Estimating spillovers from sampled connections*

Kieran Marray[1]

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

Submission to 2024 European Summer Meeting of the Econometric Society

Abstract

Empirical papers often measure spillover effects using sampled networks containing too few or too many links between agents. Here, we show that undersampling or oversampling of links biases spillover estimates from ordinary least-squares and spatial autoregressive models, often upwards. Biases can be larger than true spillover effects, and cause hypothesis tests for non-zero spillovers to over-reject the null. We introduce a simple rescaling procedure to construct unbiased estimators from sampled networks. Estimates should be rescaled based on the mean number of missing links. Rescaled estimators are unbiased, and tests based on limiting distributions have a correct size. Without knowing the mean number of missing links, researchers can also bound true effect sizes or determine how many links would have to be missing for debiased spillover estimates to be insignificant. In simulations, debiased estimators perform well even in cases where standard estimators are heavily biased. We apply our results to re-estimate the propagation of climate shocks between US public firms in Barrot and Sauvagnat (2016). Rejection of the null hypothesis of no shock propagation is very sensitive to the number of missing suppliers. Conservative assumptions on the number of missing suppliers suggest that reported point estimates are 3-4 times too large.

Keywords— Networks, spillovers, sampling
JEL Codes: C21

Empirical papers measuring spillovers often sample too few or too many links between the agents in their sample. For lack of an alternative, or because exactly sampling all links between agents is too resource-intensive, researchers assume that the sampled network approximates the true underlying network 'well enough' and construct estimates of spillovers on the true network using the sampled links. For example, Oster and Thornton (2012) ask girls to name at most three friends, and then use these sampled friendships to measure spillovers in adoption of sanitary products. Over two thirds of subjects name the maximum number of friends, suggesting that Oster and Thornton (2012) undersample the number of friends each girl has.

We show that such under-or-over sampling systematically biases spillover estimates from ordinary least squares and spatial autoregressive models. The most common form, undersampling links, leads to large upward biases in the magnitude of estimates of spillovers. Estimates of positive spillovers are too positive; estimates of negative spillovers are too negative. Estimates are biased even when treatment is perfectly randomised and the network structure is strictly exogenous. Test sizes are hugely distorted, as sampling bias also causes hypothesis tests for non-zero spillovers to over-reject the null. Limit distributions are

*We are grateful to Lina Zhang, François Lafond, Stanislav Avdeev, Sander de Vries, Jos van Ommeren and seminar participants at Vrije Universiteit Amsterdam and Tinbergen Institute for comments, and 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.

not centered on zero, and undersampling links also reduces the variance of the limit distribution. The result is confidence intervals that are both incorrectly centered and too tight.

This bias occurs because the sampling distorts the observed spillovers into some agents more than others depending on their true degree. Undersampling reduces the observed spillovers into agents with more links more than those with fewer links. Oversampling increases observed spillovers into agents with fewer links more than those with more links. So the size of the bias depends on the difference between the mean degree of the sampled network and the true network, and the covariance of the treatment and probability of sampling a link. The second is a concern in observational data when weights might be determined endogenously.

To assess the size of bias in practice, we simulate data from networks sampled as in popular datasets. We find that biases are economically significant. Under these common sampling designs, biases are of similar magnitude or larger than the true spillover effect. For example, in one simulation we sample at most five links from each agent as in the popular Add Health dataset (Harris, 2009). The mean ordinary least-squares estimate of the spillover effect is over one and a half times the true spillover effect. The mean spatial autoregressive estimate is nearly double the true spillover effect. Two-sided T tests for non-zero spillovers incorrectly reject the null hypothesis of no spillovers in 96.6% of cases at a 5% significance level.

To remedy this, we introduce a rescaling procedure to construct debiased estimators from sampled network data. Ordinary least-squares estimates should be scaled based on the mean number of missing links. Researchers need to rescale both the second stage estimate and instruments in two-stage least squares estimators for spatial autoregressive models as standard instruments are not valid. If the distribution of treatment across agents is independent from the distribution of links, the researcher only needs to know the mean number of missing links. In observational data where the size of the bias may be correlated with treatment, the researcher needs to estimate the covariance of the treatment with the probability of sampling a true link. The correction does not depend on any assumption about how agents form links, or require fitting any auxillary model for the network. Even when researchers cannot ascertain the mean number of missing links, we show how researchers can use the results to ascertain how many links would need to be missing for the true estimate to fall above or below a certain threshold (e.g to infer that spillovers are non-zero at a preferred significance level), and to construct bounds for unbiased estimates given the sampled data.

Debiased estimators perform well in simulated data sampled as in commonly used datasets. The debiased estimates are tightly centered around true parameter values, despite small sample sizes and standard estimators being heavily biased.

A benefit of our debiasing procedure is that it is easy for researchers to implement in practice. If treatment is independently distributed from links between agents researchers only need to know the mean number of incorrectly sampled links to debias estimates. Collecting such data only requires one more survey question – "how many friends do you have?". The mean number of unobserved links is also an aggregate quantity, so easy for observational data providers to disclose without violating privacy.

We then apply our results to re-estimate the propagation of climate shocks between US public firms. Barrot and Sauvagnat (2016) estimate the effect of a supplier being hit by extreme weather event on the sales growth of their customers using a network built from self-reported large customers. They find that a shock to a supplier reduces subsequent sales growth by 3 percent, as much as the firm itself experiencing the shock. The supply network they use is heavily under-sampled, and Barrot and Sauvagnat (2016) suggest that this will bias estimates downwards. We use auxiliary data on mean number of suppliers between public firms from Herskovic et al. (2020) and Bacilieri et al. (2023), plus the descriptive statistics from Barrot and Sauvagnat (2016), to construct debiased estimates. Our results suggest that undersampling the production network leads to estimates that are $3-4$ times too large. We also assess how robust their finding of non-zero spillovers to the number of suppliers that they are missing. We find that their estimates are highly sensitive to undersampling suppliers. If they are missing more than 0.380 suppliers per firm on average, then the estimates would no longer be significantly different from zero at the 5% level. If they are missing more than 0.552 suppliers per firm on average, then the estimates would no longer be significantly different from zero at the 5% level. These numbers are much lower than best estimates of how many suppliers are missing per firm in the data that they use (Herskovic et al., 2020). In further work, we also assess the bias in prominent studies of spillovers in development

economics and economics of education (e.g Banerjee et al., 2013).

Related econometric literature There is a small existing literature on estimating parameters from sampled networks (Chandrasekhar and Lewis, 2016; Hseih et al., 2018; Lewbel et al., 2022; Yauck, 2022; Zhang, 2023). Most assume that some nodes in the network are unobserved but all links between observed nodes are observed. We instead deal with the more common case when all nodes are observed but some links between these nodes are unobserved. This allows us to correct estimates without fitting a network formation model. Therefore our results apply under less stringent assumptions. Our results for spatial autoregressive models also nest those in Lewbel et al. (2022); Griffith (2022) for the cases mentioned in those papers.

In the treatment evaluation literature, our results are also closely related to Borusyak and Hull (2023), who present a debiased estimator for treatment effects when exposure to treatment is correlated with the error term. We show that mismeasurement of exposure introduces bias even when exposure is not correlated with the error term in the regression.

Related empirical papers Researchers in many areas of applied economics try to estimate spillovers between agents when they cannot collect true links by constructing proxies. In economics of innovation, researchers use technological similarity to proxy connections between inventors and between firms (e.g Jaffe, 1986; Bloom et al., 2013). In economics of education, researchers either ask individuals to name a certain number of friends (e.g Bifulco et al., 2011) or assuming all individuals in a classroom influence each other (e.g Chetty et al., 2011). In development and environmental economics, researchers assume technology diffuses between individuals based on proximity (e.g see Nauze, 2023). Other examples include neighbourhood spillovers in crime (Glaeser et al., 1996), the transmission of climate shocks in production networks (Barrot and Sauvagnat, 2016), and deworming in children (Miguel and Kremer, 2004). Our results suggest these existing estimates are biased, and provide a way to correct for the unavoidable mismeasurement of links.

Our results are also relevant for applied papers using differing exposure of individual agents to a collection of exogenous shocks to identify causal effects (Borusyak and Hull, 2023). Examples include shift-share and market-access instruments. Viewing shock exposure as a spillover on a bipartite network, our results imply that mismeasurement of exposure biases these estimators even if uncorrelated with the value of the shocks. Our results allow researchers to construct unbiased estimators of treatment effects under mismeasurement without knowing the exact measurement error.

