Estimating Spillover Effects from Sampled Connections*

Kieran Marray[1]

[1]School of Business and Economics and Tinbergen Institute, Vrije Universiteit Amsterdam

July 2025

(Link to most recent version)

Abstract

Empirical researchers often estimate spillover effects by fitting linear or non-linear regression models to sampled network data. We show that common sampling schemes bias these estimates, potentially upwards, 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 US public firms from self-reported supply links, building a new dataset of county-level incidence of large climate shocks.

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

1 Introduction

Empirical researchers measuring spillovers often use data that samples too few or too many links between individuals (Newman, 2010). In economics of education and development economics, researchers often collect network data through surveys asking subjects to name up to a certain number of links (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, researchers often use technological similarity or physical distance as a proxy for whether firms are connected (e.g Jaffe, 1986; Foster and Rosenzweig, 1995; Bloom et al., 2013). When studying supply networks, researchers often observe only large supply relationships between firms (e.g see Atalay et al., 2011; Barrot and Sauvagnat, 2016) or payments from one firm to another collected by a specific bank or credit rating firm (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-September 2024. Out of the 30 papers measuring spillovers, 21 (70%) use such proxies for links between individuals.

*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 usual disclaimer applies.

[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).

A common empirical strategy is to estimate the spillover effect from some shock or treatment by regressing the (weighted) sum of treatments of sampled neighbours on outcomes. Here, we show that sampling links can bias these estimates of spillover effects even in design-based estimation strategies or randomised experiments where treatment is independently distributed across individuals. By sampling links, the researcher can creates an omitted variable – the weighted sum of treatments of neighbours through unobserved links – that covaries with the observed spillover from treatment through the sampling scheme and affects outcomes. A sufficient condition for sampling links to cause bias is that all links have the same sign, and that the sampled network contains either too few or too many links between individuals. This is common in social and economic network datasets (Harris, 2009; Banerjee et al., 2013; Barrot and Sauvagnat, 2016).

Unlike attenuation bias from classical measurement error, estimates can be biased upwards or downwards. We show in simulations that biases can be economically significant. 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 true spillover effects on average.

Sampling too few or too many links is often unavoidable in practice as sampling networks is highly resource intensive (Newman, 2010; Beaman et al., 2021). So, we derive bias-corrected estimators for spillover effects in linear an non-linear social network models that use the sampled network data plus statistics of the true degree distribution (the distribution of the true number of links between individuals). Researchers must adjust estimates to account for the expected dependence between the weighted sum of neighbours' treatments on observed and unobserved links given their sampling rule. When the distribution of treatment does not depend on the network structure, as in a randomised controlled trials or quasi-experimental designs, researchers can construct estimates using the average number of missing links, or the average number of missing links given observed links when the number or strength of missing links depends on the number of strength of reported links. These are aggregate network statistics. Adjustment does not require knowledge of who is linked to whom. So, if the researcher collects network data through surveys, they only need to include one more survey question – "How many friends do you have?". When researchers cannot sample the network themselves, they might use network statistics from studies that survey a specific type of network in detail (e.g see Jackson et al. (2022) for study partnerships at universities, Bacilieri et al. (2023) for firm-level supply relationships) under the assumption that the their network is similar enough. In cases where treatment assignment itself depends upon the network structure, we show how researchers can still construct a bias-corrected estimator by modeling the dependence between treatment assignment and network structure using a copula.

If researchers cannot ascertain the relevant network statistics, we show how they instead can determine the robustness of results to missingness and construct bounds for the true spillover effect given sampled data. Bias-corrected estimators perform well in simulation under common sampling rules, while standard estimators are heavily biased. Researchers can construct standard errors for the bias-corrected estimates using a bootstrap.

For demonstration, we apply our results to estimate the propagation of large climate shocks between US public firms using a popular dataset containing self-reported supply links. There are too few links in the reported the supply network, because firms are only mandated to report customers that make up more than 10% of sales. We construct a new dataset of the county-level incidence of large weather shocks in the US $2004 - 2019$, estimate of the spillover effect using the sampled network, and then bias-correct the estimates using more complete production network statistics from Bacilieri et al. (2023); Herskovic et al. (2020). Bias-corrected estimates are $0.53 - 0.56$ times the standard regression estimates. In the appendix, we additionally show that undersampling the frequency of study partnerships between high and low-ability students can account for part of the difference between estimated and realised peer effects in Carrell et al. (2013).

Our paper relates to a literature on the effect of misspecified networks econometric models (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 paper differs by focusing on the case where outcomes depend on sums of shocks that are distributed independently from links, which allows us to construct bias-corrected estimators without imposing parametric assumptions about the network generation process (e.g Breza et al., 2020; Boucher and Houndetoungan, 2025; Herstad, 2023), or assuming constant link missingness rates (Lewbel et al., 2025). Researchers might worry about conditioning results on such assumptions in practice. By contrast, our bias-corrected estimators only depend upon quantities that researchers can sample directly. (Chandrasekhar and Lewis, 2016) suggests that researchers simply drop observations that might be incorrectly sampled in regression estimators. Our results do not require individuals to drop observations, which is especially useful when researchers do not know which proxies for links might be accurate or not. Our results nest those in Griffith (2022) for the specific case of fixed choice designs, which he analyses in detail. The idea of using additional network data is similar to Lewbel et al. (2023); Zhang (2023). The difference is that we do not require researchers to collect an entire different measure of the same network in detail. 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 in regression estimates.

A related literature on unobserved networks assumes that researchers do not observe any links between individuals, and seeks to estimate the 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). A researcher could take a similar approach when trying to estimate spillover effects on a mismeasured network – first, attempt to measure the true network given the mismeasured network, and then estimate a spillover effect parameter using the estimated network in place of the mismeasured one. Such an approach would require more data – typically a short time series of outcomes per individual – and stronger structural assumptions on the data generating process than we require here (e.g see Battaglini et al., 2021; De Paula et al., 2024). The researcher may also worry that measurement error in the estimated network biases their 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 estimators by simulation. Finally, Section 6 presents our empirical examples. All proofs and additional results are given 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). In 2.1–2.2, we introduce our setting and then show that a regression of sampled spillovers on outcomes gives a biased estimates of spillover effects when unobserved spillovers depend on sampled spillovers through the sampling rule. To illustrate when this happens, we provide a sufficient condition than covers popular sampling rules in studies of social and economic networks.

Then, in 2.3, we introduce a bias-corrected estimator, and show how researchers can construct it in practice. Our main result is that, under independence of treatment from links and plausible assumptions on the sampled network, researchers can construct a bias-corrected estimator using statistics of the degree distribution of the network. Finally, we derive the asymptotic distribution of these estimators and how researchers can use the results to assess robustness of estimators to sampling links. In the subsequent sections, we extend this approach to non-linear estimators and cases where treatment depends on links.

