Estimating Spillover Effects from Sampled Connections*

Kieran Marray[1]

[1]Vrije Universiteit Amsterdam and Tinbergen Institute

October 2025

(Link to most recent version)

Abstract

Empirical researchers often estimate spillover effects by fitting linear or non-linear regression models using sampled network data. We show that common sampling schemes bias these estimates, potentially upwards or downwards, and derive biased-corrected estimators that researchers can construct from aggregate network statistics. Our results apply under different assumptions on the relationship between observed and unobserved links, allow researchers to bound true effect sizes, and to determine robustness to mismeasured links. As an application, we estimate the propagation of climate shocks between U.S. public firms from self-reported supply links, building a new dataset of county-level incidence of large climate shocks. Corrected estimates are half the size of standard regression estimates.

Keywords— Networks, Sampling, Peer Effects
JEL Codes: C21

1 Introduction

Empirical researchers measuring spillover effects often observe networks imperfectly, sampling either too few or too many links between individuals (Newman, 2010). In economics of education

*We thank Stanislav Avdeev, Vasco Carvalho, Jonathan Dingel, Alan Griffith, Eyo Herstad, Chih-Sheng Hseih, Max Kasy, François Lafond, Xiaodong Liu, Jos van Ommeren, Xun Tang, Sander de Vries, Lina Zhang, and participants at the Network Science in Economics Conference 2025, European Summer Meeting of the Econometric Society 2024, University of Warwick, Vrije Universiteit Amsterdam and Tinbergen Institute for comments. The author also thanks the Smith School of Enterprise and the Environment at the University of Oxford for hospitality while preparing initial parts of this draft. The R package `spillest` implements estimators in the paper. The usual disclaimer applies.

and development economics, for instance, researchers often collect network data through surveys that ask subjects to name up to a certain number of friends or contacts (e.g Rapoport and Horvath, 1961; Harris, 2009; Calvó-Armengol et al., 2009; Conley and Udry, 2010; Oster and Thornton, 2012; Banerjee et al., 2013; Shakya et al., 2017). In industrial organisation and economics of innovation, technological similarity or geographic proximity are often used to proxy firm connections (e.g Jaffe, 1986; Foster and Rosenzweig, 1995; Bloom et al., 2013). When studying production networks, researchers often observe only large supply relationships between firms (e.g see Atalay et al., 2011; Barrot and Sauvagnat, 2016) or payments recorded by a specific bank or credit rating agency (e.g Carvalho et al., 2020).[1] To illustrate the prevalence of this, we surveyed articles published in the American Economic Review, Econometrica, or Quarterly Journal of Economics from January 2020 to September 2024. Out of the 30 papers measuring spillover effects, 21 (70%) sample the network imperfectly.

A common empirical strategy for estimating spillover effects from some treatment is to regress outcomes on the (weighted) sum of treatments of sampled neighbours, given that treatment is independent of the strength of links between individuals. Examples include randomised controlled trials on networks, natural experiments, and other design-based estimation strategies that are increasingly prevalent in applied research (Borusyak et al., 2024). Our first main result is to derive a simple closed-form expression showing how these estimates of spillover effects are biased, and that the direction and magnitude of the bias is determined in predictable ways by the sampling scheme. Estimates are biased because standard sampling schemes create an omitted variable – the (weighted) sum of treatments of unobserved neighbours – that covaries with the (weighted) sum of treatments of sampled neighbours because of how links are sampled, and that affects outcomes directly.

One important consequence is that, unlike attenuation bias from classical measurement error, estimates can be biased upward or downward. Furthermore, simulations suggest these that biases can be economically significant in both directions. For example, applying the sampling rule for females friends from the popular National Longitudinal Adolescent Health Data Set (Harris, 2009) to simulated networks leads to estimates that are over one and a half times the true spillover effects on average. We further show that a sufficient condition for sampling links to cause bias is that all links have the same sign, and that the researcher systematically samples too few or too many links – typical in social and economic network datasets (Harris, 2009; Banerjee et al., 2013; Barrot and Sauvagnat, 2016).

To address this issue, our second main result is a simple bias-correction for estimates from linear and non-linear regression models, which only depends upon average numbers of missing links. This is useful because collecting exact network data is very difficult in practice (Newman, 2010; Beaman et al., 2021), but researchers can collect or estimate average numbers of missing links relatively easily. For example, they can including a single additional survey question (e.g.,

[1]Other examples of researchers using proxies for links between individuals include in estimates of neighbourhood spillovers in crime (Glaeser et al., 1996), the role of social networks in labour markets (Munshi, 2003; Beaman, 2011), and the effect of deworming on educational outcomes (Miguel and Kremer, 2004).

"How many friends do you have?") or use external datasets that survey comparable networks more comprehensively (e.g see Jackson et al. (2022), Bacilieri et al. (2023)). The bias-corrected estimators are consistent, asymptotically normally distributed, and perform well in simulation under common sampling rules where the uncorrected estimators are severely biased. We further show how researchers can construct standard errors accounting for the uncertainty in the necessary network statistics using a bootstrap. When the necessary network statistics are not observable, we show how researchers can bound spillover effects and assess robustness of estimates to measurement error instead. Furthermore, we demonstrate how researchers can extend our results to cases where treatment assignment depends upon the network structure by modeling the dependence between treatment assignment and network structure using a copula. This is useful for researchers evaluating observational data where treatment might have been targeted to a maximise a network-based outcome, or individuals may have re-adjusted links in response to treatment.

As an application, we estimate how large climate shocks propagate between U.S. public firms using a popular dataset containing self-reported supply links (Atalay et al., 2011). As firms are only mandated to report customers making up more than 10% of sales, the dataset under-samples their supply relationships. We combine the dataset with a newly constructed county-level measure of exposure to large weather shocks, and then estimate spillover effects correcting for sampling bias using network statistics from Bacilieri et al. (2023); Herskovic et al. (2020).

We find that how links between firms are sampled biases estimates upwards. Corrected estimates are half the size of standard regression estimates. Consequently, the indirect effects of climate change though firm supply chains are lower than a conventional regression analysis would suggest. In the appendix, we also show that undersampling study partnerships between high and low-ability students can help account for differences between estimated and realised peer effects in Carrell et al. (2013).

Our paper relates to a large literature in non-classical measurement error in econometrics in general (e.g see Heckman, 1979; Bound et al., 2001; Oster, 2019), and contributes to the nascent literature on the effect of misspecification in network econometrics in particular (Chandrasekhar and Lewis, 2016; Griffith, 2022; Lewbel et al., 2023; Yauck, 2022; Zhang, 2023; Hseih et al., 2024; Griffith and Kim, 2024; Boucher and Houndetoungan, 2025). Our main contribution is to focus on the common setting where treatment is distributed independently from link strength, and show how researchers can construct bias-corrected estimators in these settings under much weaker assumptions than in the existing literature. Our estimator does not require researchers to impute the missing links using a parametric model of network formation (as in the aggregate-relational data approach of Breza et al., 2020), impose parametric assumptions about the counterfactual distributions of links (as in Boucher and Houndetoungan, 2025; Herstad, 2023), assume constant link-missingness rates (as in Lewbel et al., 2025), or drop large parts of their sample that might contain incorrectly sampled links (as in Chandrasekhar and Lewis, 2016). By contrast, our bias-corrected estimators depend only upon quantities that researchers can directly observe. Our bias correction results are similar to those derived by Griffith (2022)

for the specific case of fixed choice designs, which he analyses in detail. Our idea of using additional network data is similar to Lewbel et al. (2023); Zhang (2023), but does not require researchers to collect an entire different measure of the network.

By deriving a simpler expression for estimator bias, our results also give applied researchers clearer guidance on when link sampling leads their estimates to be biased, the sign of the bias, and the expected magnitude of the bias. These are ambiguous from existing results (Chandrasekhar and Lewis, 2016; Breza et al., 2020; Boucher and Houndetoungan, 2025). Our bias function further allow researchers to easily tell when inclusion of control variables can reduce bias, which we discuss in Appendix A3 – another point that is unclear in existing results.

Our results are also closely related to the literature on design based estimation using linear combinations of exposures to exogenous shocks (Borusyak and Hull, 2023; Borusyak et al., 2024). Again, our approach differs by not requiring researchers to specify a counterfactual distribution of exposure to exogenous shocks to correct for bias.

A common strategy to address the mismeasurement of links is to first impute missing links before then using full and imputed links to estimate spillover effects. A related literature on unobserved networks seeks to estimate these missing links between individuals (e.g see Manresa, 2013; Lam and Souza, 2019; Battaglini et al., 2021; Higgins and Martellosio, 2023; Lewbel et al., 2023; Rose, 2023; De Paula et al., 2024; Griffith and Kim, 2024; Marray, 2025). But these estimators require much richer data - typically a short panel of individual outcomes - and stronger structural assumptions on the data-generating process than our approach (e.g. Battaglini et al., 2021; De Paula et al., 2024). Moreover, measurement error in the estimated network may itself bias regression estimates of spillover effects.

We proceed as follows. In Section 2, we characterise the effect of sampling links on linear regression estimates of spillover effects, and present bias-corrected estimators. Section 3 extends our results to common non-linear models, and 4 to cases when treatment depends on network structure. In Section 5, we assess performance of estimators by simulation. Finally, Section 6 presents our empirical examples. Proofs and additional results are provided in the appendix.

2 Theory for linear models

Here, we develop an econometric framework for estimating spillover effects from sampled links when outcomes are linear in the (weighted) sum of neighbours' treatments (spillovers). We present two main sets of results: on estimator bias, and on bias-correction.

First, in Section 2.1, we derive a tractable closed-form expression for the bias from sampling links in linear regression estimators. From the expression, a researcher can easily determine whether sampling will bias their estimate, and if so whether it is biased upwards or downwards. We then derive a sufficient condition for bias that covers most common sampling schemes in economic and social network research when the underlying network is binary, and illustrate this in Section 2.2 with extended examples of the three most common sampling schemes.

Second, we use the expression to construct bias-corrected estimators for spillover effects from sampled network data that only require knowledge of aggregate network statistics – not an

estimate of where the links are missing (Breza et al., 2020), multiple measures of the network (Lewbel et al., 2023), or knowledge of counterfactual network generating distributions (Herstad, 2023). The estimators require an assumption on the relationship between the observed and unobserved number of links, which will vary depending on the sampling rule that a researcher uses. So, in Section 2.3, we derive these estimators under two assumptions that cover the common set of sampling rules used in applied research. Furthermore, we derive the asymptotic distribution of these estimators, present a bootstrap estimator for the variance, and show how researchers can use the results to assess robustness of estimators to sampling links.

2.1 Setup

Let there be $\mathcal{N} = \{1, ..., N\}$ individuals situated on a simple network $\mathcal{G} = (\mathcal{N}, \mathcal{E}^{\mathcal{G}}, \mathcal{W}^{G})$ with edges $\mathcal{E}^{\mathcal{G}}$ and weights $\mathcal{W}^{G}$. We can represent these relationships with the $N \times N$ adjacency matrix G, where elements $g_{ij} \in \{0, 1\}$ if links are unweighted and $g_{ij} \in \mathbb{R}$ if links are weighted. Define the degree of individual i as $d_i = \sum_j g_{ij}$ the (possibly weighted) number of connections from all other individuals to i.

Instead of observing the true network, the researcher samples a set of edges $\mathcal{E}^{H}$ and weights $\mathcal{W}^{H}$ between individuals in $\mathcal{N}$ through the non-stochastic sampling rule $S : (\mathcal{E}^{G}, \mathcal{W}^{G}) \to (\mathcal{E}^{H}, \mathcal{W}^{H})$ such that $\mathcal{E}^{H} \cap \mathcal{E}^{G} \neq \emptyset$. We can split the adjacency matrix of the true network into the sampled part H and an unsampled part B that encodes the network of incorrectly sampled links

$$G = H + B. \tag{1}$$

This decomposition is straightforward yet useful, as it yields a simple closed-form expression for estimator bias.

Let $\mathcal{B}$ denote the set of nodes with at least one (incoming) link sampled incorrectly.[2] Further, define the sampled degree of node i – the total (weighted) number of sampled connections from all other individuals to i – as $d_i^H = \sum_j h_{ij}$, and the unobserved degree of node i – the total (weighted) number of connections from all other individuals to i that are not sampled – as $d_i^B = \sum_j g_{ij} - \sum_j h_{ij}$.

Consider the problem of estimating the causal effect or structural parameter β – the 'spillover effect' of an additional neighbour being treated on outcomes – in the model

$$y_i = \beta \sum_j g_{ij} x_j + \epsilon_i. \tag{2}$$

Outcomes y_i are linear in the (weighted) sum of treatment x_i of neighbours on the network (we refer to this sum as 'spillovers').[3] We assume treatment is independently and identically distributed across nodes, and distributed independently of link strength.

[2] Equivalently, $\mathcal{B}$ is the index set of rows of B with at least one non-zero entry.

[3] Formally, $((g_{ij})_{j=1}^{N}, x_i, \epsilon_i)_{i=1}^{N}$ can be described with some joint distribution that we do not restrict here.

Assumption 1. Distribution of treatment x_i.

A: $x_i \sim$ i.i.d. F_X – treatment is drawn i.i.d. from a common distribution,

B: $x_j \perp\!\!\!\perp g_{ij}, h_{ij}\ \forall i, j \in \mathcal{N}$ – treatment is distributed independently of true and sampled link strength.

Further, assume that (2) is specified correctly

Assumption 2. Distribution of structural shocks. $E(\epsilon_i) = 0$, $\sum_j g_{ij}x_j, \sum_j h_{ij}x_j \perp\!\!\!\perp \epsilon_i$.

and that the expectation of the square of observed spillovers is finite.[4]

Assumption 3. Finite second moment of observed spillovers. $E((\sum_j h_{ij}x_j)^2) < \infty$.

The researcher only observes the (weighted) sum of treatments of sampled neighbours

$$\sum_j h_{ij}x_j = \begin{cases} \sum_j g_{ij}x_j & \text{if } i \notin \mathcal{B}, \\ \sum_j g_{ij}x_j - \sum_j b_{ij}x_j & \text{if } i \in \mathcal{B}, \end{cases} \tag{3}$$

as opposed to those of the true neighbours. Suppose they estimate spillover effects using their sampled analogue of the data generating process – regressing outcomes on sampled spillovers:

$$y_i = \beta \sum_j h_{ij}x_j + \xi_i. \tag{4}$$

This regression model is misspecified. Using our decomposition in (1), we can express the misspecification in a very simple form:

$$\xi_i = \beta \sum_j b_{ij}x_j + \epsilon_i.$$

Sampling links inadvertently creates an omitted variable – spillovers on unobserved links – that enters the error term. Thus, we see that the linear regression estimator is biased, and with the familiar bias function (MacKinnon and Smith, 1998)

Proposition 1. Make Assumptions 1-A, 2, 3. The ordinary least-squares estimator for (4) $\hat{\beta}^{\text{OLS}}$ is biased, with bias function

$$\hat{\beta}^{\text{OLS}} = \beta\Big(1 + \frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) + \frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)\epsilon_i}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}.$$

This bias function tells us three things. First, sampling biases the regression estimator if observed spillovers covary with sampled spillovers due to the sampling rule,

$$\begin{aligned} &E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) \neq 0 \\ \implies &E(\hat{\beta}^{\text{OLS}}) = \beta\Big(1 + E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)\Big) \neq \beta \end{aligned} \tag{5}$$

[4] In the general case with an intercept, controls etc in Appendix A.3, this is the familiar assumption that regressors have finite variance (Cameron and Trivedi, 2005).

even though treatment is assigned independently across individuals as in a controlled or natural experiment.

Second, unlike classical measurement error where estimates are always attenuated, sampling can bias estimates upwards or downwards depending on whether the dependence between observed and unobserved spillovers is positive or negative.[5] Indeed, the sign of the bias is knowable in advance in many cases. The estimator is upwards biased if $E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)) > 0$, and downwards biased if $E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)) < 0$.

Third, all we need to model to obtain unbiased and consistent estimates of spillover effect is the dependence between observed and unobserved spillovers $E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)$. The researcher does not necessarily need to know where the missing links are, to impute the missing links (Breza et al., 2020), or to know anything else about the distribution of unobserved links (Borusyak and Hull, 2023; Herstad, 2023; Boucher and Houndetoungan, 2025).

Some discussion of our assumptions is warranted before proceeding. We consider the simplest data generating process here without loss of generality. Results apply to functional forms including an intercept, controls, and panel data (see Appendix A.3), as well as alternative specifications where researchers construct a dummy variable for at least one neighbour being treated (e.g Barrot and Sauvagnat, 2016, see Appendix A.5), and non-linear social network models in Section 3.

Assumption 1-B is strong assumption, but commonly satisfied in applied research. It corresponds to experiments where the researcher directly assigns treatment (e.g Miguel and Kremer, 2004; Oster and Thornton, 2012; Conley and Udry, 2010), the increasingly popular 'design-based' estimation strategies that leverage different unit-level exposures to exogenous shocks (e.g Barrot and Sauvagnat, 2016; Carvalho et al., 2020; Borusyak et al., 2024), or cases where treatment and network formation are determined by different processes (e.g Coleman et al., 1957; Calvó-Armengol et al., 2009). In Section 4, we discuss how to extend our results when Assumption 1-B does not hold, such as when treatment is targeted by a planner based on network structure or individuals endogenously adjust links based on treatment.

