Estimating Spillovers from Sampled Connections*

Kieran Marray[1]

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

Current draft: June 2024

Abstract

Empirical researchers often estimate spillover effects by fitting linear or non-linear models to network data that contains a subset or superset of links between individuals. Here, we show that such spillover estimates are biased, often upwards. We then derive an unbiased estimator by rescaling estimates based on the mean number of missing links. Our results can be used to bound true effect sizes, determine robustness of estimates to missingness, and construct estimates when missingness depends on treatment. As an application, we re-estimate the propagation of climate shocks between public firms in Barrot and Sauvagnat (2016). Under conservative assumptions, reported point estimates are 3-4 times too large. Rejection of no spillovers is sensitive to small numbers of missing suppliers.

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

1 Introduction

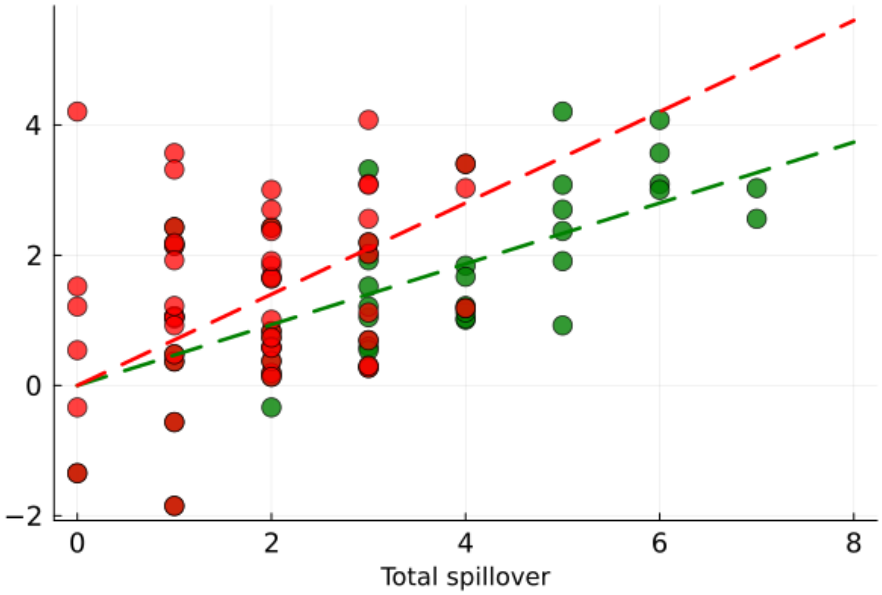
Empirical researchers measuring spillovers often use data that samples too few or too many links between individuals (Newman, 2010). This can occur for multiple reasons. Surveying all possible links in moderately-sized networks can be prohibitively resource intensive (e.g see Beaman et al., 2021). Individuals may only disclose certain links to preserve privacy (Atalay et al., 2011; Barrot and Sauvagnat, 2016). Explicit network data may not exist, forcing researchers to assume that all individuals within certain categories such as geographic location or technology classes are connected (e.g Foster and Rosenzweig, 1995; Munshi, 2003; Miguel and Kremer, 2004; Bloom et al., 2013). Due to these limitations, researchers often estimate spillovers on the true network using a subset or superset of the links.

For example, a common strategy is to use a 'fixed-choice' survey design to elicit relationships between individuals in a sample. Researchers ask subjects to name up to a certain (typically small) number of friends, and estimate spillovers by regressing outcomes on the treatment of reported friends (Calvó-Armengol et al., 2009; Oster and Thornton, 2012; Banerjee et al., 2013). Unless all individuals have fewer friends than the sampling threshold, the researcher only uses a subset of true links. The researcher samples all links of less popular individuals, but only some of the links of more popular individuals.[1]

*We thank Lina Zhang, Xun Tang, François Lafond, Stanislav Avdeev, Sander de Vries, Max Kasy, Jos van Ommeren and seminar participants at the 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]For example, Oster and Thornton (2012) measure spillovers in adoption of sanitary products between schoolgirls who are friends. To sample friendships, they ask girls to name at most three friends who attended the same information session. 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. Addressing potential sampling bias, they state: "In addition, given the randomization, we are able to obtain an unbiased estimate of the impact of additional treatment friends even if we do not observe all of an individual's friends."

Figure 1: Relationship between sampled spillovers and outcomes under fixed choice designs



Notes: Green denotes true spillover values. Red denotes sampled values. Lines fit by ordinary least-squares.

Here, we show that estimates of spillovers from linear and non-linear models fitted to undersampled or oversampled network data are biased, even when treatment is independently and identically distributed across individuals. Estimates can be both upwards and downwards biased, and the biases can be economically significant. When we apply common sampling rules to simulated networks, biases are often of similar magnitude or larger than the true spillover effect. Biases also affect limit distributions of estimators, distorting test sizes. For example, when we apply the fixed-choice sampling rule from the popular National Longitudinal Adolescent Health Data Set (Harris, 2009), mean estimated spillover effects from linear models are over one and a half times true spillover effects. Mean estimated spillover effects from nonlinear models are nearly double true spillover effects. Under the null hypothesis of no spillovers, two-sided T-tests incorrectly reject the null in 96.6% of cases at a 5% significance level.

Sampling a subset or superset of true links results biases estimators by introducing non-classical measurement error into observed spillovers. Undersampling reduces the observed spillovers into agents with more links more than those with fewer links, biasing spillover estimates upwards in magnitude. Figure 1 plots true values (green) and sampled spillover values (red) against an outcome. Undersampling affects the true spillover values associated with higher values of the outcome more than lower values. Oversampling increases observed spillovers into agents with fewer links more than those with more links, biasing spillover estimates downwards in magnitude. Undersampling links also reduces the variance of the limit distribution, result in interval estimates that are both incorrectly centered and too tight.

Next, we show how to rescale standard estimators to construct unbiased spillover estimates from sampled links. Under the assumption that the network structure is exogenous from the distribution of treatment, ordinary least-squares estimates can be rescaled based on the mean number of missing links. Standard instruments for endogenous peer effects in nonlinear models are no longer value. So, unbiased two-stage least squares estimators for non-linear models can be constructed by adjusting instruments and then rescaling second stage estimates. If researchers cannot ascertain the mean number of missing links, we show how they can test the robustness of results to missingness, and construct bounds for the true spillover effect given the sampled data. In observational data where conditional exoogeneity of the network may fail, we present a two-step estimator where researchers model dependence between treatment and degree using a copula. The correction does not depend on any behavioural assumption about how agents form links, or require fitting any auxillary model of network formation. Our rescaled estimators have good finite-sample properties, performing well on our simulated networks.

