JOURNAL OF ECONOMIC SURVEYS

doi: 10.1111/joes.12256

CREDIBLY IDENTIFYING SOCIAL EFFECTS: ACCOUNTING FOR NETWORK FORMATION AND MEASUREMENT ERROR

Arun Advani

University of Warwick, Institute for Fiscal Studies and CAGE

Bansi Malde*

University of Kent, and Institute for Fiscal Studies

Abstract. Understanding whether and how connections between agents (networks) such as declared friendships in classrooms, transactions between firms, and extended family connections, influence their socio-economic outcomes has been a growing area of research within economics. Early methods developed to identify these *social effects* assumed that networks had formed exogenously, and were perfectly observed, both of which are unlikely to hold in practice. A more recent literature, both within economics and in other disciplines, develops methods that relax these assumptions. This paper reviews that literature. It starts by providing a general econometric framework for linear models of social effects, and illustrates how network endogeneity and missing data on the network complicate identification of social effects. Thereafter, it discusses methods for overcoming the problems caused by endogenous formation of networks. Finally, it outlines the stark consequences of missing data on measures of the network, and regression parameters, before describing potential solutions.

Keywords. Networks; Social effects; Econometrics; Endogeneity; Measurement error; Sampling design

1. Introduction

Networks – connections between agents – are an ubiquitous part of life. Student's academic achievement is influenced by their friends and classmates; employee productivity by interactions with other team members; individuals learn about new products and opportunities from their acquaintances and friends; firms cooperate and compete with other firms in developing new innovations; and so on. Understanding the nature and magnitude of the effects of networks is key to constructing meaningful models and designing effective policies. A particular interest lies in identifying *social effects* – direct spillovers from the outcomes of one agent to the outcomes of others.

Early empirical work seeking to identify social effects used data with limited information on networks, typically information on membership of mutually exclusive groups such as classrooms, neighbourhoods, or villages. Estimating social effects with this type of data suffers from two key limitations. First, identifying the social effect is complicated by the reflection problem – a form of simultaneity where it is not possible to identify who is influencing whom (Manski, 1993). Second, since more detail on interactions within a group is not available, studies (implicitly) assume that all agents within a group interact with one another in the same way. However, the composition of the group on both observed and

*Corresponding author contact email: b.k.malde@kent.ac.uk; Tel: +44 (0) 1227 816464

unobserved dimensions could influence within-group interactions, and through this the actual outcome. Ignoring variation in interactions *within* such groups can lead to misleading conclusions and policy design, as shown in recent work by Carrell *et al.* (2013).

More recently, a growing body of research within empirical economics uses data which directly measure interactions between pairs of agents (*network data* hereon) to sidestep these issues. This growth has been spurred by the increasing availability of such data, as well as the development of methods to identify and estimate social effects with such data. Starting with Bramoullé *et al.* (2009) and De Giorgi *et al.* (2010), methods have been developed to overcome the reflection problem. They show how information on network structure can be used to break the simultaneity, and obtain the necessary exclusion restrictions for parameter identification. These methods, reviewed in detail by Advani and Malde (2014), Topa and Zenou (2015) and Boucher and Fortin (2015), impose strong restrictions on the network formation process and the quality of the data.

In particular, the network is assumed to be exogenous conditional on observed agent- and network-level characteristics, and to be fully and perfectly observed by the researcher. Both assumptions are unlikely to hold in practice. In a schooling context, for example, personality traits which are rarely observed by a researcher might influence both a child's choice of friends and her schooling performance. Estimates of the influence of a child's friends' outcomes on her outcomes will be biased if her choice of friends is not accounted for. Similarly, accurately collecting fine-grained information on all connections is very costly and logistically challenging, making it rare to observe a complete, perfectly measured network. This has important implications for identification of social effects using restrictions based on the network structure: for example, the methods proposed by Bramoullé *et al.* (2009) and De Giorgi *et al.* (2010) rely on information of who is not connected with whom to provide exclusion restrictions. Missing or mismeasured data on link status will impair the ability of these methods to yield unbiased and consistent social effect estimates.

The issue of endogenous link formation has long been recognized in the empirical literature, while that of measurement error has received increasing attention recently. In this paper, we provide an overview of a range of econometric methods to deal with network endogeneity and measurement error when estimating *linear models* of social effects. The majority of empirical work on social effects uses linear models, motivating our focus on this class of model.[1] Whilst the process of network formation is of interest in its own right, our interest is in identifying causal social effects, for which endogenous network formation is a clear confounder. We therefore focus on methods that deal with the endogeneity in network formation when estimating social effects, rather than all methods to model and estimate network formation processes. We draw on methods developed in a broad range of disciplines, including economics, sociology and mathematics, and express ideas in a manner that can be easily understood by economists.

We begin by laying out a general linear econometric model of social effects, separately for individual-level and network-level outcomes. The individual-level specification nests a number of economic models that have been applied in the literature. These specifications clarify the social effect parameters of interest, and allow us to illustrate the consequences of endogenous link formation and measurement error in the network on social effect estimates.