2.2 Motivating examples

To fix ideas, consider the three following examples that cover common sampling schemes used to study economic and social networks.

Example – classroom setting, fixed choice sampling rule. Suppose the individuals are children in a classroom. Links denote friendships between children, and any weights may denote time spent together. A child's degree is their number of friends. To collect network data, the researcher asks each individual to name at most m friends. This sampling rule is commonly used to collect network data through surveys (Coleman et al., 1957; Calvó-Armengol et al., 2009;

[5] This may seem to be a trivial point. But it is known for applied researchers to assume that errors sampling links will attenuate estimates like classical measurement error (for an example, see Oster and Thornton, 2012).

Oster and Thornton, 2012; Banerjee et al., 2013; Shakya et al., 2017).[6]

If a child has fewer than m friends, the researcher samples all of their possible friendships correctly. If a child has more than m friends, some are missed, and only m links are recorded ($i \in \mathcal{B}$ if $d_i > m$). Therefore, the spillovers they sample are

$$\sum_j h_{ij}x_j = \begin{cases} \sum_j g_{ij}x_j & \text{if } d_i \leq m \\ \sum_j g_{ij}x_j - \sum_j b_{ij}x_j & \text{otherwise} \end{cases}$$

– equal to true spillovers for children with fewer than m friends, but different for children with more than m friends. In expectation, the difference weakly increases with the number of friendships the child has.

The researcher estimates spillover effects by regressing each child's outcomes on their sampled spillovers. The error term ξ_i contains the treatments of additional friends of children with more than m friends. As this number is higher the more friends a child has over m, it covaries positively with sampled spillovers. Therefore the spillover estimate is upward biased.

Example – village setting, sampling based on group membership. Suppose the individuals are villagers across a set of villages. Links denote borrowing relationships between villagers, and any weights may denote amount lent to each other (as in Banerjee et al., 2013). A villager's degree is their number of individuals they have lent to (or total amount lent, if weighted). To collect network data, the researchers assume that all individuals within the same village lend to each other. This is common in observational data where researchers can tell which types of individuals might be connected, but not the exact connections (e.g Chetty et al., 2011; Bloom et al., 2013; Carrell et al., 2013).

Consider a case where villages are sized m_i, so the researchers assume that each individual is connected to the m_i others in their village. This adds links for villagers who have lent to fewer than m_i others ($i \in \mathcal{B}$ if $d_i < m_i$). Therefore, sampled spillovers are

$$\sum_j h_{ij}x_j = \begin{cases} \sum_j g_{ij}x_j & \text{if } d_i = m_i \\ \sum_j g_{ij}x_j - \sum_j b_{ij}x_j & \text{otherwise} \end{cases}$$

– equal to true spillovers for the villagers who lend to all m_i others in the village, but more than true spillovers for villagers who have not. In expectation, the difference weakly decreases with the number of links the villager has.

The researcher estimates spillover effects by regressing each individual's outcome on the treatments of all others in their village. The error term ξ_i subtracts the sum of treatments of the individuals in the village that each individual does not lend to. The more individuals they actually lend to, the closer this number is to zero, so it covaries negatively with sampled spillovers. Therefore the spillover estimate is downward biased.

[6] All comments about fixed choice designs also apply to increasingly common cases where researchers ask individuals to name up to an unspecified number of friends but individuals with more friends are less likely to name all of their friends due to time constraints or effort filling out the survey.

Example – firm supply network, high-weight links. Suppose the individuals are firms. Links denote supply relationships, and weights denote the proportion of total sales that goes to that firm. To collect network data, researchers take the links where weights are greater than a threshold – the 'most important' connections. This is common in observational data where individuals must disclose important interactions (Atalay et al., 2011; Barrot and Sauvagnat, 2016). For example, US publicly listed firms must disclose customers that make up at least 10% of their sales to the Securities and Exchange Commission.

Consider a case where researchers only sample links above some weight τ. Unless all firm supply relationships have weight greater than τ, researchers sample fewer links to firms than they actually have. Therefore, sampled spillovers are

$$\sum_j h_{ij}x_j = \begin{cases} \sum_j g_{ij}x_j & \text{if } g_{ij} > \tau \ \forall j \\ \sum_j g_{ij}x_j - \sum_j b_{ij}x_j & \text{otherwise.} \end{cases}$$

The researcher estimates spillover effects by regressing outcomes on treatments of customers. The error term contains the sum of treatments of additional customers of firms with link weights less than the threshold. Under the standard distribution used to model firm sales, this will co-vary positively with observed sales (Herskovic et al., 2020).[7] Therefore the spillover estimate is upward biased.

An obvious next question is when sampling schemes induce dependence between spillovers on sampled and unobserved links. In other words, when does sampling link lead to biased spillover estimates in general? Suppose all links on the network have the same sign, Assumption 1 holds, and expected treatment is non-zero. A sufficient condition is that the expected number of unobserved links of each individual has the same sign – so the researcher either samples a subset or superset of the true links.

Proposition 2. Make Assumption 1-A, 1-B. Further, assume that all links on the network have the same sign – either $g_{ij} \geq 0$ or $g_{ij} \leq 0 \ \forall j$ – and that $E(x) \neq 0$. Then if the expectation of unobserved degree has the same sign for all nodes with potentially unsampled links

$$E(d_i^B | d_i^H) \geq 0 \ \ \forall i \in \mathcal{B} \text{ or } E(d_i^B | d_i^H) \leq 0 \ \ \forall i \in \mathcal{B}$$

and is non-zero for at least one i, then

$$E\Big(\frac{1}{N}\sum_i (\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)\Big) \neq 0.$$

This covers the sampling schemes commonly used to study economic and social networks, including the examples of fixed choice designs, group membership, and sampling high weight links discussed below. Many social and economic networks have links with all positive or all negative signs, such as firm-level production networks (Atalay et al., 2011), information sharing networks (Banerjee et al., 2013), and friendship networks (Calvó-Armengol et al., 2009). For intuition, we give an extended example with a fixed choice design in Appendix A.2.

[7] See the corresponding simulation in section 5.

2.3 Bias-corrected estimators

Recall our bias function (MacKinnon and Smith, 1998)

$$\hat{\beta}^{\text{OLS}} = \beta + \beta \frac{\frac{1}{N}\sum_i (\sum_j h_{ij} x_j)(\sum_j b_{ij} x_j)}{\frac{1}{N}\sum_i (\sum_j h_{ij} x_j)^2} + \frac{\frac{1}{N}\sum_i (\sum_j h_{ij} x_j)\epsilon_i}{\frac{1}{N}\sum_i (\sum_j h_{ij} x_j)^2}.$$

Taking expectations and solving for β gives us a bias-corrected estimator[8]

$$\hat{\beta} = \frac{\hat{\beta}^{\text{OLS}}}{1+\eta},$$

where

$$\eta = E\Big(\frac{\frac{1}{N}\sum_i (\sum_j h_{ij} x_j)(\sum_j b_{ij} x_j)}{\frac{1}{N}\sum_i (\sum_j h_{ij} x_j)^2}\Big).$$

In words, the researcher needs to rescale their estimate of the spillover effect to adjust for the dependence between sampled spillovers and unobserved spillovers induced by the sampling rule. In practice, the researcher needs to compute[9]

$$\hat{\beta} = \frac{\hat{\beta}^{\text{OLS}}}{1+\hat{\eta}}$$

where $\hat{\eta}$ is an estimate of η. Where shocks are distributed independently of links (Assumption 1-B), the researcher can construct a good analytic approximation of $\hat{\eta}$ from aggregate statistics of the degree distribution and expected treatment. Taking the Taylor expansion of η around the mean observed and unobserved spillovers gives (Billingsley, 2012)

$$\begin{aligned} \eta &= \frac{E(\frac{1}{N}\sum_i (\sum_j h_{ij} x_j)(\sum_j b_{ij} x_j))}{E(\frac{1}{N}\sum_i (\sum_j h_{ij} x_j)^2)} + \mathcal{O}(\frac{1}{\frac{1}{N}\sum_i (\sum_j h_{ij} x_j)^4}), \\ &\approx \frac{E(\frac{1}{N}\sum_i (\sum_j h_{ij} x_j)(\sum_j b_{ij} x_j))}{E(\frac{1}{N}\sum_i (\sum_j h_{ij} x_j)^2)}. \end{aligned} \tag{6}$$

where the remainder term $\mathcal{O}(\frac{1}{\frac{1}{N}\sum_i (\sum_j h_{ij} x_j)^4})$ is negligible.[10]

[8] This approach is equivalent to controlling for the expected unsampled spillovers amongst nodes that have at least some incorrectly sampled links

$$z_i = \begin{cases} 0 & \text{if } i \notin \mathcal{B}, \\ E(\sum_j h_{ij} x_j | i \in \mathcal{B}) & \text{if } i \in \mathcal{B}. \end{cases}$$

But it does not require knowing which nodes have some incorrectly sampled links, just how many. In many cases – such as the group membership and high-weight link examples – the researcher does not know which nodes have some incorrectly sampled links. Thus, we consider the bias-corrected estimator instead.

[9] In some cases, the researcher may be able to observe η directly. For example, in population-level datasets where some links are distorted in order to preserve privacy, a data provider could in principle disclose η alongside the data. But in most cases, this dependence will be unobserved.

[10] In cases where the researcher is worried it might not be, we show how researchers can apply a higher order expansion or approximate the full expectation by simulation in Appendix A2.

Th researcher can estimate this expression from aggregate statistics of the degree distribution plus the sampled network data and expected treatment status under the assumption that treatments are distributed independently from links, plus an appropriate assumption on the dependence of sampled and unobserved degree for their sampling scheme.

First, assume that the distribution of observed degree is independent of the distribution of unobserved degree amongst nodes with some potentially incorrectly sampled links. This assumption applies to common sampling schemes used by applied economists when the underlying network is unweighted.

Assumption 4.a. Distribution of unsampled degree – $d_i^B \perp\!\!\!\perp d_i^H | i \in \mathcal{B}$.

For illustration, consider the following examples.

Example – classroom setting, fixed choice design with binary network. If there are potentially unsampled friendships to a child $i \in \mathcal{B}$, we know that the sampled (in)degree equals the threshold value $d_i^H = m$ i.e they have m friends. Therefore, the distribution of sampled degrees d_i^H given that $i \in \mathcal{B}$ has a point mass at m. All children with some unsampled friendships have m sampled friends. It follows that the distribution of the number of unsampled links d_i^B is independent of the distribution of sampled links amongst individuals where $i \in \mathcal{B}$.

Example – village setting, group membership with binary network. Assume for simplicity that all villages have an equal size m. For all i, the number of sampled neighbours equals one minus the village size $d_i^H = m - 1$ by construction. Therefore, the distribution of the degree d_i^H given that $i \in \mathcal{B}$ has a point mass at $m - 1$. It follows that the distribution of the unsampled degree d_i^B is independent of the distribution of sampled links amongst individuals where $i \in \mathcal{B}$.[11]

If no links differ in strength, we need not worry that subjects report links in an order that might violate this assumption.

In this case, we can construct $\hat{\eta}$ in terms of the mean sampled degree of nodes that have at least one potentially unsampled link, the mean missing degree of nodes that have at least one potentially unsampled link, and the expected treatment status of each node, defined as

$$\hat{d}^H = \frac{1}{\sum_{i \in \mathcal{B}} 1_i} \sum_{i \in \mathcal{B}_i, j} h_{ij}, \quad \hat{d}^B = \frac{1}{\sum_{i \in \mathcal{B}} 1_i} \sum_{i \in \mathcal{B}, j} b_{ij},$$
$$\bar{x} = \frac{1}{N} \sum_i x_i, \qquad N^B = |\mathcal{B}|.$$

The resulting bias-corrected estimator is as follows.

Proposition 3. Make Assumptions 1-A, 1-B 2, 3, 4.a. Consider the estimator

$$\hat{\beta} = \frac{\hat{\beta}^{\text{OLS}}}{1 + \hat{\eta}} \text{ where } \hat{\eta} = \frac{\frac{N^B}{N} \hat{d}^H \hat{d}^B \bar{x}^2}{\frac{1}{N} \sum_i (\sum_j h_{ij} x_j)^2} \tag{7}$$

[11] This argument extends to the case where the size of the group varies across some groups, as long as the degree of each individual within each group does not itself depend on the size of the group.

Let $\hat{\beta}_N$ denote an estimate from sample size N. $E(\hat{\beta}_N) \approx \beta$ and $\hat{\beta} \xrightarrow{p} \beta$.

The rescaling factor $\hat{\eta}$ depends only on two aggregate network statistics – the sampled mean degree and true mean degree of individuals who have at least one potentially unsampled link. It does not require knowing which specific links are missing.

Second, we can relax Assumption 4.a and allow the number or strength of unobserved links to depend on the number or strength of observed links. For example, individuals may name stronger connections first in a survey. Instead, we can adopt the weak assumption that we can model observed and unobserved link counts or strengths using a common conditional distribution.

Assumption 4.b. There exists a joint distribution over $(d_i^H, d_i)_{i\in\mathcal{B}}$ such that we can write $E(d_i^B|d_i^H) = E(d_i|d^H = d_i^H) - d_i^H \ \forall i \in \mathcal{B}$.

Example – classroom setting, fixed choice design naming stronger connections first. Assume that weighted degrees (number of friends) are drawn from a common degree distribution F_d, and for simplicity assume that all weights are positive. The researcher samples up to m friendships per child. Children list their strongest friendships first. The strength of each unobserved friendship must be less than or equal to the lowest strength of the observed friendships. Otherwise, the child would have named that friendship earlier. So, for children with at least one potentially missing friendship, $0 \leq d_i - d_i^H \leq (N-m)\min\{h_{ij}|h_{ij} > 0\}$. Accordingly, $E(d_i^B|d_i^H) = E(d_i|d_i \geq (N-m)\min\{h_{ij}|h_{ij} > 0\}) - d_i^H$.

In this case, we can construct the bias-corrected estimator using the approximation $\hat{\eta}$ in terms of each individual's sampled degree, the average number of unobserved links for individuals with given sampled degree,

$$\hat{d}^B(d^H) = \frac{1}{\sum_{i\in\mathbf{B}} \mathbf{1}(d_i^H = d^H)} \sum_{i\in\mathcal{B}} d_i^B \mathbf{1}(d_i^H = d^H),$$

and the expected treatment status. The resulting bias corrected estimator is as follows.

Proposition 4. Make Assumptions 1-A, 1-B, 2, 3, 4.b. Consider the estimator

$$\hat{\beta} = \frac{\hat{\beta}^{\text{OLS}}}{1+\hat{\eta}} \text{ where } \hat{\eta} = \frac{\frac{1}{N}\sum_{i\in\mathcal{B}} d_i^H \hat{d}^B(d_i^H)\bar{x}^2}{\frac{1}{N}\sum_i (\sum_j h_{ij}x_j)^2}$$

Letting $\hat{\beta}_N$ denote an estimate from a sample of size N, $E(\hat{\beta}_N) \approx \beta$ and $\hat{\beta} \xrightarrow{p} \beta$.

These estimators have an asymptotically normal distribution, with a variance that depends upon the estimates of mean missing degrees. Denote these as a vector θ defined as the solution to the moment conditions

$$\theta - \frac{1}{N}\sum_{i=1}^{N} \theta_i = 0.$$

Our estimators are sequential estimators, meaning the asymptotic distribution of the spillover estimate depends upon both the uncertainty in the estimates of θ and the sensitivity of $\hat{\eta}(\hat{\theta})$ to $\hat{\theta}$ (Newey, 1984).

Proposition 5. Make Assumptions 1A ,2,3, 5, A1. Define

$$\begin{pmatrix} h_1(\theta) \\ h_2(\theta,\beta) \end{pmatrix} = \begin{pmatrix} \theta - \frac{1}{N}\sum_{i=1}^{N}\theta_i \\ \frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(y_i - (1+\eta(\theta))\beta(\sum_j h_{ij}x_j)) \end{pmatrix}$$

$$\begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} = \text{plim}\,\frac{1}{N}\sum_i E\begin{pmatrix} -1 & 0 \\ -\frac{\partial\eta(\theta)}{\partial\theta}\beta(\sum_j h_{ij}x_j)^2 & -(1+\eta(\theta))(\sum_j h_{ij}x_j)^2 \end{pmatrix}$$

$$\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \text{plim}\,\frac{1}{N}\sum_i E\begin{pmatrix} h_{1i}h_{1i}' & h_{2i}h_{1i}' \\ h_{2i}h_{1i}' & h_{2i}h_{2i}' \end{pmatrix}$$

$\hat{\beta}$ is a consistent estimator of β. Furthermore,

$$\sqrt{N}(\hat{\beta}-\beta) \sim N(0, K_{22}^{-1}(S_{22} + K_{21}K_{11}^{-1}S_{11}K_{11}^{-1}K_{21}' - K_{21}K_{11}^{-1}S_{12} - S_{21}K_{11}^{-1}K_{21}')K_{22}^{-1}).$$

In practice, we propose using a bootstrap to estimate the variance of the estimator. For example, assume that we are computing $\hat{\eta}$ under Assumption 4.a. In the first step, we simulate P different possible unobserved graphs consistent with the same missing degree. In the absence of any link function that determines how likely any two individuals are to be connected given that their links are not sampled correctly, we assume that incorrectly observed links are distributed uniformly at random over all possible missing entries in B. In the second step, we construct M bootstrap estimates of $\hat{\beta}$ for each B. Similar bootstrap estimators can be derived under alternative assumptions on the network sampling process.