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. 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 that data providers can disclose without violating privacy. In many cases, researchers have explicitly collected the mean number of actual connections across datasets (e.g see the results in Herskovic et al., 2020; Bacilieri et al., 2023, for firm-level production networks).

We apply our results to re-estimate the propagation of climate shocks between public firms in the United States. 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 severe weather shock to a supplier reduces subsequent sales growth by 3 percent – similar to a shock to the firm itself. 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 1% 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).

1.1 Related econometric literature

There is a small existing literature on estimating parameters from sampled networks (Chandrasekhar and Lewis, 2016; Lewbel et al., 2022; Yauck, 2022; Zhang, 2023; Hseih et al., 2024). Most assume that some nodes in the network are unobserved but all links between observed nodes are observed. To deal with sampling error, they therefore have to condition estimates on a specific network formation model (e.g see Chandrasekhar and Lewis, 2016; Herstad, 2023). In many applied cases, an assumption on the network formation process may be quite strong. 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. Chandrasekhar and Lewis (2016) propose an 'analytical correction' in this case that involves estimating a model only on observations where we observe the true links (Herstad, 2023). This is problematic in cases where most or all nodes have some links that are possibly sampled incorrectly (e.g Miguel and Kremer, 2004; Oster and Thornton, 2012). By placing different restrictions on how links are sampled, we can construct unbiased estimates without removing any observations.

Our estimator is complimentary to the work of Lewbel et al. (2022), who deal with a case where links are missing or included at random with some probability independent of their degree. This more closely matches cases where individuals forget to name, or accidentally name, others in surveys than the systematic sampling error that we consider. Our results also nest those in Griffith (2022) for the specific case of fixed choice designs analysed in that paper.

In the treatment evaluation literature, our results are also closely related to the literature on design based estimation using linear combinations of exposure to exogenous shocks. The closest paper is 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 data generating process.

1.2 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 (Miguel and Kremer, 2004; Bloom et al., 2013; Mas and Moretti, 2004; Nauze, 2023). These proxies are often a subset or super-set of the true links. Our results

suggest that existing spillover estimates on these networks are biased, and provide a way to correct for the unavoidable mismeasurement of links.

In economics of innovation, researchers often use technological similarity or physical distance to proxy connections that lead to diffusion of technologies (e.g Jaffe, 1986; Foster and Rosenzweig, 1995; Bloom et al., 2013). In a seminal study, Coleman et al. (1957) measures the diffusion of knowledge of a new drug amongst physicians in four American cities through their social and professional networks. They ask the physicians "who they most often turn to for advice and information", who did they "most often discuss ... cases", and "the friends, ..., they saw most often socially". Each was asked to name up to three other doctors.

In economics of education, researchers often estimate peer effects between students using fixed choice surveys (e.g Rapoport and Horvath, 1961; Harris, 2009; Calvó-Armengol et al., 2009, and subsequent papers) or assuming all students in a classroom influence each other (e.g Chetty et al., 2011). In related work to ours, Carrell et al. (2013) provides evidence that the second approach leads to mismeasurement of peer effects. They first measure peer effects between freshmen cadets in the United States Air Force Academy assuming that all individuals in the same squadron are friends. Then, they conduct an experiment assigning individuals to squadrons to maximise performance based on the estimated peer effects. The experimental results have the opposite sign to the expected effect. Subsequently, they survey the cadets asking them to name up to five friends, and find out that cadets are only influenced by a subset of the others within their initial squadron.

In development economics, Banerjee et al. (2013) sample social networks in Indian villages by getting participants to name up to between five and eight others who they interact with in specific ways (those who visit the respondent's house, those who the respondent might ask to borrow money and so on). Oster and Thornton (2012) study the diffusion of technology between students by invite girls in several Nepalese schools to a meeting to discuss menstrual cups. After the meeting, each girl at the meeting to name up to three close friends who were also present at the meeting. 68% of girls list three friends. In additional data, they find that the mean number of friends per girl is 3.8. Conley and Udry (2010) measure the diffusion of farming technology by presenting pineapple farmers in Ghana with the names of seven other individuals, randomly chosen without replacement from survey participants in their own village. Farmers name individuals on the list that they have gone to for advice about their farms.

In the literature on the propagation of shocks on production networks, researchers commonly have access to either self-reported large customers from disclosure regulations, or supply relationships constructed from payments between customers of a specific bank. Carvalho et al. (2020) study the propagation of the 2011 Tōhoku tsunami through supply links between Japanese firms. They construct supply relationships between firms using payments through a single bank. Atalay et al. (2011); Barrot and Sauvagnat (2016) construct the supply relationships between public firms in the United States from mandatory disclosure of large customers in 10K forms filed with the SEC.

Other examples include 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).

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 exposure is 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.

1.3 Outline

In section 2, we characterise the effect undersampling and oversampling of links on observed spillovers. section 3 shows the effect of sampling on estimates of spillovers from linear models, and presents the debiased estimator. section 4 shows the effect of sampling on estimates of spillovers from nonlinear models, and presents the debiased estimator. In section 5, we assess the size of bias and performance of debiased estimators under common sampling schemes by simulation. In section 6, we extend our results

to cases where missingness of links may depend on treatment due to endogenous network formation. Finally, in section 7 we apply our results to re-evaluate the propagation of climate shocks in firm-level production networks. All proofs of results are given in the appendix.

1.4 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. $Y_{i,:}$ denotes the ith row of Y. $Y_{:,j}$ denotes the jth column of Y. 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.

2 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.[2] Describe the network with a (possibly weighted) adjacency matrix G^* s.t $g^*_{ij} \neq 0$ if and only if $(i, j) \in \mathcal{E}^*$. Collect some individual-specific outcomes $\{y_1, y_2, ..., y_N\}$ into the $N \times 1$ vector Y, and K individual-specific covariates $\{X_1, X_2, ..., X_n\}$ stacked into the $N \times K$ matrix X. The individuals might be children in a classroom, residents of 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 sampled 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.