Outline In section 1, we introduce undersampling and oversampling of links. Section 2 shows the effect of sampling on estimates of spillovers from linear models, and presents the debiased estimator. Section 3 shows the effect of sampling on estimates of spillovers from nonlinear models, and presents the debiased estimator. In section 4, we assess the size of bias and performance of debiased estimators under common sampling schemes by simulation. Finally, in section 5 we apply our results to re-evaluate the propagation of climate shocks in firm-level production networks.

Notation Throughout, we use the following notation. Y denotes either the $N \times 1$ vector of scalars $(y_1, ..., y_N)$ or some matrix of scalars $(Y_1, ..., Y_N)$ depending on the context. If y, x are scalars, $\frac{y}{x}$ denotes division of y by x. If Y, X are vectors or matrices, then $\frac{Y}{X}$ denotes $X^{-1}Y$. We use this notation to make results consistent for spillovers of a single and multiple variables on the same network. $\mathcal{Y}$ refers to the set $\{y_1, ...\}$. $Y \sim D$ denotes that the entries of Y are distributed according to probability distribution D. We use plim Y to denote the probability limit of Y as $N \to \infty$. We use $\xrightarrow{p}$ to denote convergence in probability, and $\xrightarrow{d}$ to denote convergence in distribution.

1 Setup

Consider $\mathcal{N} = i \in \{1, ..., N\}$ agents are situated on a 'true' weighted simple network $\mathcal{G}^* = (\mathcal{N}, \mathcal{E}^*, \mathcal{W}^*)$, where $\mathcal{E}$ is the set of edges between nodes and $\mathcal{W}$ are corresponding weights.[1] We collect some agent-specific outcomes $\{y_1, y_2, ..., y_N\}$ into the $N \times 1$ vector Y, and some agent-specific covariates $\{X_1, X_2, ..., X_n\}$ stacked into the $N \times K$ matrix X. Describe the network with a (possibly weighted) adjacency matrix G^* s.t $g^*_{ij} \neq 0$ if and only if $(i, j) \in \mathcal{E}^*$. The agents might be children in a classroom, individuals in a village, or firms in the global economy. Links may represent study groups, friendships, or supply relationships.

Instead of the true network, we observe some network $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{W})$ that we can describe with an adjacency matrix G. Either $\mathcal{E} \subset \mathcal{E}^*$ – we *undersample* links – or $\mathcal{E}^* \subset \mathcal{E}$ – we *oversample* links. The links within $\mathcal{E}$ may depend on the elements of $\mathcal{W}^*$ – sampling depends on weights – or not.

To illustrate this, consider collecting the network of friends between children in a classroom. If we get children to name m friends where m is less than the largest number of friends, then we will undersample the network. If we assume that all children in a classroom are friends, then we will oversample the network. If we ask children to name their m closest friends, then sampling depends on weights – we will obtain the highest weighted links first. Each of these are standard approaches to collecting network data.

We can summarise the effect of sampling with a matrix B such that

$$b_{ij} = g^*_{ij} - g_{ij}$$

– the difference between the true network and the sampled network. Equivalently,

$$G^* = G + B,$$

B encodes all the links that actually exist but the researcher does not observe. By definition, we can also write

$$G = G^* - B \tag{1}$$

– the adjacency matrix we observe is the true matrix minus the unobserved component. Being able to decompose the true adjacency matrix into the difference of a part that we observe and a part that we do not plus our models being linear will allow us to nicely characterise the bias in our estimators of spillovers.

A useful quantity will be the mean (weighted) degree of each network which we denote with $d^{(.)}$ e.g

$$d^{G^*} = \frac{1}{N} \sum_i \sum_j g^*_{ij}.$$

We note here that from eq. (1),

$$d^B = d^{G^*} - d^G, \tag{2}$$

The mean degree of the true network plus the mean degree of the sampled network are sufficient to recover the mean degree of the bias matrix.

Oversampling or undersampling by themselves lead to $d^B \neq 0$. If $d^B = 0$, then we end up with classical measurement error in our estimators. This is the case where we over-and-under sample links such that the average total weight of links over-and-under sampled is equal to zero. For example, we might get the identity of m friends incorrect for each student in a binary friendship network.[2]

[1] Throughout, we assume that G^* is undirected unless stated without loss of generality. This assumption simplifies notation, as we do not have to refer to both in-degree and out-degree. All results can be extended to directed networks by considering these separately.

[2] Note that links being included or dropped at random does not lead to classical measurement error unless $p(g_{ij} = 1) = \frac{d^{G^*}}{N}$.

Example of undersampling Consider a case where we record at most M links for each node i. This is a common practice when collecting network data by surveys. Surveyors ask participants some variation of the question 'please name M friends' who satisfy the relevant criterion for a link. For example, in Harris (2009) researchers ask adolescents to nominate up to five male and up to five female friends from a roster of all other individuals within their school. Sampling M friends is an example of undersampling links. If a participant has fewer than five female friends, the researcher will observe all of their female friends. But if a participant has more than five female friends, the researcher will only observe some of their female friends.

For ease, consider the case when the network is binary and undirected.[3] Then, the number of unobserved friends of agent i is

$$d_i^B = \sum_{j=1}^{N} g_{ij}^* - M$$

where $M < \sum_{j=1}^{N} g_{ij}^*$.

In the case where individuals name their 'best friends' – that is friends with the highest weight – then the weighted number of unobserved friends of agent i is

$$d_i^B = \sum_{j=1}^{N} g_{ij}^* - \sum_{k=1}^{M} g_{N-k}^*$$

where g_{N-k}^* is the $N-k$th order statistic of a sample of size N drawn from the weighted degree distribution of $\mathcal{G}^*$. We see that if the relevant network is weighted, then how sampling interacts with weights will affect the mean degree of unobserved links. Later, we will see that this affects the bias of the estimator. For simplicity, we largely leave this issue aside for now.

Example of oversampling Consider a case where the researcher assumes that every agent within some category is connected when only some are. This is common practice in observational data where researchers can tell what type of agents might be connected but not whom is connected with whom exactly. For example, Miguel and Kremer (2004) assume that parasitic worms may spread between all children within a certain geographical area. To give another example, Bloom et al. (2013) assume that there are technology spillovers between all firms that file patents in the same classes, and spillovers from competition between all firms that produce goods within the same primary and secondary industries.

For ease, consider the case when the network is binary and undirected, and there are M agents in the relevant group. Then, the number of unobserved friends of agent i will be

$$d_i^B = \sum_{j=1}^{N} g_{ij}^* - M.$$

where $M > \sum_{j=1}^{N} g_{ij}^*$.

Our structure here does not mean that agents have to be arranged into one single connected component. Agents may be arranged into many smaller groups that are disconnected from each other (e.g classrooms, villages), in which case G^* is block diagonal. Agents may be arranged into many smaller groups where most connections are within groups but some are between groups (e.g classrooms where children have friends in other classes, partially connected villages), in which case G^* is block diagonal with overlapping blocks. Our spillover vector G^*X needs to satisfy the Lindenberg conditions in section A1. This rules out networks that are 'too dense', where mean degree grows too fast relative to N. In this case, both our linear and non-linear estimators of spillovers will fail regardless of sampling.

[3] In the directed case, these will be the agent's out-degree

2 Linear models

Assume that agents' outcomes Y come from the linear data generating process

$$Y = G^* X\beta + \epsilon. \tag{3}$$

An agent's outcome depends on the (possibly weighted) sum of neighbours' treatments X. The true adjacency matrix G^* encodes who is connected to whom, and how much each agent influences each other.

A researcher tries to estimate β by taking the analogue

$$Y = GX\beta + \epsilon \tag{4}$$

where G is the adjacency matrix of the sampled network, and constructing the ordinary least-squares estimator $\hat{\beta}^{\text{OLS}} = ((GX)'GX)^{-1}(GX)'Y$. This a very common empirical strategy in many different areas of applied research. We show that this estimator is biased and inconsistent.