Next, we provide a brief overview of strategies to deal with endogenous network formation. We do not hope or attempt to provide a comprehensive review of this now large and expanding literature. Instead, we discuss, in a general way, four common approaches, using specific examples to illustrate ideas. The first approach exploits exogenous variation arising from random assignment of interventions or links. Though this provides clean identification, random assignment may not often be feasible. The second approach exploits local shocks to network structure induced by natural- and quasi-experiments such as policy rules, or unanticipated deaths of agents. When such variation is not available, a third approach – instrumental variables – may be promising. This involves finding a variable which affects the link formation decision but has no direct influence on the outcome of interest. However, such a variable may not be available in many contexts. A final strand of the literature thus jointly models link and action choices. Approaches

within this literature model these choices either sequentially, or simultaneously. In the former case, a natural solution to account for bias arising from the self-selection of link is the control function where one estimates the selection bias term, and 'controls' for it when estimating the social effect model. However, where multiple equilibria are possible, this approach requires additional assumptions about equilibrium selection.

Thereafter, we discuss the challenge posed by imperfectly measured networks. Missing data, due to the sampling method or otherwise, have important consequences for both measurement of statistics of the network, and the parameter estimates of social effect models. This is because networks consist of two interrelated objects: agents (nodes) and links. A sampling strategy over one of these objects defines the (conditional) sampling process over the other. This means that econometric and statistical methods for estimation and inference developed under classical sampling theory are often not applicable to network data. We first discuss the implications of missing data for the estimation of network statistics and regression parameters. Thereafter, we review the methods available to correct for these problems, and the conditions under which they can be applied.

Given the breadth of research in these areas alone, we naturally have to make some restrictions to narrow the scope of what we cover. We do not cover methods for estimating social effects when networks are conditionally exogenous. Surveys by Blume *et al.* (2010), Advani and Malde (forthcoming), Topa and Zenou (2015) and Boucher and Fortin (2015) more than amply cover this ground. In our discussion of endogeneity, we touch lightly on issues of network formation; a fuller treatment of network formation can be found in Advani and Malde (2014), Graham (2015), de Paula (forthcoming) and Chandrasekhar (2015). Similarly, whilst we discuss briefly models in which characteristics of the network structure are important, a fuller treatment can be found in Jackson *et al.* (2017). Finally, we do not survey findings on the size, magnitude or heterogeneity of the social effects found in applied economics: other reviews more than amply cover these, for example, Epple and Romano (2011) and Sacerdote (2011) provide surveys of peer effects in education, while Chuang and Schechter (2015) provide a survey of applied work on networks in developing countries.

The rest of the paper is organized as follows. Section 2 lays out a general linear econometric model of social effects, separately for individual- and network-level outcomes. Section 3 considers methods to deal with endogenous formation of network links. Section 4 considers the implications of measurement error in the network, and outlines some of the methods that have been proposed to account for these. Section 5 provides some concluding remarks, considers some of the limits of what is currently known about econometric methods for linear social effect models and offers some potential directions for future work.

2. Conceptual Framework

We begin by laying out a general linear econometric model of social effects, separately for individual- and network-level outcomes. These nest a number of the key empirical specifications used in the literature, and elucidate the parameters of interest. We draw on these specifications to outline some of the common assumptions imposed to identify the parameters of interest. Thereafter, we illustrate the implications of endogenous network formation and measurement error in the network.

Throughout we use the following notation. A *network* (or *graph*), $g = (\mathscr{N}_g, \mathscr{E}_g)$, is defined by a set of nodes, $\mathscr{N}_g$, and the edges (or links) $\mathscr{E}_g$ between them. The nodes represent agents (individuals, households, firms or countries), and the edges represent the links between pairs of nodes (e.g. friendship, kinship, coworking, economic transactions). We index networks by g, and nodes within a network g by $i \in \mathscr{N}_g$. The number of nodes in network g is N_g, and the number of edges is E_g. We define $\mathcal{G}_N$ as the set of all possible networks on N nodes. We consider *binary networks* where any (ordered) pair of nodes i, j is either linked, $G_{ij,g}= 1$, or not linked, $G_{ij,g}= 0$. If $G_{ij,g}= 1$ then j is described as being a *neighbour* of i. We denote by $nei_{i,g} = \{j : G_{ij,g} = 1\}$ the *neighbourhood* of node i, which contains all nodes with whom i is linked. $d_{i,g} = |\{j : G_{ij,g} = 1\}|$ is the number of neighbours, or *degree,* of i. Nodes that are

neighbours of neighbours will often be referred to as '*second degree neighbour*'. Typically it is convenient to assume that $G_{ii,g} := 0 \quad \forall i \in g$. Edges may be directed, so that $G_{ij,g}$ is not necessarily the same as $G_{ji,g}$; in this case the network is a *directed graph* (or *digraph*). The network can be represented by an $N_g \times N_g$ *adjacency matrix*, $\boldsymbol{G}_g$, with typical element $G_{ij,g}$; and whose leading diagonal is normalized to 0. We also define the *influence matrix,* $\tilde{\boldsymbol{G}}_g$, as the row-stochastised adjacency matrix.[2] Elements of this matrix are defined as $\tilde{G}_{ij,g} = d_{i,g}^{-1} G_{ij,g}$.