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

$$b_{ij} = g^*_{ij} - g_{ij} \tag{1}$$

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

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

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

– the adjacency matrix we observe is the true matrix of connections minus the unobserved component.

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. (2),

$$d^B = d^{G^*} - d^G, \tag{3}$$

The mean degree of the true network plus the mean degree of the sampled network are sufficient to recover the mean degree of the unobserved network B.

The researcher constructs a vector of spillovers

$$GX = G^*X - BX. \tag{4}$$

[2]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.

For one individual i this gives

$$(GX)_i = \begin{cases} G^*_{i,:}X - B_{i,:}X \text{ if } B_{i,:} \neq 0 \\ G^*_{i,:}X \text{ else.} \end{cases}$$

Observed spillovers equal true spillovers if the researcher samples all links between i and others; else they do not.

The error $B_{i,:}X$ reduces to classical measurement error if and only if (Cameron and Trivedi, 2005)

$$E(BX) = 0, (BX)_i \perp (G^*X)_i.$$

Oversampling links gives $B > 0$. Undersampling links gives $B < 0$. Thus, oversampling or undersampling can generate $E(BX) \neq 0, (BX)_i \not\perp (G^*X)_i$ even if X is independently and identically distributed across individuals.

Example of undersampling Consider a case where we record at most m links for each node i, known as a 'fixed choice' study (Newman, 2010). This is a common practice when collecting network data through surveys (Coleman et al., 1957; Calvó-Armengol et al., 2009; Oster and Thornton, 2012; Banerjee et al., 2013). If the maximum number of links per participant is greater than the number of possible choices, fixed choice designs undersample links between participants. If a participant has fewer than m friends, the researcher will observe all of their friends. But if a participant has more than m friends, the researcher will only observe some of their friends. Formally,

$$d_i^B = \begin{cases} \sum_{j=1}^N g^*_{ij} - m \text{ if } \sum_{j=1}^N g^*_{ij} > m, \\ 0 \text{ else.} \end{cases}$$

Therefore, observed spillovers will be lower than true spillovers for nodes with more friends than the sampling threshold, but equal to true spillovers for individuals with equal or fewer links

$$(GX)_i \begin{cases} \neq (G^*X)_i \text{ if } d_i^B > 0, \\ = (G^*X)_i \text{ else.} \end{cases}$$

If X is not centered on zero, this biases observed spillovers of individuals with more links. For example, consider a randomised controlled trial where $X_i \in \{0, 1\}$ is independently and identically distributed across individuals (e.g Oster and Thornton, 2012). Then

$$BX = \begin{cases} G^*X - mp(x_i = 1) \text{ if } d_i^B > 0, \\ 0 \text{ else.} \end{cases}$$

Then, $E(BX) > 0$, and $(BX)_i$ is positively related to $(G^*X)_i$.

Example of oversampling Consider a case where the researcher assumes that every individual within some category is connected. This is common practice in observational data where researchers can tell which types of individuals might be connected but not whom is connected with whom. For example, Miguel and Kremer (2004) assume that parasitic worms may spread between all children within a certain geographical area, but do not know which exact chiildren spread worms to which others.

Let there be m individuals in the relevant group. Then the number of unobserved connections of individual i is

$$d_i^B = \begin{cases} \sum_{j=1}^N g^*_{ij} - m \text{ if } \sum_{j=1}^N g^*_{ij} < m, \\ 0 \text{ else.} \end{cases}$$

Therefore, observed spillovers will be greater than true spillovers for nodes with fewer friends than the whole group, but equal to true spillovers for individuals that are friends with the whole group

$$(GX)_i \begin{cases} \neq (G^*X)_i \text{ if } d_i^B < 0, \\ = (G^*X)_i \text{ else.} \end{cases}$$

Again, consider a randomised controlled trial with a binary treatment $X_i \in \{0, 1\}$ that is independently and identically distributed across individuals (e.g Oster and Thornton, 2012). Then

$$BX = \begin{cases} mp(x_i = 1) - (G^*X)_i \text{ if } d_i^B < 0, \\ 0 \text{ else.} \end{cases}$$

Then, $E(BX) < 0$, and $(BX)_i$ is negatively related to $(G^*X)_i$.

Applications to design-based estimation Many 'design-based' estimators of causal effects construct node-level treatment as a linear combination of a series of exogenous shocks (Borusyak and Hull, 2023; Borusyak et al., 2024)

$$z_i = \sum_j c_{ij} x_j \tag{5}$$

where c_{ij} proxies the exposure of unit i to shock j. A classic example of this is the linear shift-share or 'Bartik' instrument (Bartik, 1991; Autor et al., 2013).

Denoting true exposures as c_{ij}^* and packing the c_{ij}^*, c_{ij} terms into a matrices C^*, C, we have that

$$CX = C^*X - BX.$$

Our results below apply to cases where either $B > 0$, $B < 0$, or $(BX)_i \not\perp (C^*X)_i$. In cases where C, C^* are binary – we construct instruments as sum of exposure to shocks – then we might worry about oversampling or undersamplng as in the network case. In the case where C denotes shares, as in linear shift-share instruments, and true exposures C^* are a matrix of shares, then we might worry that measurement error in exposure weights is correlated with the value of treatment – that is $E(BX) = 0$ but $BX \not\perp C^*X$. In that case, our results in section 6 will apply.

3 Ordinary least-squares estimators

Assume that an individual's outcome Y depends linearly on the (possibly weighted) sum of neighbours' treatments X

$$Y = \alpha + X\gamma + G^*X\beta + \epsilon. \tag{6}$$

A researcher tries to estimate the spillover effect β by taking the analogue

$$Y = \alpha + X\gamma + GX\beta + \epsilon \tag{7}$$

where G is the adjacency matrix of the sampled network. Without loss of generality, assume that $\alpha, \gamma = 0$. So, the researcher estimates β by constructing the ordinary least-squares estimator $\hat{\beta}^{\text{OLS}} = ((GX)'GX)^{-1}(GX)'Y$. 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).

Assumption 1 (OLS assumptions)**.** Assume the following about our data generating process eq. (6)

1. (Y, G^*, B, X) are independently but not identically distributed over i,
2. $E(\epsilon | G^*, X) = 0$

3. $E(G^*X_i) = \xi_i$, $V(G^*X_i) = r_i^2$, 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 $\delta > 2$,

4. $E(BX_i) = \nu_i$, $V(BX_i) = s_i^2$, 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 $\delta > 2$,

5. 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.