2.1 Setup

Let there be $\mathcal{N} = \{1, ..., N\}$ individuals situated on a simple network $\mathcal{G} = (\mathcal{N}, \mathcal{E}^{\mathcal{G}}, \mathcal{W}^{G})$ where $\mathcal{E}^{\mathcal{G}}$ is the set of edges, and $\mathcal{W}^{G}$ is the set of weights on those edges. We describe these relationships using the $N \times N$ adjacency matrix G, where elements $g_{ij} \in \{0, 1\}$ if the network is unweighted and $g_{ij} \in \mathbb{R}$ if the network is weighted. The adjacency matrix is a single draw from some network-generating distribution F_G that we leave general. Denote the degree of each individual on the network – the total (possibly weighted) connections from all other individuals to i – as $d_i = \sum_j g_{ij}$.

Instead of observing the true network, the researcher samples a set of edges and weights between $\mathcal{N}$ through the non-stochastic sampling rule $S : (\mathcal{E}^G, \mathcal{W}^G) \to (\mathcal{E}^H, \mathcal{W}^H)$ such that $\mathcal{E}^H \cap \mathcal{E}^G \neq \emptyset$. Denote this *sampled network* $\mathcal{H} = (\mathcal{N}, \mathcal{E}^{\mathcal{H}}, \mathcal{W}^H)$. We give three common sampling rules used in research on social and economic networks below.

Example – fixed choice design. The researcher ask each individual to name at most m others that they are connected to. This is common when collecting network data through surveys (Coleman et al., 1957; Calvó-Armengol et al., 2009; Oster and Thornton, 2012; Banerjee et al., 2013; Shakya et al., 2017).

Example – group membership. The researchers split individuals into groups that capture the possible connections between individuals (e.g villages, technology classes, or classrooms). t is assumed that all individuals within the same group are connected. This is common in observational data where researchers can tell which types of individuals might be connected, but not who exactly is connected with whom (e.g Chetty et al., 2011; Bloom et al., 2013; Carrell et al., 2013).

Example – high-weight links. Individuals disclose links where they interact with another with at least some intensity i.e the 'most important' connections. This is common in observational data. For example, US publicly listed firms must disclose customers that make up at least 10% of their sales to the Security and Exchange Commission. Researchers use this to construct a network of supply relationships between the firms (Atalay et al., 2011; Barrot and Sauvagnat, 2016).

We can split the adjacency matrix of the true network into a part H that the researcher samples and a part B that they do not.

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

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

Consider the problem of estimating the casual 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 (which we refer to as spillovers).[3] Results apply directly to functional forms including an intercept, controls, and panel data under the assumption that B is distributed independently from controls Results can also be extended to cases where B is not distributed independently of controls with slight adjustment. For discussion, see see Appendix A.3. Sometimes, a researcher may instead wish to construct a dummy variable denoting whether at least one neighbour is shocked, and regress outcomes on this dummy (e.g Barrot and Sauvagnat, 2016). We derive the bias and equivalent bias-corrected estimators for this case in Appendix A.5.

The researcher only observes the sampled network. So, they only observe the (weighted) sum of treatments of sampled neighbours

[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.

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

as opposed to true neighbours. These sampled spillovers only equal true spillovers for individuals with all links sampled correctly. Deviations between true and sampled spillovers depend on the sampling rule.

Example – fixed choice design. Consider a fixed-choice design where researchers collect at most m links to each individual. If an individual has fewer than m neighbours, the researcher samples all links correctly. But researchers miss some neighbours of higher-degree individuals with more than m neighbours – they only observes m links to these individuals ($i \in \mathcal{B}$ if $d_i > m$). Therefore, sampled spillovers 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 individuals with fewer than m links, but different for individuals with more than m links. In expectation, the difference is non-decreasing in the number of links the individual has.

Example – group membership. Consider a case where individuals are split into classrooms of size m, and researchers assume that all m individuals within each classroom are connected. Researchers incorrectly add more links to with fewer than m neighbours ($i \in \mathcal{B}$ if $d_i < m$). Therefore, sampled spillovers are

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

– equal to true spillovers for individuals with m friends, but more than true spillovers for individuals with fewer than m friends. In expectation, the difference is non-increasing in the number of links the individual has.

Example – high-weight links. Consider a case where researchers only sample links above some weight τ. Unless all links have weight greater than τ, researchers sample fewer links to individuals than they actually have. Therefore, sampled spillovers are

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

We focus on design-based estimation strategies for spillover effects (Borusyak et al., 2024). So, assume that treatment is independently and identically distributed across nodes, and distributed independently of the strength of links.

Assumption 1. We make the following two assumptions on the distribution of treatment x.

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 (the strength of) true and sampled links from the individual to others on the network.

Assumption 1-B is key for our construction of bias-corrected estimators in practice degree distribution in section 2.3. It is a reasonable assumption in many economic applications such as when a researcher directly assigns treatment (e.g Miguel and Kremer, 2004; Oster and Thornton, 2012; Conley and Udry, 2010), natural experiments that assign treatment to some individuals on a network and not others (e.g Barrot and Sauvagnat, 2016; Carvalho et al., 2020), or when individuals interact across a network that is fixed before treatment or by a process other than treatment assignment (e.g Coleman et al., 1957; Calvó-Armengol et al., 2009). It is not reasonable when treatment is targeted by a planner based on network structure, or individuals can endogenously adjust links based on treatment. In section 4, we discuss how to perform our bias correction when this assumption is not appropriate.

Next, assume that Eq. 2 is specified correctly

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

Finally, we make a technical assumption that the expectation of the square of observed spillovers is finite.[4]

Assumption 3. Squared observed spillovers are finite – $E((\sum_j h_{ij}x_j)^2) < \infty$.

These are both weak assumptions that allow us to use regression estimators to estimate spillover effects. For asymptotic results, we assume standard regularity conditions on $\sum_j g_{ij}x_j$ that allow us to apply standard laws of large numbers and central limit theorems for independently but not identically distributed data, and a technical regularity condition on the dependence between observed and unobserved spillovers. We list these at the beginning of appendix A.1 for brevity.

2.2 Sampling generates endogeneity

Next, we show that only using the spillovers on the sampled links can lead to biased estimates of spillover effects. The bias can be upwards or downwards, with the sign depending on the sampling scheme.