Ordinary least-squares estimators are biased and inconsistent Start by making standard assumptions for ordinary least-squares with stochastic regressors (Cameron and Trivedi, 2005), which we state in section A1. These assumptions allow the researcher to sensibly estimate eq. (4) by ordinary least-squares and characterise the asymptotic behaviour of the estimates using the Markov law of large numbers and Lindenberg-Levy central limit theorem. We make the additional assumption

Assumption 1. $BX \perp \epsilon | G^* X$

– which links are sampled does not depend directly on agents' outcomes. An example where this might fail in practice is if a researcher put more effort into sampling the links of children with higher grades, firms with higher sales, or inventors with more patents, than the links of their lower outcome peers.

$\hat{\beta}^{\text{OLS}}$, is biased and inconsistent.

Proposition 1 (Ordinary least-squares bias)**.** The ordinary least-squares estimator $\hat{\beta}^{\text{OLS}}$ is biased, with a bias of size

$$E(A^{-1}(GX)'BX\beta) \tag{5}$$

where

$$A := (GX)'(GX).$$

Furthermore, $\hat{\beta}^{\text{OLS}}$ is inconsistent with a limiting bias of size

$$\text{plim} A^{-1}((GX)'BX)\beta. \tag{6}$$

The bias in the network depends on the product of the sum of covariates of observed neighbours and the sum of the covariates of unoberserved neighbours. The covariates of the unobserved neighbours function as an omitted variable. To see this, expand out $(GX)'BX\beta$ when we only have one covariate. Then we can write the bias as a scalar where

$$A^{-1}(GX)'BX\beta = \beta \frac{\sum_{i=1}^{N}(\sum_{j=1}^{N} g_{ij}x_j)(\sum_{j=1}^{N} b_{ij}x_j)}{\sum_{i=1}^{N}(\sum_{j=1}^{N} g_{ij}x_j)(\sum_{j=1}^{N} g_{ij}x_j)}.$$

The more the 'missing spillovers' and observed spillovers covary, the more biased the ordinary least-squares estimator will be.

Next, consider the limit distribution of the ordinary least-squares estimator.

Theorem 1. Make assumptions 1 and 2. The ordinary least-squares estimator $\hat{\beta}^{\text{OLS}}$ has the limiting distribution

$$\frac{1}{\sqrt{N}}(\hat{\beta}^{\text{OLS}} - \beta) \xrightarrow{d} \mathcal{N}(\frac{1}{\sqrt{N}} M_G^{-1} M_{GB}\beta, M_{B\Omega B}),$$

where:

$$\begin{aligned} M_G &= \text{plim} N^{-1}(GX)'(GX), \\ M_{GB} &= \text{plim} N^{-1}(GX)'(BX), \text{and} \\ M_{B\Omega B} &= \text{plim} N^{-1}((G^* - B)X)'\Omega((G^* - B)X). \end{aligned}$$

The limit distribution is no centered around zero. Therefore interval estimates for β from $\hat{\beta}^{\text{OLS}}$ will be incorrect, as the interval will not be centered around β. Furthermore, the variance of the estimator can also be too large or small depending on the sampling scheme. If we oversample, $G - G^* \succeq 0$. So the asymptotic variance of the OLS estimator is larger than without sampling. If we undersample, $G - G^* \preceq 0$. So the asymptotic variance is smaller than without sampling. It also follows that standard two-sided T-tests constructed using the OLS estimator will over-reject the null hypothesis of zero spillovers. Under the null hypothesis $\beta = 0$, the incorrect asymptotic variance means that our T-statistic no longer takes the correct distribution.

Debiased estimators Our result motivates a theoretically simple debiasing procedure.

Proposition 2. Define

$$\eta = E(A^{-1}(GX)'BX).$$

The estimator

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

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

This rescaled estimator will also have the correct limiting distribution – that is, the limiting distribution of the ordinary least-squares estimator of β when we observe the true network.

Theorem 2. Consider the debiased estimator $\hat{\beta}$, and make our OLS assumptions. Then

$$\begin{aligned} \text{plim}\hat{\beta} &= \beta \\ \frac{1}{\sqrt{N}}(\hat{\beta} - \beta) &\xrightarrow{d} \mathcal{N}(0, M_{B\Omega B}), \end{aligned}$$

where

$$M_\Omega = \text{plim} N^{-1}(G^*X)'\Omega(G^*X).$$

Now, we try to characterise η. For now, make the independence assumption

Assumption 2. (G^*, B) are independent of X.

This is plausible in cases where treatment is randomly assigned across agents in the network as in Miguel and Kremer (2004), Barrot and Sauvagnat (2016), and other spillover estimates based on real or natural experiments. It may not be plausible in observational data where spillovers give individuals an incentive to form links with others based on their x_j. So we will consider the alternative case later on.

Under the independence assumption, we can rewrite our bias term to show that it depends on the mean degree of the unobserved network.

Proposition 3. Denote: the mean of column k of X as $\bar{X}_k$, the mean degree of the unobserved network as d^B, and the mean degree of the observed network as d^G. Then, the expected bias is

$$A^{-1} \begin{pmatrix} \bar{X}_1^2 \beta_1 \\ ... \\ \bar{X}_k^2 \beta_k \end{pmatrix} N d^G d^B \tag{8}$$

This implies that

$$\eta = A^{-1} \begin{pmatrix} \bar{X}_1^2 \\ ... \\ \bar{X}_k^2 \end{pmatrix} N d^G d^B$$

Therefore, to debias their estimates, the researcher only has to know the mean (weighted) number of links per individual that they do not sample. In a survey, the researcher could ascertain this by asking one more question: 'how many friends do you have?'. Data providers can also easily disclose this quantity, especially when sampling bias is introduced to preserve privacy. The provider might disclose the mean bias alongside the biased network to allow researchers to construct unbiased estimates from the sampled data without knowing exactly which connections are missing.

If the researcher is unable to get a precise estimate of d^B, then this procedure is still useful to assess the robustness of spillover estimates to sampling bias two ways.

First, a researcher can recover how many links they would need to miss to reduce the estimate below some value. For some threshold $\tau > 0$, rearranging our the formula for debiased estimate gives that

$$\begin{aligned} &\hat{\beta}^{\text{ OLS}} > \tau \\ &\text{ if and only if} \\ &d^B < \Big(\frac{1}{NA^{-1}\bar{X}^2 d^G}\Big) \frac{\hat{\beta}^{\text{ OLS}} - \tau}{\tau}. \end{aligned} \tag{9}$$

Researchers might use this to see how many links they would have to be under-or-over sampling for-spillover estimates to still be statistically significant given their preferred significance levels and estimated standard errors.[4]

Alternatively, researchers can instead bound estimates of spillover based on a plausible range of missing links. To get these bounds, the researcher can construct a set $d^B \in [d^B_{\min}, d^B_{\max}]$. Then, the bounds for the estimates are

$$\beta \in \Big[\frac{\hat{\beta}^{\text{ ols}}}{I + \eta(d^B_{\max})}, \frac{\hat{\beta}^{\text{ ols}}}{I + \eta(d^B_{\min})}\Big], \tag{10}$$

where the upper and lower bounds may flip if $\bar{d}^B < 0$. As the mean degree of an unweighted is bounded below by 0 and above by $N-1$, the widest such bounds for spillovers on unweighted networks would be $d^B \in [-d^G, (N-1) - d^G]$. These are the analogue of no assumption bounds (Manski, 1990).

Plausible upper and lower bounds could be obtained from similar networks sources in different contexts when the researcher cannot sample the number of missing links. For example, in section 5 we use reported data from observed firm-level production networks given in Bacilieri et al. (2023) to debias estimates of shock propagation on an undersampled network.