2.1 *Individual-Level Models*

Common specifications of individual-level linear social effect models can be written as a special case of the following equation:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \boldsymbol{w}_y(\boldsymbol{G},\boldsymbol{Y})\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{w}_x(\boldsymbol{G},\boldsymbol{X})\boldsymbol{\delta} + \boldsymbol{Z}\boldsymbol{\eta} + \boldsymbol{L}\boldsymbol{\nu} + \boldsymbol{\varepsilon} \tag{1}$$

$\boldsymbol{Y}$ is an $\sum_{g=1}^{M} N_g \times 1$ vector stacking individual outcomes of nodes across all networks (indexed by $g = 1, \ldots, M$). $\boldsymbol{X} = (\boldsymbol{X}_1^{'}, \ldots, \boldsymbol{X}_M^{'})^{'}$ is an $\sum_{g=1}^{M} N_g \times K$ matrix of K individual-level observable characteristics that influence a node's outcome and potentially that of others in the network. $\boldsymbol{G} = diag\{\boldsymbol{G}_g\}_{g=1}^{g=M}$ is a block-diagonal matrix with the adjacency matrices of each network along its leading diagonal, and zeros on the off-diagonal. The block-diagonal nature of $\boldsymbol{G}$ means that only the characteristics and outcomes of nodes in the same network are allowed to influence a node's outcome. $\boldsymbol{w}_y(\boldsymbol{G},\boldsymbol{Y})$ and $\boldsymbol{w}_x(\boldsymbol{G},\boldsymbol{X})$ are functions of the adjacency matrix, and the outcome and observed characteristics, respectively. These functions indicate how network features, interacted with outcomes and exogenous characteristics of other nodes in the network, influence the outcome. $\boldsymbol{Z}$ is an $\sum_{g=1}^{M} N_g \times Q$ matrix of Q network-level observed variables that influence nodes' outcomes. The matrix $\boldsymbol{L} = diag\{\boldsymbol{\iota}_g\}_{g=1}^{g=M}$ is an $\sum_{g=1}^{M} N_g \times M$ matrix where each column is an indicator for being in a particular network. $\boldsymbol{\nu} = \{\nu_g\}_{g=1}^{g=M}$ is a vector of network-specific effects, unobserved by the econometrician but known to nodes; and $\boldsymbol{\varepsilon}$ is a vector stacking the (unobservable) error terms for all nodes across all networks. In any given specification only one of $\boldsymbol{Z}$ and $\boldsymbol{L}$ can be included.

This representation nests a range of models estimated in the economics literature:

Local Average Model: This model arises when a node's outcomes are influenced by the average behaviour and characteristics of its direct neighbours.[3] This happens, for example, when social effects operate through a desire for a node to conform to the behaviour of its neighbours. This implies that $\boldsymbol{w}_y(\boldsymbol{G},\boldsymbol{Y}) = \tilde{\boldsymbol{G}}\boldsymbol{Y}$ and $\boldsymbol{w}_x(\boldsymbol{G},\boldsymbol{X}) = \tilde{\boldsymbol{G}}\boldsymbol{X}$ above. Bramoullé *et al.* (2009) and De Giorgi *et al.* (2010) provide conditions for identifying model parameters when the network is conditionally exogenously formed.

Local Aggregate Model: When there are strategic complementarities or substitutabilities between a node's outcomes and the outcomes of its neighbours, one can obtain the local aggregate model. In this case, a node's outcome depends on the aggregate outcome of its neighbours, which corresponds to $\boldsymbol{w}_y(\boldsymbol{G},\boldsymbol{Y}) = \boldsymbol{G}\boldsymbol{Y}$ in equation (1). $\boldsymbol{w}_x(\boldsymbol{G},\boldsymbol{X})$ is typically defined to be $\tilde{\boldsymbol{G}}\boldsymbol{X}$. See Calvó-Armengol *et al.* (2009), Lee and Liu (2010), Liu *et al.* (2014b), and Bramoullé *et al.* (2014) for details on identification conditions when the network is conditionally exogenously formed.

Hybrid Local Model: This class of models nests both the local average and local aggregate models, which allows the social effect to operate through both a desire for conformism and through strategic complementarities/substitutabilities. In the notation of equation (1), it implies that $\boldsymbol{w}_y(\boldsymbol{G},\boldsymbol{Y}) = [\boldsymbol{G}\boldsymbol{Y},\ \tilde{\boldsymbol{G}}\boldsymbol{Y}]$, while $\boldsymbol{w}_x(\boldsymbol{G},\boldsymbol{X})$ is typically defined to be $\tilde{\boldsymbol{G}}\boldsymbol{X}$. Liu *et al.* (2014a) provide identification conditions when the model is conditionally exogenously formed.