Imagine that a researcher attempts to estimate spillover effects by regressing outcomes on sampled spillovers i.e fitting

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

by ordinary least-squares. This regression model is misspecified. By writing the sampled network as the true network minus unobserved links, we can see that this misspecification takes a simple form. By sampling links on the network, the researcher inadvertently creates an omitted variable – spillovers on unobserved links – that enters the error term in the regression

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

We can write the estimator from the regression model

$$\hat{\beta}^{\text{OLS}} = \beta\Big(1 + \frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) + \frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)\epsilon_i}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}. \tag{5}$$

If there is dependence between spillovers on sampled links $\sum_j h_{ij}x_j$ and unobserved links $\sum_j b_{ij}x_j$, then the expectation of $\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}$ is non-zero. Therefore the estimator is biased.

Proposition 1. Make assumptions 1-A, 2, 3. If

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

then

$$E(\hat{\beta}^{\text{OLS}}) = \beta\Big(1 + E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big)\Big) \neq \beta. \tag{6}$$

[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).

Unlike attenuation bias from classical measurement error, estimates can be larger or smaller in magnitude than the true spillover effect. 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 $E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)) < 0$.[5]

An obvious next question is when sampling schemes induce dependence between spillovers on sampled and unobserved links. In the case where all links on the network have the same sign, assumption 1 holds, and the expected treatment is non-zero, a sufficient condition is that the expected number of unobserved links of each individual has the same sign. In other words, the researcher either samples a subset or superset of the true links between individuals.

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

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

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 given above. 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). To develop intuition, we give an extended example with a fixed choice design in A.2.

2.3 Bias-corrected estimators

We can write a bias function for the linear regression estimator (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[6]

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

$$\hat{\beta} = \frac{\hat{\beta}^{\text{OLS}}}{1+\eta} \tag{7}$$

is an unbiased estimator of β i.e $E(\hat{\beta}) = \beta$.

[5] Evaluating the expectation of the ratio using a Taylor expansion gives that $E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) = \frac{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)} + \mathcal{O}(\frac{1}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^4}) \approx \frac{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)}$ (Billingsley, 2012). For details, see the proof of prop. 4.

[6] 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.

To construct the estimator in practice, the researcher needs to approximate η using some $\hat{\eta}$. Here, we offer analytic approximations for η using the Taylor expansion (Billingsley, 2012)

$$E\Big(\frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) = \frac{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)} + \mathcal{O}(\frac{1}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^4}),$$
$$\approx \frac{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))}{E(\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2)}. \quad (8)$$

As the remainder term is $\mathcal{O}(\frac{1}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^4})$, it should be vanishingly small in most cases. Indeed, in Section 5, we see that estimators constructed using this approximation work well in simulation even when N is relatively small. In cases where this remainder term could be large, a researcher could either use a higher order expansion or approximate the full expectation by simulation using a bootstrap.

Next, we connect $\hat{\eta}$ to empirically estimable statistics of the degree distribution. First, assume that the distribution of observed degree is independent of the distribution of unobserved degree amongst nodes with some potentially incorrectly sampled links.

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

This assumption applies to many sampling schemes used in economic research when the underlying network is binary. To illustrate, consider the following examples.

Example – fixed choice design with binary network. If there are sample potentially unsampled links into a node i $i \in \mathcal{B}$, we know that the sampled (in)degree equals the threshold value $d_i^H = m$. Therefore, the distribution of sampled degrees d_i^H given that $i \in \mathcal{B}$ has a point mass at m. 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 – group membership with binary network. Assume that all groups within the space have an equal size m. For all i, the number of sampled neighbours equals one minus the group size $d_i^H = m-1$ by construction. Therefore, the distribution of d_i^H given that $i \in \mathcal{B}$ has a point mass at $m-1$. 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 $i \in \mathcal{B}$.[7]

If no connections are stronger or weaker than others, then we do not need to worry that subjects report underlying connections in some order that might violate this assumption.

Under assumption 4.a, we can characterise the expected dependence 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. Let

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

denote these terms. Then, using Eq. 8

$$\hat{\eta} \approx \frac{\frac{N^B}{N}\hat{d}^H\hat{d}^B\bar{x}^2}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2},$$

[7] 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.

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

$$\hat{\beta} = \frac{\hat{\beta}^{\text{OLS}}}{1+\hat{\eta}} \text{ where } \hat{\eta} = \frac{\frac{N^B}{N}\hat{d}^H \hat{d}^B \bar{x}^2}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2} \tag{9}$$

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 $\hat{\eta}$ only depends 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 a researcher to know which individual each other individual is connected to.

The requirement that researchers obtain mean missing degree relatively mild compared to conditioning on unobservable counterfactual network distributions (Breza et al., 2020; Herstad, 2023; Borusyak and Hull, 2023) or constructing multiple measures if the same network (Lewbel et al., 2023). It is an aggregate network statistics, so relatively easy to collect or estimate. In a survey, the researcher could collect the sampled mean degree and true mean degree of individuals who have at least one potentially unsampled link by including one more question: 'How many of these types of connections do you have?'. As they are aggregate quantities, data providers can easily disclose them while preserving privacy. In cases where the researcher cannot sample individuals in the network – for example when using data collected by others – researchers can plausibly construct the statistics from the mean degree of similar observed networks plus the sampling rule. Researchers could also use additional survey questions on connections to 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.

We can relax assumption 4.a and allow the number or strength of unobserved links depends on the number or strength of observed links (as when individuals name stronger connections first). Instead, we can make the weaker assumption that the number or strength of observed and unobserved links are identically distributed through some conditional distribution.

Assumption 4.b. There exists some 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 – fixed choice design naming stronger connections first Assume that we can describe the distribution of weighted degree as drawn from a single degree distribution F_d, and for simplicity that all weights are positive. The researcher samples up to m links. Individuals name their strongest links first. The strength of each unobserved link must be less than or equal to the lowest strength of the observed links. Otherwise, the individual would have named the missing link before at least one of the links that they did name. So, the degree of individuals with at least one potentially missing link must be such that $0 \leq d_i - d_i^H \leq (N-m)\min\{h_{ij}|h_{ij} > 0\}$. Imagine the weight of any unobserved link was greater than the weakest reported link. Then the individual would have named that link to the researcher in place of the weakest reported link. So, the expected unobserved degree is $E(d_i^B|d_i^H) = E(d_i|d_i \geq (N-m)\min\{h_{ij}|h_{ij} > 0\}) - d_i^H$.

Denote the sample average number of unobserved links amongst an individual with some potentially unsampled links and d^H sampled links as

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

Then, we can instead approximate

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

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 sample size N, $E(\hat{\beta}_N) \approx \beta$ and $\hat{\beta} \xrightarrow{p} \beta$.

