condition, i.e., the covariance of characteristics of those linked to an individual is invariant to the order in which those links are reported or censored. This condition mitigates the issue of endogenous selection of uncensored links, and in this sense is analogous to our assumption of randomly missing links. In comparison, we consider different settings where links are missing at random in a model with a non-zero peer effect $\lambda \neq 0$. (This is later generalized to the case with heterogeneous missing rates.) We show that the 2SLS estimand in this case contains a simple augmentation bias in peer effects (in the sense of converging to $\lambda/(1-p)$, with p being the missing rate), and no bias in other individual effects. Bias correction in our case is immediate once the missing rate is estimated using a simple approach that we provide.

Liu (2013b) estimates a social network model when the data consists of a *subset* of individuals sampled randomly from a larger group in the population. In his setting, the links and outcomes among this sampled subset of group members are perfectly measured while those of all others are not reported in the data.[6] In comparison, we do not study the inference of sampled networks; instead, we let the group memberships be fixed and known, and allow every individual in the sample to have randomly missing links. As noted above, this imperfect measure of links leads to failure of conventional 2SLS in our setting.

3 Model and Identification

Consider a DGP from which a large number of small, independent networks (groups) are drawn. Each group s consists of n_s individual members, with $n_s \geq 3$ being finite integers. In Section 3-5, we identify and estimate a linear social network model with missing links in the data as the number of groups in the sample approaches infinity. Later we consider the extension to a single growing network.

To simplify exposition, let the group sizes $n_s = n$ be fixed across groups $s = 1, ..., S$. This allows us to drop the group subscript s while presenting our identification argument. We will later add back these group subscripts and allow for variation in group sizes when we define our estimator in Section 4.

The structural form for the n-vector of individual outcomes y in each group is:

$$y = \lambda Gy + X\beta + \varepsilon, \tag{3}$$

where the peer effect λ and the direct effects β are constant parameters of interest, X is an n-by-K matrix of individual- or group-level explanatory variables, and $G \in \{0,1\}^{n\times n}$ is a network (adjacency) matrix with its (i,j)-th entry $G_{ij} = 1$ if an individual member i is

[6] In our notation, this means some rows in G, as well as their corresponding rows in Y and X, are not included in the data due to random sampling.

linked to another member j, and $G_{ij} = 0$ otherwise. The matrix G may be asymmetric with directed links (i.e., $G_{ij} \neq G_{ji}$ for some $i \neq j$), or symmetric with all links being undirected (i.e., $G_{ij} = G_{ji}$ for all $i \neq j$ almost surely).

Note that, like y, X, and ε, the adjacency matrix G varies by group, and so it too has an s subscript that has been dropped for now. Only the coefficients λ and β are constants that do not vary across groups. Assume that $(I - \lambda G)$ is invertible almost surely. A sufficient condition for this is that $||\lambda G|| < 1$ for *any* matrix norm $|| \cdot ||$. Solving equation (3) for y gives the reduced form for outcomes:

$$y = M(X\beta + \varepsilon), \text{ where } M \equiv (I - \lambda G)^{-1}. \tag{4}$$

For each group, the sample only reports a noisy measure of the adjacency matrix G, with randomly missing links. Denote this noisy measure by $H \in \{0, 1\}^{n \times n}$. Let $G_{ii} = 0$ and $H_{ii} = 0$ by convention.

3.1 Assumptions

We maintain the following conditions on the noisy measure H throughout Section 3:

$$\text{(A1) } E(H_{ij}|G, X) = E(H_{ij}|G_{ij}, X) \text{ for all } i \text{ and } j;$$

$$\text{(A2) } E(H_{ij}|G_{ij} = 1, X) = 1 - p \text{ and } E(H_{ij}|G_{ij} = 0, X) = 0 \text{ for all } i \neq j;$$

$$\text{(A3) } E(\varepsilon|X, G, H) = 0.$$

Condition (A1) states that the incidence of missing a link between two individual members i and j is conditionally independent from the state of links involving other individuals $l \notin \{i, j\}$. Condition (A2) specifies that misclassification of links is one-sided in that existent links are missing from the sample at a rate of $p \in (0, 1)$ while non-existent links are never mistakenly coded as existent. Condition (A3) rules out endogeneity in link formation, by assuming that (X, G, H) are exogenous to the structural error ε.

Conditions (A1) and (A2) hold jointly in two scenarios that are common in practice. In the first scenario, which we refer to as "*unsymmetrized measures*", each (i, j)-th entry in H is an independent measure (e.g., a survey response from an individual i about whether a link with j exists) of the corresponding (i, j)-th entry in G. An adjacency measure H constructed this way is flexible in that it allows the researcher to remain agnostic about whether the actual adjacency matrix G is symmetric or not. (Of course, this is also an intuitive way to construct H when the actual G is known to have directed links.) In this scenario, if these measures randomly miss existing links at a rate p, but never erroneously impute links between those who are not linked, then (A1) and (A2) are satisfied with a missing link rate p. To reiterate, (A1) and (A2) hold in this case, regardless of whether the actual G is symmetric or not.

In the second scenario, which we refer to as "*symmetrized measures*", the actual G is known to be symmetric with all links being undirected, and the researcher therefore chooses to symmetrize H using independent measures of entries in G. For example, the researcher may ask two individuals i and j whether they share an undirected link, and construct a symmetrized measure by setting two entries H_{ij} and H_{ji} both to 1 if *either* i *or* j responds positively, and both to 0 otherwise. Again, suppose the responses from i or j independently miss an existing undirected link between them at a rate of $\varphi > 0$ (say, due to idiosyncratic recall errors) but never erroneously report one if no such link exists between i and j in G. In this case, (A1) and (A2) hold with each entry in H_{ij} having a missing rate of $p = \varphi^2$.

On the other hand, the conditions (A1) and (A2) rule out another (third, empirically much less plausible) scenario, in which the actual G is asymmetric with directed links but researchers mistakenly impose symmetrized measure H as in the second scenario above. In this case, (A1) fails because $E(H_{ij}|G_{ij} = 1, G_{ji} = 1) = 1 - \varphi^2$ while $E(H_{ij}|G_{ij} = 1, G_{ji} = 0) = 1 - \varphi$, and (A2) fails because $E(H_{ij}|G_{ij} = 0) = (1 - \varphi)q_0 \neq 0$ with $q_0 \equiv E(G_{ji}|G_{ij} = 0) > 0$.

A clear advantage of the method we propose in this paper is that it allows researchers to consistently estimate social effects while being agnostic about whether the actual links in G are directed or not. The method only requires that the noisy measures H satisfy (A1)-(A3), which, as explained above, do not immediately impose the (a)symmetry of G or H. We therefore recommend a simple guideline for practitioners collecting link data: if a researcher is unsure about whether the actual links in G are directed or undirected, then a safe approach is to construct an unsymmetrized measure H as in the first scenario, and apply our method in this paper to deal with the possibility of missing links.

It is also important to note that Assumptions (A1)-(A3) do *not* specify how the actual links in G are formed. These conditions do not impose any known information about the actual adjacency matrix, except for its exogeneity in (A3). Nor do they impose any structure that can be used to derive a conditional likelihood for the actual adjacency matrix, that is, $Pr\{G|H,X\} = \frac{Pr(H|G,X)Pr(G|X)}{\int Pr(H|G,X)dF(G|X)}$. Constructing such a conditional likelihood would require specifying the likelihood of the actual network $Pr\{G|X\}$, which we avoid in this paper. Our method is therefore qualitatively different from existing alternative methods, which either use graphical reconstructions such as Chandrasekhar and Lewis (2011), or require knowledge of the distribution of actual adjacency matrix such as Boucher and Houndetoungan (2020).

Under (A1) and (A2), we can write:

$$E(H|G,X) = (1-p)G. \tag{5}$$

In the next subsection we show how this property, along with condition (A3), leads to a simple interpretation of the 2SLS estimand despite missing links.

3.2 Augmentation bias in two-stage least squares

In place of equation (1), we write a *reparametrized* structural form using H instead of G:

$$y = \lambda^* Hy + X\beta + \underbrace{\varepsilon + (\lambda G - \lambda^* H)\, y}_{\equiv v}, \text{ where } \lambda^* \equiv \lambda/(1-p). \tag{6}$$

Note the peer effect λ is replaced with a rescaled version λ^* in (6). Lemma 1 shows how this replacement restores the exogeneity of X with the new composite error v in (6).

Lemma 1. *Under (A1), (A2), and (A3),* $E(v|X,G) = 0$.

Lemma 1 may seem rather surprising *ex ante*, because one would expect (X,G) to be generically correlated with the composite error v which depends on y. The intuition for this result is as follows. Once we condition on the actual network G and explanatory variables in X, the randomness in individual outcomes y is solely due to the actual structural errors ε, which are uncorrelated with both X and (H, G) under (A3). As a result, any potential correlation between v and (X, G) could only be due to the reparametrized measurement error $\lambda G - \lambda^* H$. But equation (5) implies that $\lambda G - \lambda^* H$, and consequently the reparametrized error v, are mean independent from (X, G).

As discussed earlier, even with Lemma 1 establishing exogeneity of X in (6) by replacing λ with λ^*, there is still endogeneity in the term Hy because $E[(Hy)' v] \neq 0$ in general.[7] We therefore next investigate the estimand from 2SLS given appropriate instruments for Hy.

Based on Lemma 1, nonlinear functions of X can serve as instruments, if the usual rank (instrument relevance) condition is satisfied. However, nonlinear functions of X might not be relevant, or might be weak as instruments, since the structural model is linear in X. To deal with this possibility, we later show in Section 3.3 and 3.4 that it is also possible to use the noisy measures H to construct instruments just from linear functions of X. For instance, in Section 3.3, we show $H'X$ can be a valid instrument, meaning $E[(H'X)' v] = 0$, when H consists of directed links as in the case with unsymmetrized measures.

More generally, let ζ be a generic n-by-L matrix of instruments for Hy. Denote $R \equiv (Hy, X)$, $Z \equiv (\zeta, X)$ so that $E(R'v) \neq 0$ while $E(Z'v) = 0$. Assume instruments satisfy the following rank condition:

(IV-R) $E(Z'R)$ and $E(Z'Z)$ have full rank.

Let $\Pi \equiv [E(Z'Z)]^{-1} E(Z'R)$. By (6) and Lemma 1,

$$\begin{aligned} \Pi' E(Z'y) &= \Pi' E(Z'R)(\lambda^*, \beta')' + \Pi' E(Z'v) \\ \Rightarrow (\lambda^*, \beta')' &= \left[\Pi' E(Z'R)\right]^{-1} \left[\Pi' E(Z'y)\right]. \end{aligned} \tag{7}$$

[7] To see this, note $E(H'G|G, X) = (1-p)G'G$ under (A1) and (A2), but $E(H'H|G, X) \neq (1-p)^2 G'G$ under the same conditions. This is because the i-th diagonal entry in $H'H$ is $\sum_k H_{ki}^2 = \sum_k H_{ki}$ while its (i, j)-th off-diagonal entry is $\sum_k H_{ki} H_{kj}$. This inequality holds even under a stronger condition (A4) introduced in Section 3.3. It then follows from (A3) and the law of iterated expectation that $\lambda^* E(y'H'Hy) \neq \lambda E(y'H'Gy)$ in general.

We formalize this result in the next proposition.

Proposition 1. *Suppose (A1), (A2), and (A3) hold, and that (IV-R) holds for instruments Z. The two-stage least-squares estimand using Z for (6) is then $(\lambda^*, \beta')'$.*

Proposition 1 shows that when links are missing at random in the sample, 2SLS estimation using valid instruments leads to *augmentation bias* in the peer effect, because 2SLS estimates λ^* instead of λ. To reiterate, this result holds under the maintained conditions, regardless of whether the actual G is symmetric or not. Intuitively, when links are missing in the sample, their contribution to peer effects are erroneously attributed to the remaining observed links, thereby exaggerating the magnitude of peer effects attributed to the observed links. In contrast to peer effects, the individual effects β are consistently estimated by 2SLS (with valid instruments) despite missing links.