6. $\text{plim}\frac{1}{N}((G^* - B)X)'\epsilon\epsilon'((G^* - B)X)$ 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 the researcher to estimate eq. (7) 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.

Note that $3, 4$ rule out networks and sampled networks that are 'too dense' – where mean degree grows too fast relative to N. In this case, spillovers grow explosively with N and both our linear and non-linear estimators of spillovers will fail regardless of sampling. Otherwise, we do not have to impose additional structure on the network.

Also make the assumption

Assumption 2. $BX \perp \epsilon|G^*X$

– which links are sampled does not depend directly on Y. 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.

The ordinary least-squares estimator of spillovers from the analogue eq. (7) 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

$$E(A^{-1}(GX)'BX\beta) \tag{8}$$

where

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

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

$$\text{plim}A^{-1}((GX)'BX)\beta. \tag{9}$$

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, we can rewrite the bias as

$$A^{-1}(GX)'BX\beta = \beta\frac{\text{Cov}(GX, BX)}{\text{Var}(GX)}.$$

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

Bias also affects the limit distribution of the estimate

Theorem 1. Make assumption 1 and assumption 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 not 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. To see this, note that the correct limiting variance is

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

If we oversample, $G^* - B \leq G^*$. So the asymptotic variance of the estimator is too large. If we undersample, $G^* - B \geq G^*$. So the asymptotic variance is of the estimator is too small. It also follows that standard two-sided z/t-tests constructed using the ordinary least-squares estimator will over-reject the null hypothesis of zero spillovers.

For example, our z-test statistic

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

instead of $\mathcal{N}(0, 1)$ as usual.

Debiased estimators Our result motivates a simple theoretical debiasing procedure.

Proposition 2. Define

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

Make assumption 1 and assumption 2. The estimator

$$\hat{\beta} = \frac{\hat{\beta}^{\text{ OLS}}}{I + \eta} \tag{10}$$

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 assumption 1 and assumption 2. 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

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

The researcher does not observe B. So, to implement the debaising procedure, the researcher needs a way to characterise η without observing B. For now, make the independence assumption

Assumption 3. (G^*, B) are independent of X.

This is plausible in cases where treatment is (conditionally) randomly assigned across agents in the network as in real or natural experiments (e.g Miguel and Kremer, 2004; Oster and Thornton, 2012; Barrot and Sauvagnat, 2016). It may not be plausible in observational data where spillovers give individuals an incentive to form links with others based on their x_j. We consider this case in section 6.

Under assumption 3, η only depends on the mean number of missing links.

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

$$E(\hat{\beta}^{\text{OLS}} - \beta) = A^{-1} \begin{pmatrix} \bar{X}_1^2 \beta_1 \\ ... \\ \bar{X}_k^2 \beta_k \end{pmatrix} N d^G d^B \tag{11}$$

This implies that

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

To construct unbiased estimates of β from $\hat{\beta}^{\text{OLS}}$, the researcher only has to know the mean (weighted) number of links per individual that they do not sample. This quantity is relatively easy to ascertain. In a survey, the researcher could include 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.

Robustness to undersampling/oversampling If the researcher is unable to get a precise estimate of d^B, the researcher can still assess robustness of spillover estimates to sampling bias two ways.

First, the researcher can recover the mean number of missing links needed to reduce the estimate below some value. For some threshold $\tau > 0$, rearranging 10) and substituting in

$$\hat{\beta}^{\text{ OLS}} > \tau$$
$$\text{if and only if}$$
$$d^B < \Big(\frac{1}{N A^{-1} \bar{X}^2 d^G}\Big) \frac{\hat{\beta}^{\text{ OLS}} - \tau}{\tau}. \tag{12}$$

Researchers can use this to see how many links per individual would have to be erroneously missing/included for spillover estimates to still be statistically significant given their preferred significance levels and estimated standard errors.[3]

Second, researchers can bound spillover based on a plausible range $d^B \in [d^B_{\min}, d^B_{\max}]$. Then, for $d^B > 0$, the true spillover estimate is contained in the range

$$\beta \in \Big[\frac{\hat{\beta}^{\text{ ols}}}{I + \eta(d^B_{\max})}, \frac{\hat{\beta}^{\text{ ols}}}{I + \eta(d^B_{\min})}\Big], \tag{13}$$

where the upper and lower bounds may flip if $\bar{d^B} < 0$. As the mean degree of an unweighted simple network 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 observed networks. For example, in section 7 we use reported data from different observed firm-level production networks given in Bacilieri et al. (2023) to debias estimates of shock propagation on the firm-level production network between public firms in the United States.

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

4 Nonlinear estimators

We can extend the approach in section 3 to the nonlinear social network models often used to estimate spillovers in the peer effects literature (e.g see Blume et al., 2015, and references therein). Specifically, we can construct debiased estimators of spillover effects by constructing two-stage least-squares estimates with recaled instruments, and then adjusting these estimates.[4]

Assume that each individual's outcome depends on a linear combination of the outcome of their neighbours

$$Y = \lambda G^* Y + X\beta + \epsilon. \tag{14}$$

Without loss of generality, we focus on the case without contextual effects G^*X 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).

A researcher tries to estimate λ, β using the sampled network G. Make the standard assumptions (Kelejian and Prucha, 1998; Bramoullé et al., 2009; Blume et al., 2015).

Assumption 4 (SAR assumptions)**.** 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. The sequence of networks $\{G^*, B\}_N$ are uniformly bounded simple networks.
5. $|\lambda| < \frac{1}{||G||}, \frac{1}{||G^*||}$ for any matrix norm $||.||$.

The standard approach is to estimate λ, β by two-stage least-squares using sampled friends of sampled friends as instruments. The first stage is

$$GY = X\gamma_1 + GX\gamma_2 + G'GX\gamma_3 + ... + \eta.$$

In the second stage, they then estimate

$$Y = \lambda \hat{GY} + X\beta + \epsilon,$$

where $\hat{GY}$ is the fitted value from the first stage.

Two-stage least-squares estimates are biased and inconsistent Undersampling or oversampling links between individuals causes two-stage least squares estimators of λ, β to be biased and inconsistent.

The first stage estimator of $\hat{GY}$ is biased because instruments constructed only using G are invalid.