Models with Network Statistics: Networks may influence node outcomes (and consequently aggregate network outcomes) through statistics of the network beyond those depending on direct neighbours only.[4] For instance, the DeGroot (1974) model of social learning implies that an individual's eigenvector centrality, which measures a node's importance in the network by how important its neighbours are, determines how influential it is in affecting the behaviour of other nodes.

Denoting a specific network statistic by ω^r, where r indexes the statistic, some possible specialisations of $\boldsymbol{w}_y(\boldsymbol{G},\boldsymbol{Y})\boldsymbol{\beta}$ in equation (1) for node i in network g include:

- $\sum_{r=1}^{R}\omega_{i,g}^{r}\beta_r$: R different network statistics, without any reference to outcomes (e.g. Banerjee *et al.*, 2013; Cruz *et al.*, forthcoming); or
- $\sum_{r=1}^{R}\sum_{j\neq i}\tilde{G}_{ij,g}y_{j,g}\omega_{j,g}^{r}\beta_r$: the average of neighbours' outcomes weighted by R different network statistics (e.g. Cai *et al.*, 2015); or
- $\sum_{r=1}^{R}\sum_{j\neq i}G_{ij,g}y_{j,g}\omega_{j,g}^{r}\beta_r$: the sum of neighbours' outcomes weighted by R different network statistics.

Analogous definitions can be used for $\boldsymbol{w}_x(\boldsymbol{G},\boldsymbol{X})\boldsymbol{\delta}$.

The social effect parameter in equation (1) is $\boldsymbol{\beta}$: the effect of a function of a node's neighbours' outcomes (e.g. an individual's friends' schooling performance) and the network. This is also known as the *endogenous effect*, to use the term coined by Manski (1993). This parameter is often of policy interest since the presence of endogenous effects implies there is a social multiplier: the aggregate effects of changes in $\boldsymbol{X}$, $\boldsymbol{w}_x(\boldsymbol{G},\boldsymbol{X})$ and $\boldsymbol{Z}$ are amplified beyond their direct effects, captured by $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$. The parameter $\boldsymbol{\delta}$, capturing the effect of neighbours' characteristics, is known as the *exogenous or contextual effect*, while $\boldsymbol{\eta}$ and $\boldsymbol{\nu}$ capture a *correlated effect*, common to everyone in the same network.

Identification of the social effect parameter depends on the restrictions imposed on the relationship between the error terms, $\boldsymbol{\nu}$ and $\boldsymbol{\varepsilon}$, and the right-hand side variables in equation (1). These restrictions reflect assumptions on common unobserved shocks and on the network formation process. For example, $\mathbb{E}[\nu_g|\,\boldsymbol{X}_g,\boldsymbol{Z}_g,\boldsymbol{G}_g]=0\,\forall\, g\in\{1,\ldots,M\}$ implies nodes sort into networks exogenously, conditional on individual-level and network-level observables, while $\mathbb{E}[\varepsilon_{i,g}|\,\boldsymbol{X}_g,\boldsymbol{Z}_g,\boldsymbol{G}_g]=0\,\forall\, i\in\mathcal{N}_g;\ g\in\{1,\ldots,M\}$ implies that the network is exogenous, conditional on individual-level and network-level observable characteristics of all nodes in network g.

The former assumption can be relaxed when data on a large number of networks are available: unobservable characteristics determining sorting into networks can be accounted for using network-level fixed effects, as in panel data specifications. A number of methods, that rely primarily on variation in network structure, have been developed to identify the social effect parameters in such models using observational data and under the assumption that the network is conditionally exogenous and well-measured. The interested reader is directed to Advani and Malde (forthcoming), Topa and Zenou (2015) and Boucher and Fortin (2015) for more details.

2.2 *Network-Level Models*

Researchers might also be interested in aggregate network-level outcomes, in which case the following specification is typically estimated:

$$\bar{\boldsymbol{y}}=\phi_0+\bar{\boldsymbol{w}}_{\bar{y}}(\boldsymbol{G})\boldsymbol{\phi}_1+\bar{\boldsymbol{X}}\boldsymbol{\phi}_2+\bar{\boldsymbol{w}}_{\bar{X}}(\boldsymbol{G},\boldsymbol{X})\boldsymbol{\phi}_3+\boldsymbol{u} \tag{2}$$

where $\bar{\boldsymbol{y}}$ is an $(M\times 1)$ vector stacking the aggregate outcome of the M networks, $\bar{\boldsymbol{w}}_{\bar{y}}(\boldsymbol{G})$ is a matrix of $\bar{R}$ network statistics (e.g. average number of links per node, also known as average degree) that directly influence the outcome, $\bar{\boldsymbol{X}}$ is an $(M\times K)$ matrix of network-level characteristics, and $\bar{\boldsymbol{w}}_{\bar{X}}(\boldsymbol{G},\boldsymbol{X})$ is a term interacting the network-level characteristics with the network statistics.[5] $\boldsymbol{\phi}_1$ captures how the network-level aggregate outcome varies with specific network features while $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_3$ capture, respectively, the effects of the network-level characteristics and these characteristics interacted with the network statistic(s) of interest on the outcome.