Based on Proposition 1, we have two main requirements for estimating the model. First, we need to construct valid instruments for 2SLS. Lemma 1 implies nonlinear functions of regressors $\zeta(X)$ may serve as instruments, provided they satisfy the rank condition in (IV-R) either through the reduced form of y or their correlation with the link formation in G. Alternatively, instruments could also be constructed from functions of the noisy network measure H (such as $\zeta(X, H) = H'X$ when H is unsymmetrized), as we show later in Section 3.3 and 3.4. Second, we need to estimate the missing rate p in order to convert the 2SLS estimate of λ^* into an estimate of λ. We address this question in Section 3.5.

3.3 Constructing instruments from a noisy network measure

We return to the question about how to construct instruments using a noisy network measure H. Assume:

(A4) Conditional on (G, X), H_{ij} and H_{kl} are independent whenever $(i, j) \neq (k, l)$.

This condition states the incidences of missing two different links are independent conditional on actual link status. Note that if $G_{ij} = G_{kl} = 0$, then (A2) implies H_{ij} and H_{kl} are both fixed at 0, which is consistent with (A4).

Condition (A4) does not restrict whether the actual network G is symmetric or not. For example, (A4) is consistent with H being an *unsymmetrized* measure of G, which is defined in the first scenario under (A1)-(A2) in Section 3.1). In this case, (A4) holds when H_{ij} and H_{ji} are independent measures of G_{ij} and G_{ji} respectively, *regardless of* whether the actual G is symmetric or not.

On the other hand, (A4) does not hold when H is a *symmetrized* measure of an actual G that is known to be symmetric with undirected links. This is the second scenario under (A1)-(A2) in Section 3.1. In this case, H and G are *both* symmetric with $(H_{ij}, G_{ij}) = (H_{ji}, G_{ji})$ for all $i \neq j$ almost surely. It then follows that $E(H_{ik}H_{ki}|G, X) = (1-p)G_{ik}$, which differs from $E(H_{ij}|G, X)E(H_{ji}|G, X) = (1-p)^2 G_{ik}$. This violates (A4). To deal with this case of symmetrized measures when the actual G is symmetric, we later give an alternative method for constructing instruments in Section 3.4. The method requires that the sample contain *multiple* symmetrized measures $H^{(1)}, H^{(2)}$.

Under (A4), we can construct instruments using H and X as follows.

Proposition 2. *Suppose (A1), (A2), (A3), and (A4) hold. Then* $E(Z'v) = 0$*, where* $Z \equiv (H'X, X)$.

There is a simple interpretation of the instruments $H'X$: the i-th component (row) of $H'X$ is the sum of characteristics of all individuals who report links with i in the sample.

Recall that GX would be valid instruments for Gy if G were perfectly observed in the sample. Therefore, one may wonder why we use $H'X$ instead of HX as instruments here. To understand this, note the composite error v in (6) contains the reparametrized measurement error $(\lambda G - \lambda^* H)$, and so in particular contains H. Hence, even under (A1)-(A4), HX is correlated with this reparametrized measurement error in v through H. In contrast, using a *transpose* of H in $H'X$ removes such correlation, because under (A4) the events of missing links between different pairs of individuals are conditionally independent. Therefore, $H'X$ satisfies instrument exogeneity while HX does not.

To apply 2SLS, the instruments need to satisfy the rank conditions in (IV-R). The next proposition specifies sufficient conditions for $H'X$ to satisfy (IV-R). These conditions are primitive, i.e., in terms of moments of functions of (X, G).

Proposition 3. *Suppose (A1), (A2), (A3), and (A4) hold, and $E(X'X)$ is non-singular. Then (IV-R) holds for $Z \equiv (H'X, X)$ if*

$$\begin{pmatrix} E(X'X) & E(X'M^{-1}X) \\ E(X'MX) & E(X'X) \end{pmatrix} \text{ and } \begin{pmatrix} E(X'G^2X) & E(X'GX) \\ E(X'GX) & E(X'X) \end{pmatrix} \text{ are non-singular.} \tag{8}$$

The rank conditions in (8) hold generically for random link formation models. Our simulations show that these conditions hold even for very restrictive cases where links are i.i.d. Bernoulli and independent from X. Violations of these conditions in (8) do exist in special cases. One such example is the linear-in-means social interactions model where G^k is proportional to a square matrix of ones for all positive integers k. It is worth noting that such an example of linear-in-means model also violates the rank condition for identifying social effects in Bramoullé et al. (2009), which requires I, G, and G^2 be linearly independent.

3.4 Instruments based on multiple symmetric measures

The method for constructing instruments in Section 3.3 assumes the sample reports *unsymmetrized* network measure H. In this section, we provide an alternative, complementary method for constructing instruments when the sample provides two (or more) *symmetrized* measures of an actual G that is known to be symmetric with all links being undirected.

For example, Banerjee et al. (2013) provide multiple measures of undirected links between households in rural villages across the State of Karnataka, India. For each pair of households, the survey asks which households you visited, and which ones visited you. Banerjee et al. (2013) symmetrize each of these two measures, yielding symmetric matrices we call $H^{(1)}$ and $H^{(2)}$. These two matrices are both measures of the same underlying symmetric network G. However, as we show later, these two matrices empirically differ substantially, indicating that they are different noisy measures of G.

Suppose we observe two symmetrized measures of the adjacency matrix, $H^{(1)}$ and $H^{(2)}$, which satisfy (A1), (A2), (A3), and

(A4') Conditional on (G, X), $H_{ij}^{(1)}$ and $H_{kl}^{(2)}$ are independent for all (i, j) and (k, l).

These two measures $H^{(1)}$ and $H^{(2)}$ have their own, different missing link rates, denoted $p^{(1)}$ and $p^{(2)}$ respectively. Condition (A4') is plausible when these distinct measures are independently collected and symmetrized, say, using responses from separate surveys.

Using either measure $H^{(1)}$ or $H^{(2)}$, we can construct a feasible structural form. That is, for $t = 1, 2$,

$$y = \tfrac{\lambda}{1-p^{(t)}} H^{(t)} y + X\beta + v^{(t)}, \text{ where } v^{(t)} = \varepsilon + \lambda \left[G - \tfrac{H^{(t)}}{1-p^{(t)}} \right] y. \tag{9}$$

Under (A1)-(A3) and (A4') and by an argument similar to Proposition 2, we can show that $H^{(2)}X$ satisfies the condition of instrument exogeneity with regard to $v^{(1)}$:

$$E\left[(H^{(2)}X)' v^{(1)} \right] = 0.$$

By a symmetric argument, analogous exogeneity holds for $H^{(1)}X$ and $v^{(2)}$. (See Appendix A for details.) We can therefore use $H^{(1)}X$ as instruments in equation (9) with $t = 2$, and $H^{(2)}X$ as instruments in (9) with $t = 1$. In Section 4, we discuss how to construct 2SLS estimators using these multiple, symmetrized network measures.

3.5 Recovering peer effects and missing rates

To remove the augmentation bias and recover the peer effect λ from the 2SLS estimand λ^*, we need to estimate the unknown missing link rate p. Here we provide two methods for identifying p in different and complementary cases.

First, consider a case where the sample reports an unsymmetrized noisy measure H, where the actual G is *known* to be symmetric with all links being undirected. This is a special case of the first scenario (of unsymmetrized measures) defined in Section 3.1. For example, the sample may collect self-reported survey responses about undirected links, with individual i reporting H_{ik} for $k \neq i$ and individual j reporting $H_{jk'}$ for $k' \neq j$.

In this case, we can construct a symmetrized measure $\tilde{H}$ with each element defined as $\tilde{H}_{ij} = \max\{H_{ij}, H_{ji}\}$. Under our assumptions, if the missing link rate for each element in H is p, then the missing link rate for $\tilde{H}$ will be p^2. Let $\psi(H) \in \mathbb{R}$ denote the average of all off-diagonal components in a network measure H. By the implication of randomly missing links in (5) and the linearity of $\psi(\cdot)$,

$$E[\psi(H)] = (1-p)E[\psi(G)] \text{ and } E[\psi(\tilde{H})] = (1-p^2)E[\psi(G)].$$

Hence, with $E[\psi(G)] \neq 0$, we can identify the missing rate as $p = E[\psi(\tilde{H})]/E[\psi(H)] - 1$.

Second, consider another case where the sample reports two independent measures, $H^{(1)}$ and $H^{(2)}$, with unknown missing rates $p^{(1)}$ and $p^{(2)}$ respectively. These measures could either be unsymmetrized as in the first scenario in Section 3.1 (where researchers may not know whether the actual G is symmetric or not), or symmetrized as in the second scenario in Section 3.1 (where researchers know the actual G are symmetric with undirected links).

In this case, we can construct a third measure with each element defined as $H_{ij}^{(3)} = \max\{H_{ij}^{(1)}, H_{ij}^{(2)}\}$. The implied missing rate for each element in $H^{(3)}$ is $p^{(3)} = p^{(1)} \times p^{(2)}$. By equation (5) we have

$$E[H^{(t)}] = (1-p^{(t)})E\left(G\right) \text{ for } t = 1, 2, 3.$$

By the linearity of ψ, we therefore get $E[\psi(H^{(t)})] = (1-p^{(t)})E[\psi(G)]$for $t = 1, 2, 3$. Hence, with $E[\psi(G)] \neq 0$, we can identify the missing rates $p^{(1)}$ and $p^{(2)}$ as:

$$p^{(1)} = \frac{E[\psi(H^{(3)})] - E[\psi(H^{(1)})]}{E[\psi(H^{(2)})]} \text{ and } p^{(2)} = \frac{E[\psi(H^{(3)})] - E[\psi(H^{(2)})]}{E[\psi(H^{(1)})]}.$$

Once the missing rates are recovered, we can use them to remove the augmentation bias in the 2SLS estimand in (6). Equivalently, we can use these rates to rescale the endogenous peer outcomes as $Hy/(1-p)$ so that 2SLS can then estimate $(\lambda, \beta')'$ consistently.

In each of the two cases above in Section 3.5, the matrix we construct, either $\tilde{H}$ or $H^{(3)}$, is a more accurate measure of G than the original H or $H^{(1)}$ and $H^{(2)}$ under maintained assumptions, in the sense of having a lower rate of missing links. However, direct estimation using these constructed matrices in place of G would still be biased and inconsistent due to the missing links. The estimators we propose in the next section do not directly use these constructed matrices (other than to estimate missing rates as above), but the estimators do make use of the information involved in such construction. For instance, our estimator in the second case uses both $H^{(1)}$ and $H^{(2)}$ that were used to construct $H^{(3)}$.

Some remarks about how we use the multiple, noisy network measures in Section 3.4 and 3.5 are in order. There is a broad and growing econometrics literature that uses repeated noisy measures to estimate nonlinear models with errors in variables, e.g., Li (2002), Chen et al. (2005) and Hu and Sasaki (2017) or unobserved heterogeneity, e.g., Hu (2008) and Bonhomme et al. (2016). More recently, Hu and Lin (2018) use repeated measurement to estimate a binary choice model with misclassification and social interactions. These papers typically apply mathematical tools such as deconvolution, and eigenvalue or LU decomposition to the joint distribution of repeated measures.

Unlike these papers, we use the repeated measures in two simple and intuitive steps that do not require any deconvolution or matrix/spectrum decomposition tools. Our focus on linear social networks allows us to exploit the identifying power from repeated measures through a standard 2SLS in Section 3.4. We are also able to apply a simple algebraic argument to recover the missing rates in Section 3.5. Both steps are constructive, and yield estimation based on the analog principle.

4 Two-Step Estimation

Section 3.5 provides two ways to recover the missing link rates, based on either observing a single unsymmetrized measure H, or observing two independent measures $H^{(1)}, H^{(2)}$. In this section, we propose 2SLS estimators for both cases.