Proposition 4. The instrument set $H^{\text{2SLS}} = [X, GX, G'GX, ...]$ constructed from the sampled adjacency matrix G are not valid instrument for $Z^{\text{2SLS}} = [GY, X]$.

We see this from expanding out our first-stage equation

$$GY = G(I - \lambda G)^{-1}(X\beta + \epsilon) + G(I - \lambda G)^{-1}\lambda B(I - \gamma G^*)^{-1}(X\beta + \epsilon). \tag{15}$$

The sampled friend of sampled friend 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.

[4] We leave the equivalent procedure for the quasi-maximum likelihood estimator to further research.

Just using valid instruments does 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

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

assuming that $\hat{GY}$ is an unbiased estimate of GY from our first stage. Applying the results we saw in section 3, we see that we still end up with a bias

$$E(\hat{\lambda}^{\text{ ss}}) = \lambda + ((GY)'GY)^{-1}(GY)'BY\lambda \tag{16}$$

as before. In section A4 we further characterise the bias in the estimated indirect effects.

We formalise this result in a proposition.

Proposition 5. Make assumption 2 and assumption 4. Let P denote a projection matrix, $Z^{\text{2SLS}} = [GY, X]$, $H^{\text{2SLS}} = [X, GX, G'GX, ...]$. 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.

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 3 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 6. 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 7. Define

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

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

$$\hat{\lambda} = \frac{\hat{\lambda}^{\text{ SS}}}{I + \eta} \tag{17}$$

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

To construct sample analogues of each stage of these estimators, we use the expectation d^BG in place of BG.

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^*. We can follow the results in section 3 to construct η from d^B, plus assess the robustness of our estimates to oversampling or undersampling.

5 Simulation results

Next, we evaluate the magnitude of bias induced by oversampling or undersampling links, and the performance of our rescaled 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 Section section 6, 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.[5] Agents have some outcome y_i that depends on a binary treatment x_i, that we assign to agents $X \sim$ Bernoulli(0.3).[6] 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 – fixed choice design First, we sample the networks using a fixed choice design as in the National Longitudinal Adolescent Health Data Set. This is a dataset of friendships between high-school students in the United States (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. To simplify, we will focus on friends of one gender – the case where students 'name up to five friends'.[7] 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 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 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.

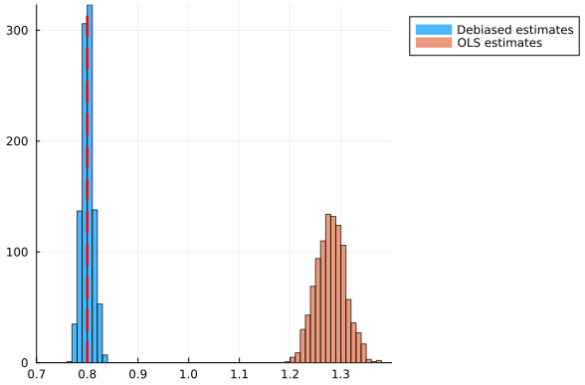
[5] We use a uniform distribution and sample neighbours uniformly at random from the population here to emphasise that the size of the bias that we find is not driven by tail behaviour of the degree distribution or preferential attachment-type mechanisms. Similar results hold when node degrees are sampled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

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

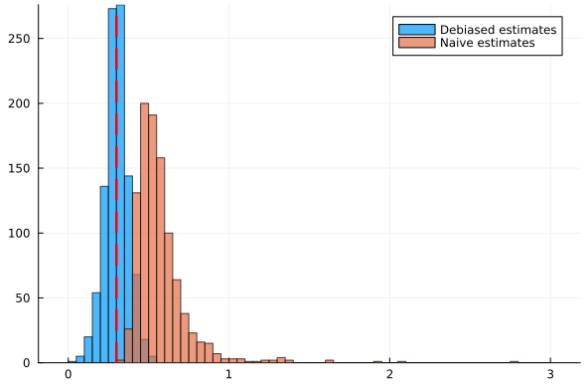
[7] Given the observed homophily by gender in the dataset, this is not too much of a simplification.

Figure 2: 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.

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. Given that treatment is independently and identically distributed across nodes, this is equivalent 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.

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 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.

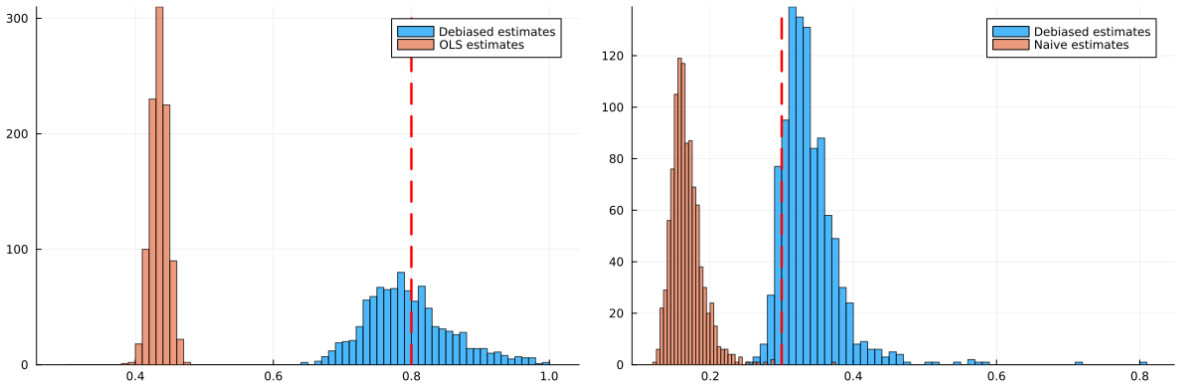
Case 3 – sampling links with highest weights Third, we sample only links between agents that are over a certain weight. This is a common example of undersampling in self-reported network data. Here, we sample the network as in the commonly-used Compustat United States firm-level production network dataset (Atalay et al., 2011). 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 4 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.