The key parameter of interest is typically $\boldsymbol{\phi}_1$: the effect of a network statistic, such as network density, on the aggregate network outcome. The key identification assumption is that $\mathbf{E}[\boldsymbol{u}_g|\boldsymbol{G}_g,\bar{\boldsymbol{X}}_g]=0$, which

will not hold if there are unobserved variables in $\boldsymbol{u}$ that affect both the formation of the network and the outcome $\bar{\mathbf{y}}$; or if the network statistics are mismeasured.

2.3 *Implications of Network Endogeneity and Measurement Error*

The assumption that the network is conditionally exogenous implies, first, that there are no unobserved (to the econometrician) agent-specific factors influencing both an agent's choice of connections and the outcome of interest; and second, that agents do not take into account the influences of their neighbours on the outcome of interest when choosing their links. Both of these are very strong requirements. To see this more easily, consider the following example. Suppose we have observational data on farming practices amongst farmers in a village, and want to identify the factors that influence take-up of a new, potentially risky technology. The data might show that more connected farmers are also more likely to adopt the technology. However, without further analysis we cannot necessarily interpret this as being *caused* by the network. There could be some underlying unobserved variable that is correlated with both the outcome and the network. For example, more risk-loving people, who might be more likely to adopt the technology, may also be more sociable, and thus have more connections. Alternatively, more connected farmers might also be more interested in learning about innovative practices, and choose to have more connections for this reason! Both of these violate the condition that $\mathbb{E}[\varepsilon_{i,g}|\, \boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}_g] = 0 \;\forall\, i \in g;\; g \in \{1, \ldots, M\}$ in Equation (1). Section 3 describes potential solutions to this endogeneity problem in more detail.

Measurement error in $\boldsymbol{G}$ can also invalidate the assumption that $\mathbb{E}[\varepsilon_{i,g}|\, \boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}_g] = 0 \;\forall\, i \in g;\; g \in \{1, \ldots, M\}$, and hence bias parameter estimates. Suppose the observed network, $\boldsymbol{G}^*$, is a noisy measure of the true underlying network, $\boldsymbol{G}$, such that $\boldsymbol{G}^* = \boldsymbol{G} + \boldsymbol{\xi}(\boldsymbol{G})$. Estimation of equation (1) would be based on the mismeasured network, $\boldsymbol{G}^*$, with the measurement error term (or a function of it) subsumed into the error term, $\boldsymbol{\varepsilon}$, in equation (1). Clearly, then $\mathbb{E}[\varepsilon_{i,g}|\, \boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}^*_g] \neq 0$, leading to bias in the social effect parameter estimates. Moreover, the measurement error in the network is likely to be non-classical, so that it is not independent of the true network.

A simple example illustrates this. Surveys often place an upper limit, ψ, on the number of links a node can report, leading to some links of agents with many connections to be recorded as not existing. In the absence of other error, the number of misclassified links for node i can be expressed as $\sum_j \boldsymbol{\xi}(\boldsymbol{G})_{ij} = \max\{0, \sum_j G_{ij} - \psi\}$. Thus, the measurement error necessarily depends on the structure of the true network, making it non-classical. The consequences of measurement error on parameter estimates will thus be quite complex. Section 4 considers this in more detail, and outlines some potential solutions.

3. Dealing with Endogeneity of Network Formation

We now discuss approaches taken to identify social effects whilst relaxing the assumption that the network is exogenous. Specifically, we allow for the possibility that network links are chosen, and that these choices might be related to the unobservables determining individuals' outcomes.[6] We discuss four approaches taken in the literature to deal with this form of endogeneity, providing examples of where they have been used, and discussing their limitations.

3.1 *Random Assignment*

The first method is random assignment, either of some intervention provided to a subset of nodes in the network, or of links in the network. Random assignments of interventions have been used to study a wide range of questions, including the diffusion of innovations in social networks (Aral and Walker, 2012; Oster and Thornton, 2012; Cai *et al.*, 2015; among others), social learning (Godlonton and Thornton, 2012), sharing of resources and savings (Comola and Prina, 2017; Angelucci *et al.*, forthcoming), peer

effects in exercise (Babcock *et al.*, 2015), peer effects in education (Angelucci *et al.*, 2010; Babcock and Hartman, 2010) and peer monitoring (Breza and Chandrasekhar, 2015).