Algorithm 1 Bootstrap estimator for $\hat{s}(\hat{\beta})$ under 4.a

procedure BOOTSTRAP (d^B, H, $\{\mathcal{E}_i^{\mathcal{H}}\}_{i=1}^N$, x, y)
  for $j \in 1, ..., M$ **do**
    Draw $\{B_{ik} | k \notin \mathcal{E}_i^{\mathcal{H}}\} s.t \sum_{\{B_{ik}|k\notin\mathcal{E}_i^{\mathcal{H}}\}} B_{ik} = N\bar{d}^B$.
    Construct $\{\hat{\beta_{kj}}\}_{k=1}^P$ by a regression bootstrap from B^j, H, x, y.
  end for
  $\bar{\beta}_{kj} = \frac{1}{MP}\sum_{k,j}\hat{\beta}_{kj}$.
  $\hat{s}(\hat{\beta}) = \sqrt{\frac{1}{MP}\sum_{k,j}(\hat{\beta}_{kj} - \bar{\beta}_{kj})^2}$
end procedure

A main benefit of our approach over existing estimators is that the estimators depend only upon aggregate network statistics researchers can sample empirically or collect from networks with similar properties. This means that researchers can construct consistent and approximately unbiased estimates under relatively mild assumptions compared to conditioning on unobservable counterfactual networks (Breza et al., 2020; Herstad, 2023; Borusyak and Hull, 2023) or constructing multiple entire network measures (Lewbel et al., 2023). In a survey, the researcher can collect the data required to implement the estimator by including one more question: 'How many of these types of connections do you have?'. If data providers sample networks to preserve

anonymity, they can disclose these quantities while preserving individual privacy. In cases where the researcher cannot sample individuals in the network – for example when using data collected by others – researchers can use the statistics from similar, better-sampled networks. Additional survey questions could also help estimate the mean missing degree under relatively weak assumptions. For example, a researcher could use the question "How many of your friends smoke?" plus an assumption on the distribution of smokers in the population to recover mean missing degree in a friendship network (closer to the common aggregate relational data approach, but still requiring less stringent assumptions on the network generating process Breza et al., 2020).

A potential concern with using a first-order approximation $\hat{\eta}$ in our bias correction is that the approximated $\hat{\eta}$ will be far from the true η and lead to estimates that are far from the true spillover effect. While a concern in theory, this does not appear to be an issue in practice. To assess this, we carry out extensive simulations in Section 5 and the Appendix on simulated and real networks under the most common sampling schemes, comparing estimates to true spillover values and bias-corrected estimates using the true η. The rescaled estimators perform very well in each case in relatively small sample sizes, and are very close to estimates constructed using the true η.

2.4 Robustness

In addition to constructing bias-corrected estimators, the researcher can use our bias function to assess the robustness of spillover estimates to sampling bias in two ways. First, they can recover the value of η needed to reduce the spillover estimate below some decision threshold τ. Examples include thresholds relevant for optimal policy decisions or values required for test statistics to cross critical values at preferred significance levels.

Proposition 6. Make Assumptions 1-A, 2, 3. Then

$$\begin{aligned} &\beta > \tau \text{ if and only if} \\ &\eta < \frac{\hat{\beta}^{\text{OLS}} - \tau}{\tau}. \end{aligned}$$

Second, if the researcher can bound the dependence of observed and unobserved spillovers $\eta \in [\eta_{\min}, \eta_{\max}]$, then the true spillover effect is bounded as

$$\beta \in \Big[\frac{\hat{\beta}^{\text{ OLS}}}{1+\eta_{\max}}, \frac{\hat{\beta}^{\text{ OLS}}}{1+\eta_{\min}}\Big].$$

Under Assumption 4.a, these results depend only on the mean missing degree among individuals with at least one missing link. In this case, the decision threshold can be rewritten as a function of the mean number of missing links among individuals with at least one missing link:

$$\begin{aligned} &\hat{\beta}^{\text{ OLS}} > \tau \text{ if and only if} \\ &\hat{d}^B < \Big(\frac{\frac{1}{N}\sum_i (\sum_j h_{ij}x_j)^2}{\frac{N^H}{N}\bar{x}^2 \hat{d}^H}\Big)\frac{\hat{\beta}^{\text{ OLS}} - \tau}{\tau}. \end{aligned} \tag{8}$$

Thus, spillover effects exceed a threshold – for example, actually passing the critical threshold to determine if we can reject the hypothesis of zero spillover effects –if and only if the researcher is missing fewer than a certain number of links. Moreover, the bounds depend on the minimum and maximum mean number of missing links

$$\eta_{max} = \eta(\hat{d}^B_{max}),\ \eta_{min} = \eta(\hat{d}^B_{min}).$$

3 Theory for nonlinear social network models

We can also extend our bias-correction approach to models where outcomes are linear in spillovers of indirect neighbours. To show this, we consider a non-linear specification commonly used in research on social networks (Bramoullé et al., 2009; Calvó-Armengol et al., 2009). Here, sampling bias also affects the standard instruments used to account for endogeneity in lagged spillovers. Thus, researchers must both correct instruments and bias-correct the resulting estimates.

3.1 Setup

An alternative model often used to measure spillover effects specifies outcomes as linear in the sum of indirect spillovers across all paths through the network, rather than just direct spillovers (e.g Calvó-Armengol et al., 2009; Carvalho et al., 2020). Formally

$$\begin{aligned} y &= \lambda G y + x\beta + \epsilon \\ &= (I - \lambda G)^{-1}(x\beta + \epsilon). \end{aligned} \tag{9}$$

where $y = (y_1, y_2, ..., y_n)$ is the $N \times 1$ vector of individual outcomes, and $x = (x_1, ..., x_n)$ is the $N \times 1$ vector of treatments.[12] The inverse

$$(I - \lambda G)^{-1} = \sum_{k=1}^{\infty} \gamma^k G^k$$

sums spillovers across all paths of length $k = 1, 2, ...$ through the network.

In this setting, sampling the network generates more complex misspecification than in the linear model. Comparing the true paths of length k to sampled paths of length k using our decomposition (1) gives

$$\begin{aligned} G^k &= (H + B)^k \\ &= H^k + H^{k-1}B + ... + B^k. \end{aligned}$$

[12]Without loss of generality, we focus on the case without contextual effects Gx here for ease. Our results extend to estimates of contextual spillover effects. Then, researchers also need to account for the identification problems raised in Manski (1990); Blume et al. (2015).

True paths include paths through only sampled links, paths through only unobserved links, and paths created by combining sampled and unobserved links. Estimator bias therefore depends on the covariance between treatment transmitted along sampled paths and treatment transmitted along additional paths

A researcher estimates structural parameters β, λ – the effect of treatment on outcomes, and the spillover effect of one individual's outcomes on others' – using the sampled network by fitting

$$y = \lambda Hy + x\beta + \xi. \tag{10}$$

Using our decomposition (1), we see that by sampling the network the researcher creates an omitted variable By that enters the error term

$$\xi = \lambda By + \epsilon.$$

The standard approach is to estimate this model by two-stage least squares, constructing instruments using the treatment of sampled friends of sampled friends (Bramoullé et al., 2009). We focus on this estimator, rather than the maximum-likelihood estimator. We adopt the standard assumptions used for this estimator (Kelejian and Prucha, 1998; Bramoullé et al., 2009; Blume et al., 2015), spelled out in Appendix A.7. Denote our regressors as $z^* = \Big(Gy, x\Big)$, $z = \Big(Hy, x\Big)$. Call $z_B = z^* - z = \Big(By, 0\Big)$, and denote instruments built from the sampled network as $J = H(I-H)^{-1}x = \Big(Hx \quad H^2x \quad ...\Big)$, and the corresponding projection matrix as P_J. The two-stage least squares estimator is

$$\begin{pmatrix} \hat{\lambda}^{\text{ 2SLS}} \\ \hat{\beta}^{\text{ 2SLS}} \end{pmatrix} = (z^{'} P_J z)^{-1} z^{'} P_J y.$$

As with the linear model, this estimator can be biased by sampling.

Proposition 7. Make Assumption 2 and the standard Assumption in A.7. Let P denote a projection matrix, $z = \Big(Gy, x\Big)$, $J = \Big(x, Hx, H'Hx, ...\Big)$. There exist H, B such that the two-stage least-squares estimator

$$\hat{\theta}^{\text{ 2SLS}} = \begin{pmatrix} \hat{\lambda}^{\text{ 2SLS}} \\ \hat{\beta}^{\text{ 2SLS}} \end{pmatrix} = (z^{'} P_J z)^{-1} z^{'} P_J Y.$$

is biased and inconsistent.

To see why, note that the instrument exogeneity condition is

$$\begin{aligned} E(J'\xi) &= E\Big(\Big(x, Hx, H'Hx, ...\Big)^{'} (\lambda By + \epsilon)\Big) \\ &= E\Big(\Big(x, Hx, H'Hx, ...\Big)^{'} (\lambda By)\Big) + E\Big(\Big(x, Hx, H'Hx, ...\Big)^{'} \epsilon)\Big). \end{aligned}$$

The second term is the instrument exogeneity condition if H is the true network. Expanding the first term gives

$$E\Big(\Big(x, Hx, H'Hx, ...\Big)^{'} (\lambda By)\Big) = E\Big((H(I-\lambda H)^{-1}x)(\lambda B(I-\lambda(H+B))^{-1}(x\beta+\epsilon))\Big),$$

the product of spillovers between two individuals on paths only containing sampled links, and spillovers between two individuals on paths containing either unsampled links alone or both sampled links and unsampled links. The estimator fails when these two covary.[13]

3.2 Bias-corrected estimator

We construct bias-corrected estimators using the logic in Section 2. First, we must construct correct instruments for spillovers through the sampled network. Once instruments are correct, the second stage is a linear regression of instrumented spillovers on the sampled network on outcomes. So, we can apply the same bias correction to the estimated coefficient in the second stage regression to account for the omitted By term.

To construct correct instruments for Hy, we account for the expected number of missing paths between individuals. Following (Kelejian and Prucha, 1998), we use that

$$\begin{aligned} E(Hy) &= E(H(I + \lambda(H + B))^{-1}x\beta) \\ &= E(H(H + B + (H + B)^2 + (H + B)^3 + ...)x\beta). \end{aligned}$$

Therefore we use instruments

$$J^* = \Big(Hx, \quad E(Bx), \quad E(HBx|H), \quad ...\Big)$$

Implementation requires computing the expected number of unobserved paths of length k between nodes through the network given the sampled network, using knowledge of the sampling scheme and an assumption on the distribution of missing links given observed links. To give an example, make the following assumption on the distribution of missing links.

Assumption 5. The distribution of unobserved links is independent of the distribution of observed links for individuals with at least some unobserved links – $B_{ij} \perp\!\!\!\perp H_{jk}\ \forall i \in \mathcal{B}, H_{ij} \perp\!\!\!\perp B_{jk}\ \forall j \in \mathcal{B}$.

This assumption applies for networks under the common sampling schemes given above when all links are drawn from a common distribution. Other assumptions may be needed if, for example, some individuals are systematically more popular or name links in an order.

Expected numbers of walks then depend on sampled walks and powers of the mean missing degree. Considering the case of paths of length 2 for simplicity, and imagining that there are m possible incorrect entries in column j of H, we have

[13]Here, we make no assumption on the fraction of links that are incorrectly sampled. Lewbel et al. (2024) show that, in this setting, if the fraction of links that are incorrectly sampled falls quadratically in sample size, the two-stage least-squares estimator remains consistent. The first term in our instrument exogeneity condition vanishes as the sample size becomes larger. But, for the common sampling schemes listed above we would not expect the fraction of links incorrectly sampled to fall with sample size. Additionally, we see large finite-sample biases in simulations of common sampling schemes on networks.

$$
\begin{aligned}
E(HB|H)_{ik} &= E(\sum_j H_{ij}B_{jk}|H), \\
&= \sum_j E(H_{ij}B_{jk}|H) \text{ by linearity of E}, \\
&= \sum_j H_{ij}E(B_{jk}|H) \text{ by 5}, \\
&= \sum_j H_{ij}\frac{d_j^B}{|\mathcal{N}|-m}.
\end{aligned}
$$

Then the researcher can proxy the numbers of missing paths through the network $H^{k-1}B, ...$ with the expected number of missing paths through the network given the sampled adjacency matrix and missing mean degree. Formally, this gives

Proposition 8. Under Assumption 5, the variables $J^* = \begin{pmatrix} Hx, & d^B Hx, & H^2 x, & ... \end{pmatrix}$ are valid instruments for Hy conditional on By.

The endogeneity problem from the missing By in the second stage regression remains

$$
\begin{aligned}
y &= \lambda \hat{H} y + \xi, \\
\xi &= By + \epsilon.
\end{aligned}
$$

As in Section 2, the researcher can bias-correct estimates to deal with the omitted term By.

Proposition 9. Define

$$\hat{\theta}^{SS} = (z^{'} P_{J^*} z)^{-1} z^{'} P_{J^*} y,\ \hat{z} = P_{J^*} z,\ \hat{\eta} = (N^{-1} z' P_{J*} z)^{-1} N^{-1} \hat{z}' z_B.$$

The estimator

$$\hat{\theta} = (I + \hat{\eta})^{-1} \hat{\theta}^{SS} \tag{11}$$

is an unbiased estimator of $\theta = \begin{pmatrix} \lambda \\ \beta \end{pmatrix}$.

Let $\hat{\theta}_N$ denote an estimate from sample size N. $E(\hat{\theta}_N) \approx \theta$ and $\hat{\theta} \xrightarrow{p} \theta$.

This estimator is also asymptotically normal. We derive these results in appendix A.7, and show in Appendix A.8 that the estimator performs well in finite samples by simulation.

4 Extension – treatment dependent on network structure

In some cases, researchers may wish to estimate spillover effects when treatment depends on links. For example, treatment may be targeted by a planner based on network structure (e.g Galeotti

et al., 2020), or individuals may form their links based on treatment status (for examples, see Calvó-Armengol et al., 2009; Jackson, 2010).

If Assumptions 1.A, 2,3 still hold, and the data is drawn from (2) as before, we can still construct unbiased estimates using the bias function in Proposition 1. But we can no longer model $\hat{\eta}$ using purely aggregate statistics of the degree distribution under Assumption 1.B. Instead, we need to model the distribution of unobserved links and treatment. Formally, unobserved links b_{ij} are dependent on x_j, and some additional dependence parameters θ. Thus,

$$\eta \approx \frac{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}(x_i,x_j,\theta)x_j))}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)}.$$

To compute $\hat{\eta}$, we need a way of modeling the expected treatment of the observed and unobserved neighbours given observed and unobserved links.

One possible route is to fit a parametric model for the joint distribution of links on the network and treatment as in Borusyak and Hull (2023) and Herstad (2023). Instead, we consider the case where the researcher does not want to impose a parametric model for joint distribution of treatment and links ex-ante, but they are willing to model treatment as dependent upon some network statistic. This is a weaker assumption, as the researcher does not have to place restrictions upon many other features of the network, as in a full parametric model. We propose using a copula, as copulas allow us to flexibly model the dependence structure between two distributions preserving their marginal distributions.

Here, for simplicity, assume that treatment x_j depends upon in-degree d_j. Denote the observed distribution of treatment as F_X, and the distribution of the in-degree as F_D. The pairs (x_i, d_i) are distributed according to some unknown joint density function $G()$ with marginal distributions F_X, F_D. The researcher can flexibly model the joint density of treatment and this network statistic from empirical marginal distributions using a copula (Nelsen, 2006; Trivedi and Zimmer, 2007).

Definition 1. A bivariate copula is a quasi-monotone function $C()$ on the unit square $[0,1] \times [0,1] \to [0,1]$ such that there exists some a_1, a_2 such that $C(a_1, y) = C(x, a_2)$, and $C(1, y) = y, C(x, 1) = x \ \forall x, y \in [0,1]$.

From Sklar's theorem (Nelsen, 2006), we can represent the joint density $G()$ using a copula $C(F_X(x), F_D(d), \theta)$. Given a fitted copula with dependence parameter $\hat{\theta}$, we can compute expected degree given a treatment status

$$\begin{aligned} E(d_i|x,\hat{\theta}) &= \int_0^1 F_D^{-1}(p(u_d < U_d|F_X(x)))dU_d, \\ &= \int_0^1 F_D^{-1}(\frac{\partial C(u_x, u_d; \hat{\theta})}{\partial u_x}|_{u_x = F_X(x)})dU_d. \end{aligned}$$

Therefore, the researcher can compute, for each i

$$E(\sum_j b_{ij}x_j|x_j) = \sum_j E(b_{ij}|x_j,\hat{\theta})x_j,$$

allowing the researcher to compute $\hat{\eta}(\hat{\theta})$.