Figure 3: Spillover estimates from oversampled network

(a) OLS estimates

(b) SAR estimates

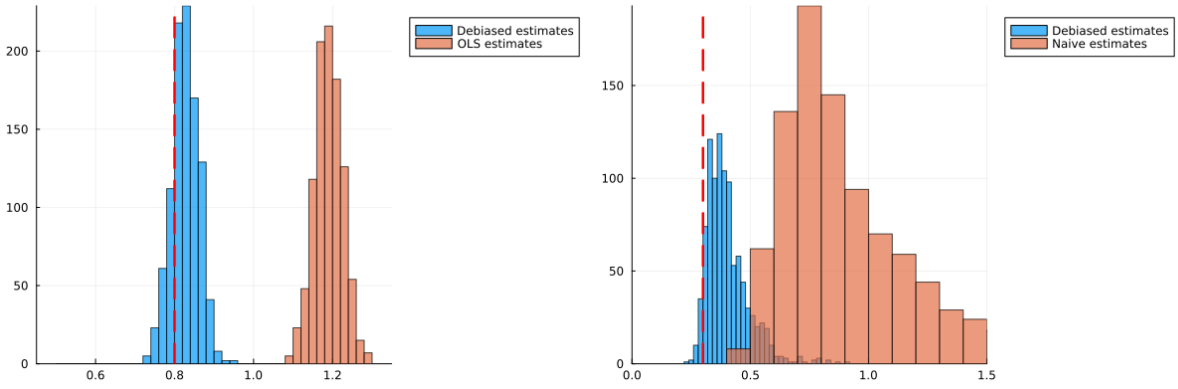


Notes: Red line denotes true parameter values of 0.8 and 0.3 respectively. Data in each case is simulated from a linear/spatial autoregressive model on the true network with $N = 1000$ and single binary treatment drawn i.i.d Bernoulli(0.3) across nodes. The true network has degree distributed $U(0, 10)$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

Figure 4: Spillover estimates from undersampled networks with undersampling based on weights

(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 as listed in text.

6 Extension – dependence between sampling and covariates

Above, we compute η assuming that (G^*, B) are independent of X so that

$$\begin{aligned} E(BX) &= E(\sum_j b_{ij}x_j) \\ &= \sum_j E(b_{ij}(X)|x_j)E(x_j) \\ &= d^B E(X) \text{ by independence of B, X.} \end{aligned}$$

If (G^*, B) are not independent of X, then we instead need to compute

$$E(BX) = \sum_j E(b_{ij}(X)|x_j)E(x_j)$$

directly to rescale estimates.
From eq. (1)

$$\begin{aligned} \sum_j E(b_{ij}(X)|x_j)x_j &= \sum_j E(g^*_{ij}(X) - g_{ij}(X)|x_j)x_j \\ &= \sum_j (E(g^*_{ij}(X)|x_j) - E(g_{ij}(X)|x_j))x_j. \end{aligned}$$

To simplify the interpretation, we make the assumption that $g^*_{ij}(X) = g^*_{ij}(x_j)$, $g_{ij}(X) = g_{ij}(x_j)$ without loss of generality. Then,

$$\sum_j E(b_{ij}(X)|x_j)x_j = \sum_j (E(g^*_{ij}(x_j)) - E(g_{ij}(x_j))x_j.$$

The expected bias depends on the dependence between the presence of a true link and treatment, and the dependence between sampling a link and treatment. If there is no dependence between either, then we can use the results in section 3.

For example, consider a model of strategic network formation with linear-quadratic utility (for examples of this structure, see Jackson, 2010). Then, if each individual's utility is increasing in the treatment of their connections, agent degree is dependent on their treatment.

$$d_i = f(x_i).$$

If there is dependence between either the presence of a true link or sampling a link and treatment, then we need to model the dependence between (G^*, X) to rescale estimates.[8] If we are willing to assume a specific model for network formation, we can fit the model and use this to correct for the bias as in Herstad (2023).

A natural tool for modelling this dependence is a copula (Nelsen, 2006; Trivedi and Zimmer, 2007). Denote the observed distribution of treatment as F_X, and the distribution of the relevant statistic of the true network as F_D. In our example, F_D is the degree distribution of the network. The pairs (x_i, d_i) are distributed according to some unknown joint density function $G()$ with marginal distributions F_X, F_D.

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, we can represent any joint density using a copula.

[8] We do not need further assumptions to compute the additional term GX, because we directly observe G.

Theorem 3 (Sklar's theorem (Nelsen, 2006))**.** Let G be a joint distribution function with marginals F_X, F_D. Then there exists a copula C such that for all x, d in the domain of G

$$H(x,d) = C(F_X(x), F_D(d)).$$

The copula is invariant under strictly increasing transformations, and thus encodes all dependence between variables (Nelsen, 2006).

Given a fitted copula, we can compute the expected network statistic of a node for a given treatment status x

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

Thus, we can compute the expectations $E(g_{ij}^*(x_i)|x_j), E(g_{ij}(x_i)|x_j)$ by fitting a copula conditional on the marginals and then sampling from the copula conditional on x_i, x_j.

This motivates a two-step estimator.

1. Fit relevant copulas $C(F_X^{-1}, F_G^{-1}, \theta_1)$ to compute $\hat{BX}$.

2. Compute debiased estimator $\hat{\beta}$ given $\hat{BX}$.

Consider the example where degree depends on treatment. Denote the observed distribution of treatment as F_X, and the degree distribution of the true network as F_D. This is a natural case of endogeneity to consider. If individuals form links strategically, and there are positive spillovers, then agents should prefer to form links to individuals who get a higher treatment.

Further, assume that the researcher samples at most m links. Then, we have

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

As $\sum_j E(g_{ij}^*|x_i) = d_i$, we need to model the dependence between treatment and degree. So, we fit a copula $C(F_X, F_D, \theta)$. We model the dependence between mean degree and treatment using the 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. As θ gets larger, a higher degree is more associated with a higher treatment.

Then, given an estimate of the dependence parameter $\hat{\theta}$, we can compute the expected degree of node i as

$$E(d_i|x_i, \theta) = \int_0^1 F_D^{-1}(\frac{\partial C(u_x, u_d; \theta)}{\partial u_x}|_{u_x F_X^{-1}(x_i)})dF_D.$$

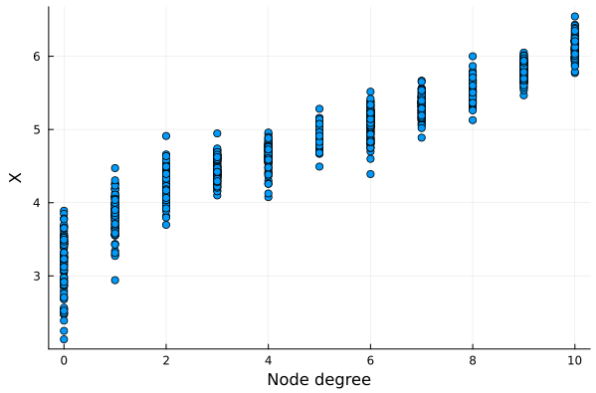
Numerically, we implement this using forward-mode autodifferentiation.