As before, constructing unbiased estimates only requires knowing aggregate network statistics, rather that who exactly is linked to whom. The researcher can construct an estimate of the distribution of missing degree given sampled degree using the empirical distribution of (the strength of) total missing links given the sampled links. As under assumption 4.a., they could collect this by adding an additional question to a survey. Data providers could disclose this as it is an aggregated quantity. Researchers could also approximate this from the degree distribution of similar fully sampled networks plus a knowledge of the sampling rule.

2.4 Asymptotic distribution

Next, we characterise the asymptotic distribution of our estimator. First, consider the case where a sample value of η is known, computed from a finite sample of size N. An example of this might be when a surveyor is able to collect the number of unreported friends per individual, or a data provider can disclose it. For sample size N, let

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

Then the bias-corrected estimator is consistent and asymptotically normal.

Proposition 5. Make assumptions 1,2,3, 5, A1. Then

$$\sqrt{N}(\hat{\beta}-\beta) \sim N(0, \frac{1}{(1+\eta)^2}\Omega).$$

This result follows from the usual asymptotic arguments plus the normal product rule (Cameron and Trivedi, 2005). Next, consider the more interesting case where we instead have to estimate $\hat{\eta}(\hat{\theta})$ as a function of some finite vector of parameters θ that we can describe as the solutions to the moment conditions

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

In our examples above, these are mean unobserved degrees. In this case, 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 $\hat{\theta}$.

Proposition 6. Make assumptions 1,2,3, 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 $\mathrm{Var}(\hat{\beta})$. 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 different assumptions on the network sampling process discussed.

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

```
procedure BOOTSTRAP ($d^B$, $H$, $\{\mathcal{S}_i\}_{i=1}^N$, $x$, $y$)
    for $j \in 1, ..., M$ do
        Draw $\{B_{ik}|k \notin \mathcal{S}_i\} s.t \sum_{\{B_{ik}|k \notin \mathcal{S}_i\}} B_{ik} = N\bar{d}^B$.
        Construct $\{\widehat{\beta_{kj}}\}_{k=1}^P$ 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
```

2.5 Robustness

In addition to constructing bias-corrected estimators, the researcher can use Theorem 1 to assess 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 might be decision thresholds for optimal policy, or values needed for test-statistics to pass critical values at preferred significance levels.

Proposition 7. Make assumptions 1-A, 2, 3. Then

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

Second, if they can bound the dependence of observed and unobserved spillovers $\eta \in [\eta_{\min}, \eta_{\max}]$, then they can bound true spillover effects 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 purely on the mean missing degree amongst individuals with at one missing link. In this case, we can rewrite our decision threshold as depending upon the mean number of missing links amongst individuals with at least one missing link

$$\hat{\beta}^{\text{ OLS}} > \tau \text{ if and only if}$$
$$\hat{d}^B < \Big(\frac{\frac{1}{N}\sum_i (\sum_j h_{ij} x_j)^2}{\frac{N^H}{N} \bar{x}^2 \hat{d}^H}\Big) \frac{\hat{\beta}^{\text{ OLS}} - \tau}{\tau}. \tag{10}$$

Spillover effects pass a certain threshold if and only if this quantity is less than some value. Further, 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 certain models that are linear in higher powers of the adjacency matrix. To show this, we consider a parallel non-linear specification common in work on social networks.

3.1 Setup

An alternative model that researchers often use to measure spillover effects is that outcomes are linear in the sum of indirect spillovers across all paths through a network as opposed to just spillovers from direct neighbours (e.g Calvó-Armengol et al., 2009; Carvalho et al., 2020). Formally

$$\begin{aligned} y &= \lambda G y + x\beta + \epsilon \\ &= (I - \lambda G)^{-1}(x\beta + \epsilon). \end{aligned} \tag{11}$$

where $y = (y_1, y_2, ..., y_n)$ is the $N \times 1$ vector stacking individual outcomes, and $x = (x_1, ..., x_n)$ is the $N \times 1$ vector stacking individual treatments.[8] The inverse

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

sums up the spillovers through all the indirect paths across the network.

Here, sampling the network generates a more complicated form of misspecification than the simple linear model given above. Comparing the true paths of length k to sampled paths gives

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

– paths through the sampled network, plus paths through the unobserved links plus spillovers on paths that can created by adding the unobserved links to observed links. So, estimator bias depends on the dependence between treatment that transmits through paths through the sampled network, and treatment that transmits on these additional paths.

Make the standard assumptions (Kelejian and Prucha, 1998; Bramoullé et al., 2009; Blume et al., 2015) that we spell out in Appendix A.7. A researcher tries to estimate structural parameters β, λ – the effect of treatment on outcomes, plus a spillover effect of an individual's outcomes on others through the network – using the sampled network by estimating.

$$y = \lambda H y + x\beta + \epsilon. \tag{12}$$

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). So we focus on this case. 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 projection matrix P_J. The two-stage least squares estimator is

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

The standard instrumental variables estimator for the non-linear social network model is biased.

[8] 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).

Proposition 8. 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 this, write out the reduced-form equation corresponding to the two-stage least squares estimator

$$\begin{aligned} &Y = \lambda(H(I - \lambda H)^{-1} x\beta) + x\beta + \zeta, \text{ where} \\ &\zeta = H(I - \lambda H)^{-1}\epsilon + \lambda By + H(I - \lambda H)^{-1}\lambda B(I - \lambda(H + B))^{-1}(x\beta + \epsilon). \end{aligned}$$

Unless H is orthogonal to B, sampling bias now leads to two sources of endogeneity. The first, which is novel here, comes from the omission of paths through the network in the instrument (the third term in ζ). The second, as in the simple linear model, is the dependence of the Hy term that we are instrumenting with By. Unless H is orthogonal to B, any network-based instrument will correlate with both By and Hy. So constructing correct instruments would not solve this endogeneity problem.

In other terms, we see that the exclusion restriction for the instrument $J = (I - H)^{-1}x$ is that

$$\text{Cov}(H(I - \lambda H)^{-1}x, \zeta) = 0.$$

fails unless H is orthogonal to B.[9]

3.2 Bias-corrected estimator

Using the same logic as for the model where outcomes are linear in spillovers, we can correct for both of these sources of endogeneity. The instrument covaries with two components of η

$$H(I - \lambda H)^{-1}\lambda B(I - \gamma G)^{-1}x\beta, \text{ and } By.$$

So, we can construct an unbiased estimator by constructing instruments that are exogenous to the first component conditional on By. After this, we can apply the bias correction in section 2 to rescale estimates for the dependence between By and the fitted Hy in the second stage. To construct correct instruments, we need to compute the expected number of missing paths through the network. To see how we might do this, 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 probabilities of all links are drawn from a common distribution. A researcher may need to make different assumptions if, for example, some individuals are systematically more popular than others due to some characteristics.