Next, consider the case where assumption 2 fails and g_{ij} may depend upon the covariate x_j. The economic intuition for why this might occur is that agents choose who they are linked with, and differing covariates change incentives for agents to create a link. Students might prefer to study with higher ability peers, for example. Then, to compute η, we need to be able to evaluate the terms $E(GX), E(BX)$. We can expand these out as

[4]Of course, the researcher would have to keep in mind that the standard errors are likely also biased, as noted above.

$$
\begin{aligned}
E(G_{i,:}X) &= E(\sum_j x_j g_{ij}), \\
&= \sum_j E(x_j g_{ij}), \\
&= \sum_j (\text{cov}(x_j, g_{ij}) - E(x_j)E(g_{ij})), \\
&= \sum_j (\text{cov}(x_j, g_{ij})) - \sum_j E(x_j)E(g_{ij})), \\
&= \sum_j (\text{cov}(x_j, g_{ij})) - N\bar{x}\bar{d}^G.
\end{aligned}
$$

Similarly, we can write

$$
E(BX) = \sum_j (\text{cov}(x_j, b_{ij})) - N\bar{x}\bar{d}^B.
$$

Therefore, we can plug this into the expectation of the bias to give

$$
\begin{aligned}
\eta =& A^{-1}\Big(\begin{pmatrix} \bar{X}_1^2\beta_1 \\ ... \\ \bar{X}_k^2\beta_k \end{pmatrix} N^{-1}\bar{d}^G\bar{d}^B - \begin{pmatrix} \sum_j(\text{cov}_N(x_{j1}, g_{ij}))N\bar{x}\bar{d}^B\beta_1 + \sum_j(\text{cov}_N(x_{j1}, b_{ij}))N\bar{x}\bar{d}^G\beta_1 \\ ... \\ \sum_j(\text{cov}_N(x_{jk}, g_{ij}))N\bar{x}\bar{d}^B\beta_k + \sum_j(\text{cov}_N(x_{jk}, b_{ij}))N\bar{x}\bar{d}^G\beta_k \end{pmatrix} \\
&+ \begin{pmatrix} \sum_j(\text{cov}_N(x_{j1}, g_{ij}))\sum_j(\text{cov}_N(x_{j1}, b_{ij}))\beta_1 \\ ... \\ \sum_j(\text{cov}_N(x_{jk}, g_{ij}))\sum_j(\text{cov}_N(x_{jk}, b_{ij}))\beta_k \end{pmatrix} \Big).
\end{aligned}
$$

Now, we get the same bias as before plus the inner product of the covariance of the covariates with the observed and unobserved components of the network minus the product of each covariance and the mean degree of the observed/unobserved network times the mean of the covariate. In this case, our OLS estimates will no longer be unbiased in cases where the sum of the weights of the observed network equals the sum of the weights of the true network. The degree of the bias depends on the covariance of each individual weight with the corresponding node's covariate. As an example, the mean ability of the student that a given student studies with might be higher than the average ability of the classroom (all the students might study with the smartest student). Furthermore, students might weigh advice from their higher ability peers more highly than their lower ability peers.

We have two additional terms compared to before, $\text{cov}_N(x_j, b_{ij})$, and $\text{cov}_N(x_j, g_{ij})$. If we can construct a good estimate of these, then we can construct the debiased estimates as before. Of course, this will depend upon the sampling scheme used to construct the network. We can estimate β using the two-step estimator

1. Construct estimates $\hat{\text{cov}}(x_j, g_{ij})$, $\hat{\text{cov}}(x_j, b_{ij})$, and corresponding estimate of OLS bias $\hat{\eta}$

2. Estimate β as

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

The uncertainty from the estimate of the covariance should also affect the limiting variance of the estimator. In further work, we apply the delta method to characterise the limit distribution of this estimator.

3 Nonlinear models

We can extend our approach to assess the bias in the spatial autoregressive models often used to estimate spillovers in the social networks literature (e.g see Blume et al., 2015, and references therein). We use that we can estimate parameters by two-stage least-squares – applying an ordinary-least-squares estimator twice – to construct debiased estimators.[5] To construct a debiased estimator, we need to rescale our instruments and then rescale our second stage estimator as we do in linear models.

Assume that our data is generated by the spatial autoregressive process

$$Y = \lambda G^* Y + X\beta + \epsilon. \tag{11}$$

An agent's outcome depends on the (possibly weighted) sum of neighbours' outcomes Y. The true adjacency matrix G^* encodes who is connected to whom, and how much each agent influences each other.

A researcher tries to estimate λ, β using the sampled network G. Make the standard assumptions, given in section A3 (Kelejian and Prucha, 1998). The two-stage least-squares estimator of λ, β using the sampled network G is

$$\begin{aligned} Y &= \lambda \hat{GY} + X\beta + \epsilon, \\ \hat{GY} &= X\gamma_1 + GX\gamma_2 + G'GX\gamma_3 + ... + \eta. \end{aligned}$$

Two-stage least-squares estimates are biased and inconsistent

Proposition 4. Let $Z^{\text{2SLS}} = [GY, X]$, $H^{\text{2SLS}} = [X, GX, G'GX, ...]$. The two-stage least-squares estimator

$$\begin{pmatrix} \hat{\lambda}^{\text{ 2SLS}} \\ \hat{\beta}^{\text{ 2SLS}} \end{pmatrix} = (Z^{\text{2SLS'}} P_{H^{\text{2SLS}}} Z^{\text{2SLS}})^{-1} Z^{\text{2SLS}'} P_{H^{\text{2SLS}}} Y$$

is biased and inconsistent.

Here, network sampling generates two sources of bias. Our first stage estimator of $\hat{GY}$ will be biased because instruments constructed only using G will be invalid.

To see this, rewrite our data generating process eq. (11) as

$$Y = \lambda(G + B)Y + X\beta + \epsilon. \tag{12}$$

Treating G as the true network and applying the standard transformation gives

$$Y = (I - \lambda G)^{-1}(X\beta + \epsilon) + (I - \lambda G)^{-1}\lambda BY.$$

Pre-multiplying by G gives the first stage for our standard two-stage least squares estimator

$$GY = G(I - \lambda G)^{-1}(X\beta + \epsilon) + G(I - \lambda G)^{-1}\lambda BY.$$

Substituting in our data generating process for Y gives

$$GY = G(I - \lambda G)^{-1}(X\beta + \epsilon) + G(I - \lambda G)^{-1}\lambda B(I - \gamma G^*)^{-1}(X\beta + \epsilon). \tag{13}$$

We see immediately that the standard instruments $(I - G)^{-1}X$ are correlated with the error term in the first stage regression. Therefore our instrumental variable estimates of β, λ will be biased.

Just using valid instruments would not suffice to produce unbiased and consistent estimates of λ. Without loss of generality, ignore our controls $X\beta$ and consider our second stage estimate

$$\begin{aligned} \hat{\lambda}^{\text{ ss}} &= (\hat{GY}'\hat{GY})^{-1}\hat{GY}'Y \\ &= (\hat{GY}'\hat{GY})^{-1}\hat{GY}'(\lambda GY + \lambda BY + \epsilon) \end{aligned}$$

[5] Consequently, we do not consider a debiasing procedure for the quasi-maximum likelihood estimator.

assuming that we have an unbiased estimate $\hat{GY}$ of GY from our first stage regression. Applying the results we saw in section 2, we see that we still end up with a bias

$$E(\hat{\lambda}^{\text{ss}}) = \lambda + ((GY)'GY)^{-1}(GY)'BY\lambda \tag{14}$$

as before.

Debiased estimators To construct unbiased estimators for λ, β we therefore need to do two things: construct valid instruments for GY and correct the bias in the second stage estimator from the unobserved component of the network. Given that we have valid instruments, we can apply our results in section 2 with $\hat{GY}$ in place of GX.

To construct valid instruments,pre-multiply the true data generating process by G to get

$$\begin{aligned} GY &= G(I - \lambda G^*)^{-1}(X\beta + \epsilon) \\ &= G(I - \lambda(G + B))^{-1}(X\beta + \epsilon). \end{aligned}$$

We see immediately that $G(I-(G+B))^{-1}X$ are valid instruments. We formalise this in a proposition.

Proposition 5. The variables $H = [X, \alpha, GBX, G^2X, ...]$ are valid instruments for GY.

We also have to deal with the omitted term BY in our second stage. We can now apply the same correction as for spillover estimates in linear models

Proposition 6. Define

$$\eta = A^{-1}(\hat{GY})'BY.$$

where $A = (\hat{GY})'(\hat{GY})$ and $\hat{GY}$ is an unbiased estimate of GY. The estimator

$$\hat{\lambda} = \frac{\hat{\lambda}^{\text{SS}}}{I + \eta} \tag{15}$$

is an unbiased estimator of λ. Furthermore, $\hat{\lambda}$ is a consistent estimator of λ.

The result follows from proposition 2. From theorem 2, the resulting estimator has the same limit distribution as the 2SLS estimator if we knew G^*.

4 Simulation results

Next, we evaluate the magnitude of bias induced by oversampling or undersampling links, and the performance of our estimators, by Monte-Carlo simulation. To do so, we simulate different sampling schemes commonly used in empirical work. In each case, the mean of our debiased estimator is close to the true parameter value even when the ordinary least-squares and two-stage least-squares estimators are severely biased.