This motivates a two-step estimator. In the first stage, the researcher estimates the copula from the empirical distribution of network statistics on a set of $M \leq N$ observations by picking the dependence parameter $\hat{\theta}$ that sets the score equal to zero

$$\frac{1}{M}\sum_{i=1}^{M}\frac{\partial \ln C_i(F_x^{-1}, F_G^{-1}, \theta)}{\partial \theta} = 0.$$

Given a value $\hat{\theta}$, the researcher then estimates the unobserved spillovers from the copula

$$\hat{\eta}(\hat{\theta}) = \frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j \hat{b}_{ij}x_j(\theta))}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}.$$

and then constructs bias-corrected estimates as

$$\hat{\beta} = \frac{\hat{\beta}^{\text{OLS}}}{1+\hat{\eta}(\hat{\theta})}.$$

This estimator is also consistent and asymptotically normal (Newey, 1984; Smith, 2003).[14]

Proposition 10. Make Assumptions 1A ,2,3, 5, A1. Define

$$\begin{pmatrix} h_1(\theta) \\ h_2(\theta,\beta) \end{pmatrix} = \begin{pmatrix} \frac{1}{M}\sum_{i=1}^{M}\frac{\partial \ln C_i(F_x^{-1},F_G^{-1},\theta)}{\partial\theta}. \\ \frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(y_i-(1+\eta(\theta))\beta(\sum_j h_{ij}x_j)) \end{pmatrix}.$$

$$\begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} = \text{plim}\,\frac{1}{N}\sum_i E\begin{pmatrix} \frac{\partial \ln C_i(F_x^{-1},F_G^{-1},\theta)}{\partial\theta} & 0 \\ -\frac{\partial\eta(\theta)}{\partial\theta}\beta(\sum_j h_{ij}x_j)^2 & -(1+\eta(\theta))(\sum_j h_{ij}x_j)^2 \end{pmatrix}$$

$$\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \text{plim}\,\frac{1}{N}\sum_i E\begin{pmatrix} h_{1i}h_{1i}' & h_{2i}h_{1i}' \\ h_{2i}h_{1i}' & h_{2i}h_{2i}' \end{pmatrix}$$

$\hat{\beta}$ is a consistent estimator of β. Furthermore,

$$\sqrt{N}(\hat{\beta}-\beta) \sim N(0, K_{22}^{-1}(S_{22}+K_{21}K_{11}^{-1}S_{11}K_{11}^{-1}K_{21}' - K_{21}K_{11}^{-1}S_{12} - S_{21}K_{11}^{-1}K_{21}')K_{22}^{-1}).$$

Of course, implementing this estimator requires that the researcher can fit the copula and construct a link between the sampled degree statistic and the unobserved link weights. How the researcher might implement this depends on the treatment assignment rule and the sampling rule. For example, if the network is sampled using a fixed choice design, there exist some (low degree) nodes where treatment status and in-degree are fully observed. The researcher can fit the copula on this subset of individuals, under the assumption that the dependence between treatment and degree is the same for low and high degree nodes.[15] The dependence parameter

[14] This result can be complicated when the researcher also has to estimate the underlying distributions that go into the copula. For a discussion of estimation issues and consistency here, see Choroś et al. (2010).

[15] Then, the researcher may use a truncation invariant copula. See Nelsen (2006) for further discussion.

of the copula is an aggregate statistic. So a data provider could disclose this from a full data source without violating individual privacy. This approach is particularly suited for evaluating a treatment is assigned by a planner through an assignment rule known to the researcher that they can model through the copula. If the researcher is unsure how the degree statistic maps onto the unsampled link weights, they could use the result to construct bounds on the estimated spillover effect by sampling η under different assumptions.

Example – classroom setting, fixed choice sampling rule. Suppose that children are assigned a continuous treatment, that we can describe with a marginal distribution $x_i \sim N(5, 1)$. In an effort to maximise the effect of the educational intervention, school administrators have given larger doses to children the more friends that they have. Therefore, treatment status depends upon the network structure – the child's degree. The researcher samples friendships using a fixed choice sampling design (asking them to name up to m friends) with $m = 5$.

The researcher observes the treatment status of each child, and their sampled number of friends. To construct $\hat{\eta}$, the researcher needs to estimate the number of expected treated friends for each child with five sampled friends by fitting the copula. In the first step, the researcher can fit the copula on the subsample of children whose degree and treatment are fully observed. There are the children with fewer than five friends. Specifically, the researcher can model the marginal distributions of child's treatments and degrees as coupled through a bivariate Gumbel copula

$$C(F_X^{-1}(x), F_D^{-1}(d); \theta) = \exp(-((-\ln F_X^{-1}(x))^\theta + (-\ln F_D^{-1}(d))^\theta)^{\frac{1}{\theta}})$$

where $\theta \in [1, \infty]$ controls the degree of dependence between treatment and degree. Fitting the copula gives an estimate of the dependence parameter $\hat{\theta}$. Then, we can use the fitted copula to estimate $\hat{\eta}$ using

$$\sum_j E(b_{ij}(x_i, \hat{\theta})|x_j)x_j = \sum_j (E(g_{ij}^*|x_i, \hat{\theta}) - m)\bar{x}.$$

To assess how well this strategy performs in finite sample, we provide simulation results for this case in Appendix A.7.

5 Simulation experiments

Next, we evaluate the bias introduced by common sampling schemes and the performance of our rescaled estimators by Monte-Carlo simulation. Standard regression estimators can be heavily biased. The size of the bias depends on how much the sampling scheme alters the true network. Bias-corrected estimators perform well in finite samples. The distribution of bias-corrected estimates using $\hat{\eta}$ is close to the distribution of bias-corrected estimates under the true η, which is unbiased. In the appendix, we also simulate the performance of our estimator on real-life economic network that has been completely sampled – the co-authorship network of economists in Ductor et al. (2014).

5.1 Simulated networks

In each simulation, there are $N = 1000$ individuals who receive a binary treatment $x_i \sim$ Bernoulli(0.3). In each case, outcomes are drawn from (2) with $\beta = 0.8$, $\epsilon_i \sim N(0,1)$. We consider five networks and sampling schemes.

1. **Fixed choice design**. Each individual draws an in-degree from a discrete uniform distribution $d_i \sim U(1,15)$.[16] We form a binary directed simple network by connecting each individual with others uniformly at random from the population. We then sample links coming into each individual using a fixed choice design with reporting thresholds $m \in 1, ..., 14$.

2. **Sampling based on groups**. Each individual belongs to a single group (e.g high school class). There are 20 groups of 25 individuals, 10 groups of 20 individuals, and 20 groups of 15 individuals. The researcher samples each individual as linked to every other individual in their group. True degrees are drawn $U(m_i - k, m_i - 5 - k)$, where m_i is their group size and $k \in \{1,2,3,4,5\}$.

3. **Link weight thresholds**. Each individual draws interaction intensities with others from $w_{ij} \sim$ LogNormal$(1,15)$.[17] Then, we construct a weighted network where $g_{ij} = \frac{w_{ij}}{\sum_k w_{ik}}$. We sample links where g_{ij} exceeds a threshold $\tau \in \{0.025, 0.05, ..., 0.2\}$.

4. **Fixed choice design with weights**. Each individual draws an in-degree from a discrete uniform distribution $d_i \sim U(1,15)$. We form a weighted directed simple network connecting individuals with others uniformly at random. Weights are $g_{ij} = \frac{1}{d_i}$ – individuals who have more friends allocate less weight to each friend. Therefore reported weighted degree depends on number of friends. We then sample weighted links coming into each individual using a fixed choice design with reporting thresholds $m \in 1, ..., 14$.

5. **Sampling based on groups, true degree depends on group size**. Each individual belongs to one group (e.g high school class). There are 20 groups of 25 individuals, 10 groups of 20 individuals, and 20 groups of 15 individuals. The researcher samples each individual as linked to every other individual in their group. Their degrees are drawn $U(25 - 3k, 20 - 3k)$, $U(20 - 2k, 15 - 2k)$, $U(15 - k, 10 - k)$ for each group respectively, where $k \in \{1,2,3,4,5\}$.

In the first three cases, Assumption 4.a holds. In the final two cases, only Assumption 4.b holds. In each case, we construct estimates of β

1. by regressing outcomes on spillovers on the sampled network (**??**),

[16] We use a uniform distribution and sample neighbours uniformly at random from the population here to emphasise that the size of the bias that we find is not driven by tail behaviour of the degree distribution or preferential attachment-type mechanisms. Similar results hold when node degrees are sampled from more natural degree distributions like a discrete Pareto distribution (Clauset et al., 2009).

[17] The exact setting is calibrated similarly to the model of the US public-firm production network in Herskovic et al. (2020).

2. using the bias-corrected estimator given the true η (A-12), and

3. using the bias-corrected estimator estimating $\hat{\eta}$ using the results from section 2.3.

We run 1000 simulations per estimator, and report average values across each simulation. Additional experiments, including simulations for non-linear models and using the co-authorship network of economists, are given in Appendix A.8.

Below, we compare mean estimates across simulations, and plot the distribution of a representative set of estimates from each setting.

Simply regressing outcomes on sampled spillovers yields biased estimates. As expected, estimates are too large when we sample a subset of the true links between individuals (cases 1, 3, and 4) and too small when we sample a superset (cases 2 and 5). Bias can be substantial. For example, in the case of a fixed choice design sampling at most five links per individual (as for within-gender friendships in the Ad-Health dataset Harris, 2009), the average spillover effect estimates is 1.28 – 1.6 times the true effect. Thresholding links based at 10% of total flows (similar to how supply links are sampled between U.S. public firms Atalay et al., 2011), gives average spillover effect estimates of 1.63 – double the true effect.

Our bias-corrected estimators recover the true spillover effect well in finite samples. With η known, estimators are almost always centered on the true spillover value. With $\hat{\eta}$ estimated under Assumption 4.a or Assumption 4.b, estimates are centered very close to the true value. Bias-corrected estimators perform well when either assumptions hold, particularly under fixed choice sampling designs (cases 1 and 4).

6 Propagation of climate shocks in production networks

As an example, we use our estimator to measure how climate shocks propagate across supply links between public firms in the United States using self-reported supply relationships. Appendix A.9, also considers peer effects in education (Carrell et al., 2013).

There is a consensus that a central effect of climate change is an increase in extreme weather events (e.g see Robinson, 2021, and references therein). Whether these types of idiosyncratic shocks propagate between firms matters for the effect of climate change on economic output (Barrot and Sauvagnat, 2016). If firms can easily substitute away from suppliers hit by extreme weather shocks, the impact remains limited to those suppliers. If, however, shocks propagates from suppliers to customers, supply chains amplify the direct effect of these shocks (e.g see Carvalho et al., 2020).

6.1 Balance-sheet and supply-chain data

Data on supply links between U.S. public firms come from the popular Compustat Supply Chain dataset (Atalay et al., 2011). Since 1997, SFAS regulation No. 131 has required public firms to report customers that account for more than 10% of sales in 10-K filings with the Securities and Exchange Commission. Firms may report other customers voluntarily. Compustat collects

Number sampled	OLS	η	$\hat{\eta}$
3	1.67	0.800	0.800
4	1.46	0.800	0.800
5	1.28	0.800	0.800
6	1.14	0.800	0.800
7	1.08	0.800	0.800
8	1.00	0.800	0.800
9	0.950	0.800	0.800
10	0.900	0.800	0.800

Table 1: Mean spillover estimates using fixed choice design, by threshold

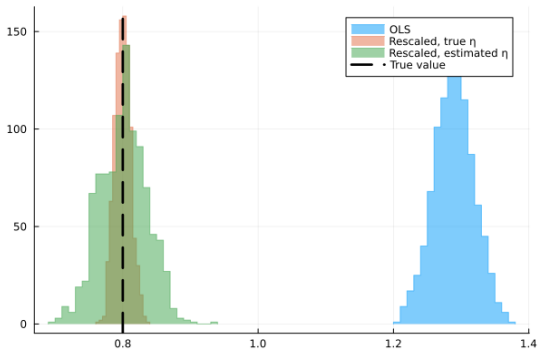


Figure 1: Distribution of spillover estimates using fixed choice design with threshold of 5

k	OLS	η	$\hat{\eta}$
1	0.700	0.800	0.770
2	0.660	0.800	0.780
3	0.630	0.800	0.780
4	0.590	0.800	0.770
5	0.550	0.800	0.770

Table 2: Mean spillover estimates sampling based on groups, by K

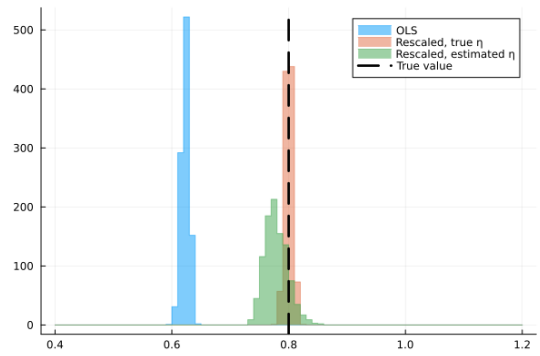


Figure 2: Distribution of spillover estimates sampling based on groups, $k = 3$

Threshold	OLS	η	$\hat{\eta}$
0.200	1.90	0.800	0.780
0.175	1.92	0.810	0.750
0.150	1.78	0.800	0.730
0.120	1.63	0.790	0.710
0.100	1.54	0.810	0.710
0.075	1.46	0.820	0.710
0.050	1.36	0.800	0.690
0.025	1.31	0.810	0.690

Table 3: Spillover estimates using fixed choice design, by threshold

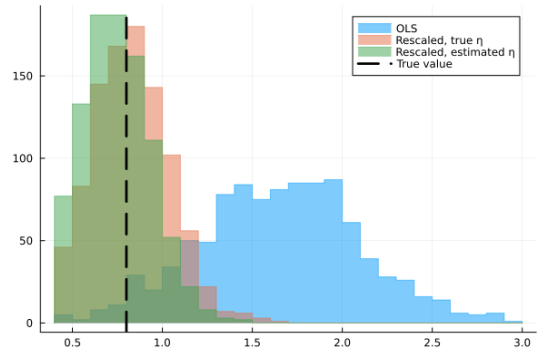


Figure 3: Distribution of spillover estimates using fixed choice design with threshold of 5

Number sampled	OLS	η	$\hat{\eta}$
3	0.990	0.800	0.800
4	0.960	0.800	0.800
5	0.920	0.800	0.800
6	0.880	0.800	0.800
7	0.880	0.800	0.800
8	0.860	0.800	0.800
9	0.850	0.800	0.800
10	0.830	0.800	0.800

Table 4: Mean spillover estimates from a fixed choice design with weights, by number sampled

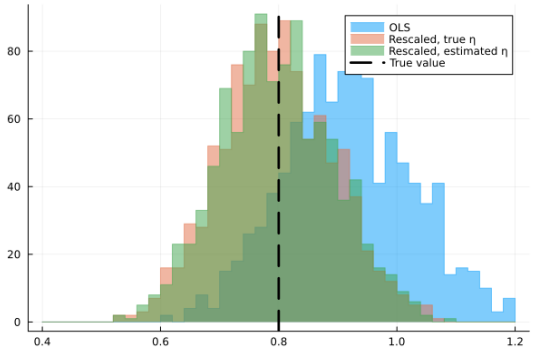


Figure 4: Distribution of spillover estimates using fixed choice design with threshold of 5

k	OLS	η	$\hat{\eta}$
1	0.650	0.800	0.780
2	0.560	0.800	0.770
3	0.470	0.800	0.760
4	0.380	0.800	0.730
5	0.290	0.800	0.710

Table 5: Mean spillover estimates sampling based on groups when true degree depends on group size, by k

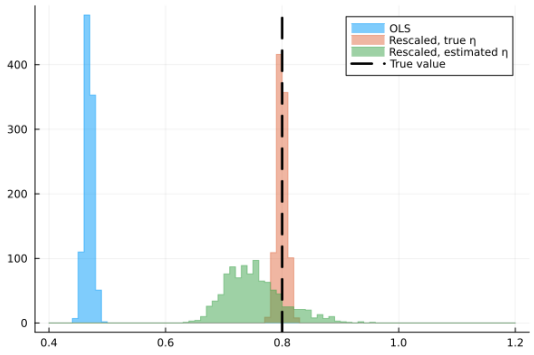


Figure 5: Distribution of spillover estimates using fixed choice design with threshold of 5

all of these self-reported links, which are understood to be a subset of the true supply links (Herskovic et al., 2020; Bacilieri et al., 2023).[18] In 2017 , the mean number of reported suppliers in is 1.36, and the median is 0.00. This is far fewer than researchers see in complete transactions data, and is an example where researchers can only sample high-weight links.[19] As described earlier, this sampling scheme satisfies Assumption 4.a.[20] So, we can construct rescaled estimates based on the mean missing degree amongst firms with at least one missing link. We use the mean degree of the (more complete) Factset production network in (Bacilieri et al., 2023), and value accounting for truncation in Herskovic et al. (2020). The resulting values are $2.7, 2.56$, corresponding to a mean missing degree of $d^B = 1.2, 1.34$.