Consider a sample of S independent groups, with each group s consisting of n_s members. (Later in Section 6 we consider extensions to a single growing network instead of many independent groups.) For each group s, the sample reports an n_s-by-1 vector of individual outcomes y_s, an n_s-by-K matrix of regressors X_s, and either an n_s-by-n_s unsymmetrized measure H_s, or two n_s-by-n_s symmetrized measures $H_s^{(1)}$ and $H_s^{(2)}$.

Let's begin with the *first case* in Section 3.5, where the actual adjacency matrices G_s are known to be symmetric (with all links being undirected), and the sample reports one *unsymmetrized* measure H_s for each group s. We first the missing rate p. Let $\tilde{H}_s$

denote a new *symmetrized* measure with its (i,j)-th component constructed as $\tilde{H}_{s,ij} = \max\{H_{s,ij}, H_{s,ji}\}$. Define:

$$\psi_s \equiv \psi(H_s) \text{ and } \widetilde{\psi}_s \equiv \psi(\widetilde{H}_s).$$

We estimate the missing rate p by

$$\widehat{p} = \frac{\frac{1}{S}\sum_{s=1}^{S}\widetilde{\psi}_s}{\frac{1}{S}\sum_{s=1}^{S}\psi_s} - 1.$$

Assume $\frac{1}{S}\sum_{s=1}^{S} E[\psi(G_s)]$ converges to a finite constant as $S \to \infty$. With (A1) and (A2) holding for each group s, $\frac{1}{S}\sum_{s=1}^{S} E(\psi_s)$ and $\frac{1}{S}\sum_{s=1}^{S} E(\widetilde{\psi}_s)$ also converge. Denote their limits by μ_ψ and $\mu_{\widetilde{\psi}}$ respectively. Let $\chi_s \equiv (\widetilde{\psi}_s - E[\widetilde{\psi}_s], \psi_s - E[\psi_s])'$. Suppose $\frac{1}{S}\sum_{s=1}^{S} E(\chi_s \chi_s') \to \Sigma_\psi$ as $S \to \infty$, with a finite Σ_ψ. By the Delta Method,

$$\sqrt{S}\,(\widehat{p} - p) \xrightarrow{d} \mathcal{N}(0, \mathcal{R}\Sigma_\psi\mathcal{R}'),$$

where $\mathcal{R} \equiv \left(\frac{1}{\mu_\psi}, -\frac{\mu_{\widetilde{\psi}}}{\mu_\psi^2}\right)$.

We then use $\widehat{p}$ to adjust the 2SLS estimator, yielding consistent estimation of λ. To simplify exposition, let $n_s = n$ for all $s = 1, ..., S$. Derivation for a general setting where n_s varies across the groups is similar, and only differs by requiring versions of the Law of Large Numbers and the Central Limit Theorem for independent and heterogeneous arrays indexed by s. In our application in Section 8, we allow variation in group (village) sizes.

For each group s and a generic $\tilde{p} \in (0,1)$, define:

$$W_s(\tilde{p}) \equiv \left(\frac{1}{1-\tilde{p}} H_s y_s, X_s\right) \text{ and } Z_s \equiv (H_s' X_s, X_s)\,.$$

Let Y denote an nS-by-1 vector that stacks y_s for $s \le S$. Similarly, let $\mathbf{W}(\widehat{p})$ be an nS-by-$(K+1)$ matrix that stacks $W_s(\widehat{p})$ for $s \le S$, and $\mathbf{Z}$ an nS-by-$2K$ matrix that stacks Z_s for $s \le S$. Our estimator for $\theta \equiv (\lambda, \beta')'$ is:

$$\widehat{\theta} \equiv \left(\mathbf{A}'\mathbf{B}^{-1}\mathbf{A}\right)^{-1} \mathbf{A}'\mathbf{B}^{-1} (\mathbf{Z}'Y)\,, \tag{10}$$

where

$$\mathbf{A} \equiv \mathbf{Z}'\mathbf{W}(\widehat{p}) \text{ and } \mathbf{B} \equiv \mathbf{Z}'\mathbf{Z}.$$

The next proposition characterizes the limiting distribution of $\widehat{\theta}$ as $S \to \infty$. Define $\Sigma_0 \equiv \left(A_0' B_0^{-1} A_0\right)^{-1} A_0' B_0^{-1}$ with $B_0 \equiv E(Z_s' Z_s)$ and $A_0 \equiv E\left[Z_s' W_s(p)\right]$, where p is the actual missing rate that generates the sample data. Let $\xi_s \equiv Z_s' v_s - F_0 \chi_s$, where v_s is the n-by-1 vector of composite errors in (6), and F_0 is a $2K$-by-1 vector defined as:

$$F_0 \equiv E\left[Z_s' \nabla W_s(p)\theta\right] = \tfrac{\lambda}{(1-p)^2} E(Z_s' H_s y_s), \text{ with } \nabla W_s(p) \equiv \tfrac{dW_s(\tilde{p})}{d\tilde{p}}|_{\tilde{p}=p} = \left(\tfrac{H_s y_s}{(1-p)^2}, 0\right).$$

Intuitively, F_0 illustrates how the moment condition in 2SLS depends on the missing rate p, and $-F_0\chi_s$ is the adjustment in the influence function that accounts for the first-stage estimation error in $\widehat{p}$.

Proposition 4. *Suppose (A1), (A2), (A3), and (A4) hold, and (IV-R) is satisfied with* $Z \equiv (H'X, X)$. *Then*

$$\sqrt{S}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_0 E(\xi_s \xi_s') \Sigma_0'),$$

under the regularity conditions (REG) in Appendix B.

The conditions (*REG*) in Appendix B are standard conditions for applying the Law of Large Numbers, the Central Limit Theorem, and the Delta Method.

Standard errors for $\widehat{\theta}$ are calculated by replacing A_0, B_0, F_0, and $E(\xi_s \xi_s')$ with their sample analogs:

$$\widehat{\mathbf{A}} = \tfrac{1}{S}\sum_s \mathbf{Z}_s' \mathbf{W}_s(\widehat{p}),\ \widehat{\mathbf{B}} = \tfrac{1}{S}\sum_s \mathbf{Z}_s' \mathbf{Z}_s,\ \widehat{F} = \tfrac{1}{S}\tfrac{\widehat{\lambda}}{(1-\widehat{p})^2}\sum_s Z_s' H_s y_s,\ \widehat{\xi}_s = Z_s'\widehat{v}_s - \widehat{F}\widehat{\tau}_s,$$

where

$$\widehat{v}_s = y_s - \mathbf{W}_s(\widehat{p})\widehat{\theta},\ \widehat{\tau}_s = \left(\tfrac{1}{\overline{\psi}}, -\tfrac{\overline{\widetilde{\psi}}}{(\overline{\psi})^2}\right)\left(\widetilde{\psi}_s - \overline{\widetilde{\psi}}, \psi_s - \overline{\psi}\right)',$$

with $\overline{\psi}, \overline{\widetilde{\psi}}$ being averages of $\psi_s, \widetilde{\psi}_s$ over $s \leq S$.

For the rest of Section 4, we explain how to apply a similar idea for estimation under the *second case* in Section 3.5. In this case, the sample reports two independent network measures, denoted as $H_s^{(1)}$ and $H_s^{(2)}$, for each group s, with missing rates $p^{(1)}$ and $p^{(2)}$ respectively. These measures may either be symmetrized or unsymmetrized. To reiterate, when $H_s^{(1)}$ and $H_s^{(2)}$ are *unsymmetrized*, our estimation method applies, *regardless of* whether the actual matrix G is known to be symmetric or not.

As noted in Section 3.4, these lead to two feasible structural forms, depending on which network measure is used:

$$y_s = W_s^{(t)}\theta + v_s^{(t)} \text{ for } t = 1, 2, \tag{11}$$

where $\theta \equiv (\lambda, \beta')'$, $W_s^{(t)} \equiv \left(\frac{H_s^{(t)} y_s}{1-p^{(t)}}, X_s\right)$, and $v_s^{(t)} \equiv \varepsilon_s + \lambda\left(G_s - \frac{H_s^{(t)}}{1-p^{(t)}}\right) y_s$. The exogenous instruments for these two systems are respectively $Z_s^{(1)} \equiv (H_s^{(2)} X_s, X_s)$ and $Z_s^{(2)} \equiv (H_s^{(1)} X_s, X_s)$. Let's write:

$$\tilde{Z}_s \equiv \begin{pmatrix} Z_s^{(1)} & 0 \\ 0 & Z_s^{(2)} \end{pmatrix}; \tilde{y}_s \equiv \begin{pmatrix} y_s \\ y_s \end{pmatrix}; \tilde{W}_s \equiv \begin{pmatrix} W_s^{(1)} \\ W_s^{(2)} \end{pmatrix}.$$

Instrument exogeneity implies the following moments:

$$E\left[\tilde{Z}_s'(\tilde{y}_s - \tilde{W}_s\theta)\right] = 0.$$

This moment condition identifies θ, provided $E(\tilde{Z}_s'\tilde{W}_s)$ has full rank. Using arguments similar to Proposition 3 in Section 3.3, we can derive analogous sufficient conditions for this rank condition. We omit the details here for brevity.

We define a system, or stacked, two-stage least squares (S2SLS) estimator as follows. Let $\tilde{\mathbf{Z}}$ denote a $2nS$-by-$4K$ matrix that is constructed by vertically stacking S matrices $(\tilde{Z}_s)_{s\leq S}$. Likewise construct a $2nS$-by-$(K+1)$ matrix $\tilde{\mathbf{W}}$ by stacking $(\tilde{W}_s)_{s\leq S}$ (with $p^{(t)}$ replaced by its analog estimates $\hat{p}^{(t)}$) and a $2nS$-by-1 vector $\tilde{\mathbf{y}}$ by stacking $(\tilde{y}_s)_{s\leq S}$. The S2SLS estimator is

$$\tilde{\theta} \equiv [\tilde{\mathbf{W}}'\tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\tilde{\mathbf{W}}]^{-1}\tilde{\mathbf{W}}'\tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\tilde{\mathbf{y}}. \tag{12}$$

This provides us with a single estimator that exploits both sets of instruments in the two structural forms in (11). Similar to $\hat{\theta}$ in (10), we can readily construct the standard error for $\tilde{\theta}$ that accounts for estimation error in $\hat{p}^{(1)}, \hat{p}^{(2)}$. We omit details here for brevity.

5 Extensions

We now extend the baseline method in Section 3 to more general settings with contextual effects, heterogeneous missing rates, or group fixed effects. In each case we focus on extending the ideas for constructive identification. Estimation in each of these cases follows from constructive identification arguments and analogous estimation steps in Section 4.

As before, to simplify exposition, let group sizes $n_s = n$ be fixed throughout the remainder of this section. This also allows us to suppress group subscripts s in notation.

5.1 Contextual effects

Suppose the structural form is:

$$y = \lambda Gy + X\beta + GX\gamma + \varepsilon,$$

where γ are contextual effects showing how individual outcomes are directly influenced by the characteristics of others linked to the individual. The reduced form is

$$y = M(X\beta + GX\gamma + \varepsilon),$$

where M is defined as in (4). The noisy structural form based on H is:

$$y = \lambda \frac{Hy}{1-p} + X\beta + \frac{HX}{1-p}\gamma + \eta,$$

where the composite error η is defined as

$$\eta \equiv \varepsilon - \lambda \left(\tfrac{H}{1-p} - G \right) y - \left(\tfrac{H}{1-p} - G \right) X\gamma.$$

Under the same conditions and by the same arguments as in the case with no contextual effects (Section 3.2), rescaling H by $(1-p)$ yields a new composite error η that is mean-independent from (X, G). We can similarly construct instruments using network measures H as before. Our next proposition establishes these results. For generality, let $\zeta(X) \in \mathbb{R}^{n \times L}$

be any generic function of X with $L \geq K$.

Proposition 5. *Suppose (A1), (A2), and (A3) hold. Then* $E(\eta|X,G) = 0$. *If in addition (A4) holds, then* $E\{[H'\zeta(X)]'\eta\} = 0$.