Then, the expected number of walks of a given length through the network given the sampled network depends on the number of walks through the sampled network and powers of the mean number of missing links. 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

[9] 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. For the common sampling schemes listed above, we would not expect this to hold. Indeed, we see large finite-sample biases in simulations of common sampling schemes on networks.

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

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. This allows the researcher to construct unbiased estimators and instruments following the same logic as for the linear models above.

In the example of the non-linear social network model, the researcher can construct the expected treatments through paths of length zero, one, two and so on given the observed network and use these in liu of the missing instruments.

Proposition 9. The variables $J^* = [Hx, d^B Hx, H^2x, ...]$ are valid instruments for Hy conditional on By.

Finally, the researcher still needs to deal with endogeneity from the missing By in the second stage. But now, the problem is linear as in 2. So, the researcher can bias-correct estimates to deal with the omitted term By in the same manner as before.

Proposition 10. Define

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

The estimator

$$\hat{\theta} = (I+\eta)^{-1}\hat{\theta}^{SS} \tag{13}$$

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

We show that this estimator is also consistent and derive the asymptotic distribution in appendix A.7.

4 Extension – treatment dependent on network structure

In some cases, researchers may wish to estimate spillover effects when treatment has been assigned to individuals in a way that depends on the strength of links between them and others on the sampled network i.e assumption 1-B is violated. Example include cases treatment is targeted by a planner based on the network (e.g Galeotti et al., 2020), or where individuals endogenously adjust links based on treatment (for examples, see Calvó-Armengol et al., 2009; Jackson, 2010). Here, we briefly discuss how researchers can construct bias-corrected estimators in these cases by modelling the dependence between treatment status and network statistics through a copula (Nelsen, 2006; Smith, 2003). The benefit of doing this as opposed using estimators that require the researcher to specify counterfactual shock exposure processes (e.g Borusyak and Hull, 2023; Herstad, 2023) is that researchers can flexibly fit copulas from marginal distributions without having a specify a parametric model of the network formation and shock assignment process.

Assume that assumptions 1.A, 2,3 hold and the data is drawn from 2 as before. Then, theorem 1 still gives us a way to construct an unbiased estimator of the spillover effect. But, assumption 1-B is not appropriate. Instead, to compute $\hat{\eta}$, we need a way of modelling the expected treatment of the observed and unobserved neighbours given observed and unobserved links to obtain the expected dependence between observed and unobserved spillovers

$$\frac{1}{N}\sum_i \Big(p(i \in \mathcal{B})E\Big((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)|i \in \mathcal{B}\Big)\Big).$$

To give an example, individuals wih more connections may also have a higher treatment status, making an unobserved link more likely to higher treatment individuals. So, to construct the bias-corrected estimator, we need to model dependence between (G, x). One possible route is to fit a parametric model for the joint distribution of links on the network and treatment. This relates to the approach taken in Borusyak and Hull (2023); Herstad (2023). Instead, we consider the case where the researcher does not have a good parametric model for joint distribution of treatment and links ex-ante. This is likely to be the case in observational data where there is no targeting rule for treatment known to the researcher. However, the researcher can compute (empirical) marginal distributions for a relevant network statistic and treatment from the data. With this, we can flexibly model the joint density of treatment and degree from these marginal distributions using a copula (Nelsen, 2006; Trivedi and Zimmer, 2007). The researcher can then use this joint density to estimate the error correction term η given the dependence parameter of the copula θ.

To illustrate this approach, we consider an example where a planner assigns a continuous treatment x_j across nodes j based on in-degree $d_j = \sum_i g_{ij}$ (e.g Galeotti et al., 2020). This has a natural interpretation in the linear model that the planner wants to assigns treatment to maximise the total spillovers subject to some constraints on who must get some treatment. We assume for simplicity of exposition that all nodes are equally likely to be connected to each other given degree e.g $p(g_{ij} = 1) = \frac{\sum_k g_{kj}}{N}$. Then, we only need to model the dependence between the in degree and treatment.

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. Next, we define a copula.

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) (stated explicitly in the appendix), we can represent the joint density $G()$ using a copula $C(F_X(x), F_D(d), \theta)$. The researcher observes the treatment status of each individual. Given a fitted copula with dependence parameter $\hat{\theta}$, we can compute expected individual 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

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

and therefore $\hat{\eta}(\hat{\theta})$ using Eq 8.

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 that sets the score equal to zero

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

How the researcher might do this exactly depends on 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. 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

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

The quality of estimates depends on the choice of copula and network statistic. For a discussion of the choices researchers could make in this context, see Smith (2003) and references therein.

We can cast this estimator as a two-step M estimator (Newey, 1984) with moment conditions

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

For a given choice of copula, consistency and asymptotic normality follow from results on the asymptotic distribution of parameter estimates from a particular family of copulas (for more details on Archimedian families, see Smith, 2003), and the results in Newey (1984). To assess how well this strategy performs n practice in finite sample, we provide simulation results for estimators constructed using a bivariate Gumbel copula in the appendix A.8. The estimator using the copula performs well in finite sample.

5 Simulation experiments

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

5.1 Simulated networks

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

1. **Fixed choice design**. Each individual draws an in-degree from a discrete uniform distribution $d_i \sim U(1,15)$.[10] We form a binary directed simple network G 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 is placed within a single group (e.g high school class). There are 20 groups containing 25 individuals, 10 groups containing 20 individuals, and 20 groups containing 15 individuals. The researcher samples each individual as being linked to every other individual in their group. Degrees are drawn $U(m_i - k, m_i - 5 - k)$ for group size m_i. We let $k \in \{1,2,3,4,5\}$.

[10] 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).

3. **Link weight thresholds**. First, each individual draws a set of interaction intensities with other individuals from the distribution $w_{ij} \sim \text{LogNormal}(1, 15)$.[11] Then, we construct a weighted network where weight is the proportion of the individual's interactions carried out through that link $g_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$. We then sample links if the value is above 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 weighted directed simple network G connecting each individual with others uniformly at random from the population. The weights on the links are $g_{ij} = \frac{1}{d_i}$ – individuals who have more friends allocate less weight to the each friend that they have. Therefore the reported weighted degree depends on the 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 is placed within a single group (e.g high school class). There are 20 groups containing 25 individuals, 10 groups containing 20 individuals, and 20 groups containing 15 individuals. The researcher samples each individual as being 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. Therefore, the mean number of unobserved connections depends on group size. We let $k \in \{1, 2, 3, 4, 5\}$.

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

1. by regressing spillovers on the sampled network on outcomes (Eq. A-14),

2. the bias-corrected estimator given the true η (Eq. 7), and

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

We run 1000 simulations per estimator, and report the average value of each estimator across each simulation. Additional simulation experiments are given in Appendix A.8.