Firm-level balance-sheet information for 1711 U.S. public firms in 2017 comes from the Compustat Fundamentals Quarterly North America dataset. Continuous variables are winsorized at the 99th and 1st percentiles. As firms may relocate headquarters, we locate firms using addresses reported in their 10-K forms instead of the location reported in Compustat (Gao et al., 2021).

6.2 Climate shocks

To determine which firms receive weather shocks, we construct a dataset of the county-level incidence of severe weather events in the United States 2004-2019.[21] Events comes from the US National Oceanic and Atmospheric Administration Billion-Dollar Weather and Climate Disasters project, which lists all weather events causing over $1 billion in damages (2024 dollars) between 1980 and 2024.[22] We match each weather event to county-level emergency declarations from the Federal Emergency Management Agency.

A county is coded as affected by a disaster if is in a state affected by the disaster and they have declared a state of emergency from that type of natural disaster (e.g a flood, a storm) in the days around the event given by the US National Oceanic and Atmospheric Administration. This yields a dataset of each county affected by a 'billion-dollar' natural disaster by month.

Table 1. lists the extreme weather events in the United States in 2017. There are six disasters in our dataset within this year: three hurricanes, two outbreaks of tornadoes, and one case of significant flooding. They affect firms in nine states over five months of the year. Total estimated damages range between $1.2 billion and $160 billion per disaster.

[18] Before the introduction of the regulation in 1997, firms would self-report certain customers. Some firms also report additional customers. For more details, see Bacilieri et al. (2023).

[19] For example, the mean number of suppliers in Belgian production network data is ≈ 30 (Dhyne et al., 2021), in Chilean data is ≈ 20 (Hunneus, 2020), and in Ecuadorian data is ≈ 33 (Bacilieri et al., 2023). The degree distribution is shifted to the left compared to true networks from VAT data, that shows similar patterns across countries (Bacilieri et al., 2023). Furthermore, Bacilieri et al. (2023) analyse a larger sample of self-reported network from 2012-2013, and find that 27 percent of firms have no listed suppliers, and 30 percent have no listed customers. The high amount of isolated firms suggests that some paths between firms are missing entirely.

[20] As in Barrot and Sauvagnat (2016), we treat the underlying network as binary. Further research could account for the effect of weights.

[21] The dataset is available on request.

[22] See `https://www.ncei.noaa.gov/access/billions`

Table 6: Major climate disasters in the United States, 2017

Disaster	Date	Damages (Billions, 2024 Dollars)	States declaring states of emergency
Southern Tornado Outbreak	January 20-22	1.4	GA, MS
Missouri and Arkansas Flooding	April 25–May 7	2.2	AR, MO
North Central Severe Weather and Tornadoes	May 15-18	1.2	OK
Hurricane Harvey	August 25-31	160.0	TX, LA
Hurricane Irma	September 6-12	64.0	FL, GA, SC
Hurricane Maria	September 19–21	115.2	GA

Notes: Events come from the NOAA Billion Dollar Weather and Climate Disasters Project. Affected states are those in which at least one county declares a state of emergency associated with the disaster as listed in the FEMA Disaster Declarations Dataset. Events that last longer than one month, or where no county declared a state of emergency, are excluded.

Following Barrot and Sauvagnat (2016), a firm is classified as hit by a shock in a given quarter if it is headquartered in a county affected by the disaster in that quarter. 14.9% of firms are hit with at least one weather shock within the year, and 11.3% have at least one reported supplier affected. There is strong evidence that firms do not choose suppliers based on the distribution of weather shocks across space (Barrot and Sauvagnat, 2016). So we can treat the distribution of these shocks as independent of the distribution of supply links between firms.

6.3 Estimation

We estimate the effect of an additional shock to a firm's supplier over a year on that year's sales growth, accounting for shocks to the firm itself, using the regression model

$$\Delta \ln \text{Sales}_{it,t-4} = \alpha + \beta_1 \sum_j h_{ij} \text{Shocked}_{jt,t-4} + \beta_2 \text{Shocked}_{t,t-4} + X\gamma + \epsilon_i.$$

We construct bias-corrected estimates of β_1 using the estimator (7) under the assumption that missingness is independent of controls.

Table 7: Estimates of propagation of climate shocks between US public firms over 2017

	$\Delta \ln$ Sales			
Estimator	OLS	OLS	Rescaled (Factset)	Rescaled (Herskovic et al.)
Suppliers shocked	−0.00675	−0.0248	−0.0140	−0.0132
	(0.00303)	(0.0114)	(0.01)	(0.01)
Shocked	0.0460	0.0650	0.0650	0.0650
	(0.0464)	(0.0608)	(0.0608)	(0.0608)
Size	Yes	Yes	Yes	Yes
Industry Fixed Effects	No	Yes	Yes	Yes
State Fixed Effects	No	Yes	Yes	Yes
Obs	1711	1243	1243	1243
R^2	0.001	0.103	0.103	0.103

Notes: Standard errors for non-rescaled estimates clustered by county (the level of shock assignment). Standard errors for rescaled estimates bootstrapped with 10000 draws. Firm-level controls are size (ppentq) and industry (4-digit NAICS fixed effects).

Table 7 reports results. In line with the existing literature (Barrot and Sauvagnat, 2016), the uncorrected estimate suggests that a shock to an additional supplier within the year leads to a 2.48% fall in yearly sales growth. After bias-correction, spillover effects are 53-56% of the initial estimates. Almost half of the naive spillover effect appears to come from bias due

to measurement error in links. We cannot reject the null hypothesis that spillover are zero at standard significance levels. Looking at robustness of the estimates to sampling bias using (8) suggests that estimates are very sensitive to missing links. Estimates fall to less than a 1.5% percent drop in yearly sales growth if we are missing at least one link on average, and a 1% fall if we are missing at least 2.25 links on average. Economically, these results may reflect the short duration of most weather shocks. Customers may be able to smooth out these short-term disruptions using inventories. We would expect the effects of these types of shocks to be smaller than those of larger natural disasters that cause long term disruption (e.g Carvalho et al., 2020).

7 Conclusion

We first show that sampling links between individuals can lead to substantial, economically significant bias in spillover estimates from linear and nonlinear models. Unlike classical measurement error, which generates downward bias, sampling can create either upward or downward bias depending on the scheme. Simulations demonstrate that popular sampling schemes lead to economically significant biases in estimates.

To solve this, we introduce bias-corrected estimators that rescale linear and nonlinear regressions to account for dependence between spillovers on observed and unobserved links. In experimental and quasi-experimental settings, researchers can implement these estimators using only aggregate statistics of the degree distribution, which are relatively easy to sample. Our estimators perform well in simulations. To illustrate, we estimate the propagation of climate shocks among US public firms in 2017 using sampled supply links.

For tractability, we rely on the linearity of the estimators in the sampled and unsampled networks for our results. Applied economists commonly fit complex structural models to sampled network data where parameters are non-linear functions of sampled networks (Badev, 2021; Lim, 2024, e.g see). Future work could extend our results to moment-based estimators in these settings. Our findings underscore that careful treatment of network sampling is essential for credible empirical estimates of spillovers.

References

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press.

Atalay, E., Hortaçsu, A., Roberts, J., and Syverson, C. (2011). Network structure of production. *Proceedings of the National Academy of Sciences*, 108(13):5199–5202.

Bacilieri, A., Borsos, A., Astudillo-Estevez, and Lafond, F. (2023). Firm-level production networks: What do we (really) know?

Badev, A. (2021). Nash equilibria on (un)stable networks. *Econometrica*, 89(3):1179–1206.

Banerjee, A., Chandrasekhar, A., Duflo, E., and Jackson, M. (2013). The Diffusion of Microfinance. *Science*, 341(1236498):363–341.

Barrot, J.-N. and Sauvagnat, J. (2016). Input Specificity and the Propagation of Idiosyncratic Shocks in Production Networks. *The Quarterly Journal of Economics*, 131(3):1543–1592.

Battaglia, L., Christensen, T., Hansen, S., and Sacher, S. (2025). Inference for Regression with Variables Generated by AI or Machine Learning. *Mimeo*.

Battaglini, M., Crawford, F., Patacchini, E., and Peng, S. (2021). A graphical lasso approach to estimating network connections: the case of us lawmakers. *Mimeo*.

Beaman, L. A. (2011). Social Networks and the Dynamics of Labour Market Outcomes: Evidence from Refugees Resettled in the U.S. *The Review of Economic Studies*, 79(1):128–161.

Beaman, L. A., BenYishay, A., Magruder, J., and Mobarak, A. M. (2021). Can network theory-based targeting increase technology adoption? *American Economic Review*, 111(6):1918–1943.

Billingsley, P. (2012). *Probability and measure*. John Wiley & Sons.

Bloom, N., Schankerman, M., and Van Reenen, J. (2013). Identifying technology spillovers and product market rivalry. *Econometrica*, 81(4):1347–1393.

Blume, L., Brock, W., Durlauf, S., and Jayaraman, R. (2015). Linear social interactions models. *Journal of Political Economy*, 123(2):444–496.

Borusyak, K. and Hull, P. (2023). Nonrandom Exposure to Exogenous Shocks. *Econometrica*, 91(6):2155–2185.

Borusyak, K., Hull, P., and Jaravel, X. (2024). Design-based identification with formula instruments: A review. *The Econometrics Journal*.

Boucher, V. and Houndetoungan, E. A. (2025). Estimating peer effects using partial network data. *Mimeo*.

Bound, J., Brown, C., and Mathiowetz, N. (2001). Chapter 59 - measurement error in survey data. volume 5 of *Handbook of Econometrics*, pages 3705–3843. Elsevier.

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

Breza, E., Chandrasekhar, A. G., McCormick, T. H., and Pan, M. (2020). Using aggregated relational data to feasibly identify network structure without network data. *American Economic Review*, 110(8):2454–84.

Calvó-Armengol, A., Patacchini, E., and Zenou, Y. (2009). Peer effects and social networks in education. *The Review of Economic Studies*, 76(4):1239–1267.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, London.

Carrell, S. E., Sacerdote, B. I., and West, J. E. (2013). From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica*, 81(3):855–882.

Carvalho, V. M., Nirei, M., Saito, Y. U., and Tahbaz-Salehi, A. (2020). Supply Chain Disruptions: Evidence from the Great East Japan Earthquake. *The Quarterly Journal of Economics*, 136(2):1255–1321.

Chandrasekhar, A. and Lewis, R. (2016). Econometrics of sampled networks. *Mimeo*.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star *. *The Quarterly Journal of Economics*, 126(4):1593–1660.

Choroś, B., Ibragimov, R., and Permiakova, E. (2010). Copula estimation. In Jaworski, P., Durante, F., Härdle, W. K., and Rychlik, T., editors, *Copula Theory and Its Applications*, pages 77–91. Springer Berlin Heidelberg.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 4:661–703.

Coleman, J., Katz, E., and Menzel, H. (1957). The diffusion of an innovation among physicians. *Sociometry*, 20(4):253–270.

Conley, T. G. and Udry, C. R. (2010). Learning about a new technology: Pineapple in ghana. *American Economic Review*, 100(1):35–69.

De Paula, A., Rasul, I., and Souza, P. (2024). Identifying network ties from panel data: theory and an application to tax competition. *Review of Economic Studies*, 00:1–39.

Dhyne, E., Kikkawa, K., Mogstad, M., and Tintlenot, F. (2021). Trade and domestic production networks. *The Review of Economic Studies*, 88(2):643–668.

Ductor, L., Fafchamps, M., Goyal, S., and van der Leij, M. J. (2014). Social networks and research output. *Review of Economics and Statistics*, 96(5):936–948.

Foster, A. D. and Rosenzweig, M. R. (1995). Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of Political Economy*, 103(6):1176–1209.

Galeotti, A., Golub, B., and Goyal, S. (2020). Targeting interventions in networks. *Econometrica*, 88(6):2445–2471.

Gao, M., Leung, H., and Qiu, B. (2021). Organization capital and executive performance incentives. *Journal of Banking and Finance*, (123):106017.

Glaeser, E. L., Sacerdote, B., and Scheinkman, J. A. (1996). Crime and social interactions. *The Quarterly Journal of Economics*, 111(2):507–548.

Griffith, A. (2022). Name your friends, but only five? the importance of censoring in peer effects estimates using social network data. *Journal of Labour Economics*, 40(4):779–805.

Griffith, A. and Kim, J. (2024). The impact of missing links on linear reduced-form network-based peer effects estimates. *Mimeo.*

Harris, K. M. (2009). The national longitudinal study of ad-olescent to adult health (add health), waves i and ii, 1994–1996. *Carolina Population Center, University of North Carolina at Chapel Hill.*

Heckman, J. (1979). Sample selection bias as specification error. *Econometrica*, 47(1):153–161.

Herskovic, B., Kelly, B., Lustig, H., and Van Nieuwerburgh, S. (2020). Firm volatility in granular networks. *Journal of Political Economy*, 128(11):4097–4162.

Herstad, E. I. (2023). Estimating peer effects and network formation models with missing links. *Mimeo.*

Higgins, A. and Martellosio, F. (2023). Shrinkage estimation of network spillovers with factor-structured errors. *Journal of Econometrics*, 233(1):66–87.

Hseih, C.-S., Hsu, Y.-C., Ko, S., Kovářík, J., and Logan, T. (2024). Non-representative sampled networks: Estimation of network structural properties by weighting.

Hunneus, F. (2020). Production network dynamics and the propagation of shocks. *Mimeo.*

Jackson, M. O., Nei, S. M., Snowberg, E., and Yariv, L. (2022). The dynamics of networks and homophily. Working Paper 30815, National Bureau of Economic Research.

Jackson, O. M. (2010). *Social and Economic Networks.* Princeton University Press, New Jersey.

Jaffe, A. (1986). Technological opportunity and spillovers of research-and-development - evidence from firms patents, profits, and market value. *American Economic Review*, 76(5):984–1001.

Kelejian, H. H. and Prucha, I. R. (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.

Lam, C. and Souza, P. (2019). Estimation and selection of spatial weight matrix in a spatial lag model. *Journal of Business and Economic Statistics*, 38(3):693–710.

Lewbel, A., Qu, X., and Tang, X. (2023). Social Networks with Unobserved Links. *Journal of Political Economy*, 131(4):898–946.

Lewbel, A., Qu, X., and Tang, X. (2024). Ignoring Measurement Errors in Social Networks. *The Econometrics Journal*, 27(2):171–187.

Lewbel, A., Qu, X., and Tang, X. (2025). Estimating Social Network Models with Link Misclassification. *Mimeo*.

Lim, K. (2024). Endogenous Production Networks and the Business Cycle. *Mimeo*.

MacKinnon, J. G. and Smith, A. A. (1998). Approximate bias correction in econometrics. *Journal of Econometrics*, 85(2):205–230.

Manresa, E. (2013). Estimating the structure of social interactions using panel data. *Mimeo*.

Manski, C. F. (1990). Nonparametric Bounds on Treatment Effects. *American Economic Review*, 80(2):319–323.

Marray, K. (2025). Estimating unobserved networks from heterogeneous characteristics, with an application to the swing riots.

Miguel, E. and Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1):159–217.

Munshi, K. (2003). Networks in the modern economy: Mexican migrants in the u. s. labor market. *The Quarterly Journal of Economics*, 118(2):549–599.

Nelsen, R. (2006). *An Introduction to Copulas*. Springer Series in Statistics, New York.

Newey, W. (1984). A method of moments interpretation of sequential estimators. *Economics Letters*, 14:201–206.

Newman, M. (2010). *Networks*. Oxford University Press, Oxford.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

Oster, E. and Thornton, R. (2012). Determinants of technology adoption: Peer effects in menstrual cup take-up. *Journal of the European Economic Association*, 10(6):1263–1293.

Rapoport, A. and Horvath, W. J. (1961). A study of a large sociogram. *Behavioral Science*, 6(4):279–291.

Robinson, W. A. (2021). Climate change and extreme weather: A review focusing on the continental united states. *Journal of the Air & Waste Management Association*, 71(10):1186–1209.

Rose, C. (2023). Identification of spillover effects using panel data. *Mimeo*.

Shakya, H. B., Stafford, D., Hughes, D. A., Keegan, T., Negron, R., Broome, J., McKnight, M., Nicoll, L., Nelson, J., Iriarte, E., Ordonez, M., Airoldi, E., Fowler, J. H., and Christakis, N. A. (2017). Exploiting social influence to magnify population-level behaviour change in maternal and child health: study protocol for a randomised controlled trial of network targeting algorithms in rural honduras. *BMJ Open*, 7(3).

Smith, M. (2003). Modelling sample selection using archimedian copulas. *Econometrics Journal*, 6:99 – 123.

Trivedi, P. K. and Zimmer, D. (2007). Copula modeling: an introduction for practitioners. In *Foundations and Trends in Econometrics*. Now Publishers.