This proposition implies that $H'\zeta(X)$ satisfies instrument exogeneity for generic functions of X. In fact, a stronger result holds under (A1)-(A4): $E(H\eta|G,X) = 0$. The intuition is the same as in Proposition 2. Thus we can apply 2SLS as before to consistently estimate $(\lambda, \beta', \gamma')'$ using $(H'X, X, H'\zeta(X))$ as instruments for $W \equiv \left(\frac{Hy}{1-p}, X, \frac{HX}{1-p}\right)$, provided appropriate rank conditions hold.

5.2 Heterogeneous missing rates

We now extend our methods to allow the missing link rate p to vary with individual characteristics X. To focus on the main idea, we return to the case with no contextual effects as in (6). The generalization to include contextual effects, using the results of the previous sub-section, is immediate.

Suppose we replace (A2) with a more general condition:

$$\text{(A2')}\ E(H_{ij}|G_{ij} = 1, X) = 1 - p_{ij}(X) \text{ and } E(H_{ij}|G_{ij} = 0, X) = 0\ \forall i \neq j.$$

Under (A2'), $E(H|G,X) = Q \circ G$, where Q is an n-by-n matrix with its (i,j)-th component $Q_{ij} \equiv 1 - p_{ij}(X)$ and "$\circ$" denotes the Hadamard product. We suppress the dependence of Q on X for simplicity. By the law of iterated expectation,

$$E(H|X) = Q \circ E(G|X).$$

To recover $p_{ij}(\cdot)$, we can apply methods similar to Section 3.5 by focusing on single links and conditioning on X. For example, consider the second case in Section 3.5, where the sample reports two noisy measures with missing rates $p_{ij}^{(1)}(X)$ and $p_{ij}^{(2)}(X)$ respectively. Under (A2'), $E\left(H_{ij}^{(t)}\middle| X\right) = \left[1 - p_{ij}^{(t)}(X)\right] E(G_{ij}|X)$ for any $i \neq j$ and $t = 1, 2$. As before, we can construct a third measure $H_{ij}^{(3)} = \max\{H_{ij}^{(1)}, H_{ij}^{(2)}\}$ for each pair $i \neq j$, and then identify the missing rates as

$$p_{ij}^{(1)}(X) = \frac{E\left(H_{ij}^{(3)} - H_{ij}^{(1)} \middle| X\right)}{E\left(H_{ij}^{(2)} \middle| X\right)} \text{ and } p_{ij}^{(2)}(X) = \frac{E\left(H_{ij}^{(3)} - H_{ij}^{(2)} \middle| X\right)}{E\left(H_{ij}^{(1)} \middle| X\right)}.$$

In practice, we can mitigate the curse of dimensionality by specifying the missing rates $p_{ij}(X)$ and link formation probability $E(G_{ij}|X)$ as functions of X_i and X_j alone.

With knowledge (or estimates) of the heterogeneous missing rates, we can use 2SLS to consistently estimate $(\lambda, \beta')'$. Let $\tilde{Q}$ denote a "pointwise inverse" of Q, with the (i, j)-th entry being $\tilde{Q}_{ij} \equiv 1/(1 - p_{ij})$. With the missing rates identified, we can transform the structural form in (6) as

$$y = \lambda \left(\tilde{Q} \circ H\right) y + X\beta + \underbrace{\varepsilon + \lambda[G - \left(\tilde{Q} \circ H\right)]y}_{v^*}.$$

Under (A2') and (A3),

$$\begin{aligned} & E(v^*|G, X) = \lambda\{GE(y|G, X) - E\left[\left(\tilde{Q} \circ H\right) y \middle| G, X\right]\} \\ = \; & \lambda[GMX\beta - \tilde{Q} \circ E(H|G, X)MX\beta] = \lambda(G - \tilde{Q} \circ Q \circ G)MX\beta = 0. \end{aligned} \quad (13)$$

Let $W^* \equiv \left(\left(\tilde{Q} \circ H\right) y, X\right)$ and $Z^* \equiv (\zeta(X), X)$ where $\zeta(X) \in \mathbb{R}^{n \times L}$ is a nonlinear function of X with $L \geq K$ (e.g., $\zeta(X) \equiv X \circ X$). Then (13) implies $E(Z^{*\prime}v^*) = 0$. If $E(W^{*\prime}Z^*)$ and $E[Z^{*\prime}Z^*]$ have full rank, we can use 2SLS to consistently estimate λ and β.[8]

5.3 Group fixed effects

Suppose each group has an unobserved fixed effect α, so that the structural form is:

$$y = \lambda Gy + X\beta + \alpha + \varepsilon.$$

[8] With heterogeneous missing rates, $H'X$ does not satisfy instrument exogeneity, because $H(\widetilde{Q} \circ H) \neq \widetilde{Q} \circ H^2$ in general.

Let $\overline{G}$ denote an n-by-n matrix with identical rows, each of which equals the average of all rows in G. Define $\overline{H}$ and $\overline{X}$ analogously. Applying *within* transformation (as in fixed-effect estimation of a linear panel data model) using the group mean $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$, we eliminate α and get

$$y - \overline{y} = \frac{\lambda}{1-p}\left(H - \overline{H}\right) y + (X - \overline{X})\beta + v - \overline{v},$$

where

$$v - \overline{v} = \varepsilon - \overline{\varepsilon} + \lambda\left[G - \frac{H}{1-p} - \left(\overline{G} - \frac{\overline{H}}{1-p}\right)\right] y.$$

Because $\overline{G}$ and $\overline{H}$ are linear functions of G and H respectively, the same argument as Lemma 1 in Section 3 applies to show that $E(v - \overline{v}|X, G) = 0$. We can therefore use 2SLS to estimate $(\lambda, \beta')'$ exactly as before, after applying the within transformation.

6 A Single Large Network

So far we have focused on cases where the sample consists of many small, fixed-sized groups, where no links exist between members of different groups.

In this section we show how our method can be applied to settings with interdependence between *all* individuals in a sample. Specifically, we consider two settings in which some forms of weak dependence exist between individuals that are "far enough" from each other. For both settings, our proposed 2SLS estimators, when pooling observations over a single large network in the sample, remain consistent and asymptotically normal.

6.1 Nearly block-diagonal (NBD) networks

In this section, we consider a setting in which the sample can be partitioned into well-defined, *approximate groups*, which we henceforth refer to as "*blocks*". Links within each block are dense (i.e., the probability of forming links between two individuals within the same block does *not* diminish as the sample size increases). Links between individuals from different blocks exist but are sparse, so the probability of forming links across blocks diminishes as the number of blocks increases. Measurement issues arise because of two reasons. First, as before, links that exist within each block are missing randomly from the sample at a fixed rate. Second, those sparse cross-block links that now exist are never reported in the sample.

Formally, we partition the individuals in the sample into S blocks. Each block s consists of n_s members, with $n_s \geq 3$ being finite integers. To reiterate, links between individuals *within the same block* are reported in the sample, but are missing at a rate $p \in (0, 1)$ due to measurement errors. Links between individuals *across different blocks* are not reported in the sample. The sample size is $N \equiv \sum_{s=1}^{S} n_s$. Let G_N and H_N denote the true and noisy, *unsymmetrized* measures of N-by-N adjacency matrices respectively, which span the S blocks observed in the sample.

To facilitate derivation of the asymptotic properties of our 2SLS estimators, let $\widetilde{G}_N$ be a hypothetical *block-diagonal approximation* of G_N, which perfectly reports all within-block links but drops all cross-block links. That is, for all individual i,

$$\widetilde{G}_{N,ij} = G_{N,ij} \text{ if } j \in s(i); \; \widetilde{G}_{N,ij} = 0 \text{ if } j \notin s(i),$$

where $s(i)$ indicates the block that i belongs to. By construction, all elements outside the diagonal blocks in $\widetilde{G}_N$ are zeros. We maintain the following assumptions on the measurement errors in H_N:

$$\text{(N1) } E(H_{N,ij}|\widetilde{G}_N, X) = E(H_{N,ij}|\widetilde{G}_{N,ij}, X) \; \forall i \neq j, \text{ and}$$

$$\text{(N2) } E(H_{N,ij}|\widetilde{G}_{N,ij} = 1, X) = 1 - p \text{ and } E(H_{N,ij}|\widetilde{G}_{N,ij} = 0, X) = 0 \; \forall i \neq j.$$

Furthermore, we maintain that the block-specific random arrays, $H_{N,s}$, $\widetilde{G}_{N,s}$, $X_{N,s}$, $\epsilon_{N,s}$ (with $H_{N,s}, \widetilde{G}_{N,s}$ being n_s-by-n_s matrices), are drawn independently across the blocks.

We provide conditions under which, in this setting of a single, large network, our 2SLS consistently estimates structural parameters up to augmentation bias, which as before can be removed using estimates of $1 - p$. We return to the model with no contextual effects, so that the structural form is

$$y_N = \lambda G_N y_N + X_N \beta + \varepsilon_N, \tag{14}$$

where y_N, ε_N are N-by-1 vectors and X_N is N-by-K matrix of individual characteristics.

In Online Appendix A, we show a structural form using the noisy network measure is

$$y_N = \tfrac{\lambda}{1-p} H_N y_N + X_N \beta + v_N + u_N, \tag{15}$$

where $u_N \equiv \left(I_N - \lambda \frac{H_N}{1-p}\right)\left(I_N - \lambda \widetilde{G}_N\right)^{-1} \lambda \Delta_N y_N$ with $\Delta_N \equiv G_N - \tilde{G}_N$ and

$$v_N \equiv \varepsilon_N + \lambda \left(\widetilde{G}_N - \tfrac{H_N}{1-p}\right) \widetilde{y}_N \text{ with } \widetilde{y}_N \equiv (I_N - \lambda \widetilde{G}_N)^{-1}(X_N \beta + \varepsilon_N).$$

Note that we decompose composite errors in (15) into u_N and v_N, which are both vectorizations of block-specific vectors $u_{N,s}$ and $v_{N,s}$. While $v_{N,s}$ are independent across the blocks, $u_{N,s}$ are correlated across the blocks because of interdependence between $y_{N,s}$ due to sparse links between the blocks. This difference requires us to apply separate tactics to characterize their contribution to the estimation errors.

This decomposition of the composite error is useful for illustrating two main steps for deriving the asymptotic result. To see this, recall the 2SLS estimator that uses $Z_N \equiv (H_N' X_N, X_N)$ as instruments for $R_N \equiv (H_N y_N, X_N)$ is:

$$\widehat{\theta}_a = \left(A_N' B_N^{-1} A_N\right)^{-1} A_N' B_N^{-1} Z_N' y_N,$$

where $A_N \equiv Z_N' R_N$ and $B_N \equiv Z_N' Z_N$. By definition,

$$\widehat{\theta}_a - \theta_a = \left(A_N' B_N^{-1} A_N\right)^{-1} A_N' B_N^{-1} Z_N' (v_N + u_N),$$

where $\theta_a \equiv \left(\frac{\lambda}{1-p}, \beta'\right)'$ with the subscript a being a reminder that this estimand has augmentation bias. Thus the asymptotic property of the estimator depends on that of $Z_N' v_N$ and $Z_N' u_N$, which we investigate sequentially.

First, we characterize the order of $Z_N' v_N$, using the fact that $v_{N,s}$ are independent across blocks s. To see why such independence holds, recall that $H_{N,s}$, $\widetilde{G}_{N,s}$, $X_{N,s}$, $\epsilon_{N,s}$ are assumed independent across blocks s. By construct, $\widetilde{G}_N$, H_N, and $(I-\lambda\widetilde{G}_N)^{-1}$ are all block-diagonal. Hence we can write $\widetilde{y}_N$ as a vectorization of *independent*, hypothetical reduced forms. That is, $\widetilde{y}_N = vec([\widetilde{y}_{N,1}, \widetilde{y}_{N,2}, ..., \widetilde{y}_{N,S}])$, where $\widetilde{y}_{N,s} = (I_s - \lambda\widetilde{G}_{N,s})^{-1}(X_{N,s}\beta + \varepsilon_{N,s})$ are independent across s.[9] It then follows that $v_{N,s} = \varepsilon_{N,s} + \lambda\left(\widetilde{G}_{N,s} - \frac{H_{N,s}}{1-p}\right)\widetilde{y}_{N,s}$, and are independent across s.