Below, we compare mean spillover effect estimates within each set of simulations, and plot the distribution of a particular set of estimates from each setting.

In each case, regressing sampled spillover on outcomes yields biased estimates. As expected, estimates are too large when we sample a subset of the true links between individuals (cases one, three, and four) and too small when we sample a superset of the true links between individuals (cases two and five). Bias can be very large. For example, in the case of a fixed choice design sampling at most five links per individual (as friendships within gender are sampled in the popular Ad-Health dataset Harris, 2009), the average estimate of the spillover effect is $1.28 - 1.6$ times the true effect. In the case of thresholding links based on weights with a threshold of 10% of total flows (similar to how supply links are sampled between US public firms Atalay et al., 2011), the average estimate of the spillover effect is 1.63 – double the true effect.

Our bias-corrected estimators perform well at recovering the true spillover effect in finite sample. With η known, estimators are almost always centered on the true spillover value. When we construct $\hat{\eta}$ under assumption 4.a or assumption 4.b, estimators are centered very close to the true spillover value. Bias-corrected estimators perform well in cases that satisfy both assumptions 4.a and 4.b, and particularly well under fixed choice sampling designs (cases 1 and 4).

6 Propagation of climate shocks in production networks

As an example, we use our bias-corrected estimator to measure how extreme weather shocks propagate across supply links between US public firms accounting for the fact that data on supply links between US public firms only contains large supply relationships. In Appendix A.9, we also consider peer effects in education in Carrell et al. (2013).

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

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

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

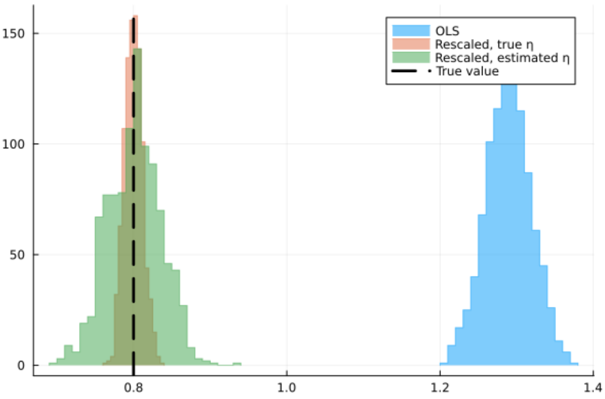


Figure 2: Distribution of spillover estimates using fixed choice design with threshold of 5

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

Figure 3: Mean spillover estimates sampling based on groups, by K

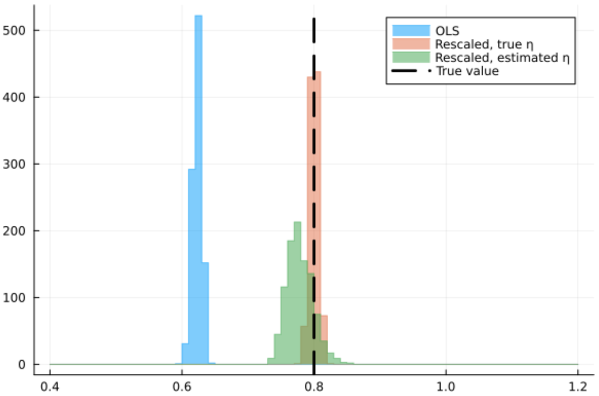


Figure 4: Distribution of spillover estimates sampling based on groups, $k = 3$

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

Figure 5: Spillover estimates using fixed choice design, by threshold

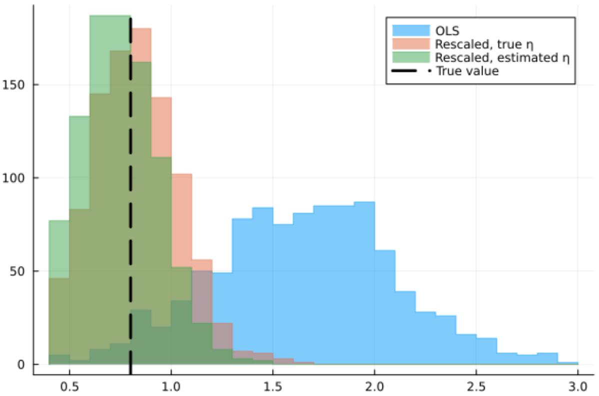


Figure 6: Distribution of spillover estimates using fixed choice design with threshold of 5

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

Figure 7: Mean spillover estimates from a fixed choice design with weights, by number sampled

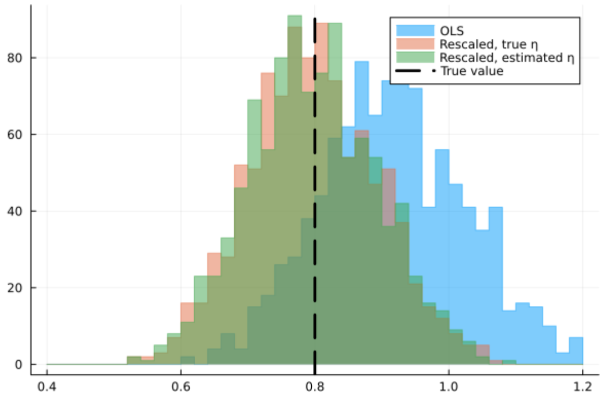


Figure 8: Distribution of spillover estimates using fixed choice design with threshold of 5

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

Figure 9: Mean spillover estimates sampling based on groups when true degree depends on group size, by k

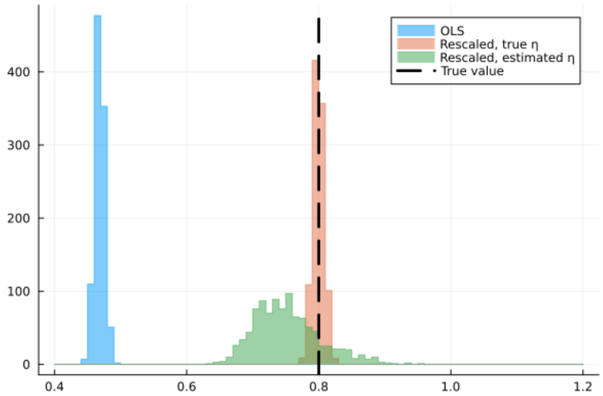


Figure 10: Distribution of spillover estimates using fixed choice design with threshold of 5

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

6.1 Balance-sheet and supply-chain data

Data on supply links between US public firms comes from the popular Compustat Supply Chain dataset (Atalay et al., 2011). Under SFAS regulation No. 131, US public firms have been required since 1997 to report customers making up more than 10% of their sales to investors in 10K forms filed with the Securities and Exchange Commission. They may also self-report other customers, but are not obliged to. The Compustat Supply Chain dataset collects all of these self-reported links between US public firms – which is understood to be a subset of the true supply links between these firms (Herskovic et al., 2020; Bacilieri et al., 2023).[12] The mean number of reported suppliers in 2017 is 1.36. The median is 0.00. This is many fewer than researchers see in complete transactions data.[13] This an example of researchers only sampling high-weight links. As described earlier, this sampling scheme satisfies assumption 4.a.[14] So, we can construct rescaled estimates based on the mean missing degree amongst firms with at least one missing links. So, we construct the mean missing degree using the mean degree of the (more complete) Factset production network in (Bacilieri et al., 2023), and the analysis accounting for truncation in Herskovic et al. (2020). We get values of $2.7, 2.56$, corresponding to a mean missing degree of $d^B = 1.2, 1.34$.