Yauck, M. (2022). On the estimation of peer effects for sampled networks.

Zhang, L. (2023). Spillovers of program benefits with missing network links.

Appendix

A1 Proofs

We make the following standard assumptions for asymptotic results (Cameron and Trivedi, 2005)

Assumption 6. The matrix with entries $\text{plim}\,\frac{1}{N}\sum_i \epsilon_i^2 \sum_j g_{ij}x_j \sum_j g_{kj}x_j$ exists and is finite positive definite. Furthermore

$$\begin{aligned}
\text{plim}\,\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2 &= E((\sum_j h_{ij}x_j)^2)\\
\text{plim}\,\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j) &= E((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)),\\
\exists\delta>0 \text{ s.t } E(|\sum_j g_{ij}x_j \sum_j g_{kj}x_j|^{1+\delta}) &\leq \infty \;\forall k,i\\
\exists\delta>0 \text{ s.t } E(|\epsilon_i^2|^{1+\delta}) &\leq \infty \;\forall k,i\\
\exists\delta>0 \text{ s.t } E(|\epsilon_i^2\sum_j g_{ij}x_j \sum_j g_{kj}x_j|^{1+\delta}) &\leq \infty \;\forall k,i
\end{aligned}$$

Note that these may fail if the network has a degree distribution that is heavy tailed (see Newman, 2010). Examples include a power-law degree distribution. We do not address this, as this is separate to the focus of this paper. In this setting, estimation of spillover effects by regression models would need further justification in general.

Here, we also note that there are edge cases where linear regression estimator for the spillover effects are not identified in the limit when a researcher includes covariates or an intercept. In the case where $g_{ij} = \frac{1}{d_i}\ \forall i,j$ and $d_i = d_j \forall i,j$ i.e all individuals have the same degree and weight all contacts equally, the spillover effect will become colinear with the intercept in the asymptotic limit (and similar for any cases where the spillovers become colinear with a covariate in expectation). These, however, are not relevant cases, as in these cases it does not make sense for the researcher to use a regression estimator with covariates in general. So, we do not discuss these further.

Proof of proposition 1

Proof. We derive the form of the bias function first. Using (2), (4) we get

$$\begin{aligned}
\hat{\beta}^{\text{OLS}} &= \frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)y_i}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\\
&= \frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\beta(\sum_j h_{ij}x_j + \sum_j b_{ij}x_j)+\epsilon_i)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\\
&= \beta\Big(1+\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) + \frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)\epsilon_i}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}.
\end{aligned}$$

To show that this does cause bias, take expectations

$$E(\hat{\beta}^{\text{OLS}}) = \beta E\Big(1 + \frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) + E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)\epsilon_i}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big).$$
$$= \beta + \beta E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) + E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}E(\epsilon_i|\sum_i h_{ij}x_j)\Big).$$

Under assumption 2

$$E(\epsilon_i|\sum_i h_{ij}x_j) = E(\epsilon_i)$$
$$= 0.$$

By assumption,

$$E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) \neq 0.$$

The proposition follows.

□

Proof of proposition 2

Proof.

$$E\Big(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)\Big) = \frac{1}{N}\sum_i E\Big((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)\Big) \text{ by linearity of } E(),$$
$$= \frac{1}{N}\sum_i \Big(p(i \notin \mathcal{B})E\Big((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)|i \notin \mathcal{B}\Big),$$
$$+ p(i \in \mathcal{B})E\Big((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)|i \in \mathcal{B}\Big)\Big)\text{splitting those with no incorrectly s}$$
$$= \frac{1}{N}\sum_i \Big(0 + p(i \in \mathcal{B})E\Big((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)|i \in \mathcal{B}\Big)\Big)$$
$$= \frac{1}{N}\sum_i p(i \in \mathcal{B})(E(x)^2E((\sum_j h_{ij})(\sum_j b_{ij})|i \in \mathcal{B}\Big)\Big)\text{under assumption 1,}$$
$$= \frac{1}{N}\sum_i p(i \in \mathcal{B})(E(x)^2E((\sum_j h_{ij})E(\sum_j b_{ij}|\sum_j h_{ij})|i \in \mathcal{B}\Big)\Big)\text{conditioning,}$$
$$= \frac{1}{N}\sum_i p(i \in \mathcal{B})(E(x)^2E(d_i^H E(d_i^B|d_i^H)|i \in \mathcal{B}\Big)\Big).$$

We look for the cases when this term is non-zero. Assume that $E(x) \neq 0,$ and $p(i \in \mathcal{B}) \neq 0.$ Then, it is equivalent to

$$\sum_i E\Big(d_i^H E(d_i^B|d_i^H)|i \in \mathcal{B}\Big) \neq 0.$$

Assume that d_i^H has the same sign for each i. Then a sufficient condition for this to be non-zero is that $E(d_i^B|d_i^H)$ is either non-negative or non-positive for each i such that $i \in \mathcal{B}$.

□

Proof of theorem 1 Here, we introduce a (somewhat trivial) theorem proving that our theoretical bias-corrected estimator with the true rescaling factor is unbiased and consistent. We introduce it to use this in the subsequent proofs.

Theorem 1. Define $\eta = E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)$. Make Assumptions 1-A, 2, 3. The estimator

$$\hat{\beta} = \frac{\hat{\beta}^{\text{OLS}}}{1+\eta} \tag{A-12}$$

is an unbiased and consistent estimator of β i.e $E(\hat{\beta}) = \beta,$ and $\hat{\beta} \xrightarrow{p} \beta$.

Proof.

$$\begin{aligned}
E(\hat{\beta}) &= E(\frac{\hat{\beta}^{\text{OLS}}}{1+\eta}) \\
&= \frac{1}{1+E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)} E(\hat{\beta}^{\text{OLS}}) \\
&= \frac{1}{1+E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)} E\Big(\beta\Big(1+E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)\Big) + \frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)\epsilon_i}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) \text{(prop} \\
&= \beta + E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)\epsilon_i}{(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)(1+E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big))}\Big) \\
&= \beta + 0 \text{ from assumption 2.}
\end{aligned}$$

□

Now, we prove consistency.

Proof. Our estimator is

$$\hat{\beta} = \frac{1}{1+E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)}\Big(\beta\Big(1+\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)\Big)+\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)\epsilon_i}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big).$$

First, consider the term

$$\text{plim}\,\frac{1}{1+E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)}\beta\Big(1+\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)\Big)$$

Then, applying Slutsky's lemma

$$\text{plim}\,\beta\Big(1+\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)\Big) = \beta + \beta\frac{E((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))}{E((\sum_j h_{ij}x_j)^2)}.$$

Consider the Taylor expansion of $E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)$ around $E((\sum_j h_{ij}x_j)^2), E((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j$

$$E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) = \frac{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)} - \frac{\mathrm{Cov}(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j), \frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)}{(E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2))^2} + \frac{\mathrm{Var}(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)}{E((\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)^3)} + ...,$$

From assumption 6, $\mathrm{Var}(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2) \to 0$, and $\mathrm{Cov}(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j), \frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2) \to 0$. Therefore

$$\mathrm{plim}\, E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) = \frac{E((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))}{E((\sum_j h_{ij}x_j)^2)}.$$

Combining these results, we have that

$$\mathrm{plim}\, \frac{1}{1+E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)}\beta\Big(1+\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)\Big) = \beta.$$

Next, consider the second term

$$\mathrm{plim}\, \frac{1}{1+E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)}\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)\epsilon_i}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}.$$

Under assumptions 1,2

$$\mathrm{plim}\, \frac{1}{N}\sum_i(\sum_j h_{ij}x_j)\epsilon_i = 0.$$

Again applying Slutksy's lemma plus assumption A1 gives

$$\mathrm{plim}\, \frac{1}{1+E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)}\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)\epsilon_i}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2} = 0.$$

Combining our two intermediate results by Slutsky's lemma gives

$$\mathrm{plim}\, \hat{\beta} = \beta + 0.$$

□

A1.1 Proofs of proposition 3

Proof. As before

$$\eta = E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)$$

First, we want to show that we can approximate

$$E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) \approx \frac{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)}.$$

From taking the Taylor expansion of this fraction around the point μ_A, μ_B, we can in general evaluate (Billingsley, 2012)

$$E\Big(\frac{A}{B}\Big) = \frac{\mu_A}{\mu_B} - \frac{\mathrm{Cov}(A,B)}{\mu_B^2} + \frac{\mathrm{Var}(B)\mu_A}{\mu_B^3} + \Delta.$$

Substituting

$$A = \frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j),$$
$$B = \frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2,$$

and solving gives

$$\begin{aligned}
E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) &= \frac{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)} \\
&\quad - \frac{\mathrm{Cov}(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j), \frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)}{(E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2))^2} \\
&\quad + \frac{\mathrm{Var}(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)}{E((\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)^3)} + ..., \\
&= \frac{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)} + \mathcal{O}\Big(\frac{1}{(\sum_i\sum_j h_{ij}x_j)^4}\Big).
\end{aligned}$$

where we disregard the final terms as they are vanishingly small. Next, we want to evaluate the top given that we do not observe B.

As in the proof of proposition 2, we can write

$$E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)) = \frac{1}{N}\sum_i p(i\in\mathcal{B})(E(x)^2 E(d_i^H E(d_i^B|d_i^H)|i\in\mathcal{B}).$$

Now, applying assumption 4a,

$$E(d_i^H E(d_i^B|d_i^H)|i\in\mathcal{B}) = E(d_i^H|i\in\mathcal{B})E(d_i^B|i\in\mathcal{B})$$

Substituting back in, we have

$$E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)) = \frac{1}{N}\sum_i p(i\in\mathcal{B})(E(x)^2E(d_i^H|i\in\mathcal{B})E(d_i^B|i\in\mathcal{B}).$$

Substituting in the sample analogues and then applying Theorem 1 gives the results.

□

A1.2 Proof of proposition 4

Proof. From the proof of proposition 3,

$$E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)) = \frac{1}{N}\sum_i p(i\in\mathcal{B})(E(x)^2E(d_i^HE(d_i^B|d_i^H)|i\in\mathcal{B}).$$

From assumption 4b

$$\begin{aligned}E(d_i^HE(d_i^B|d_i^H)|i\in\mathcal{B}) &= E(d_i^HE(d_i|d^H=d_i^H)-d_i^H|i\in\mathcal{B})\\ &= E(d_i^HE(d_i^B|d^H=d_i^H)|i\in\mathcal{B}).\end{aligned}$$

Substituting in the sample analogues and then applying Theorem 1 gives the results. □

A1.3 Proofs of proposition 5, 10

Proof. We first now derive the asymptotic distribution of the estimator when η is known.

As in proof of prop 1., we have

$$\begin{aligned}\frac{\hat{\beta}^{\text{OLS}}}{1+\eta} &= \frac{1}{1+\eta}\Big(\beta(1+\eta)+\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)\epsilon_i}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big),\\ &= \beta+\frac{1}{1+\eta}\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)\epsilon_i}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big).\end{aligned}$$

Define the matrices

$$\begin{aligned}M_{XX} &= \text{plim}\,\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2\\ M_{X\Omega X} &= \text{plim}\,\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j h_{ij}x_j)\epsilon_i^2\end{aligned}$$

Under the maintained assumptions, we can apply the standard proof of the asymptotic distribution of the OLS estimator from Cameron and Trivedi (2005). This yields

$$\sqrt{N}\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)\epsilon_i}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) \sim N(0, M_{XX}^{-1}M_{X\Omega X}M_{XX}^{-1}).$$

Now, applying the normal product rule, we get

$$\sqrt{N}\Big(\frac{\hat{\beta}^{\text{OLS}}}{1+\eta} - \beta\Big) \sim N(0, (\frac{1}{1+\eta})^2 M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}).$$

Now, to prove proposition 5, note that we can write our estimator as a two-step M estimator (Newey, 1984).

$$\begin{aligned}\begin{pmatrix} h_1(\theta) \\ h_2(\theta,\beta) \end{pmatrix} &= \begin{pmatrix} \theta - \frac{1}{N}\sum_{i=1}^{N}\theta_i \\ \frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(y_i - (1+\eta(\theta))\beta(\sum_j h_{ij}x_j)) \end{pmatrix}, \\ &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}.\end{aligned}$$

Define

$$\begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} = \text{plim}\,\frac{1}{N}\sum_i E\begin{pmatrix} -1 & 0 \\ -\frac{\partial\eta(\theta)}{\partial\theta}\beta(\sum_j h_{ij}x_j)^2 & -(1+\eta(\theta))(\sum_j h_{ij}x_j)^2 \end{pmatrix}$$
$$\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \text{plim}\,\frac{1}{N}\sum_i E\begin{pmatrix} h_{1i}h_{1i}' & h_{2i}h_{1i}' \\ h_{2i}h_{1i}' & h_{2i}h_{2i}' \end{pmatrix}$$

Assume that

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_i h_{1i}(\eta) &\xrightarrow{d} N(0, S_{11}(\eta)), \\ \frac{1}{\sqrt{N}}\sum_i h_{2i}(\eta,\beta) &\xrightarrow{d} N(0, S_{22}(\eta,\beta)).\end{aligned}$$

We have just shown the second. Assume the first. Then, applying the results in Newey (1984), we know that therefore

$$\Omega = \text{Var}(\hat{\beta}) = K_{22}^{-1}(S_{22} + K_{21}K_{11}^{-1}S_{11}K_{11}^{-1}K_{21}' - K_{21}K_{11}^{-1}S_{12} - S_{21}K_{11}^{-1}K_{21}')K_{22}^{-1}, \qquad \text{(A-13)}$$

and

$$\sqrt{N}(\hat{\beta} - \beta) = N(0, \Omega).$$

□

Proposition 11 follows by the same logic by noting that our estimates for the copula parameters will also satisfy the assumptions for applying the two-step M estimator under standard regularity conditions - see Smith (2003); Choroś et al. (2010) and references therein.

A1.4 Proof of proposition 6

Proposition 7 follows by simply rearranging

$$\frac{\hat{\beta}^{\text{OLS}}}{1+\eta} > \tau$$

for $\hat{\beta}^{\text{OLS}}$.

Results for the non-linear social network model are presented in a separate section later

A2 Alternative approximations for the rescaling factor

As noted in the main text, the first order expansion of the rescaling factor performs well in finite sample in our simulation experiments. Therefore, though we have no analytic guarantee of how close the first order approximation is to the true rescaling factor, we should expect it to be close. In cases where a researcher doubts the performance of the first order approximation, however, we provide two alterative methods to compute the rescaling factor.

The first is a second-order expansion

$$E\left(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\right) \approx \frac{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)}$$
$$\Big(1 - \frac{\text{Cov}(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)), \frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)} + \frac{\text{Var}(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j$$

To compute the expansion, we need to evaluate the terms in the brackets. Denote

$$S_h = \sum_i h_{ij}x_j,$$
$$S_b = \sum_i b_{ij}x_j.$$

Then

$$-\frac{\text{Cov}(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)), \frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)} = -\frac{\frac{1}{N^4}d_h E(x)\Big(E(S_b^7) - E(S_b^3)E(S_b^4)\Big)}{\frac{1}{N^2}d_h E(x)E(S_b)E(S_b^2)},$$
$$= -\frac{\frac{1}{N^2}\Big(E(S_b^7) - E(S_b^3)E(S_b^4)\Big)}{E(S_b)E(S_b^2)}.$$
$$= \frac{1}{N^2}\frac{E(S_b^3)E(S_b^4) - E(S_b^7)}{E(S_b)E(S_b^2)}.$$

$$\frac{\text{Var}(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)^2} = \frac{E(S_h^4) - E(S_h^2)^2}{E(S_h^2)^2}$$
$$= \frac{E(S_h^4)}{E(S_h^2)^2} - 1.$$

Therefore,

$$1 - \frac{\text{Cov}(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)), \frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)} + \frac{\text{Var}(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)^2}$$
$$= \frac{1}{N^2}\frac{E(S_b^3)E(S_b^4) - E(S_b^7)}{E(S_b)E(S_b^2)} + \frac{E(S_h^4)}{E(S_h^2)^2}.$$

The second method for computing $\hat{\eta}$ is by simulation. In this case, the researcher first constructs M possible matrices of missing links $\{B^{(m)}\}_{m=1}^M$ based on either the number of total missing links, or the number of missing links by individual. Then, we can estimate

$$\hat{\eta}_M = \frac{1}{M}\sum_m \frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j \sum_j b_{ij}^{(k)}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}.$$

The bias-corrected estimator is then

$$\hat{\beta} = \frac{\hat{\beta}^{\text{OLS}}}{1 + \hat{\eta}_M}.$$

Then, the researcher must compute the variance of the estimator using a bootstrap.

A3 Detailed example with fixed choice design

Example – fixed choice design. To fix ideas, consider the case of a binary network $h_{ij}, b_{ij} \in \{0, 1\}$ where $x_j = 1\ \forall j$. The logic extends to the more general case without loss of generality.