We maintain exogeneity and independence conditions which are analogous to (A3) and (A4) for the case with small groups in Section 3:

(N3) $E(\varepsilon_{N,s}|X_{N,s}, G_{N,s}, H_{N,s}) = 0$ for all s;

(N4) Conditional on (G_N, X_N), $H_{N,ij} \perp H_{N,kl}$ for all $(i,j) \neq (k,l)$.

Under these conditions, $E(v_{N,s}|X_{N,s}, H_{N,s}) = 0$. The independence between $v_{N,s}$ mentioned above then allows us to apply the law of large numbers (Lemma A3 in the Online Appendix) to show that

$$\frac{1}{S} Z_N' v_N = \frac{1}{S}\sum\nolimits_s Z_{N,s}' v_{N,s} = O_p(S^{-1/2}).$$

Second, for analyzing the large-sample property of $Z_N' u_N$, we exploit the fact that it takes the form of $\mathcal{C}_N \Delta_N y_N$, where both $\mathcal{C}_N$ and y_N are uniformly bounded under mild regularity conditions (Lemma A2 in the Online Appendix). Hence the order of $\frac{1}{S} Z_N' u_N$ is bounded above by the expected number of missing links across the blocks, which are assumed to be sparse in the following sense:

$$\text{(S-LOB)} \sum\nolimits_{i=1}^{N} \sum\nolimits_{j \notin s(i)} E\left(|\Delta_{N,ij}|\right) = O(S^{\rho}) \text{ for some } \rho < 1.$$

This condition holds trivially in the many-group setting considered in Section 3, where by construction the total number of cross-block links is zero. It also holds if, for individuals in each block s, cross-block links only exist with a finite number of nearby blocks, and if the

[9] We refer to $\widetilde{y}_N$ as a *hypothetical* reduced form, because it is based on the block-diagonal approximation $\widetilde{G}_N$ rather than the actual G_N.

probability for forming such links q_S diminishes fast enough as the sample size grows (that is, $q_S = O(S^{-\alpha})$ with $\alpha > 1$).[10] Therefore, with $\mathcal{C}_N$ and y_N bounded, we can establish that $\frac{1}{S}Z_N' u_N = O_p(S^{\rho-1})$ under (S-LOB) (Lemma A1 in the Online Appendix.) This sparsity condition ensures A_N, B_N converge in probability to deterministic arrays in large samples (Lemma A3 in the Online Appendix.) Regularity conditions used for deriving asymptotic properties of $\widehat{\theta}_a - \theta_a$ are collected as Condition (S-REG) in Online Appendix A.

Putting these pieces together, we show that a feasible 2SLS estimator, which uses a noisy measure H and ignores all links between different blocks, consistently estimates the (augmented) structural parameter $\theta_a \equiv \left(\frac{\lambda}{1-p}, \beta'\right)'$ at a rate that is governed by the order of sparse, cross-block links. This result is formalized in the next proposition.

Proposition 6. *Suppose (N1), (N2), (N3) and (N4) hold. If Assumptions (S-LOB) and (S-REG) hold, then*

$$\widehat{\theta}_a - \theta_a = O_p(S^{-1/2} \vee S^{\rho-1}).$$

If in addition $\rho < 1/2$, then

$$\sqrt{S}\left(\widehat{\theta}_a - \theta_a\right) \xrightarrow{d} \mathcal{N}(0, \Omega),$$

where $\Omega \equiv \left(A_0' B_0^{-1} A_0\right)^{-1} A_0' B_0^{-1} \omega_0 B_0^{-1} A_0 \left(A_0' B_0^{-1} A_0\right)^{-1}$ *with* A_0, B_0, ω_0 *being non-stochastic arrays defined in the Online Appendix A.*

This proposition implies $\widehat{\theta}_a \xrightarrow{p} \theta_a$ because $\rho < 1$. Furthermore, if $\rho < 1/2$, the asymptotic distribution is determined by the leading term of $\frac{1}{\sqrt{S}}Z_N' v_N$, and hence matches the case of S independent small groups.

To estimate the missing rate p and remove the augmentation bias, one can apply the same method as the first step in Section 4, which remains valid because of independence of $H_{N,s}$ across the blocks $s = 1, 2, ..., S$.

[10] To see this, suppose all blocks have identical size $n_s = n < \infty$. Then the expected number of the cross-block links is $S(S-1) \times n^2 \times q_S = O(S^{2-\alpha})$, which satisfies (S-LOB).

6.2 Networks with near-epoch dependence (NED)

Here we obtain another consistency result for a different setting, in which the data consists of a single network that does *not* admit any definition of "approximate groups", but does include some notion of "distance" between individuals on the network. The main working assumption in this case is that the dependence between two individuals weakens as the distance between them increases, which is reminiscent of the notion of weak dependence in time series models.

Using this primitive condition, we show that observed outcomes satisfy a notion of near-epoch dependence (NED) as used in Jenish and Prucha (2012). Hence a form of the law of large numbers and the central limit theorem can be applied to sample averages over individual outcomes and covariates. We also show that the adjusted-2SLS estimator, when pooling observations over a single large network in the sample, converges in probability to the structural parameters, where the augmentation bias can be removed as before, once missing rates are estimated. Details are in the Online Appendix B and C of this paper.

7 Simulation

In this section we use monte carlo simulation to investigate the finite sample performance of our two-step 2SLS estimator in Section 4. Recall that the structural form of the data-generating process is:

$$y_s = \lambda G_s y_s + X_s \beta + \varepsilon_s,\ s = 1, 2, ..., S.$$

We fix each group size to be $n_s = 20$ in our simulation. In our design, each member i in each group s has two individual characteristics $X_{s,i} \in \mathbb{R}^2$, which are drawn independently across i and s. The first component $X_{s,i,1}$ is uniformly distributed over a finite support $\{-1, 1, 2\}$ while the second component $X_{s,i,2}$ is standard normal $N(0, 1)$. We consider three designs, corresponding to small, medium, and large peer effects, in which the true parameters are:

$$\lambda \in \{0.20, 0.35, 0.60\} \text{ while } (\beta_1, \beta_2) = (-1.5, 2).$$

The formulation of *undirected* links in the data-generating process is specified as follows. First, each individual sends invitations to two other individuals who are drawn randomly from the same group without replacement. An undirected link exists between two group members if either of them sends an invitation to the other. No links are formed across the groups. This generates each G matrix. Each H matrix is then constructed by dropping existing links randomly at the rate $p = 1/2$.

The size of a sample is defined as the number of independent groups in that sample. For each fixed sample size $S \in \{100, 400, 900\}$, we generate $T = 200$ samples. (Our simulated samples do not contain networks that are singular, which would violate regularity conditions.) By applying our two-step 2SLS estimator from Section 4 in each sample $t = 1, 2, ...T$, we record the empirical distribution of these estimates of $(\lambda, \beta_1, \beta_2)$. Table 1 below reports the average bias, sample variance, and mean-squared errors (MSEs) based on this empirical distribution.

Table 1. Two-step 2SLS Estimator Performance in Simulated Samples

S		$\lambda_{=0.2}$	β_1	β_2	$\lambda_{=0.35}$	β_1	β_2	$\lambda_{=0.6}$	β_1	β_2
100	avg. bias	0.000	0.014	0.002	0.009	0.055	-0.089	-0.303	0.734	-0.679
	variance	0.000	0.009	0.008	0.015	0.175	0.404	0.173	1.672	1.694
	m.s.e.	0.000	0.009	0.008	0.015	0.178	0.412	0.265	2.211	2.155
400	avg. bias	0.000	0.003	0.002	0.006	0.016	-0.037	-0.142	0.361	-0.235
	variance	0.000	0.002	0.002	0.002	0.033	0.083	0.176	0.780	0.606
	m.s.e.	0.000	0.002	0.002	0.002	0.033	0.084	0.196	0.910	0.661
900	avg. bias	0.000	0.001	0.000	0.005	0.004	-0.019	-0.056	0.162	-0.147
	variance	0.000	0.001	0.001	0.001	0.013	0.035	0.072	0.382	0.410
	m.s.e.	0.000	0.001	0.001	0.001	0.013	0.036	0.074	0.408	0.431

Table 1 shows that the mean-squared errors diminish as the sample size increases. For each parameter, the rate of decrease in MSE is roughly proportional to the rate of increase in the sample size. This offers evidence for the root-n convergence of our 2SLS estimator. It is also clear that estimator variance accounts for a major portion of the MSEs. For a fixed sample size and design, both the bias and variance of the peer effect λ are smaller than those for the individual effect (β_1, β_2). We also note that, as the peer effects λ increase, the MSEs increase for all parameters. This might be related to the fact that the variance of the estimator depends on the variation of y_s, which is scaled by $(1 - \lambda G)^{-1}$.

8 Application: Microfinance Participation in India

We apply our method to study how peer effects influence household decisions to participate in a microfinance program in India. The sample was collected by Banerjee et al. (2013) using survey questionnaires from the State of Karnataka, India between 2006-2007. Banerjee et al. (2013) impute a social network structure in the sample by aggregating several network measures that were inferred from the survey responses. They studied how the dissemination of information about a microfinance program, Bharatha Swamukti Samsthe, or *BSS*, depended on the network position of the households that were the first to be informed about the program. Banerjee et al. (2013) use a binary response model with social interactions to disentangle the effect of information diffusion from the peer effects, a.k.a. *endorsement* effects. In contrast, we use two of the multiple measures in Banerjee et al. (2013) as noisy proxies for an actual network, and apply our method to estimate peer effects in a linear social network model.

8.1 Institutional background and data

The sample was collected by Banerjee et al. (2013) through survey questionnaires from 43 villages in the State of Karnataka, India.[11] These villages are largely linguistically homogeneous but heterogeneous in terms of caste. The sample contains information about the socioeconomic status and some demographic characteristics of 9,598 households. On average, there were about 223 households in each village, with a minimum of 114, a maximum of 356, and a standard deviation of 56.2.

We merge the information from a full-scale household census and an individual-level

[11] The data are publicly available at: http://economics.mit.edu/faculty/eduflo/social.

survey in Banerjee et al. (2013). The household census gathered demographic information and data on a variety of amenities, such as roofing material, type of latrine, and quality of access to electric power. The individual survey was administered to a randomly selected sub-sample of villagers, which covered 46% of all households in the census. Individual questionnaires collected demographic information, such as age, caste and sub-caste, education, language, and having a ration card or not, but did not include explicit financial information. We merge the information about the head of household from the individual survey with the household information from the census. This yields a sample of 4,149 households. Table 2(a) reports summary statistics for the dependent variable ($y = 1$ if participates in the microfinance program) as well as a few continuous and binary explanatory variables. Summary statistics for additional categorical variables, such as religion, caste, property ownership, access to electricity, etc, are reported in Table 2(b).

Table 2(a): Summary of Dependent and Explanatory Variables

Variable	definition	obs.	mean	s.d.	min	max
y	dummy for participation	4149	0.1894	0.3919	0	1
room	number of rooms	4149	2.4389	1.3686	0	19
bed	number of beds	4149	0.9229	1.3840	0	24
age	age of household head	4149	46.057	11.734	20	95
edu	education of household head	4149	4.8383	4.5255	0	15
lang	whether to speak other language	4149	0.6799	0.4666	0	1
male	whether the hh head is male	4149	0.9161	0.2772	0	1
leader	whether it has a leader	4149	0.1393	0.3463	0	1
shg	whether in any saving group	4149	0.0513	0.2207	0	1
sav	whether to have a bank account	4148	0.3840	0.4864	0	1
election	whether to have an election card	4149	0.9525	0.2127	0	1
ration	whether to have a ration card	4149	0.9012	0.2985	0	1

The individual-level survey in Banerjee et al. (2013) also collected information about social interactions between households, such as (i) individuals whose homes the respondent visited, and (ii) individuals who visited the respondent's home. The sample in Banerjee et al. (2013) contains two *symmetric* measures for the latent network, based on the responses to (i) and (ii) respectively.[12] These two measures, reported as "visitGo" and "visitCome" matrices, are denoted as $H^{(1)}$ and $H^{(2)}$ in our notation.