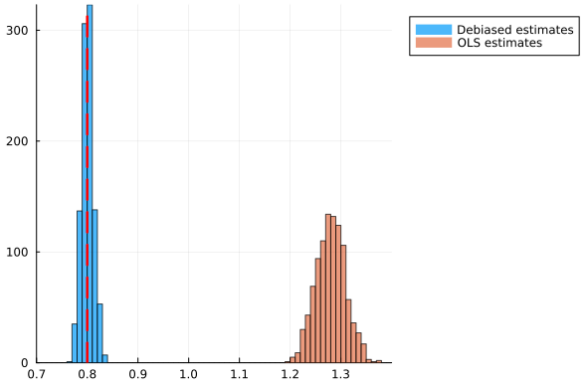
For now, we simulate cases when the sampled network is independent of treatment. Relevant empirical cases include randomised controlled trials with spillovers, and design-based estimates of spillovers. In further work, we also assess the performance of debiased estimators when sampling covaries with treatment.

Setup Throughout, we simulate networks of $N = 1000$ agents, where each agent draws a degree d_i where $D \sim U(0, 10)$ and is connected with d_i other agents uniformly at random from the population.[6]

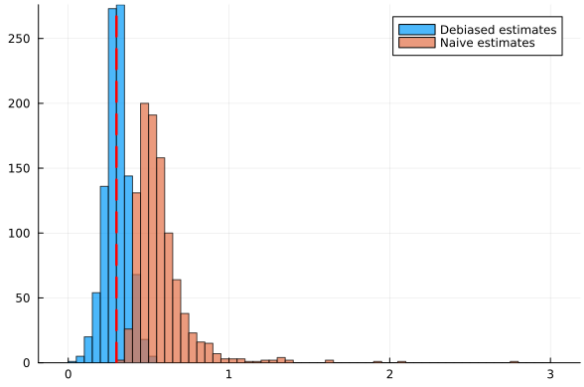
[6]We use a uniform distribution and sample neighbours uniformly at random from the population here to emphasise that the size of the bias that we find is not driven by tail behaviour of the degree distribution or preferential attachment-type mechanisms. Similar results hold when node degrees are sampled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

Figure 1: Spillover estimates from undersampled networks as in Add Health

(a) OLS estimates



(b) SAR estimates



Notes: Red line denotes true parameter values of 0.8 and 0.3 respectively. Data in each case is simulated from a linear/spatial autoregressive model on the true network with $N = 1000$ and single binary treatment drawn i.i.d Bernoulli(0.3) across nodes. The true network has degree distributed $U(0, 10)$ and receiving nodes sampled uniformly at random from the population. Sampled network generated by sampling 5 links per agent uniformly at random from their true links, or all if degree is less than 5.

Agents have some outcome y_i that depends on a binary treatment x_i, that we assign to agents $X \sim \text{B}(0.3)$.[7] For linear models, our true data generating process is

$$Y = G^* X\beta + \epsilon$$

with $\beta = 0.8$. For nonlinear models, our true data generating process is

$$Y = \lambda G^* Y + X\beta + \epsilon$$

with $\lambda = 0.3, \beta = 0.8$. In both cases, $\epsilon \sim N(0, 1)$. We run 1000 simulations per estimator, starting each set with the same random seed. In all cases, debiased estimators are constructed using the mean missing degree d^B and not the true unobserved network B, as researcher would not observe the second in practice.

Case 1 – sampling M links First, we sample the networks as in the Add Health dataset. This is a dataset of friendships between high-school students in the US (Harris, 2009). The dataset it very popular in the literature on social networks (for examples, see Jackson, 2010; Badev, 2021, is a recent example). Surveyors ask students to 'name up to five female friends', and 'name up to five male friends' from a list of all individuals in their school. Similar sampling schemes are also common in other datasets (e.g see Banerjee et al., 2013). To simplify, we will focus on friends of one gender – the case where students 'name up to five friends'.[8] If the agent's true degree is less than or equal to five, we sample all of the agent's links. If the agent's true degree is greater than five, we sample five of their links uniformly at random.

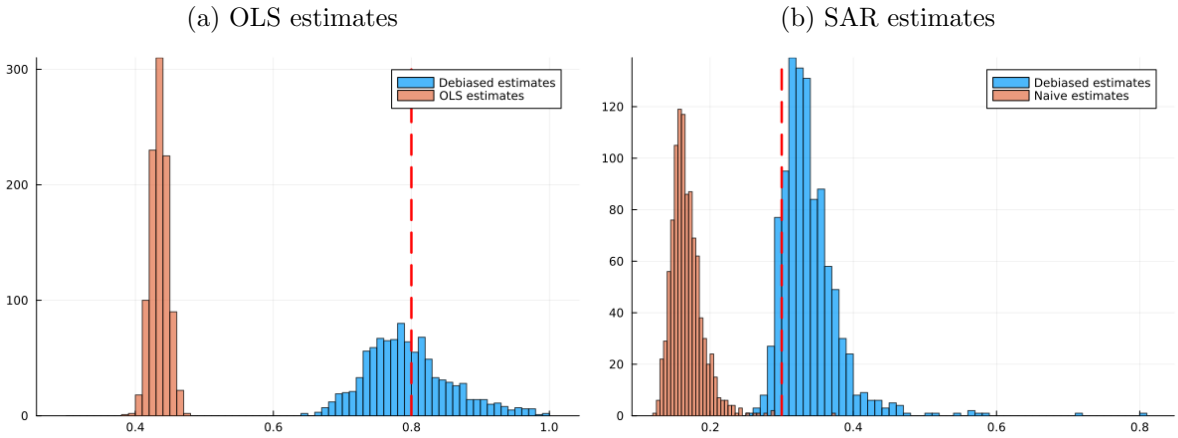
Figure 1 plots the distribution of the estimates from standard and debiased estimators of spillovers for the linear and non-linear models. The standard estimators are heavily upwards biased. The mean ordinary least-squares estimate of 1.29 is over one and a half times the true spillover effect of 0.8. The mean spatial autoregressive estimate of 0.57 is nearly double the true spillover effect of 0.3. The mean debiased estimates, 0.800 and 0.300, are close to the true spillover value and the estimates are tightly centered around it.

Furthermore, undersampling severely distorts the size of tests. We simulate standard hypothesis tests for non-zero spillovers in the linear case under the null of $\beta = 0$. Under null $\beta = 0$, hypothesis tests for $H_0 : \beta = 0$ reject null 96% of the time at 5% significance level.

[7] This is for simplicity. In simulations, similar results hold with multiple treatments that spill over the same network and treatments sampled from continuous distributions.

[8] Given the observed homophily by gender in the dataset, this is not too much of a simplification.

Figure 2: Spillover estimates from oversampled network



Notes: Red line denotes true parameter values of 0.8 and 0.3 respectively. Data in each case is simulated from a linear/spatial autoregressive model on the true network with $N = 1000$ and single binary treatment drawn i.i.d Bernoulli(0.3) across nodes. The true network has degree distributed $U(0, 10)$ and receiving nodes sampled uniformly at random from the population. Sampled network generated by sampling $10 - d_i$ additional links per agent i uniformly at random from the population.