The researcher samples up to m links into each individual. For illustration, let $m = 5$ (as for same-sex friends in the Ad Health dataset Harris, 2009). If an individual has five or fewer connections, the researcher samples all of their connections. Sampled spillovers equal observed spillovers. If an individual has more than five connections, the researcher does not sample some of their spillovers. So they have some positive unobserved spillovers. As they have the maximum number of sampled links, their spillovers are also higher. Individuals with more than five links have a sampled spillover of 5, greater than or equal to individuals with five or fewer friends (whose spillovers are in $\{0, 1, 2, 3, 4, 5\}$). Thus, sampling based on generates positive dependence between observed and unobserved spillovers.

Formally, we can derive the expected dependence between observed and unobserved spillovers under a fixed choice design as:

$$
\begin{aligned}
E\Big(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)\Big) &= \frac{1}{N}\sum_i E\Big((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)\Big) \text{ by linearity of } E(),\\
&= \frac{1}{N}\sum_i (p(d_i \le m)E\Big((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)|d_i \le m\Big),\\
&+ (1-p(d_i \le m))E\Big((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)|d_i > m\Big)),\\
&= \frac{1}{N}\sum_i (1-p(d_i \le m))E\Big((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)|d_i > m\Big) \text{ as } b_{ij}=0\forall j \text{ if } d_i \le\\
&= \frac{1}{N}\sum_i (1-p(d_i \le m))mE(d_i - m|d_i > m) \text{ from the sampling rule},\\
&= \frac{1}{N}\sum_i (1-p(d_i \le m))m(E(d_i|d_i > m) - m) > 0.
\end{aligned}
$$

Therefore, under this sampling design, estimates are upwards biased ($|\hat{\beta}^{\text{OLS}}| > |\beta|$).

A4 Extension to models with covariates

Here, we derive our results in matrix notation to allow for arbitrary covariates. This allows us to extend the results to general linear regression models, and regression models for panel data. Let

$$
Z = \begin{pmatrix} Hx \\ W \end{pmatrix}.
$$

Our model in matrix form is

$$
y = \begin{pmatrix} Gx \\ W \end{pmatrix}' \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \epsilon. \tag{A-14}
$$

The OLS estimator solves

$$
\begin{pmatrix} \hat{\beta}^{\text{ OLS}} \\ \hat{\gamma}^{\text{ OLS}} \end{pmatrix} = (Z'Z)^{-1}Z'y
$$

Solving yields

$$
\begin{aligned}
\hat{\gamma}^{\text{OLS}} &= (W'(I - P_{Hx})W)^{-1}W'(I - P_{Hx})y,\\
\hat{\beta}^{\text{OLS}} &= ((Hx)'(I - P_W)Hx)^{-1}(Hx)'(I - P_W)y.
\end{aligned}
$$

Let $(\widetilde{A})$ denote $(I - P_W)A$. For readability, write

$$
\hat{\beta}^{\text{OLS}} = ((\widetilde{Hx})'\widetilde{Hx})^{-1}(\widetilde{Hx})'\widetilde{y}.
$$

Substituting (A-14) for y ,

$$\hat{\beta}^{\text{OLS}} = \beta + ((\widetilde{Hx})'\widetilde{Hx})^{-1}(\widetilde{Hx})'(\widetilde{Bx}\beta + \widetilde{\epsilon}).$$

Taking expectations

$$E(\hat{\beta}^{\text{OLS}}) = (I + E((\widetilde{Hx})'\widetilde{Hx})^{-1}(\widetilde{Hx})'\widetilde{Bx})\beta.$$

Therefore the multiplicative bias is

$$E((\widetilde{Hx})'\widetilde{Hx})^{-1}(\widetilde{Hx})'\widetilde{Bx}).$$

Equivalents of proposition 1, theorem 1 follow immediately.

Under the same Taylor approximation as in the proof of proposition 4,

$$E(((\widetilde{Hx})'\widetilde{Hx})^{-1}(\widetilde{Hx})'\widetilde{Bx}) \approx E(((\widetilde{Hx})'\widetilde{Hx})^{-1})E((\widetilde{Hx})'\widetilde{Bx}),$$

giving the results in section 2.4 for the mean degree of the sampled network projected onto the orthogonal complement of the space of the column space of covariates W and the mean number of missing links after projection onto the orthogonal complement of the space of the column space of covariates W.

If we further assume that measurement errors and spillovers are distributed indepedently of covariates throughout

$$Bx, Gx \perp\!\!\!\perp W$$

then the results in section 2.4 apply identically.

In practice, it is important to consider whether this assumption holds or not before bias-correcting the estimator. If it does not, the researcher needs to apply the results using

$$\widetilde{d^H} = \frac{1}{\sum_{i\in\mathcal{B}} 1_i} \sum_{i\in\mathcal{B},j} \widetilde{h}_{ij}$$
$$\widetilde{d^B} = \frac{1}{\sum_{i\in\mathcal{B}} 1_i} \sum_{i\in\mathcal{B},j} \widetilde{b}_{ij}.$$

In practice, the researchers could construct these by regressing reported number of links/missing links on covariates amongst all individuals, removing the expectation given the covariates for all individuals, and then taking the mean for individuals with at least some missing links.

We brush over it in the main text for reasons similar to Battaglia et al. (2025) – considering it directly dilutes the main point of the paper.

In certain cases, including controls can lead to $E(\widetilde{Bx}) = 0$. In this case, the linear regression estimator is not biased, and correction would be erroneous. Our approach gives a transparent way to see when adding controls will also account for measurement error. An example is a panel data regression with individual fixed effects with constant sampling error by node. In this case

$$\begin{aligned}\widetilde{BX}_{it} &= (d_{it}^B - \bar{d_i^B})E(X) \\ &= (d_{it}^B - d_{it}^B)E(X) \\ &= 0\end{aligned}$$

as by construction $d_{it}^B = \bar{d_i^B}$.

A5 Dummy variable estimators

Again, assume that we can describe the underlying data generating process with (2). Instead of estimating the direct spillover effect β, the researcher wants to estimate the average effect of at least one neighbour being treated on outcomes[23]

$$\gamma = E(\beta \sum_j g_{ij}x_j | \sum_j g_{ij}x_j > 0).$$

For example, the researcher wants to estimate the effect of at least one supplier experiencing a shock on sales (Barrot and Sauvagnat, 2016). A common estimation strategy is to construct a dummy variable that encodes whether at least one sampled neighbour is treated

$$d_i = \begin{cases} 1 & \text{if and only if} \sum_j h_{ij}x_j \geq 1 \\ 0 & \text{else} \end{cases}$$

and regress on outcomes on this dummy plus an intercept (e.g specifications in Oster and Thornton, 2012; Barrot and Sauvagnat, 2016) [24] By splitting spillovers into observed and unobserved components, we see that this estimator recovers (Angrist and Pischke, 2009)

$$\begin{aligned}\hat{\gamma}^{\text{ OLS}} &= E(\beta \sum_j g_{ij}x_j | \sum_j h_{ij}x_j > 0) - E(\beta \sum_j g_{ij}x_j | \sum_j h_{ij}x_j = 0) \\ &\neq E(\beta \sum_j g_{ij}x_j | \sum_j g_{ij}x_j > 0).\end{aligned}$$

where the second term may be non-zero.

Again, we can construct an unbiased estimator by rescaling based on the mean number of missing links on the network.

Proposition 11. Make assumptions 1,2,3. Consider the estimator

$$\hat{\gamma} = \frac{\frac{E(d_i^H)+E(d_i^B)}{p(\sum_j g_{ij}x_j>0)}}{\frac{E(d_i^H)}{p(\sum_j h_{ij}x_j>0))} + E(d_i^B|\sum_j h_{ij}x_j > 0) - E(d_i^B|\sum_j h_{ij}x_j = 0)} \hat{\gamma}^{\text{ OLS}}.$$

[23] Note that this is a different estimand to the spillover effect β, though the two are sometimes conflated (Barrot and Sauvagnat, 2016). Different degree distributions of the true underlying network can deliver different γ for the same β.

[24] We omit controls here without loss of generality.

$\hat{\gamma}$ is an unbiased estimator of γ.

Proof

Proof. By definition,

$$\gamma = \frac{\gamma}{\hat{\gamma}^{\text{ OLS}}}\hat{\gamma}^{\text{ OLS}}.$$

Therefore, $\frac{\gamma}{\hat{\gamma}^{\text{ OLS}}}\hat{\gamma}^{\text{ OLS}}$ is an unbiased estimator of γ.
Now, we simplify this fraction. Given that outcomes follow (2),

$$\begin{aligned}
\gamma &= E(\sum_j g_{ij}x_j + \epsilon_i | \sum_j g_{ij}x_j > 0) - E(\sum_j g_{ij}x_j + \epsilon_i | \sum_j g_{ij}x_j = 0) \\
&= E(\sum_j g_{ij}x_j | \sum_j g_{ij}x_j > 0) + E(\epsilon_i | \sum_j g_{ij}x_j > 0) - E(\epsilon_i | \sum_j g_{ij}x_j = 0), \\
&= E(\sum_j g_{ij}x_j | \sum_j g_{ij}x_j > 0) \text{ by assumption 2}, \\
&= \frac{E(\sum_j g_{ij}x_j)}{p(\sum_j g_{ij}x_j > 0)} \\
&= \frac{E(x)E(\sum_j h_{ij} + \sum_j b_{ij})}{p(\sum_j g_{ij}x_j > 0)} \text{ by assumption 1} \\
&= \frac{E(x)E(d_i^H + d_i^B}{p(\sum_j g_{ij}x_j > 0)} \\
&= \beta E(x)\frac{(E(d_i^H) + E(d_i^B))}{p(\sum_j g_{ij}x_j > 0)}.
\end{aligned}$$

Similarly

$$\begin{aligned}
\hat{\gamma}^{\text{ OLS}} &= E(\beta\sum_j g_{ij}x_j + \epsilon_i | \sum_j h_{ij}x_j > 0) - E(\beta\sum_j g_{ij}x_j + \epsilon_i | \sum_j h_{ij}x_j = 0) \\
&= E(\beta\sum_j g_{ij}x_j | \sum_j h_{ij}x_j > 0) - E(\beta\sum_j g_{ij}x_j | \sum_j h_{ij}x_j = 0) + E(\epsilon_i | \sum_j h_{ij}x_j > 0) - E(\epsilon_i | \sum_j h_{ij}x_j = 0) \\
&= \beta(E(\sum_j h_{ij}x_j + \sum_j b_{ij}x_j | \sum_j h_{ij}x_j > 0) - E(h_{ij}x_j + \sum_j b_{ij}x_j | \sum_j h_{ij}x_j = 0)), \\
&= \beta(E(\sum_j h_{ij}x_j | \sum_j h_{ij}x_j > 0) + E(\sum_j b_{ij}x_j | \sum_j h_{ij}x_j > 0) - E(\sum_j b_{ij}x_j | \sum_j h_{ij}x_j = 0)) \\
&= \beta E(x)(E(\sum_j h_{ij} | \sum_j h_{ij}x_j > 0) + E(\sum_j b_{ij} | \sum_j h_{ij}x_j > 0) - E(\sum_j b_{ij} | \sum_j h_{ij}x_j = 0)) \text{ by assumption 1} \\
&= \beta E(x)\Big(\frac{E(\sum_j h_{ij})}{p(\sum_j h_{ij}x_j > 0))} + E(\sum_j b_{ij} | \sum_j h_{ij}x_j > 0) - E(\sum_j b_{ij} | \sum_j h_{ij}x_j = 0)\Big) \text{ by assumption 1}, \\
&= \beta E(x)\Big(\frac{E(d_i^H)}{p(\sum_j h_{ij}x_j > 0))} + E(d_i^B | \sum_j h_{ij}x_j > 0) - E(d_i^B | \sum_j h_{ij}x_j = 0)\Big)
\end{aligned}$$

Therefore

$$\gamma = \frac{\gamma}{\hat{\gamma}^{\text{ OLS}}}\hat{\gamma}^{\text{ OLS}},$$

$$= \frac{\beta E(x)\frac{(E(d_i^H)+E(d_i^B))}{p(\sum_j g_{ij}x_j>0)}}{\beta E(x)\Big(\frac{E(d_i^H)}{p(\sum_j h_{ij}x_j>0))} + E(d_i^B|\sum_j h_{ij}x_j>0) - E(d_i^B|\sum_j h_{ij}x_j=0)\Big)}\hat{\gamma}^{\text{ OLS}}$$

$$= \frac{\frac{E(d_i^H)+E(d_i^B)}{p(\sum_j g_{ij}x_j>0)}}{\frac{E(d_i^H)}{p(\sum_j h_{ij}x_j>0))} + E(d_i^B|\sum_j h_{ij}x_j>0) - E(d_i^B|\sum_j h_{ij}x_j=0)}\hat{\gamma}^{\text{ OLS}}.$$

□

Sample analogues for $E(d_i^H)$, $\frac{E(d_i^H)}{p(\sum_j h_{ij}x_j>0))}$ are directly computable from observed H, x. The other missing terms – the expected number of unobserved links, and difference in the the expected number of unobserved links between individuals with at least one sampled treated neighbour and individuals with no sampled treated neighbours – are again aggregate network statistics. The researchers can construct sample analogues for the other terms. They can do this by asking each individual how many connections they have in a survey, disclosed by data providers without violating privacy, or approximated from detailed sampling of similar datasets.

A6 Equivalence to control function approach

Writing out our data-generating process again, we have that

$$\begin{aligned} y_i &= \beta \sum_j g_{ij}x_j + \epsilon_i \\ &= \beta \sum_j h_{ij}x_j + \xi_i \end{aligned}$$

where

$$\xi_i = \sum_j b_{ij}x_j\beta + \epsilon_i.$$

A model for the error under assumption 1 is

$$E(\xi_i) = d_i^B E(x_j).$$

The resulting regression model would be

$$y_i = \beta \sum_j g_{ij}x_j + \gamma d_i^B E(x_j).$$

which gives the same regression estimator as in the main text. Of course, this requires knowing which individuals have at least some incorrectly sampled links. Thus, it is only implementable for a subset of the sampling schemes used to study economic networks (e.g fixed choice designs, but not assuming that all individuals in the same group are connected).

A7 Results for non-linear social network models

Make the standard assumptions (Kelejian and Prucha, 1998; Bramoullé et al., 2009; Blume et al., 2015).

Assumption A2 Assume that

1. (y, G, B, x) are independently but not identically distributed over i,
2. $E(\epsilon|G, x) = 0$
3. ϵ are independent and not identically distributed over i such that for some $\delta > 0$ $E(|u_i^2|^{1+\delta}) < \infty$ with conditional variance matrix

$$E(\epsilon\epsilon'|(G - B)x) = \Omega$$

which is diagonal.

4.

$$\begin{aligned} \text{plim } N^{-1}z'P_{J*}z &= Q_{ZZ} \\ \text{plim } N^{-1}z'P_{J*}z_B &= Q_{ZB} \\ \text{plim } N^{-1}z'P_{J*} &= Q_{HJ} \end{aligned}$$

which are each finite nonsingular.

5. $|\lambda| < \frac{1}{||H||}, \frac{1}{||G||}$ for any matrix norm $||.||$.

The estimator for non-linear social network models given in Section 3 is consistent and asymptotically normal.

Theorem 2. Consider the debiased estimator $\hat{\theta}$, and make assumption A7. Then plim $\hat{\theta} = \theta$ and

$$\frac{1}{\sqrt{N}}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, N(0, \sigma^2(I + Q_{ZZ}^{-1}Q_{ZB})^{-1}Q_{ZZ}^{-1}Q_{HJ}((I + Q_{ZZ}^{-1}Q_{ZB})^{-1}Q_{ZZ}^{-1})'),$$

where

$$\begin{aligned} \text{plim } N^{-1}Z'P_{J*}Z &= Q_{ZZ} \\ \text{plim } N^{-1}Z'P_{J*}Z_B &= Q_{ZB} \\ \text{plim } N^{-1}Z'P_{J*} &= Q_{HJ} \end{aligned}$$

Proof. Let $z^* = \Big(Gy, x\Big)$, $z = \Big(Hy, x\Big)$. Call $z_B = z^* - z = \Big(By, 0\Big)$. Finally, denote the projection matrix onto the space spanned by our instruments $P_{J*} = J^*(J^{*'}J^*)^{-1}J^{*'}$.