In these designs, also known as partial population experiments, (Moffitt, 2001), researchers randomly assign a subset of nodes in a network to receive a treatment. Untreated nodes in the network will be indirectly exposed to the treatment through their interactions with treated nodes. This indirect exposure will vary with the position of the untreated nodes in the pre-treatment network relative to nodes that were randomly assigned the treatment. Since the treatment is randomly assigned, conditional on their network position the exposure levels of untreated nodes will be orthogonal to the network structure. Thus, a reduced form social effect can be identified by comparing the outcomes of untreated nodes with the same network position but different levels of exposure to the treatment.[7] The identified reduced form social effect need not solely capture the spillover of neighbours' outcomes on a node's own outcome: it may also capture other channels through which the intervention may influence those neighbours. For example, in the case of the adoption of innovations, a treatment such as providing information to a subset of the network could influence innovation take-up through both diffusion of information, as well as through the adoption decisions of the initially informed nodes (see Banerjee *et al.*, 2013), making it difficult to separately identify the endogenous social effect without further modelling.

Randomly assigned treatments can only be used to identify social effects if the treatment does not also change the social network. Recent work by Comola and Prina (2017), Delavallade *et al.* (2016) and Dupas *et al.* (forthcoming) shows that interventions may alter the network of interactions, so that a randomly assigned treatment will not be orthogonal to the final network structure. Use of the pre-treatment network does not solve the problem: treatment effects identified based on the pre-treatment network may be misleading since they ignore the effects on network structure. This is shown by Comola and Prina (2017), who extend the local average model to allow for the network to change in response to a treatment. This extended model allows for the recovery of both the total treatment effect, and the social effect. However, the randomly assigned treatment can no longer be used to identify the social effect. To recover this parameter, Comola and Prina (2017) propose to, first, exploit the panel dimension of their network data to account for time-invariant unobserved variables that influence both network formation and the outcome of interest. Second, to account for time-varying unobservables, they use predicted changes in the network (partly due to the treatment) as an instrument for the actual changes. This is similar to the strategy in König *et al.* (2014), described in detail in Section 3.3.

A third set of designs relies on variation arising from randomly assigned links. While this strategy has been widely applied in laboratory experiments of network effects, recent work has applied this to real-life contexts, or exploited real-life contexts where this occurs, including classrooms (Carrell *et al.*, 2009), dorm rooms (Sacerdote, 2001), sport partners (Guryan *et al.*, 2009) and among firm managers (Fafchamps and Quinn, 2016). Random assignment to a group is likely to increase interactions among those assigned to the same group, and through this affect the social effect of interest. Social effect parameters identified using this variation would thus not be subject to biases associated with endogenous network formation.

Nonetheless, researchers still need to account for unobserved network shocks in order to obtain consistent estimates of the social effect.[8] To account for these confounders, existing studies use pre-randomization, rather than contemporaneous, values of outcomes and characteristics. In particular, they estimate reduced-form specifications of the following type:

$$\boldsymbol{Y}_{post} = \alpha\tilde{\boldsymbol{\iota}} + \boldsymbol{w}_y(\boldsymbol{G},\boldsymbol{Y}_{pre})\tilde{\boldsymbol{\beta}} + \boldsymbol{X}_{pre}\tilde{\boldsymbol{\gamma}} + \boldsymbol{w}_x(\boldsymbol{G},\boldsymbol{X}_{pre})\tilde{\boldsymbol{\delta}} + \tilde{\boldsymbol{\varepsilon}}_{post} \tag{3}$$

where the subscript *post* indicates variables measured after random assignment to the network, and *pre* indicates variables measured before random assignment. When shocks are i.i.d., the pre-randomization outcome $\boldsymbol{Y}_{pre}$, will be uncorrelated with current unobserved shocks, allowing for identification of the reduced form social effect parameter, $\tilde{\boldsymbol{\beta}}$. This need not solely capture the spillover of peers' outcomes on a node's own outcome. It will also capture other channels through which past peer outcomes may

influence the node's current outcome, so that $\tilde{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$ in equation (1). For example, in a classroom setting, a teacher may put in more effort to teach a class with higher past performance, leading to $\tilde{\boldsymbol{\beta}} > \boldsymbol{\beta}$.

There are two further limitations to this approach. First, forced creation of links is very difficult to achieve in practice: links can only be encouraged (or discouraged) by the random assignment rule. The formation of more complex network structures such as transitive or intransitive triads is not currently well understood, making it difficult to use this method to generate exogenous variation in these. Second, the identified parameter will capture a local, rather than average, effect.[9] In particular, the experiment allows researchers to study the effect of altering an agent's randomly chosen group members on his outcome. If agents form links only with a subset of group members, and make this choice non-randomly (e.g. they choose those that provide the highest net value), these estimates will not be very informative about the likely social effect when the group is constructed in another way, making it difficult to draw credible policy recommendations.