Firm-level balance-sheet information for 1711 public firms headquartered in the United States in 2017 comes from the Compustat Fundamentals Quarterly North America dataset. All continuous variables are winzorized at the 99th and 1st percentiles. A significant fraction of firms move headquarters over time. So, we use the location of firm headquarters reported in their $10K$ forms to locate firms instead of the location reported in Compustat (Gao et al., 2021).

6.2 Climate shocks

To recover large weather shocks to firms, we construct a dataset of the county-level incidence of severe weather events in the United States $2004 - 2019$.[15] Data on weather events comes from the US National Oceanic and Atmospheric Administration Billion-Dollar Weather and Climate Disasters project.[16] This details all weather events that cause over $1 billion in total damages in 2024 dollars between 1980 and 2024. To match each weather event to a list of affected counties, we match each disaster to a dataset of county-level emergency declarations from the Federal Emergency Management Agency. The dataset records all states of emergency declared in response to man-made or natural disasters at the county level. We say that a county is affected by a disaster enough for a firm to experience a shock 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

[12] 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).

[13] 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.

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

[15] The dataset is available on request.

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

Table 1: 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.

Administration. The result is a dataset that records each county affected by a natural disaster that caused at least a billion dollars in total damages by month.

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

As in Barrot and Sauvagnat (2016), we classify a firm being hit by a shock in a given quarter if they are 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. 11.3% of firms have at least one reported supplier hit with a weather shock within the year. 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 weather shocks across firms as independent of the distribution of supply links between firms.

6.3 Estimation

We estimate the effect of an additional shock to 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 Eq. 9.

Table 2: Estimates of propagation of climate shocks between US public firms over 2017

		Δ 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 2 reports results. In line with existing results, the uncorrected estimator suggests that a shock to an additional supplier within the year leads to a 2.48% fall in yearly sales growth. When we perform bias-correction, spillover effects are around $53 - 56\%$ of the intial estimates. We cannot reject the null hypothesis that spillover are zero at standard significance levels. Looking at robustness of the estimates to sampling bias using equation 10 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, we might explain these results by most of these weather shocks lasting for short periods of time. Customers may be able to smooth out these short-term disruptions in supply using inventories. Effects of these shocks were not as large on local economies as larger natural disasters that have well-documented spillover effects down the supply chain (Carvalho et al., 2020).

7 Conclusion

Here, we show that sampling links between individuals can lead to, possibly economically significant, biases in spillover estimates from linear and non-linear models. Unlike classical measurement error, which causes downwards biases, biases can be either upwards or downwards depending on the sampling scheme. In simulations, we show that the sampling schemes used in popular network datasets would induce large biases in estimated spillover effects.

We then introduce bias-corrected estimators for spillover effects, that rescale linear and non-linear regression estimators for the effect of dependence between spillover on the observed and unobserved links. In experimental and quasi-experimental settings, researchers can compute the bias-corrected estimator using only aggregate statistics of the degree distribution that are relatively easy to sample. Our estimators perform well in simulations. To illustrate our results, we construct estimates of the propagation of climate shocks amongst US public firms in 2017 from sampled large 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 complicated structural models to sampled network data (Badev, 2021; Lim, 2024, e.g see) Thus, further work could extend results to moment-based estimators of such models.

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*.

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.

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.*

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. 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 assumptions for asymptotic results

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

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

$$\text{plim}\,\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j) = E((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)),$$

$$\exists\delta > 0 \text{ s.t } E(|\sum_j g_{ij}x_j \sum_j g_{kj}x_j|^{1+\delta}) \leq \infty \; \forall k, i$$

$$\exists\delta > 0 \text{ s.t } E(|\epsilon_i^2|^{1+\delta}) \leq \infty \; \forall k, i$$

$$\exists\delta > 0 \text{ s.t } E(|\epsilon_i^2 \sum_j g_{ij}x_j \sum_j g_{kj}x_j|^{1+\delta}) \leq \infty \; \forall k, i$$

Note that these may fail if the network has a degree distribution that is heavy tailed, like a power-law degree distribution. We do not address this, as this is separate to the focus of this paper.

Proof of proposition 1

Proof.

$$\hat{\beta}^{\text{OLS}} = \beta\Big(1 + \frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j)}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}\Big) + \frac{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)\epsilon_i}{\frac{1}{N}\sum_i(\sum_j h_{ij}x_j)^2}. \tag{A-14}$$

Therefore,

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

Under assumption 2

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

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.

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

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

Proof.

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

□

A1.1 Proofs of propositions 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)

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

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

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

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

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

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,

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

From assumption 4b

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

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

A1.3 Proofs of proposition 5, 6

First, 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

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

Then, applying Slutsky's lemma

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

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 $E((\sum_j h_{ij}x_j)^2), E((\sum_j h_{ij}x_j)(\sum_j b_{ij}x_j))$

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

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

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

Combining these results, we have that

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

Next, consider the second term

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

Under assumptions 1,2

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

Again applying Slutksy's lemma plus assumption A1 gives

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

Combining our two intermediate results by Slutsky's lemma gives

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

□

Given consistency, we now need to derive the asymptotic distribution of the estimator.

Proof. 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

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

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

Proposition 5 is a simple application of the normal product rule to the usual derivation of the asymptotic distribution of the ordinary least-squares estimator. See Cameron and Trivedi (2005) for details.

To prove proposition 6, 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

$$\Omega = \text{Var}(\hat{\beta}) = K_{22}^{-1}(S_{22} + K_{21}K_{11}^{-1}S_{11}K_{11}^{-1}K'_{21} - K_{21}K_{11}^{-1}S_{12} - S_{21}K_{11}^{-1}K'_{21})K_{22}^{-1}, \tag{A-15}$$

and

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

□

Proposition 7 follows by simply rearranging

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

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

Subsequent propositions are proved in appendix A7.