Case 2 – assuming that groups are fully connected Second, we sample networks where we assume that agents are connected to more others than they actually are. This is similar to assuming that everyone in nearby locations is connected, as common in empirical work (e.g see Miguel and Kremer, 2004, or the other papers listed above). We sample each agent's links as if they were connected to ten others. If the agent's true degree is ten, we sample all of the agent's links. If the agent's true degree is less than ten, we sample additional links uniformly at random.[9]

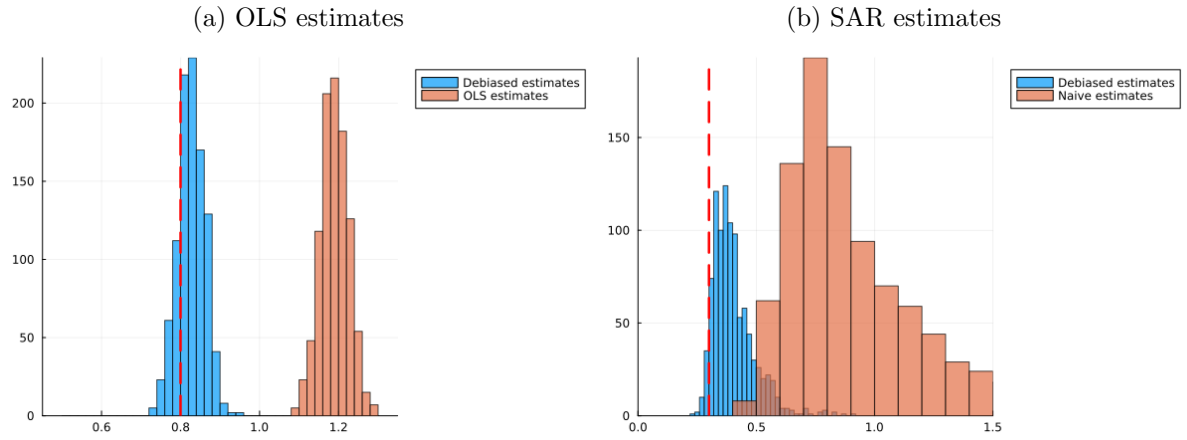
Figure 2 plots the distribution of the estimates from standard and debiased estimators of spillovers for the linear and non-linear models. The standard estimators are heavily downwards biased. The mean ordinary least-squares estimate of 0.438 is approximately half the true spillover effect from that simulation of 0.8. The mean spatial autoregressive estimate of 0.168 is just over half the true spillover effect from that simulation of 0.3. The mean debiased estimates, 0.798 and 0.33, are close to the true spillover value and the estimates are centered around it.

Case 3 – sampling links with highest weights Third, we sample only links between agents that are over a certain weight. This is a common example of undersampling in self-reported network data. Here, we sample the network as in the commonly-used Compustat US firm-level production network dataset. There, firms self-report customers that make up more than ten percent of their total sales. So, we associate each link with a strength that we draw from a LogNormal$(0, 1)$ distribution. We then sample only those links between an agent and others whose strength makes up more than ten percent of the total strength of all the agent's links. When simulating data and estimating parameters we treat both the true and sampled networks as unweighted networks – we only observe whether a link is there, and not its weight.

Figure 3 plots the distribution of the estimates from standard and debiased estimators of spillovers for the linear and non-linear models. The standard estimators are heavily upwards biased. The mean ordinary least-squares estimate of 1.19 is approximately one and a half times the true spillover effect from that simulation of 0.8. The mean spatial autoregressive estimate of 1.2 is four times the true spillover effect from that simulation of 0.3. The mean debiased estimates, 0.827 and 0.4 are close to the true spillover value. Our debiased spatial autoregressive estimators perform worse in this case than in the others, as estimates take longer to converge to the true value in this case than others.

[9]In further work, we simulate data from a sparse stochastic block model where agents connect to others based on a latent location.

Figure 3: Spillover estimates from undersampled networks with undersampling based on weights



Notes: Red line denotes true parameter values of 0.8 and 0.3 respectively. Data in each case is simulated from a linear/spatial autoregressive model on the true network with $N = 1000$ and single binary treatment drawn i.i.d Bernoulli(0.3) across nodes. The true network has degree distributed $U(0, 10)$ and receiving nodes sampled uniformly at random from the population. Sampled network generated as listed in text.

5 Propagation of climate shocks in production networks

Finally, we apply our results to re-evaluate estimates of the propagation of idiosyncratic shocks in firm-level production networks. We use the results in Barrot and Sauvagnat (2016), who estimate how extreme weather shocks to US public firms affect the sales of their customers using self-reported data on major customers (the Compustat production network mentioned above).[10]

Their sample contains 2051 US public firms from 1978–2013. As idiosyncratic shocks, they use major natural disasters that cause damages over $1 billion in 2013 dollars, and last fewer than 30 days. One example is Hurricane Sandy, which hit the eastern US in 2012. A firm is affected by the disaster in a given year if the disaster leads to a FEMA emergency warning in the same county that they are headquartered in.

Barrot and Sauvagnat (2016) estimate how much shocks propagate from customers to suppliers by running the regression

$$\Delta\text{SALES}_{it,t-4} = \alpha + \beta\ \text{SUPPLIER\_HIT}_{it-4} + X_i\gamma + \epsilon_{it},$$

where $\text{SUPPLIER\_HIT}_{it-4}$ is a dummy for whether one supplier of firm i is affected by a natural disaster in quarter $t-4$, $\Delta\text{SALES}_{it,t-4}$ is the sales growth of firm i over the next year, and X_i are multiple controls including individual level fixed effects. Depending on the different control variables that they include, they find that a shock to a supplier leads to a $2-3$ percent fall in sales growth over the subsequent year. The effect size is as large as the effect of firm itself being hit by the shock – a striking finding. We pick the coefficient estimate of -0.031 from Table 5 in their paper as a representative example of the effect that they find.

As firms have on average ≈ 1 supplier in their data, we treat the regression as

$$\Delta\text{SALES}_{it,t-4} = \alpha + \beta \sum_j g_{ij}\ \text{SHOCK}_{jt-4} + X_i\gamma + \epsilon_{it},$$

where $\text{SHOCK}_{j,t-4}$ is a dummy variable for whether firm j is located in a county containing a weather shock in $t-4$.[11]

[10] In the appendix, Barrot and Sauvagnat (2016) discuss measurement error in links, and assert that measurement error in the network will bias their estimates downwards in magnitude compared to the true effect

[11] As very few firms are hit by the shocks in the dataset and the mean degree is very low, this is a good approximation. But note that our estimates lower-bound the bias in the Barrot and Sauvagnat (2016) results, as they also ignore that a firm might have multiple suppliers hit by the same shock.

Network sampling Barrot and Sauvagnat (2016) construct a network of supply relationships between these firms using the firms' self-reported large suppliers from their filings to the SEC. From 1978-1997, firms could self-report customers in filings to the SEC, but had no obligation to do so. Under regulation SFAS 131, issued in 1997, public firms are mandated to report customers that make up more than 10 percent of their total sales to the SEC within their 10K filing. Firms may also report additional customers as before, but they have no obligation to do so.

The self-reported network is heavily under-sampled. The mean number of suppliers is 1.38, with a median of 0.000, many fewer than researchers see in complete transactions data. 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.

To assess how sampling bias affects estimates, we first, we construct debiased estimates based on the results reported in Barrot and Sauvagnat (2016), plus different plausible estimates of the mean number of missing suppliers. Then, we compute how many suppliers per firm Barrot and Sauvagnat (2016) would have to be missing to no longer reject the null hypothesis of no spillovers at standard significance levels if these missing links were included.

Debiased estimates We first construct debiased estimates of the propagation of climate shocks by constructing η from results in the paper and adjusting the reported estimates. We assume that the structure of the network G is independently distributed from the weather shocks. Barrot and Sauvagnat (2016) present evidence that supplier choice does not depend on extreme weather events. So our independence assumption is plausible in this case. Then

$$\eta = E(A^{-1}) p_{\text{shock}}^2 N d^G d^B,$$

and our debiased estimates are

$$\hat{\beta}(d^B) = \frac{-0.031}{1 + \eta(d^B)}.$$

From the descriptive statistics in the paper, we have that: $N = 80,574, p_{\text{shock}} = 0.016, d^G = 1.38$. We construct an estimate of $E(A^{-1})$ using a bootstrap. We simulate networks with degree distributions that match the percentiles of the distribution of number of customers in their dataset as reported in Table 2 of Barrot and Sauvagnat (2016), using the sampler of Clauset et al. (2009).[12] We then compute estimates of A^{-1} from each simulated network. Taking the mean gives $E(A^{-1}) = 0.07$

Table 1: Debiased estimates of the propagation of climate shocks between US public firms

	Barrot and Sauvagnat (2016)	Factset	Herskovic et al. (2020)	Belgium
d^B	0	1.2	1.32	26.27
Estimate	-0.031	- 0.009	-0.0085	-0.0006

We take different plausible values for d^B from three other sources. First, we use the mean degree of firms in the Factset production network dataset (Bacilieri et al., 2023) – a more complete dataset of the supply relationships between the same kinds of large firms that are included in Barrot and Sauvagnat (2016)'s sample. Using the mean degree in this dataset as a proxy for the true mean degree in Barrot and

[12] We use the sampler from Clauset et al. (2009) as it matches known properties of the degree distribution of production networks – see Bacilieri et al. (2023) for more.