This is demonstrated in the work of Carrell *et al.* (2013), who use peer effects estimated in an earlier paper (Carrell *et al.*, 2009) to 'optimally assign' a random sample of Air Force Academy students to squadrons, with the intention of maximizing the achievement of lower ability students. In fact, test performance in the 'optimally assigned' squadrons turned out to be worse than in the unconditionally randomly assigned squadrons! The authors suggest that this finding is driven by a failure to account for the choice of links formed by individuals within squadrons.[10]

3.2 *Quasi-Experimental Approaches*

A second approach exploits natural or quasi-experiments that generate *local shocks* in network structure that can be argued to be independent of nodes' network formation propensities as well as of common network-level unobserved variables.[11] Examples include unanticipated deaths of individuals (Patnam, 2013, for board members; Mohnen, 2016, for super-star scientists), policy-based reassignments of students to schools (Hoxby and Weingarth, 2005), the Nazi expulsion of Jewish scientists (Waldinger, 2010, 2012) and natural disasters such as the 2011 Great East Japan earthquake (Carvalho *et al.*, 2016). This method recovers a social effect parameter by comparing outcomes of agents affected by a shock to their local network with those of agents with similar pre-shock characteristics (including local network structure) who do not face a shock to their local network. The key underlying assumption is that agents with similar pre-shock observed characteristics and local network structure would have faced a similar trend in their outcomes in the absence of the shock.

In addition, this method also requires that agents choose not to directly respond to the shock.[12] Importantly, non-response in this case includes both, not adjusting links in response to the shock, and not *ex ante* choosing links strategically to (unobservably) insure against the probabilistic exogenous link destruction process. This can be difficult to satisfy in practice: in the case of the unanticipated deaths of board members, for example, the former restriction would imply that company boards do not immediately fill the emerging vacancy with a similarly connected new board member, while the latter restriction would imply ignoring the board member's age and health status when hiring. Finally, if there is heterogeneity in the social effect, this approach provides only a local social effect, based on an average over the links that change as a result of the shock. This may not be representative of the average social effect if, for example, older board members have more influence and are more likely to die.

3.3 *Instrumental Variables*

An alternative approach is to use instrumental variables: variable(s) correlated with the endogenous network covariate, $\boldsymbol{w}_y(\boldsymbol{G},\boldsymbol{Y})$ in equation (1) but excluded from the equation itself. Applications of

this approach include Munshi and Myaux (2006), Hoxby (2000), Mihaly (2009), König *et al.* (2014), Acemoglu *et al.* (2015), Patacchini and Zenou (2016) and Cohen-Cole *et al.* (forthcoming).

As ever with instrumental variables, their effectiveness as a solution to endogeneity relies on having a good instrument: a variable which has strong predictive power for the network covariate but does not enter the outcome equation directly. This will generally be easiest to find when there are some exogenous constraints that make particular edges much less likely to form than others, despite their strong potential benefits. For example, when studying fertility in rural Bangladesh, Munshi and Myaux (2006) exploit strong social norms that prevent the formation of cross-religion edges even where these might otherwise be very profitable. The restrictions on cross-religion connections mean that having different religions is a strong predictor that two women are not linked.

Another approach in the education literature, pioneered by Hoxby (2000), and applied by Bifulco *et al.* (2011) and Patacchini and Zenou (2016), makes use variation in the composition of peers in different cohorts in the same grades in a school. The underlying argument is that parents may choose a school based on the observed average composition of a cohort, but they will not know the actual composition of a new cohort: differences between the average and realized composition are an 'unexpected shock'. Similarly, cohort composition is not subject to biases arising from schools assigning students of different types to specific classrooms or teachers. Thus, the unexpected variation in cohort composition can be used as an instrument for the composition of a child's peers. A concern with this strategy is that cohort composition could affect achievement through other channels, for example, by changing teachers' behaviour. Hoxby (2000) offers a useful test for this. Specifically, if there are multiple groups (e.g. race), and the effects of group composition on achievement operate solely through an endogenous social effect, then the effects of changing the share of say black students should be the same as that of changing the share of Asian students, given the average achievement of each race group.

Alternatively, secondary motivations for forming edges that are unrelated to the primary outcome could be used to obtain independent sources of variation in edge formation probabilities. An application of this approach is Cohen-Cole *et al.* (forthcoming), who consider multiple outcomes of interest, but where agents can form only a single network which influences all of these. Recent work by König *et al.* (2014) instead makes use of instruments based on the network adjacency matrix predicted from a dyadic network formation model. In their study of spillovers from R&D collaborations between firms connected by a web of collaboration agreements (and who also might compete with one another), link formation is modelled as a function of variables that do not otherwise affect the outcome. Specifically, they use indicators for having collaborated on R&D in the past, having a common collaborator in the past, and lagged measures of firms' technological proximity.

Importantly, this type of solution can only be employed when the underlying network formation model has a unique equilibrium, so there is only one network structure consistent with the characteristics (observed and unobserved) of the agents and environment. When multiple equilibria are possible – generally the case when the incentives for a pair of agents to link depend on the state of the other potential links – instrumental variable solutions cannot be used without imposing some equilibrium selection rule. Issues of uniqueness in network formation models, and how one might estimate the these models, are discussed in Advani and Malde (2014). Care must also be taken when interpreting the estimated social effect, particularly in the presence of effect heterogeneity, since instrumental variables generally identify a local social effect. In particular, the estimated $\hat{\beta}_{IV}$ will be a weighted average of individual-specific β_i's, with more weight given to agents for whom the network covariate of interest is induced to change most by the instrument. Hence, the estimated social effect would be larger than the unweighted average social effect if these agents are also those whose outcomes are most responsive to those of their peers (or vice versa).