Given the expected degree, we have

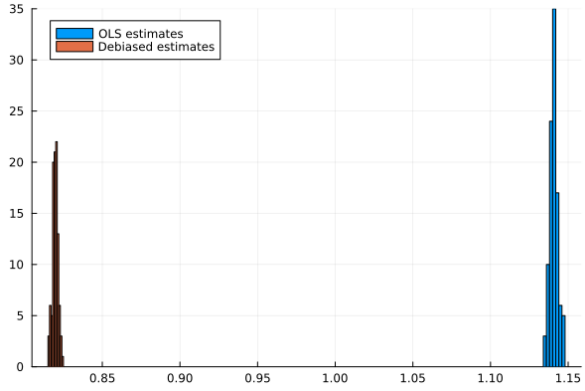
$$\sum_j E(b_{ij}(x_i)|x_j)x_j = \begin{cases} (d^i(\hat{\theta}) - m)\bar{x} \text{ if } d^i(\hat{\theta}) > m, \\ 0 \text{ else.} \end{cases}$$

Figure 5: 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.

6.1 Simulation results

Next, we assess the performance of this estimator in finite sample. As above, we simulate networks of $N = 1000$ agents, where each agent draws an in-degree d_i with the marginal distribution $D \sim U(0, 10)$ and is connected with d_i other agents uniformly at random from the population. Each agent draws a treatment with the marginal distribution $X \sim N(5, 1)$. Now, the marginal distributions are coupled through a bivariate Gumbel copula with $\theta = 10$. The left panel of fig. 5 plots an example joint distribution of treatment and in-degree. We see that degree and treatment are dependent. Higher treatment nodes have higher degree. The network is then sampled as in the National Longitudinal Survey of Adolescent Health Data Set. The researcher observes at most five incoming links per node.

We estimate spillovers using the two-step estimator we describe above. In the first step, we estimate the dependence between treatment and degree using a Gumbel copula by maximum likelihood. In order that sampling does not affect the estimate of dependence, we estimate the copula using only observations with fewer than five links. 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.

The quality of our estimates depends on the choice of copula, and assumptions on the marginal distribution of the network statistic and our variable. The distributional assumption is similar to the assumption on the distribution of the shock process over space needed to compute unbiased estimates in Borusyak and Hull (2023). This approach to modelling dependence is also similar to control function approaches to left-hand side selection in the sample selection literature (Heckman, 1979; Smith, 2003).

For some forms of dependence and sampling schemes, there are natural assumptions that we can make on the relevant marginal distributions that allow us to model the dependence between unobserved links and spillovers. For example, consider the firm-level production network of the United States. As we detail later in section 7, firms self-report customers who make up more than 10 percent of sales. Further, degree distributions of firm-level production networks tend to have similar shapes (Bacilieri et al., 2023), and Herskovic et al. (2020) estimate the tail exponent for the degree distribution of a firm-level production network for the United States accounting for sampling bias. If the relevant dependence is between the degree and treatment, we could then model the dependence on the marginal degree distribution and the observed marginal distribution of treatment.

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

$$\hat{\beta}^{\text{OLS}} > \tau$$
$$\text{if and only if}$$
$$(GX)'BX < A\frac{\hat{\beta}^{\text{OLS}} - \tau}{\tau}. \tag{18}$$

The senstivity 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.

7 Application: climate shocks in production networks

Finally, we apply our estimator to re-evaluate the propagation of idiosyncratic shocks in firm-level production networks. Specifically, we re-evaluate the results in Barrot and Sauvagnat (2016), who estimate how extreme weather shocks to public firms in the United States affect the sales of their customers using self-reported data on large customers.[9]

Barrot and Sauvagnat (2016)'s sample contains 2051 public firms in the United States 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 United States 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.

They then 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 $t-4$.[10]

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 Securities and Exchange Commission. From 1978-1997, firms could self-report customers in filings to the Securities and Exchange Commission, 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, but are not obliged to.

Compared to existing complete datasets, the self-reported network appears to undersample links between firms. The mean number of suppliers is 1.38, with a median of 0.000, many fewer than researchers

[9] 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

[10] 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.

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).[11] 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 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 public firms in the United States 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

[11] 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.

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 The other networks that we use to construct d^B may not be a good proxy for the supply network of public firms in the United States. So, we also 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 3, rearranging our bias formula for $\beta < 0$ gives

$$\begin{aligned} &\hat{\beta}^{\text{OLS}} < \tau \\ &\text{if and only if} \\ &d^B < \Big(\frac{1}{NA^{-1}p_{\text{shock}}^2 d^G}\Big)\frac{-0.031-\tau}{\tau}. \end{aligned} \tag{19}$$

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 3 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. 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 spillover estimates.

8 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. In the case where sampling is correlated with treatment, we show that researchers can construct a two-step estimator for spillover values from marginal distributions using copulas. Finally, we use our results to characterise how undersampling of the production network amongst public firms in the United States 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

Aronow, P. M. and Samii, C. (2021). Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics*, 11(4):1912–1947.

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.

Autor, D. H., Dorn, D., and Hanson, G. H. (2013). The china syndrome: Local labor market effects of import competition in the united states. *American Economic Review*, 103(6).

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.

Bartik, T. J. (1991). *Who benefits from state and local economic development policies?* WE Upjohn Institute for Employment Research.

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.

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*.

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

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.

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

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.

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.*

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

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

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

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, 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.

Mas, A. and Moretti, E. (2004). Peers at Work. *American Economic Review*, 99(1):112–145.

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.

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.

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

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.

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

Contents

A1 Potential outcomes notation

Here, we derive results for a simple version of our model in the potential outcomes notation of Aronow and Samii (2021).

Consider the case of an experiment on a network with a single binary treatment $X_i \in \{0, 1\}$. Assume without loss of generality that treatment only affects individuals through their neighbours. Consider a set of possible sampling rules $s \in \mathcal{S}$.