By definition, an adjacency matrix based on (i) should be identical to that based on (ii), simply because household A visited household B also means B was visited by A. In the sample, the network matrices are symmetrized and, therefore, if it weren't for measurement errors, $H^{(1)}$ would be identical to $H^{(2)}$, with both measuring the same underlying G. However, as Table 3 shows, there is substantial discrepancy between these two measures in the sample. This suggests both $H^{(1)}, H^{(2)}$ are noisy proxies of G, with their zero entries possibly indicating missing links.[13]

Table 3 reports the empirical distribution of the degrees of $H^{(1)}$ and $H^{(2)}$. Because these measures are symmetric, there is no distinction between the degrees of in-bound or out-bound links. We pool all households across 43 villages into a single, large network. There are no links between households from different villages in the sample, so the network structure is block-diagonal.

Table 3 indicates large variation in the number of connections the households have. To reiterate, if there were no missing links in these reported measures, we would expect the two matrices $H^{(1)}$ and $H^{(2)}$ to be identical, and therefore have exactly the same degree distribution. The fact that they differ substantially is indicative of many missing links, possibly due to the respondents' recall errors, or to differences in how they interpret the visiting question. Thus, the two measures $H^{(1)}, H^{(2)}$ lend themselves to application of our method using two symmetric, noisy network measures in Section 3.4.

[12] Two households i and j are considered connected by an *undirected* link if an individual from either household mentioned the name of someone from the other household in response to the question in (i). Likewise, a second symmetric network measure is constructed based on responses to (ii).

[13] Banerjee et al. (2013) aggregate responses from 12 questions, including (i) and (ii), to construct a single symmetric network, which they consider to be, without any errors, an actual relevant adjacency matrix G. In contrast, we take a different approach by interpreting responses to questions (i) and (ii) as two different noisy measures of a true underlying latent network.

Table 2(b): Summary of Category Variables

Variable	definition	obs.	per.	Variable	definition	obs.	per.
religion				*latrine*			
-	Hinduism	3943	95.04	-	Owned	1195	28.80
-	Islam	198	4.77	-	Common	20	0.48
-	Christianity	7	0.19	-	None	2934	70.72
roof				*own*	property ownership		
-	Thatch	82	1.98	-	Owned	3727	89.83
-	Tile	1388	33.45	-	Owned & shared	32	0.77
-	Stone	1172	28.25	-	Rented	390	9.40
-	Sheet	868	20.92				
-	RCC	475	11.45				
-	Other	164	3.95				
electricity	electricity provision			*caste*			
				-	Scheduled caste	1139	27.54
-	Private	2662	64.18	-	Scheduled tribe	221	5.34
-	Government	1243	29.97	-	OBC	2253	54.47
-	No power	243	5.86	-	General	523	12.65

Table 3: Degree Distribution in Two Network Measures

Degree	0	1	2	3	4	5	6	7	8	9	10
$H^{(1)}$	2	21	110	227	357	505	526	546	506	379	269
$H^{(2)}$	4	24	112	245	384	522	534	577	491	386	255
Degree	11	12	13	14	15	16	17	18	19	20	≥ 21
$H^{(1)}$	224	145	90	74	54	33	27	15	9	6	24
$H^{(2)}$	179	137	102	59	46	28	22	13	9	3	17

8.2 Empirical strategy for estimating peer effects

We use the following specification for the feasible structural form:

$$y = \lambda \left(\frac{H^{(t)}}{1 - p^{(t)}} \right) y + X\beta + villageFE + v^{(t)} \text{ for } t = 1, 2, \tag{16}$$

where y is a binary variable indicating whether the household participated in the microfinance program (BSS), X is a matrix of household characteristics, and $villageFE$ are village fixed effects. Definition and summary statistics of regressors in X are listed in Table 2. Note that (16) provides *two* different feasible structural forms (of the same underlying true structural model), corresponding to $t = 1, 2$ respectively.

To implement an adjusted-2SLS estimator, we first estimate the missing rates $p^{(t)}$ for $t = 1, 2$, and use them to rescale the endogenous regressors as in Section 4. Following Section 3.5, we construct $H^{(3)} = \max\{H^{(1)}, H^{(2)}\}$ and estimate the missing rates as

$$\widehat{p}^{(1)} = \frac{\psi(H^{(3)}) - \psi(H^{(1)})}{\psi(H^{(2)})} = 0.1681, \text{ and } \widehat{p}^{(2)} = \frac{\psi(H^{(3)}) - \psi(H^{(2)})}{\psi(H^{(1)})} = 0.1909,$$

where $\psi(H)$ is the mean of off-diagonal entries in H. We replace $p^{(1)}$ and $p^{(2)}$ in equation (16) with $\widehat{p}^{(1)}$ and $\widehat{p}^{(2)}$ respectively, and then apply the 2SLS estimators in Section 4.

The results are reported in Table 4. The columns of Table 4 are all 2SLS estimates, defined as follows:

Column (a) ignores missing links in $H^{(1)}$, and so treats $H^{(1)}$ as if it were the true adjacency matrix G, by putting (unscaled) $\lambda H^{(1)} y$ on the right-hand side, and using $H^{(1)}X$ as the instruments for $H^{(1)}y$ in 2SLS.

Column (b) estimates the structural form for $t = 1$ in (16), using $H^{(2)}X$ as instruments for $\left(\frac{H^{(1)}}{1-\widehat{p}^{(1)}}\right) y$ in adjusted 2SLS.

Column (c) is identical to Column (a), except for using $H^{(2)}$ instead of $H^{(1)}$ everywhere, and so treats $H^{(2)}$ as if it were the true matrix G for 2SLS estimation

Column (d) is identical to column (b), except for switching the roles of the matrices $H^{(1)}$ and $H^{(2)}$. So the feasible structural model in (16) is written in terms of $t = 2$, and $H^{(1)}X$ is used as instruments for $\left(\frac{H^{(2)}}{1-\widehat{p}^{(2)}}\right) y$.

Column (e) applies the S2SLS estimator defined in (12) in Section 4. This estimator combines (stacks) the 2SLS moments used in Columns (b) and (d) above, and so combines the moments generated by both of the feasible structural models and their associated IVs into a single estimator.

In summary, the estimators in (a) and (c) are what a researcher would do if he or she ignored the missing links problem and treated either $H^{(1)}$ or $H^{(2)}$, respectively, as if it were the true adjacency matrix G, applying the standard 2SLS estimator that is proposed in the literature. In contrast, the corresponding adjusted-2SLS estimators in (b), (d) and (e) are estimators that we propose to remove the augmentation bias in 2SLS resulting from missing links.[14] Column (e) in particular combines the information used to construct the estimators in both columns (b) and (d), and so is our preferred estimator.

8.3 Empirical results

Table 4 reports that our adjusted 2SLS estimates for the peer effect $\widehat{\lambda}$ are 0.0456 when using $H^{(1)}y$ in the structural form (column (b)), 0.0484 using $H^{(2)}y$ (column (d)), and 0.0461 using both measures and S2SLS (column (e)). These estimates are all significant at the 1% level, and the differences between them are small relative to the standard errors. These estimates imply the likelihood of a household to participate in the microfinance program is increased by about 4.6% when the household is linked to one more participating household on the network (note for this calculation that our model does not row-normalize the network measures). With the average participation rate being 18.9% in the sample, these estimates suggest that peer effects, called "endorsement effects" in Banerjee et al. (2013), are economically substantial.

The signs of estimated marginal effects by individual or household characteristics are plausible. Column (e) suggests the head of household being a "leader" (e.g. a teacher, a leader of a self-help group, or a shopkeeper) increases the participation rate by around 3.9%. These households with "leaders" were the first ones to be informed about the program, and were asked to forward information about the microfinance program to other potentially interested villagers. These leaders had received first-hand, detailed information about the program from its administrator, which could be conducive to higher participation rates. Households with younger heads are more likely to participate, but the magnitude of this

[14] We need two noisy network measures in this particular context because the available reported measures are symmetric. As we show in Section 3.3, our method can also be used if the sample reports a single yet *asymmetric* noisy measure of the network.

age effect is less substantial. Being 10 years younger increases the participation rate by 1.7%. Having a ration card increases the participation rate by around 4.3%. Compared to households using private electricity, households using government-supplied electricity have a 3.4% higher participation rate. These two factors indicate that, holding other factors equal, households in poorer economic conditions are more inclined to participate in the microfinance program.

Table 4 also shows that, if we had ignored the issue of missing links in network measures, and had done 2SLS using $H^{(t)}X$ as instruments for the (unscaled) endogenous peer outcomes $H^{(t)}y$, then the estimator would have been considerably biased upward. In (a), where we use $H^{(1)}X$ as instruments for $H^{(1)}y$, the estimate for λ is 0.0498. In comparison, in (b), where we correct for missing link bias by using $H^{(2)}X$ as instruments for $\frac{H^{(1)}y}{1-\hat{p}^{(1)}}$, the estimated λ is 0.0456. The upward bias resulted from ignoring the missing links is about 9.2% (as 0.0498/0.0456=1.092). Likewise, in (c) where we erroneously use $H^{(2)}X$ as instruments for $H^{(2)}y$, we get a proportionally almost the same upward bias in the peer effect estimate compared with the correct estimate in (d) (as 0.0529/0.0484=1.093).

The over 9% upward bias in (a) and (c) is a manifestation of two factors at work. First, with missing links the instruments $H^{(t)}X$ are invalid because of the correlation between $H^{(t)}X$ and the composite errors $v^{(t)}$. Second, even if these instruments were valid, the augmentation bias, as defined in Section 3.2, would be present without rescaling the endogenous peer outcomes $H^{(t)}y$ by $1 - p^{(t)}$.

The magnitude of this upward bias is determined by the magnitude of $p^{(t)}$ and by the correlation between the composite error and the invalid instruments. The microfinance survey data in Banerjee et al. (2013) is considered to have high quality social network information. In other empirical environments, we may expect even larger bias when missing links are not accounted for in estimation. The method we propose in this paper provides an easy remedy for this issue.

Table 4: **Two-stage Least Square Estimates**

	(a)	(b)	(c)	(d)	(e)
r.h.s. endogeneity	$H^{(1)}y$	$\frac{H^{(1)}}{1-\widehat{p}_1}y$	$H^{(2)}y$	$\frac{H^{(2)}}{1-\widehat{p}_2}y$	$\frac{H^{(t)}}{1-\widehat{p}}y$
IV used	$H^{(1)}X$	$H^{(2)}X$	$H^{(2)}X$	$H^{(1)}X$	Combined
$\widehat{\lambda}$	0.0498***	0.0456***	0.0529***	0.0484***	0.0461***
	(0.0076)	(0.0096)	(0.0092)	(0.0087)	(0.0075)
$leader$	0.0378**	0.0364**	0.0418**	0.0405**	0.0387**
	(0.0185)	(0.0186)	(0.0182)	(0.0182)	(0.0183)
age	-0.0016***	-0.0017***	-0.0016***	-0.0017***	-0.0017***
	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
$ration$	0.0441**	0.0435**	0.0423**	0.0413**	0.0426**
	(0.0201)	(0.0201)	(0.0195)	(0.0194)	(0.0197)
$electricity - gov$	0.0343**	0.0333**	0.0352**	0.0341**	0.0339**
	(0.0157)	(0.0157)	(0.0156)	(0.0155)	(0.0156)
$electricity - no$	0.0223	0.0229	0.0237	0.0247	0.0236
	(0.0297)	(0.0297)	(0.0300)	(0.0298)	(0.0298)
$caste - tribe$	-0.0285	-0.0272	-0.0275	-0.0257	-0.0268
	(0.0312)	(0.0309)	(0.0305)	(0.0300)	(0.0305)
$caste - obc$	-0.0520**	-0.0490**	-0.0486**	-0.0441***	-0.0473***
	(0.0217)	(0.0212)	(0.0215)	(0.0206)	(0.0210)
$caste - gen$	-0.0734***	-0.0698***	-0.0688***	-0.0628**	-0.0673***
	(0.0239)	(0.0242)	(0.0241)	(0.0234)	(0.0239)
$religion - Islam$	0.0980***	0.0955***	0.0893***	0.0849***	0.0910***
	(0.0323)	(0.0323)	(0.0343)	(0.0344)	(0.0332)
$religion - Chri$	0.1434	0.1420	0.1466	0.1452	0.1438
	(0.130)	(0.1287)	(0.1314)	(0.1300)	(0.1293)
$Controls$	√	√	√	√	√
$VillageFE$	√	√	√	√	√
R^2	0.1332	0.1345	0.1350	0.1365	0.1353
Obs	4134	4134	4134	4134	4134

Note: s.e. in parentheses. ***, **, and * indicate 1%, 5%, and 10% significant.