Our two-stage least squares estimates with our unbiased instruments J^* are

$$\begin{aligned}\hat{\theta}^{2sls} &= ((P_{J*}z)'P_{J*}z)^{-1}(P_{J*}z)'y, \\ &= ((P_{J*}z)'P_{J*}z)^{-1}(P_{J*}z)'(z^*\theta + \epsilon) \\ &= (z'P_{J*}z)^{-1}(P_{J*}z)'(z\theta + z_B\theta + \epsilon) \\ &= \theta + ((z'P_{J*}z)^{-1}(P_{J*}z)'z_B\theta + (z'P_{J*}z)^{-1}(P_{J*}z)'\epsilon.\end{aligned}$$

Therefore,

$$\hat{\theta} = (I + (z'P_{J*}z)^{-1}(z'P_{J*}z_B))^{-1}\hat{\theta}^{2sls} = \theta + (I + (z'P_{J*}z)^{-1}z'P_{J*}z_B)^{-1}(z'P_{J*}z)^{-1}(P_{J*}z)'\epsilon.$$

Note that

$$z'P_{J*}z_B = \begin{pmatrix} 0 & (Hy)'P_{J*}By \\ 0 & x'P_{J*}By \end{pmatrix}.$$

First, we show the consistency of this estimator. As per assumption A7

$$\begin{aligned}\text{plim } N^{-1}z'P_{J*}z &= Q_{ZZ} \\ \text{plim } N^{-1}z'P_{J*}z_B &= Q_{ZB} \\ \text{plim } N^{-1}z'P_{J*} &= Q_{HJ}\end{aligned}$$

which are each finite nonsingular.
Therefore

$$\begin{aligned}\text{plim } \hat{\theta} &= \text{plim } (\theta + (I + (N^{-1}z'P_{J*}z)^{-1}N^{-1}z'P_{J*}z_B)^{-1}(N^{-1}z'P_{J*}z)^{-1}(N^{-1}P_{J*}z)'\epsilon) \\ &= \theta + (I + Q_{ZZ}^{-1}Q_{ZB})^{-1}Q_{ZZ}^{-1}\text{plim}N^{-1}z'P_{J*}\epsilon \text{ by Slutsky's lemma}\end{aligned}$$

Finally, we need to characterise the properties of

$$\text{plim}N^{-1}z'P_{J*}\epsilon.$$

$$N^{-1}z'P_{J*} = \begin{pmatrix} N^{-1}(P_{J*}Gy)'\epsilon \\ N^{-1}(P_{J*}x)'\epsilon \end{pmatrix}.$$

We can characterise the behaviour of the second row using a standard weak law of large numbers. But, the vector Gy involves a sum of random variables y. So, here, we need to apply a law of large numbers for triangular arrays. From assumption A7, it follows that the array $G_{1,1}y_1, G_{1,2}y_2, ...$ is a triangular array (Kelejian and Prucha, 1998). So, the term $GY)'\epsilon$ is the sum of

$$(G_{1,1}y_1, G_{1,2}y_2, ...)\epsilon_1 + (G_{2,1}y_1, G_{2,2}y_2, ...)\epsilon_2 + ...$$

which is itself a triangular array. Call this triangular array W. Assume that $\sup_N E_N(W^2) < \infty$ for all N. Then we can apply a weak law of large numbers for triangular arrays to W to say that

$$\text{plim } N^{-1}(P_{J*}Gy)'\epsilon = E((P_{J*}Gy)'\epsilon)_i) = 0.$$

Therefore our estimator is both unbiased and consistent.

Next, we need to characterise the asymptotic distribution of the estimator.

$$\sqrt{N}(\hat{\theta} - \theta) = (I + (N^{-1}z'P_{J*}z)^{-1}N^{-1}z'P_{J*}z_B)^{-1}(N^{-1}z'P_{J*}z)^{-1}(\frac{1}{\sqrt{N}}P_{J*}z)'\epsilon)$$

Again, applying Slutsky's lemma, all terms on the right hand side except

$$\frac{1}{\sqrt{N}}(P_{J*}z)'\epsilon$$

will converge to finite limits. To characterise the distribution of this term, we need to apply a law of large numbers for triangular arrays. We use the central limit theorem for triangular arrays from (Kelejian and Prucha, 1998).

Theorem 3 (CLT for triangular arrays)**.** Let ϵ, $P_{J*}Hy$ be triangular arrays of identically distributed random variables with finite second moments. Denote $\text{Var}(\epsilon) = \sigma^2$. Assume that plim $N^{-1}(P_{J*}Hy)'P_{J*}Hy = Q_{HJ}$ is finite and nonsingular. Then

$$\frac{1}{\sqrt{N}}(P_{J*}z)'\epsilon \xrightarrow{d} N(0, \sigma^2 Q_{HJ}).$$

Applying this result, we have that

$$\frac{1}{\sqrt{N}}(P_{J*}z)'\epsilon \xrightarrow{d} N(0, \sigma^2 Q_{HJ}).$$

Therefore, by Slutky's lemma

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2(I + Q_{ZZ}^{-1}Q_{ZB})^{-1}Q_{ZZ}^{-1}Q_{HJ}((I + Q_{ZZ}^{-1}Q_{ZB})^{-1}Q_{ZZ}^{-1})').$$

□

A8 Additional simulations

A8.1 Real-data simulation

We further test the performance of our estimator on a real network - the co-author network of economists from Ductor et al. (2014). This is the complete network of co-authorships between economists on papers published in journals in the EconLit database. As in Ductor et al. (2014),

we use co-authorships over a three-year window – here 1996-1998 – to account for lags in publications. This gives us across 44,776 economists and 57,407 links between them. Note that the network is very sparse. The mean degree is 1.28. The 95th percentile of the degree distribution is 4 collaborations.

We simulate the effect of a treatment across this network as above. In each simulation, each economist draws a binary treatment $x_i \sim$ Bernoulli(0.3). Outcomes are drawn from (2) with $\beta = 0.8$, $\epsilon_i \sim N(0,1)$.

We sample the network using a fixed choice design with thresholds $k \in \{1,2,3,4,5,6\}$. Next, we sample based on groups. We then construct spillover estimates using the sampled network, and using our debiased estimator under assumption 4.a.

Number sampled	OLS	η	$\hat{\eta}$
1	1.02	0.799	0.798
2	0.902	0.800	0.799
3	0.871	0.800	0.800
4	0.847	0.799	0.799
5	0.834	0.800	0.800
6	0.823	0.799	0.799

Figure A8.1: Mean spillover estimates using fixed choice design, by threshold

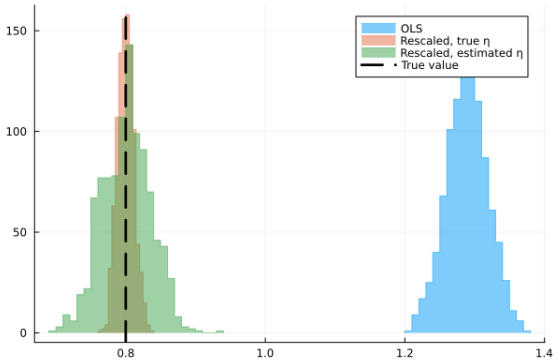


Figure A8.2: Distribution of spillover estimates using fixed choice design with threshold of 3

As in our simulated networks, we see that linear regression of outcomes on sampled spillovers leads to biased estimates. The bias is relatively small because the true network is so sparse. With a threshold of 3, 90% of individuals maintain all of their true links. Our error corrected estimate performs still perform very well.

A8.2 Simulations for nonlinear social network models

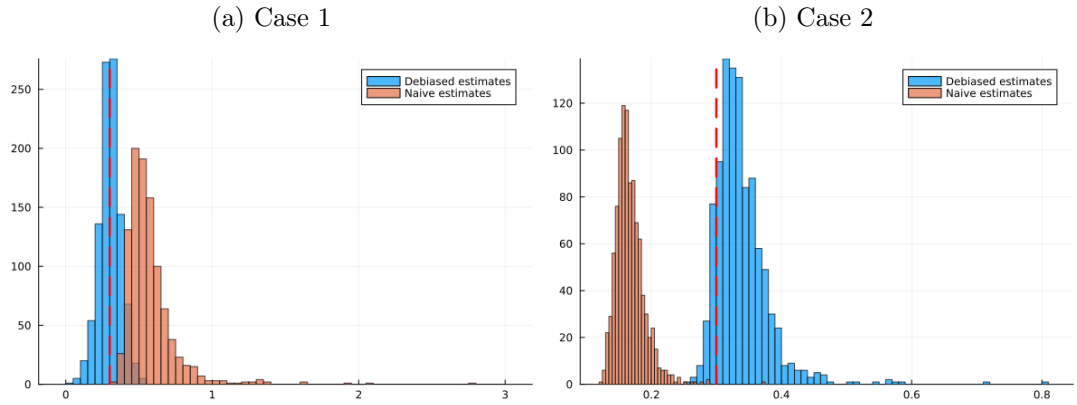
We test the performance of our estimator in Section 3 in finite sample. As in the experiments in the main text, we simulate $N = 1000$ individuals who draw a true degree $d_i \sim U(0,10)$ and are then connected with others uniformly at random from the population.

Our data generating process is

$$y = \lambda G y + x\beta + \epsilon$$

with $\lambda = 0.3, \beta = 0.8$. In all cases, $\epsilon \sim N(0,1)$. We run 1000 simulations per estimator, starting each set with the same random seed. Bias corrected estimators are constructed using the mean missing degree d^B under Assumption 5 for cases 1 and 2 in Section 5.

Figure A8.3: Spillover estimates from nonlinear social network models



Notes: Red line denotes true parameter value. Sampled network in left panel generated by sampling 5 links per agent uniformly at random from their true links. Sampled network in right panel generated by sampling $10 - d_i$ additional links per agent i uniformly at random from the population.

As for linear models, we see that naive two-stage least squares estimators using the sampled network are heavily biased. Our bias-corrected estimators recover the true spillover effect well in finite samples.

A8.3 Copula-based estimator

We assess the performance of an example of this estimator in section 4 in finite sample. As above, we simulate $N = 1000$ individuals who draw a true degree $d_i \sim U(0, 10)$ and are then connected with others uniformly at random from the population.

Each agent draws continuous treatment from the marginal distribution $X_i \sim N(5, 1)$. Marginal distributions of treatment and degree are coupled through a bivariate Gumbel copula

$$C(F_X^{-1}(x), F_D^{-1}(d); \theta) = \exp(-((-\ln F_X^{-1}(x))^\theta + (-\ln F_D^{-1}(d))^\theta)^{\frac{1}{\theta}})$$

where $\theta \in [1, \infty]$ controls the degree of dependence between treatment and degree. We set $\theta = 10$. The left panel of figure A8.4 plots an example joint distribution. Higher treatment nodes have higher degree. Researchers sample networks using a fixed choice design sampling $m = 5$ links per node as in the National Longitudinal Survey of Adolescent Health Data Set. Then

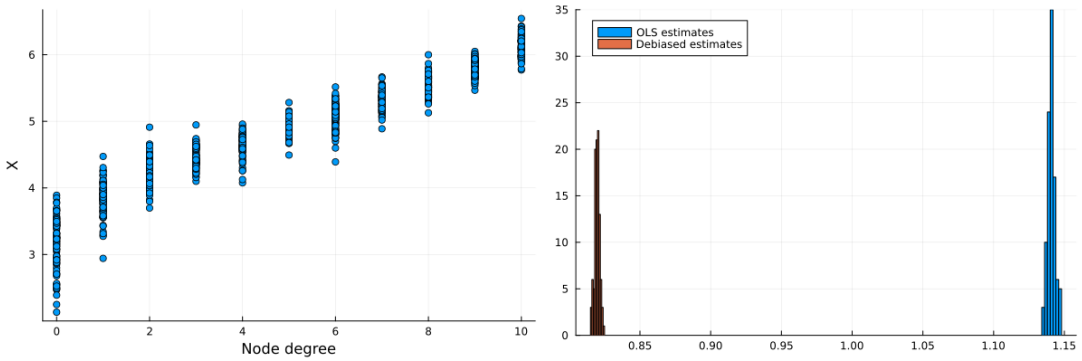
$$\sum_j E(b_{ij}(x_i)|x_j)x_j = \sum_j (E(g_{ij}^*|x_i) - m)\bar{x}.$$

We estimate spillovers using the two-step estimator we describe above. In the first step, we estimate the dependence between treatment and degree by fitting a Gumbel copula by maximum likelihood using only the observations where we correctly sample the network. In the second stage, we then construct a spillover estimate $\hat{\beta}$, constructing BX by sampling from the copula.

Our two-step estimator performs well even though the ordinary least-squares estimator does not. The mean debiased estimate of 0.813 is close to the true spillover value.

Figure A8.4: Spillover estimates when degree depends on treatment

(a) Joint distribution of degree and treatment (b) Spillover estimates



Notes: Red line denotes true parameter value of 0.8. Data is simulated from a linear model on the true network with $N = 1000$. Treatment drawn from marginal $N(5, 1)$, and degree distributed $U(0, 10)$, coupled by a Gumbel copula with $\theta = 10$. Sampled network generated by sampling 5 links per agent uniformly at random from their true links, or all if degree is less than 5.

A9 Peer effects from classrooms

Carrell et al. (2013) estimate the effect of the share of (randomly assigned) high and low ability peers on student GPA at the United States Air Force Academy assuming that all individuals within a peer group (squadron) influence each other equally.

Specifically, each student i is placed within one squadron S_i with 30 other individuals. Denote whether a student has high, middle, or low predicted GPA with the dummies $\{D^H, D^M, D^L\}$, whether they have a high SAT-Verbal score with the dummy x^H, and whether they have a low SAT-Verbal score with the dummy x^L.

The sampled network of peers G is a binary network such that $G_{ij} = 1$ if and only if i and j are in the same squadron. Treatments are the high-ability and low-ability peers in the same squadron $\mathbb{1}(S_i = S_j)x_j^H$, $\mathbb{1}(S_i = S_j)x_j^L$. Students are assigned randomly to squadrons. Therefore sampled spillovers from high-low SAT-Verbal peers are

$$S_i^k = \frac{1}{|\mathcal{S}_i|-1} \sum_j G_{ij} \mathbb{1}_{S_i=S_j} x_j^k$$

for $k \in \{H, L\}$ where normalising by $\frac{1}{|\mathcal{S}_i|-1}$ give the share of that type of peer in the squadron.

Carrell et al. (2013) then estimate spillover coefficients for each predicted-GPA group using the reduced-form regression

$$GPA_i = W\gamma + \sum_l \sum_k D_l S^k \beta_{kl} + \epsilon_i.$$

They use the results to run a treatment where they assign new students to squadrons to maximise the GPA of students with the lowest GPA. Using estimated $\hat{\beta}_{HL}^{OLS}, \hat{\beta}_{LL}^{OLS} = 0.464, 0.065$ predicts a positive average treatment effect

$$\begin{aligned} \Delta S^H \times \beta^{LH} + \Delta S^L \times \beta^{LH} &= 0.0464 + 0.006600 \\ &= 0.053 > 0 \end{aligned}$$

on the students with the lowest GPA, where $\Delta S^H = 0.1, \Delta S^L = 0.1015$. Surprisingly, they instead find a negative treatment effect.

One reason reassignment might have less positive effects than expected is that different types interact with different intensities. For example, students may interact less intensely with students with low SAT verbal scores than implied by their shares in the squadron, and more intensely with students with high SAT verbal scores than their shares in the squadron.

Jackson et al. (2022) survey the network of most important study partnerships between Caltech students, and compute shares of study partners across the GPA distribution. There are 36.28% more study partnerships between students above and below the median on the GPA distribution than implied by their shares in the population. To investigate how sampling of the initial network might affect the Carrell et al. (2013) results, take this as an initial prediction for missing interactions between low predicted GPA and high SAT verbal students.[25] Then, taking values from Tables 1 and 2 in Carrell et al. (2013) gives an estimate of β^{LH} of

$$\begin{aligned}\hat{\beta} &= \frac{0.464}{1 + \frac{\bar{S^H}^2 \times 0.3628}{\text{Var}(S^H)}} \\ &= 0.07709.\end{aligned}$$

Then, the predicted treatment effect would be

$$0.007709 + 0.006600 = 0.01431,$$

a null effect given the forecast standard errors reported in Table 4.

In the paper, they find a negative treatment effect. So, sampling bias cannot entirely rationalise the results. But, it goes a way to explaining how the relatively small amount of endogeneous network adjustment reported in response to treatment could explain the negative result.

A9.1 Calculations from Caltech cohort study

From Jackson et al. (2022), there are an average of 3.5 study partners for male students, and 3.3 for female students. 65.23% of the cohort are male, and 34.77% are female. So, the average number of study partners is

$$3.5 \times 0.6523 + 3.3 \times 0.3477 = 3.43.$$

893 students answered the survey in 2014. Therefore

$$893 \times 3.43 = 3063$$

[25] Note that Carrell et al. (2013) define high, medium, and low in terms of thirds of the distribution. So, these are not directly comparable. Instead, it can be viewed as a best approximation to the level of sampling bias.

study links exist between students. The study network is a simple network. Therefore, there are $\binom{893}{2} = 398278$ possible links. The number of links present per 1000 possible links is therefore

$$\frac{3063}{398278} \times 1000 = 7.69.$$

In Table 4, Jackson et al. (2022) report that there are 2.79 fewer links per 1000 potential links between pairs of students that both have above/below median GPA than pairs of students with GPA on opposite sides of the median. As there are 7.69 links on average, if links were drawn uniformly at random across students there would be

$$\frac{7.69}{2} = 3.845$$

links within and across the GPA categories. The results imply that instead there are

$$3.845 - \frac{2.79}{2} = 2.45$$

links within the GPA categories, and

$$3.845 + \frac{2.79}{2} = 5.24$$

links across the GPA categories. This is

$$\frac{5.24 - 3.845}{3.845} \times 100 = 36.28\%$$

more than implied by the shares in the population.