We summarise the effects of treatment on the true and sampled networks with the exposure mappings Aronow and Samii (2021)

$e_i(G^*)$– exposure of node i through the true network, and
$e_i(G, s)$ – exposure of node i through the sampled network under sampling rule s.

Let n_i denote the count of treated neighbours. As we have a simple graph, $n_i \in \{1, ..., N-1\}$. Potential outcomes are linear in the number of treated neighbours

$$y_i^{n_i} = \alpha + \beta n_i.$$

For simplicity, assume unconfoundedness of spillovers on the true network i.e

$$E(y^{n_i}|e_i(G^*) = m_i) = E(y^{n_i}|e_i(G^*) = m_i') \ \forall n, m, m' \in \{1, ..., N-1\}.$$

This could come from randomisation of treatment across nodes, or conditioning (that we suppress for notational simplicity).

Under linear potential outcomes, the ordinary least squares estimator for beta computes the causal effect of an additional neighbour as the difference in levels

$$\begin{aligned} E(y|e_i(G) = n) - E(y|e_i(G) = n-1) &= \sum_{m=1}^{N-1} y^m p(e_i(G^*) = m|e_i(G) = n) - \sum_{m=1}^{N-1} y^m p(e_i(G^*) = m|e_i(G) = n-1), \\ &= y^n p(e_i(G^*) = n|e_i(G) = n) + \sum_{m \neq n} y^m p(e_i(G^*) = m|e_i(G) = n) \\ &- (y^{n-1} p(e_i(G^*) = n-1|e_i(G) = n-1) + \sum_{m \neq n-1} y^m p(e_i(G^*) = m|e_i(G) = n - \end{aligned}$$

Rewriting $p(e_i(G^*) = n'|e_i(G) = n') = 1 - \sum_{m \neq n'} p(e_i(G^*) = m|e_i(G) = n')$ and substituting in gives

$$E(y|e_i(G)=n) - E(y|e_i(G)=n-1) = y^n - y^{n-1} + \sum_{m\neq n} p(e_i(G^*)=m|e_i(G)=n)(y^m - y^n) - \sum_{m\neq n-1} p(e_i(G^*)=m|e_i(G)=n-1)(y^m - y^{n-1}).$$

Therefore, $E(y|e_i(G)=n) - E(y|e_i(G)=n-1) = y^n - y^{n-1}$ if and only if

$$\sum_{m\neq n} p(e_i(G^*)=m|e_i(G)=n)(y^m - y^n) = \sum_{m\neq n-1} p(e_i(G^*)=m|e_i(G)=n-1)(y^m - y^{n-1}).$$

Undersampling implies that the probablities on the left hand side are non-zero whereas the probabilities on the right hand side are non-zero. Oversampling implies the converse.

A2 Distribution of test statistics for linear models

With oversampling or undersampling the t-statistic is no longer t-distributed. We show this by showing that the sum of squared residuals no longer takes a standard normal distribution.

Define

$$\hat{e} = Y - P_{GX}Y$$

where P is the linear projection matrix.

Expanding gives

$$\hat{e} = (I - P_{GX})\epsilon - (I - P_{GX})BX\beta$$

instead of the usual expression

$$\hat{e} = (I - P_{GX})\epsilon.$$

So, squaring gives

$$\hat{e}'\hat{e} = \epsilon'(I - P_{GX})\epsilon + \beta^2(BX)'(I - P_{GX})(BX)$$

Applying the standard diagonalisation $(I - P_{GX}) = Q'\Lambda Q$ gives

$$\hat{e}'\hat{e} = (Q'\epsilon + \beta Q'(BX))'\Lambda(Q'\epsilon + \beta Q'(BX)).$$

The random vector $(P'\epsilon + \beta P'(BX))$ does not necessarily have a mean of zero, completing the proof. Indeed, from standard results on the sum of squared normals we see that

$$\hat{e}'\hat{e} \sim \text{Gamma}(\beta Q'(BX), N, 2\text{var}(\beta Q'(BX)))).$$

A3 Proofs of results for linear models

Lemma 4. 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 5. $BX \perp \epsilon | G^*X$.

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

A3.0.1 Proof of proposition 1 and theorem 1

Proof. Substituting in eq. (2), we get

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

Our ordinary least-squares 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 5. $E(((GX)'(GX))^{-1}((G^* - B)X)'\epsilon) = 0$.

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 ordinary least-squares 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 6. 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 ordinary least-squares 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 ordinary least-squares estimator. First, we establish the following lemma.

Lemma 7.

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

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

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

□

A3.0.2 Proof of theorem 2

Write the ordinary least-squares 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 ordinary least-squares estimator if we observed the true network. Following the standard derivation of the limiting distribution of the ordinary least-squares estimator e.g in Cameron and Trivedi (2005), this gives the theorem.

A4 Proofs of results for nonlinear social network models

Proof of proposition 4 To see this, rewrite eq. (14) as

$$Y = \lambda(G + B)Y + X\beta + \epsilon. \tag{A-20}$$

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

$$GY = G(I - \lambda G)^{-1}(X\beta + \epsilon) + G(I - \lambda G)^{-1}\lambda B(I - \gamma G^*)^{-1}(X\beta + \epsilon). \tag{A-21}$$

The standard instruments $(I - G)^{-1}X$ are correlated with the error term in the first stage regression.

Effect on indirect estimates We use the result that

$$(I - (A + B))^{-1} = ((I - A)(I - (I - A)^{-1}B))^{-1}.$$

We can split our estimated

$$\begin{aligned}\hat{\lambda}G &= (\lambda + \eta)(G^* - B) \\ &= \lambda G^* + H.\end{aligned}$$

Then

$$\begin{aligned}(I - \hat{\lambda}G)^{-1} &= (I - (\lambda G^* + H))^{-1} \\ &= ((I - \lambda G^*)(I - (I - \lambda G^*)^{-1}H))^{-1} \qquad = (I - (I - \lambda G^*)^{-1}H))^{-1}(I - \lambda G^*)^{-1}.\end{aligned}$$

So, sampling also creates a multiplicative bias in the indirect effect that depends on the set of unobserved walks through the network.

Proof of proposition 6 Pre-multiply the true data generating process by G to get

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

Thus suffices to show the result.