Controls include *male*, *roof*, *room*, *bed*, *latrine*, *edu*, *lang*, *shg*, *sav*, *election*, and *own*.

We conclude this section with some model validation results in Table 5, which shows how the predicted values of $E(y|X)$ fit with the sample data. The Probit and Logit models use the same set of regressors as in Table 4. We report the summary statistics of the fitted values $\widehat{E(y|X)}$ under different models. Columns (a) through (d) of Table 5 are the fitted values of the feasible structural models used in each of the corresponding columns in Table 4. Column (e) in Table 4 used two different feasible structural models to obtain S2SLS estimates. To make use of both for fitted values, in column (e) of table 5 we use the S2SLS estimates of (λ, β) and construct fitted values based on $\hat{\lambda}\frac{H^{(3)}}{1-p_1p_2}y + X\hat{\beta} + \hat{FE}$, where $H^{(3)}$ is as defined in Section 3.5.

Table 5: Model Validation: Predicted Microfinance Participation

$\widehat{E(y\|X)}$	Probit	Logit	OLS	(a)	(b)	(c)	(d)	(e)
mean	0.1894	0.1894	0.1894	0.1894	0.1894	0.1894	0.1894	0.1881
s.t.d	0.1176	0.1181	0.1151	0.1339	0.1376	0.1356	0.1409	0.1337
min	0.0103	0.0166	-0.095	-0.104	-0.108	-0.127	-0.131	-0.110
max	0.7490	0.7673	0.6895	0.7807	0.8016	0.7279	0.7576	0.8036
< 0	0%	0%	2.95%	4.67%	4.98%	4.79%	5.49%	4.84%
$I\{\widehat{E(y\|X)} > 0.5\}$								
underpredict (1 to 0)	17.76%	17.66%	18.34%	17.34%	17.17%	17.34%	17.13%	17.30%
overpredict (0 to 1)	0.92%	1.11%	0.27%	0.87%	1.02%	0.85%	0.97%	0.80%
correct	81.33%	81.23%	81.40%	81.79%	81.81%	81.81%	81.91%	81.91%

In all but one of the models in Table 5, the sample mean of the predicted participation probability $\widehat{E(y|X)}$ is 0.1894, which is equal to the sample mean of y in the 4,134 observations used in the regression. The standard deviation of the predicted participation probability varies across different models. Predictions of linear probability models (LPM), reported under the column of "OLS" and (a)-(e), are mostly within the unit interval $[0, 1]$. LPM predictions are strictly less than 1 for all observations in the sample; Only 2.95% to 5.49% of the households in the sample end up with negative LPM predictions. That is, about 95% all LPM predictions in the sample are indeed within the unit interval.

Based on $\widehat{E(y|X)}$, we use the indicator $I(\widehat{E(y|X)} > 0)$ to predict whether an individual participates in the microfinance program, and calculate prediction rates. Predictions in our linear social network models in columns (a)-(e) generally outperform the OLS, Probit and Logit models in terms of the percentage of correct predictions.

9 Conclusion

This paper proposes adjusted-2SLS estimators that consistently estimate structural parameters, which include peer, individual, and contextual effects, in social network models when actual existing links are missing randomly from the sample. By rescaling the endogenous peer outcomes and applying new instruments constructed from noisy network measures, our estimators resolve the additional endogeneity issues caused by missing links. As an intermediate step of the method, we provide methods to estimate the rates at which links are missing from noisy measures of network links. We also show that ignoring missing links generally leads to augmentation bias, i.e., peer effect estimates are generally biased upward.

We apply our method to analyze the peer (endorsement) effects in households' decisions to participate in a microfinance program in Indian villages, using the data collected by Banerjee et al. (2013). Consistent with our theoretical results, our empirical estimates show that ignoring the issue of missing links in the 2SLS estimation of the social network model leads to a substantial upward bias (over 9%) in the estimates of peer effects.

References

Advani, A. and B. Malde (2018). Credibly identifying social effects: Accounting for network formation and measurement error. *Journal of Economic Surveys 32*(4), 1016–1044.

Aigner, D. J. et al. (1973). Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics 1*(1), 49–59.

Auerbach, E. (2022). Identification and estimation of a partially linear regression model using network data. *Econometrica 90*(1), 347–365.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013). The diffusion of microfinance. *Science 341*(6144), 1236498.

Blume, L. E., W. A. Brock, S. N. Durlauf, and Y. M. Ioannides (2011). Identification of social interactions. In *Handbook of social economics*, Volume 1, pp. 853–964. Elsevier.

Bollinger, C. R. (1996). Bounding mean regressions when a binary regressor is mismeasured. *Journal of Econometrics 73*(2), 387–399.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2016). Non-parametric estimation of finite mixtures from repeated measurements. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 211–229.

Boucher, V. and A. Houndetoungan (2020). *Estimating peer effects using partial network data*. Centre de recherche sur les risques les enjeux économiques et les politiques.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009). Identification of peer effects through social networks. *Journal of econometrics 150*(1), 41–55.

Butts, C. T. (2003). Network inference, error, and informant (in) accuracy: a bayesian approach. *social networks 25*(2), 103–140.

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

Chen, X., H. Hong, and E. Tamer (2005). Measurement error models with auxiliary data. *The Review of Economic Studies 72*(2), 343–366.

Goldsmith-Pinkham, P. and G. W. Imbens (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics 31*(3), 253–264.

Graham, B. S. (2020). Network data. In *Handbook of Econometrics*, Volume 7, pp. 111–218. Elsevier.

Griffith, A. (2021). Name your friends, but only five? the importance of censoring in peer effects estimates using social network data. *Journal of Labor Economics*.

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

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics 144*(1), 27–61.

Hu, Y. and Z. Lin (2018). Misclassification and the hidden silent rivalry.

Hu, Y. and Y. Sasaki (2017). Identification of paired nonseparable measurement error models. *Econometric Theory 33*(4), 955–979.

Jenish, N. and I. R. Prucha (2012). On spatial processes and asymptotic inference under near-epoch dependence. *Journal of econometrics 170*(1), 178–190.

Kelejian, H. H. and I. R. Prucha (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics 17*(1), 99–121.

Klepper, S. (1988). Bounding the effects of measurement error in regressions involving dichotomous variables. *Journal of Econometrics 37*(3), 343–359.

Lee, L.-F. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics 140*(2), 333–374.

Lewbel, A. (2007). Estimation of average treatment effects with misclassification. *Econometrica 75*(2), 537–551.

Li, T. (2002). Robust and consistent estimation of nonlinear errors-in-variables models.

Journal of Econometrics 110(1), 1–26.

Lin, X. (2010). Identifying peer effects in student academic achievement by spatial autoregressive models with group unobservables. *Journal of Labor Economics 28*(4), 825–860.

Liu, X. (2013a). Estimation of a local-aggregate network model with sampled networks. *Economics Letters 118*(1), 243–246.

Liu, X. (2013b). Estimation of a local-aggregate network model with sampled networks. *Economics Letters 118*(1), 243–246.

Mahajan, A. (2006). Identification and estimation of regression models with misclassification. *Econometrica 74*(3), 631–665.

Molinari, F. (2008). Partial identification of probability distributions with misclassified data. *Journal of Econometrics 144*(1), 81–117.

Shalizi, C. R. and A. Rinaldo (2013). Consistency under sampling of exponential random graph models. *Annals of statistics 41*(2), 508.

Appendix

A. Identification proofs

Proof of Lemma 1. Under (A3), we have $E(Gy|X,G) = E[GM(X\beta+\varepsilon)|X,G] = GMX\beta$, and $E(Hy|X,G) = E[HME\left(X\beta+\varepsilon|X,G,H\right)|X,G] = E(H|G,X)MX\beta$.

Under (A1) and (A2), $E(H|G,X) = (1-p)G$. It follows from the definition of v in (6) that $E(v|X,G) = 0$. □

Proof of Proposition 2. Under (A1), (A2), and (A4), the conditional mean of the (i,j)-th entry in H^2 is

$$\begin{aligned} E\left[(H^2)_{ij}|G,X\right] &= E\left(\sum\nolimits_{k\neq i,j} H_{ik}H_{kj}\Big| G,X\right) = \sum\nolimits_{k\neq i,j} E\left(H_{ik}H_{kj}|\, G,X\right) \\ &= \sum\nolimits_{k\neq i,j} E\left(H_{ik}|\, G_{ik},X\right) E\left(H_{kj}|\, G_{kj},X\right) = \sum\nolimits_{k\neq i,j}(1-p)G_{ik}(1-p)G_{kj} \\ &= (1-p)^2\left(G^2\right)_{ij}. \end{aligned} \tag{17}$$

Besides, under (A1) and (A2),

$$E\left[HG|G,X\right] = E(H|G,X)G = (1-p)G^2. \tag{18}$$

It then follows that

$$\begin{aligned} E[(H'X)'v|G,X] &= E(X'H\varepsilon|G,X) + \lambda E\left[X'H\left(G-\frac{H}{1-p}\right)y\Big|\,G,X\right] \\ &= \lambda E\left[X'H\left(G-\frac{H}{1-p}\right)MX\beta\Big|\,G,X\right] \\ &= \lambda X'\left(E(HG|G,X)-\frac{E(H^2|G,X)}{1-p}\right)MX\beta = 0, \end{aligned}$$

where the first two equalities are due to (A3), and the last holds because of (17) and (18) under (A1), (A2), and (A4). □

As noted in Section 3.4, one can construct instruments from multiple symmetrized measures for G, denoted by $H^{(1)}$ and $H^{(2)}$. Suppose $H^{(1)}$ and $H^{(2)}$ both satisfy (A1), (A2), (A3), and are independent in the sense of (A4'). Then one can construct feasible structural forms as in (9), and use $H^{(2)}X$ as instruments for $v^{(1)}$, and vice versa. To see why, note that for all i and j (including the case with $i=j$):

$$\begin{aligned} &E\left[(H^{(2)}H^{(1)})_{ij}|G,X\right] = E\left(\sum\nolimits_{k\neq i,j} H^{(2)}_{ik}H^{(1)}_{kj}\Big|\,G,X\right) \\ =\;& \sum\nolimits_{k\neq i,j} E\left(H^{(2)}_{ik}H^{(1)}_{kj}\Big|\,G,X\right) = \sum\nolimits_{k\neq i,j} E\left(H^{(2)}_{ik}\Big|\,G_{ik},X\right)E\left(H^{(1)}_{kj}\Big|\,G_{kj},X\right) \\ =\;& \sum\nolimits_{k\neq i,j}(1-p^{(2)})G_{ik}(1-p^{(1)})G_{kj} = (1-p^{(2)})(1-p^{(1)})\left(G^2\right)_{ij}. \end{aligned} \tag{19}$$

Besides, under (A1) and (A2),

$$E\left[H^{(2)}G|G,X\right] = E(H^{(2)}|G,X)G = (1-p^{(2)})G^2. \tag{20}$$

It then follows that

$$\begin{aligned} E[(H^{(2)}X)'v^{(1)}|G,X] &= E(X'H^{(2)}\varepsilon|G,X) + \lambda E\left[X'H^{(2)}\left(G - \frac{H^{(1)}}{1-p^{(1)}}\right)y\middle| G,X\right] \\ &= \lambda E\left[X'H^{(2)}\left(G - \frac{H^{(1)}}{1-p^{(1)}}\right)MX\beta\middle| G,X\right] \\ &= \lambda X'\left(E(H^{(2)}G|G,X) - \frac{E(H^{(2)}H^{(1)}|G,X)}{1-p^{(1)}}\right)MX\beta = 0. \end{aligned}$$

where the first two equalities are due to (A3), and the last holds because of (19) and (20) under (A1), (A2), and (A4').