A2 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:

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

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

A3 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

$$y = \begin{pmatrix} Gx \\ W \end{pmatrix}' \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \epsilon. \tag{A-16}$$

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

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

Substituting Eq. A-16 for y ,

$$\hat{\beta}^{\text{OLS}} = \beta + ((\widetilde{H}x)'\widetilde{H}x)^{-1}(\widetilde{H}x)'(\widetilde{B}x\beta + \widetilde{\epsilon}).$$

Taking expectations

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

Therefore the multiplicative bias is

$$E((\widetilde{H}x)'\widetilde{H}x)^{-1}(\widetilde{H}x)'\widetilde{B}x).$$

Equivalents of proposition 1, theorem 1 follow immediately.

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

$$E(((\widetilde{H}x)'\widetilde{H}x)^{-1}(\widetilde{H}x)'\widetilde{B}x) \approx E(((\widetilde{H}x)'\widetilde{H}x)^{-1})E((\widetilde{H}x)'\widetilde{B}x),$$

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

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

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 that, in cases when we can correct spillover estimates from regression models for sampling bias using statistics of the degree distribution of the network.

In certain cases, including controls can lead to $E(\widetilde{B}x) = 0$. In this case, the linear regression estimator is not biased, and correction would be erroneous. A benefit of our approach is that it gives a transparent way to see if this will be the case. An example is a panel data regression with individual fixed effects with constant sampling error in links. In this case

$$\begin{aligned}
\widetilde{B}X_{it} &= (d_{it}^B - \bar{d}_i^B)E(X) \\
&= (d_{it}^B - d_{it}^B)E(X) \\
&= 0
\end{aligned}$$

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

A4 Dummy variable estimators

Again, assume that we can describe the underlying data generating process with Eq. 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[17]

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

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 this dummy on outcomes with an intercept (e.g specifications in Oster and Thornton, 2012; Barrot and Sauvagnat, 2016) [18] 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

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

$\hat{\gamma}$ 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,

[17] Note that this is a different estimand to the spillover effect β, though the two are sometimes conflated (Barrot and Sauvagnat, 2016). With homogeneous effects, $\beta = \frac{\gamma}{E(\sum_j h_{ij}x_j|\sum_j h_{ij}x_j>0)}$. Different degree distributions of the true underlying network can deliver different γ for the same β.

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

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

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

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

□

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.

A5 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.

A6 Asymptotic distribution non-linear social network model

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. $|\lambda| < \frac{1}{||G||}, \frac{1}{||G^*||}$ 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 A6. 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 A6

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

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 A6, 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

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

□

A7 Additional simulations

A7.1 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 44776 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 equation 2 with $\beta = 0.8$, $\epsilon_i \sim N(0, 1)$.

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 A7.1: Mean spillover estimates using fixed choice design, by threshold

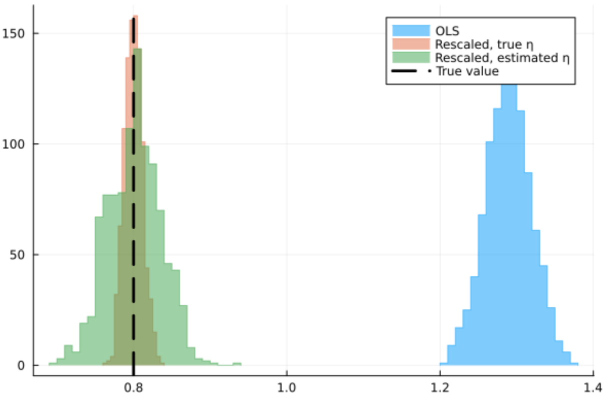


Figure A7.2: Distribution of spillover estimates using fixed choice design with threshold of 3

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.

As in our simulated networks, we see that linear regression on the sampled networks 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.

A7.2 Copula-based estimator

We assess the performance of an example of this estimator in section 3.3 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 A7.3 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

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

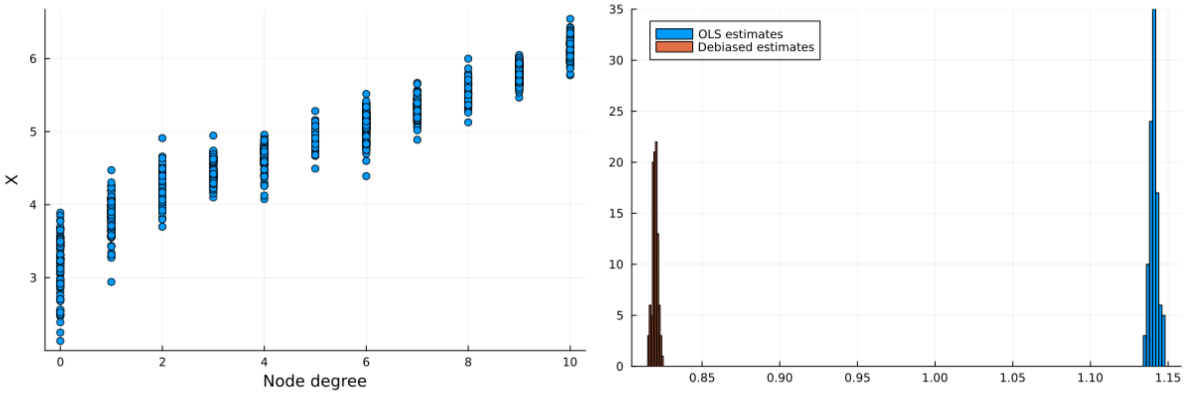
We estimate spillovers using the two-step estimator we describe above. In the first step, we estimate the dependence between treatment and degree by fitting a Gumbel copula by maximum likelihood using only the observations where we correctly sample the network. In the second stage, we then construct 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.

A7.3 Robustness to sampling

In the case where the researcher is unable or unwilling to make assumptions on the marginal distribution, they can recover how large the covariance between observed and unobserved spillovers must be to reduce the estimate below some value. For some threshold $\tau > 0$, rearranging our the formula for debiased estimates gives

Figure A7.3: 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.

$$\hat{\beta}^{\text{ OLS}} > \tau$$

if and only if

$$(GX)'BX < A\frac{\hat{\beta}^{\text{ OLS}} - \tau}{\tau}. \tag{A-17}$$

The sensitivity of estimates depends on both the value of the spillovers on the observed and unobserved components of the network plus the dependence between the two.

A8 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.[19] 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 in response to treatment that they report could explain the negative result.

A8.1 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

$$893 \times 3.43 = 3063$$

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

[19] 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.

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.