Sauvagnat (2016)'s sample gives us a true mean degree of 2.58, and $d^B = 1.2$. Second, we use the implied mean degree from Herskovic et al. (2020). Herskovic et al. (2020) estimate the tail exponents of the true degree distribution amongst US public firms accounting for the censoring induced by the reporting thresholds. Using their estimated tail exponent gives us a true mean degree of 2.69, and $d^B = 1.31$ for a network the size of the sample. Finally, for completeness, we use the implied mean degree from the Belgian production network (Dhyne et al., 2021). The Belgian production network contains all the supply relationships between firms in the country, taken from VAT data. Using the tail exponent of the degree distribution of Belgian firms gives us a true mean degree of 27.65, and $d^B = 26.27$ for a network the size of the sample. This is unlikely to be a good proxy for the true mean degree however, as it includes links between all types of firms and not just between large firms.

Table 1 presents our debiased estimates for each assumption about the mean number of missing links per individual. We see that the original estimates are between three and four times as large as in magnitude as the debased estimates under the plausible assumptions in the second and third column.

Robustness to missing links Next, we test how robust the result of the hypothesis test for non-zero spillovers is to the missingness of links. Barrot and Sauvagnat (2016) report that they can reject the null hypothesis of zero spillovers under hypothesis tests with greater than 1% significance level. As in section 2, rearranging our bias formula for $\beta < 0$ gives

$$\begin{aligned} &\hat{\beta}^{\text{OLS}} < \tau \\ &\text{if and only if} \\ &d^B < \Big(\frac{1}{NA^{-1}p_{\text{shock}}^2 d^G}\Big)\frac{-0.031-\tau}{\tau}. \end{aligned} \tag{16}$$

So we can construct levels of d^B such that estimates will only be 'significant' at the requisite level if they are missing fewer than this number of suppliers per firm on average.

Table 2: Maximum mean number of missing links required to reject null of by significance level

	Reported	1%	5%	10%
Threshold	-0.031	-0.0225	-0.01764	-0.01476
d^B	0	0.190	0.380	0.552

Conservatively, we assume that their estimated standard errors of 0.009 are correct, though our results in section 2 imply that they are overly tight. Then we can construct threshold values for the estimates $\hat{\beta}$ such that their hypothesis tests would not reject the null at significance levels of 1%, 1%, and 10%.

Table 2 gives our results. The magnitude of the spillover estimates is very sensitive to undersampling of suppliers. Adding small numbers of missing suppliers would cause Barrot and Sauvagnat (2016) to no longer reject the null of no spillover effects at standard significance levels. If there are ≈ 0.5 missing suppliers per firm, then the debiased estimate is under half of the reported estimate and we can no longer reject the null of no spillovers at the 10 percent level. So even if the true number of missing links is half of that implied by Factset and Herskovic et al. (2020), the results in table 2 suggest that undersampling suppliers inflates these estimates.

6 Conclusion

We show that oversampling or undersampling connections between agents lead to bias in spillover estimates from linear and non-linear models. Unlike classical measurement error, which causes downwards biases, undersampling can cause large economically significant upwards biases in parameter estimates. Biases can swamp true effects, and cause researchers to incorrectly reject the null of no spillovers in their

hypothesis tests. In simulations, we show that the sampling schemes used in popular network datasets would induce large biases in estimated spillover effects.

We then present debiased estimators for spillover effects from both ordinary least-squares estimators of linear models and two-stage least-squares estimators for nonlinear models. To correct for bias, researchers need an idea of the mean number of missing links per agent. If they cannot ascertain this, researchers can bound the true estimate or work out how many links they would need to miss for estimates to be spuriously significant. Finally, we use our results to characterise how undersampling of the US firm-level production network biases estimates of the propagation of climate shocks between firms.

Further work could explore how sampling bias affects generalised method of moments estimation of structural models. Biases should be larger, because all parameters are sensitive to the errors in a single moment condition.

References

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.

Bifulco, R., Fletcher, J. M., and Ross, S. L. (2011). The effect of classmate characteristics on post-secondary outcomes: Evidence from the add health. *American Economic Journal: Economic Policy*, 3(1).

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.

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

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.

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

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.

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.

Hseih, C.-S., Ko, S., Kovářík, J., and Logan, T. (2018). Non-randomly sampled networks: Biases and corrections.

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

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. (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.

Lewbel, A., Qu, X., and Tang, X. (2022). Estimating Social Network Models with Missing Links. *Mimeo*.

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

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

Nauze, A. L. (2023). Motivation Crowding in Peer Effects: The Effect of Solar Subsidies on Green Power Purchases. *The Review of Economics and Statistics*, 105(6):1465–1480.

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.

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

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

Appendix

Contents

A1 OLS assumptions

Assumption 3 (OLS assumptions)**.** Make the standard OLS assumptions (Cameron and Trivedi, 2005)

1. (Y, G^*, B, X) are independently but not identically distributed over i,
2. $Y = G^*X\beta + \epsilon$,
3. $E(\epsilon|G^*, X) = 0$
4. $E(G^*X_i) = \xi_i$, $V(G^*X_i) = r_i^2$, and $\lim \frac{\sum_{i=1}^N E(|G^*X_i - \xi_i|^{2+\delta})}{(\sum_{i=1}^N r_i^2)^{\frac{2+\delta}{2}}} = 0$ for some $\delta > 2$,
5. $E(BX_i) = \nu_i$, $V(BX_i) = s_i^2$, and $\lim \frac{\sum_{i=1}^N E(|BX_i - \nu_i|^{2+\delta})}{(\sum_{i=1}^N s_i^2)^{\frac{2+\delta}{2}}} = 0$ for some $\delta > 2$,
6. ϵ are idependent 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.
7. $\text{plim}\frac{1}{N}((G^* - B)X)'\epsilon\epsilon'((G^* - B)X)$ exists, is finite, and is positive definite. Additionally, for some $\delta > 0$ $E(|\epsilon_i^2((G^* - B)X)_{ij}((G^* - B)X)_{ik}|^{1+\delta}) < \infty$ for all j, k.

These assumptions allow us to apply the Markov law of large numbers, and the Lindenberg central limit theorem to characterise consistency and asymptotic distributions of our estimators.

Lemma 3. Assumptions 4 and 5 imply that $E(GX_i) = \mu_i$ for some μ_i, $V(GX_i) = \sigma_i^2$ for some σ_i, and $\lim \frac{\sum_{i=1}^N E(|GX_i - \mu_i|^{2+\delta})}{(\sum_{i=1}^N \sigma_i^2)^{\frac{2+\delta}{2}}} = 0$ for some $\delta > 2$.

Proof. By definition, either $\mathcal{G} \subset \mathcal{G}^*$ or $\mathcal{G}^* \subset \mathcal{G}$. From the definition of the adjacency matrix, it follows that either $G^* - G \succeq 0$ and $G_{ij} \leq G_{ij}^*$, or $|B| - G \succeq 0$ and $G_{ij} \leq |B|_{ij}$. All elements of B must also have the same sign. We need to include the absolute value sign in the second condition, because in the case where the second condition holds but the first does not, the value will be negative. In the first case, the condition implies that moments of GX are bounded in absolute value below moments of G^*X. This implies the result. In the second case, the moments of BX being bounded and all elements of B having the same sign implies the moments of $|BX|$ are bounded. Then, the moments of GX are bounded below the moments of $|BX|$. □

We use the further assumption

Assumption 4. $BX \perp \epsilon|G^*X$.