Proof of Proposition 3. Define the following K-by-K moments involving (G,X) :

$$\begin{aligned} B_1 &\equiv E(X'G^2MX),\ B_2 \equiv E(X'GMX),\ B_3 \equiv E(X'G^2X), \\ B_4 &\equiv E(X'GX),\ B_5 \equiv E(X'X). \end{aligned}$$

Under (A1), (A2), (A3), and (A4),

$$\begin{aligned} E(Z'R) &= \begin{pmatrix} E(X'H^2y) & E(X'HX) \\ E(X'Hy) & E(X'X) \end{pmatrix} = \begin{pmatrix} E[X'H^2M(X\beta+\varepsilon)] & E(X'HX) \\ E[X'HM(X\beta+\varepsilon)] & E(X'X) \end{pmatrix} \\ &= \begin{pmatrix} (1-p)^2E(X'G^2MX\beta) & (1-p)E(X'GX) \\ (1-p)E(X'GMX\beta) & E(X'X) \end{pmatrix} \equiv \begin{pmatrix} (1-p)^2B_1\beta & (1-p)B_4 \\ (1-p)B_2\beta & B_5 \end{pmatrix}. \end{aligned}$$

Suppose the $2K$-by-$(1+K)$ matrix $E(Z'R)$ does not have full rank. By definition the $2K$-by-$2K$ square matrix

$$\begin{pmatrix} (1-p)^2B_1 & (1-p)B_4 \\ (1-p)B_2 & B_5 \end{pmatrix}$$

must be singular. This implies $[B_1, B_4; B_2, B_5]$ must also be singular because

$$\det\begin{pmatrix}(1-p)^2B_1 & (1-p)B_4\\ (1-p)B_2 & B_5\end{pmatrix} = \det(B_5)\det\left[(1-p)^2B_1-(1-p)^2B_4(B_5)^{-1}B_2\right]$$

$$= \quad (1-p)^{2K}\det(B_5)\det(B_1-B_4B_5^{-1}B_2) = (1-p)^{2K}\det\begin{pmatrix}B_1 & B_4\\ B_2 & B_5\end{pmatrix}.$$

Therefore, non-singularity of $[B_1, B_4; B_2, B_5]$ implies that $E(Z'R)$ has full rank.

As $M-\lambda GM = I$, we have $GM = \frac{1}{\lambda}(M-I)$ and $G^2M = \frac{1}{\lambda}(GM-G) = \frac{1}{\lambda^2}(M-I-\lambda G)$. We can write

$$\begin{pmatrix}B_1 & B_4\\ B_2 & B_5\end{pmatrix} = \begin{pmatrix}\frac{1}{\lambda}E(X'(GM-G)X) & E(X'GX)\\ E(X'GMX) & E(X'X)\end{pmatrix}.$$

Adding the product of the 2nd row and $(-\frac{1}{\lambda})$ to the 1st row, we get:

$$\begin{pmatrix}-\frac{1}{\lambda}E(X'GX) & E(X'GX)-\frac{1}{\lambda}E(X'X)\\ E(X'GMX) & E(X'X)\end{pmatrix}.$$

Adding the product of the 2nd column and $(\frac{1}{\lambda})$ to the 1st column, we get

$$\begin{pmatrix}-\frac{1}{\lambda^2}E(X'X) & E(X'GX)-\frac{1}{\lambda}E(X'X)\\ E(X'(GM+\frac{1}{\lambda}I)X) & E(X'X)\end{pmatrix} = \begin{pmatrix}-\frac{1}{\lambda^2}E(X'X) & -\frac{1}{\lambda}E(X'M^{-1}X)\\ \frac{1}{\lambda}E(X'MX) & E(X'X)\end{pmatrix}.$$

Hence, $\begin{pmatrix}B_1 & B_4\\ B_2 & B_5\end{pmatrix}$ is non-singular iff $\begin{pmatrix}E(X'X) & E(X'M^{-1}X)\\ E(X'MX) & E(X'X)\end{pmatrix}$ is non-singular.

By the same token, (A1), (A2), and (A4) imply that

$$\begin{aligned}E(Z'Z) &= \begin{pmatrix}E(X'H^2X) & E(X'HX)\\ E(X'HX) & E(X'X)\end{pmatrix} = \begin{pmatrix}(1-p)^2E(X'G^2X) & (1-p)E(X'GX)\\ (1-p)E(X'GX) & E(X'X)\end{pmatrix}\\ &= \begin{pmatrix}(1-p)^2B_3 & (1-p)B_4\\ (1-p)B_4 & B_5\end{pmatrix}.\end{aligned}$$

Similarly, the determinant of $E(Z'Z)$ is proportional to that of $[B_3, B_4; B_4, B_5]$. Therefore, the non-singularity of $[B_3, B_4; B_4, B_5]$ implies $E(Z'Z)$ has full rank. $\square$

Proof of Proposition 5. Under (A3), we have

$$\begin{aligned} E(Gy|X,G) &= E[GM(X\beta + GX\gamma + \varepsilon)|X,G] = GM\left(X\beta + GX\gamma\right), \\ E(Hy|X,G) &= E[HME\left(X\beta + GX\gamma + \varepsilon|X,G,H\right)|X,G] = E(H|G,X)M(X\beta + GX\gamma). \end{aligned}$$

Under (A1) and (A2), $E(H|G,X) = (1-p)G$. It then follows that $E(\eta|X,G) = 0$. Note

$$\begin{aligned} E\left[\zeta(X)'HHy|G,X\right] &= \zeta\left(X\right)'E(H^2|G,X)M(X\beta + GX\gamma); \\ E[\zeta\left(X\right)'HHX|G,X] &= \zeta\left(X\right)'E(H^2|G,X)X; \\ E\left[\zeta(X)'HGy|G,X\right] &= \zeta\left(X\right)'E(H|G,X)GM(X\beta + GX\gamma); \\ E[\zeta\left(X\right)'HGX|G,X] &= \zeta\left(X\right)'E(H|G,X)GX. \end{aligned}$$

As shown in the proof of Proposition 2, under (A4), $E(H^2|G,X) = (1-p)^2G$. Because $E(H|G,X) = (1-p)G$ under (A1) and (A2), this implies $E\left[\zeta\left(X\right)'H\eta\right] = 0$. $\square$

B. Asymptotic property of two-step Estimator

In this section we sketch a proof of asymptotic distribution for $\widehat{p}$, $\widehat{\lambda}$, and $\widehat{\beta}$. We maintain the following regularity conditions:

(REG) $E(\psi_s) \neq 0$; $0 < p < 1$; $E(|Z_s'W_s(p)|) < \infty$, $E(|Z_s'Z_s|) < \infty$, $E(||\xi_s||^2) < \infty$ where ξ_s is defined below.

These conditions are needed for applying the law of large numbers, the central limit theorem, and the delta method below.

First off, by the central limit theorem,

$$\frac{1}{\sqrt{S}}\begin{pmatrix} \sum_s\left[\widetilde{\psi}_s - E(\widetilde{\psi}_s)\right] \\ \sum_s\left[\psi_s - E(\psi_s)\right] \end{pmatrix} \xrightarrow{d} \mathcal{N}(0,\Omega),$$

where Ω is the covariance matrix of $(\widetilde{\psi}_s, \psi_s)'$. The delta method implies $\sqrt{S}(\hat{p} - p) \xrightarrow{d} \mathcal{N}(0, D\Omega D')$, where

$$D = \left(\frac{1}{E(\psi_s)}, -\frac{E(\widetilde{\psi}_s)}{E(\psi_s)^2} \right).$$

The asymptotic linear presentation of $\widehat{p}$ is

$$\sqrt{S}(\widehat{p} - p) = \tfrac{1}{\sqrt{S}} \sum\nolimits_s \tau_s + o_p(1),$$

where $\tau_s \equiv D \times \left(\widetilde{\psi}_s - E(\widetilde{\psi}_s), \psi_s - E(\psi_s) \right)'$ with $E[\tau_s] = 0$.

Hence, $\sqrt{S}(\hat{p} - p) \xrightarrow{d} \mathcal{N}(0, E(\tau_s \tau_s'))$. Next, note that by construction,

$$\begin{aligned} \sqrt{S}\left(\widehat{\theta} - \theta\right) &= \sqrt{S}\left(\mathbf{A}'\mathbf{B}^{-1}\mathbf{A}\right)^{-1} \mathbf{A}'\mathbf{B}^{-1}\mathbf{Z}'\left[Y - \mathbf{W}(\widehat{p})\theta\right] \\ &= \left(A_0' B_0^{-1} A_0\right)^{-1} A_0' B_0^{-1} \tfrac{1}{\sqrt{S}} \mathbf{Z}'\left[Y - \mathbf{W}(\widehat{p})\theta\right] + o_p(1), \end{aligned} \tag{21}$$

where the second "=" holds since $\mathbf{A}/S \xrightarrow{p} A_0$, $\mathbf{B}/S \xrightarrow{p} B_0$ and $\frac{1}{\sqrt{S}}\mathbf{Z}'\left[Y - \mathbf{W}(\widehat{p})\theta\right] = O_p(1)$.

Recall the definition from the text:

$$F_0 \equiv E\left[Z_s' \triangledown W_s(p)\theta\right] = \tfrac{\lambda}{(1-p)^2} Z_s' H_s y_s, \text{ from } \triangledown W_s(p) \equiv \tfrac{dW_s(\tilde{p})}{d\tilde{p}}|_{\tilde{p}=p} = \left(\tfrac{H_s y_s}{(1-p)^2}, 0\right).$$

Let $\nabla \mathbf{W}(p)$ be nS-by-$(K+1)$ matrix that stacks $\triangledown W_s(p)$ over $s \leq S$. Then,

$$\begin{aligned} \tfrac{1}{\sqrt{S}}\mathbf{Z}'\left(Y - \mathbf{W}(\widehat{p})\theta\right) &= \tfrac{1}{\sqrt{S}}\mathbf{Z}'\left(Y - \mathbf{W}(p)\theta\right) - \left(\tfrac{1}{S}\mathbf{Z}'\triangledown\mathbf{W}(p)\theta\right)\sqrt{S}(\widehat{p} - p) + o_p(1) \\ &= \tfrac{1}{\sqrt{S}} \sum\nolimits_s Z_s'\left(y_s - W_s(p)\theta\right) - F_0\left(\tfrac{1}{\sqrt{S}} \sum\nolimits_s \tau_s\right) + o_p(1) \\ &= \tfrac{1}{\sqrt{S}} \sum\nolimits_s \underbrace{\left(Z_s' v_s - F_0 \tau_s\right)}_{\xi_s} + o_p(1). \end{aligned} \tag{22}$$

The first equality follows form a Taylor approximation around the true missing rate p; the second from $\left(\frac{1}{S}\mathbf{Z}'\triangledown\mathbf{W}(p)\theta\right) \xrightarrow{p} E[Z_s'\triangledown W_s(p)\theta]$ and from the asymptotic linear representation of the estimator $\widehat{p}$; the third from $y_s = W_s(p)\theta + v_s$. This proves the claim of limiting distribution of $\sqrt{S}(\widehat{\theta} - \theta)$ in the text.