A2 Proofs of results for linear models

$$BX \perp \epsilon | G^*X \implies E(\epsilon | G^*X, BX) = E(\epsilon | G^*X)$$

A2.0.1 Proof of proposition 1 and theorem 1

Proof. Substituting in eq. (1), we get

$$GX = (G^* - B)X.$$

Our OLS estimate is

$$\hat{\beta}^{\text{OLS}} = ((GX)'GX)^{-1}(GX)'Y$$

where

$$Y = G^*X\beta + \epsilon.$$

Expanding and taking expectations gives

$$\begin{aligned} E(\hat{\beta}^{\text{OLS}}) &= E(((GX)'(GX))^{-1}((G^*X)'(G^*X) - (BX)'(G^*X))\beta + ((GX)'(GX))^{-1}((G^* - B)X)'\epsilon) \\ &= E(((GX)'(GX))^{-1}((G^*X)'(G^*X) - (BX)'(G^*X))\beta) + E(((GX)'(GX))^{-1}((G^* - B)X)'\epsilon) \end{aligned}$$

by the linearity of the expectations operator. Next, we prove two lemmas about the second term

Lemma 4. $E(((GX)'(GX))^{-1}((G^* - B)X)'\epsilon) = 0.$

Proof. As standard, under assumption 1.4, 1.5 and lemma 3, we can write

$$\begin{aligned} E(((GX)'(GX))^{-1}((G^* - B)X)'\epsilon) &= E(((GX)'(GX))^{-1}((G^* - B)X)'E(\epsilon|G)) \\ &= E(((GX)'(GX))^{-1}((G^* - B)X)'E(\epsilon|G^* - B)). \end{aligned}$$

Now, assumption 2 implies that

$$\begin{aligned} E(\epsilon|G^* - B) &= E(\epsilon|G^*, B), \\ &= E(\epsilon|G^*) \\ &= 0. \end{aligned}$$

The lemma follows. □

Applying the lemma,

$$E(\hat{\beta}) = E(((GX)'(GX))^{-1}((G^*X)'(G^*X) - (BX)'(G^*X))\beta).$$

Now, separating out the components of the OLS estimate, we have that

$$\begin{aligned} (GX)'(GX) &= \Big((G^*X)'(G^*X) + (BX)'(BX) - (G^*X)'(BX) - (BX)'(G^*X)\Big) \\ &= (G^*X)'(G^*X) + \Gamma), \end{aligned}$$

where

$$\Gamma = (BX)'(BX) - (G^*X)'(BX) - (BX)'(G^*X).$$

Now, making the substitution $(G^*X)'(G^*X)\beta = ((G^*X)'(G^*X) + \Gamma - \Gamma)\beta$ into our expression for $E(\hat{\beta})$ gives the nicer expression

$$
\begin{aligned}
E(\hat{\beta}^{\text{OLS}}) &= \beta + E(((G^*X)'(G^*X) + \Gamma)^{-1}(-\Gamma - (BX)'(G^*X)))\beta \\
&= \beta + E(((G^*X)'(G^*X) + \Gamma)^{-1}((-(B - G^*)X)'BX)\beta) \\
&= \beta(I + E(((G^*X)'(G^*X) + \Gamma)^{-1}((GX)'BX)).
\end{aligned}
$$

We can then write this in the equivalent form

$$
\begin{aligned}
E(\hat{\beta}^{\text{OLS}}) &= \beta(I + E(((GX)'(GX))^{-1}((GX)'BX)), \\
&= \beta + \beta E(((GX)'(GX))^{-1}((GX)'BX)).
\end{aligned}
$$

Next, we need to prove asymptotic bias. To do this, we first prove another lemma

Lemma 5. plim $N^{-1}((G^* - B)X)'\epsilon = 0$.

Proof. Applying lemma 4, $N^{-1}E((G^* - B)X)'\epsilon) = N^{-1}0 = 0$. So, for the lemma, we need to establish sufficient conditions for the Markov law of large numbers for $N^{-1}((G^* - B)X)'\epsilon$. Note that we can rewrite this as $N^{-1}(GX)'\epsilon$. GX and ϵ are independently but not identically distributed from assumptions 1.1 and 1.2. Then lemma 3 plus assumptions 1.1-1.6 are sufficient for us to invoke the Markov law of large numbers as in (Cameron and Trivedi, 2005). □

Write our OLS estimate as

$$
\begin{aligned}
\text{plim}(\hat{\beta}^{\text{OLS}}) &= \text{plim}(((GX)'(GX))^{-1}((G^*X)'(G^*X) - (BX)'(G^*X))\beta + ((GX)'(GX))^{-1}((G^* - B)X)'\epsilon), \\
&= \text{plim}(N^{-1}(GX)'(GX))^{-1}(\text{plim}N^{-1}(GX)'(GX)\beta \\
&+ \text{plim}N^{-1}((GX)'BX))\beta + \text{plim}N^{-1}((G^* - B)X)'\epsilon).
\end{aligned}
$$

where we have applied Slutsky's lemma to separate out the plims. From our assumptions 1.3 and 1.4 plus lemma 3, we have that

$$
\begin{aligned}
&\text{plim } N^{-1}(GX)'(GX) = M_G, \text{ and} \\
&\text{plim } N^{-1}(GX)'(BX) = M_{GB}.
\end{aligned}
$$

Combining with lemma 5, this gives

$$
\text{plim}\hat{\beta}^{\text{OLS}} = \beta + M_G^{-1}M_{GB}\beta.
$$

Next, we need to derive the asymptotic distribution of the OLS estimator. First, we establish the following lemma.

Lemma 6.

$$
\frac{1}{\sqrt{N}}((G^* - B)X)'\epsilon \xrightarrow{d} N(0, M_{B\Omega B})
$$

where

$$
M_{B\Omega B} = \text{plim}N^{-1}((G^* - B)X)'\Omega((G^* - B)X).
$$

Proof. Split the left hand side

By assumptions 1.4 and 1.5 plus lemma 3, the left hand side satisfies the conditions for the Lindenberg-Levy central limit theorem. So we can apply the continuous mapping theorem to write

$$
\frac{1}{\sqrt{N}}((G^* - B)X)'\epsilon \xrightarrow{d} N(0, \text{plim}N^{-1}((G^* - B)X)'\Omega((G^* - B)X)
$$

□

Now, write

$$\frac{1}{\sqrt{N}}(\hat{\beta}^{\text{OLS}} - \beta) = \frac{1}{\sqrt{N}}\beta((GX)'GX)^{-1}(GX)'BX + \frac{1}{\sqrt{N}}((G^* - B)X)'\epsilon.$$

Applying lemma 6 and the continuous mapping theorem gives

$$\frac{1}{\sqrt{N}}(\hat{\beta}^{\text{OLS}} - \beta) \xrightarrow{d} N(\text{plim}\frac{1}{\sqrt{N}}\beta((GX)'GX)^{-1}(GX)'BX, \text{plim}N^{-1}((G^* - B)X)'\Omega((G^* - B)X).$$

Now, repeating the derivation of consistency above and applying Slutsky's theorem for the multiplication by $\frac{1}{\sqrt{N}}$ gives

$$\text{plim}\frac{1}{\sqrt{N}}\beta((GX)'GX)^{-1}(GX)'BX = \frac{1}{\sqrt{N}}M_G^{-1}M_{GB}\beta.$$

Substituting this in gives that

$$\frac{1}{\sqrt{N}}(\hat{\beta}^{\text{OLS}} - \beta) \xrightarrow{d} N(\text{plim}\frac{1}{\sqrt{N}}\frac{1}{\sqrt{N}}M_G^{-1}M_{GB}\beta, \text{plim}N^{-1}((G^* - B)X)'\Omega((G^* - B)X).$$

□

A2.0.2 Proof of theorem 2

Write the OLS estimator if we observed the true network as
$\hat{\beta}^{\text{OLS true}} = ((G^*X)'(G^*X))^{-1}(G^*X)'Y$.
Another way of phrasing our results from theorem 1 above is that

$$\begin{aligned}\hat{\beta}^{\text{ OLS}} &= (I + \eta)\hat{\beta}^{\text{OLS true}},\\ \hat{\beta} &= (I + \eta)^{-1}\hat{\beta}^{\text{ OLS}}\\ &= (I + \eta)^{-1}(I + \eta)\hat{\beta}^{\text{OLS true}}\\ &= I\hat{\beta}^{\text{OLS true}}.\end{aligned}$$

It immediately follows that the limiting distribution of our rescaled estimator is the limiting distribution of the OLS estimator if we observed the true network. Following the standard derivation of the limiting distribution of the OLS estimator e.g in Cameron and Trivedi (2005), this gives the theorem.

A3 Assumptions for nonlinear models

Assume that

1. (Y, G^*, B, X) are independently but not identically distributed over i,

2. $Y = \lambda G^*Y + X\beta + \epsilon$,

3. $E(\epsilon|G^*, X) = 0$

4. item ϵ are idependent 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.

5. The sequence of networks $\{G^*, B\}_N$ are uniformly bounded simple networks.

6. $|\lambda| < \frac{1}{||G||}, \frac{1}{||G^*||}$ for any matrix norm $||.||$.