3.4 *Jointly Modelling Link and Action Choices*

3.4.1 *Sequential Link and Action Choices*

Another method that has been proposed (Blume *et al.*, 2015) and implemented in recent work is the control function. Endogenous linking decisions create selectivity bias in social effect estimates. Control function methods propose to correct this by including an estimated selectivity bias term, estimated from a first stage network formation model, as an additional regressor in the main equation of interest (Heckman, 1979; Lee, 1983; Heckman and Robb, 1985). Recent work by Goldsmith-Pinkham and Imbens (2013), Arduini *et al.* (2015), Horrace *et al.* (2016) and Hsieh and Lee (2016) extends control function methods to a networks context. The selection correction term is a non-linear function of the predicted network, and thus of variables determining link choice. Identification of the social effect parameter can be achieved even in the absence of a variable that influences the outcome only through link choices (an exclusion restriction) by relying on functional form assumptions. The presence of an exclusion restriction, however, may make identification more credible.

The key challenge in operationalizing this method is specifying a sufficiently tractable first-stage model of link formation. This is a result of the size of the joint distribution of edges: for a directed binary network this is a $N(N-1)$-dimensional simplex with $2^{N(N-1)}$ points of support (potential networks).[13] Recent advances in specifying and estimating network formation models are detailed in Advani and Malde (2014), Graham (2015), Chandrasekhar (2015) and de Paula (forthcoming).

Context-specific features can potentially help simplify the first-stage model. For example, Horrace *et al.* (2016) consider the performance of a sports team, where the network is taken to be the set of players that play in the same game for one team. The team size is fixed, and relatively small, so that the network formation process can be modelled as the choice of selecting a fixed number of players from a longer list. Under the assumption that the team manager's choice of players is solely a function of a random shock he observes, but which is not observed by the researcher, parametric and semi-parametric selection correction approaches suggested by Lee (1983) and Dahl (2002) can be applied to account for endogenous link formation.[14] As explained above, identification of model parameters relies on functional form assumptions.

Other studies including Goldsmith-Pinkham and Imbens (2013), Hsieh and Lee (2016) and Arduini *et al.* (2015) use dyadic models of link formation.[15] The former two studies incorporate a 'strategic' element to network formation, whereby linking decisions are allowed to depend on the status of other links in the network. Goldsmith-Pinkham and Imbens (2013) assume that links are formed homophilously – individuals who have more similar characteristics are more likely to be friends – but they also allow network covariates to enter the link formation model. Similarity can be based on the observed characteristics, $\boldsymbol{X}$, and/or on one (binary) unobserved characteristic, ς. By imposing parametric restrictions on the distribution of the unobservable, they are able to characterize a parametric distribution for $(\boldsymbol{Y}, \boldsymbol{G})$. Likelihood estimation can then be used to recover the parameters. The presence of network covariates makes this computationally difficult to estimate directly, since the space of possible networks is large, making the denominator in the likelihood function difficult to compute. A Bayesian Markov Chain Monte Carlo (MCMC) approach is used to overcome this, by providing an estimate for the denominator based on a sample of networks.

Hsieh and Lee (2016) consider linking decisions in directed networks in a framework similar to Goldsmith-Pinkham and Imbens (2013), though crucially they allow for decisions to be affected by multiple unobserved variables. Linking decisions are assumed to be homophilous, and are influenced by dyad-specific characteristics, $\boldsymbol{C}$, individual characteristics, $\boldsymbol{X}$, and unobserved network statistics such as transitivity. Assuming that the unobservable terms in the social effects and the network formation equations are joint normally distributed, Hsieh and Lee (2016) are able to characterize the conditional distribution of $(\boldsymbol{Y}, \boldsymbol{G}|\boldsymbol{X}, \boldsymbol{C}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of model parameters from both the network formation

and social effect equations. The dyad-specific characteristics appear only in the link formation model, and thus provide exclusion restrictions for the identification of model parameters. As with Goldsmith-Pinkham and Imbens (2013), likelihood estimation using maximum likelihood is computationally difficult, necessitating the use of a Bayesian MCMC approach.

Arduini *et al.* (2015) consider two further ways of modelling the first stage: (i) a dyadic link formation model of Graham (2017), which assumes homophilous link formation and agent-specific unobserved heterogeneity, and (ii) a model where the link formation probability is a function of the node's characteristics only. The former assumption requires parametric estimation, while the latter method allows for semi-parametric estimation. They derive the asymptotic properties of the estimators, and evaluate their effectiveness in correcting for endogeneity using simulations.

3.4.2 *Simultaneous Link and Action Choices*

A final method for accounting for endogeneity also relies on jointly modelling link formation and action choices though, contrary to the control function approach, links and actions are simultaneously chosen. This approach is taken by Boucher (2016) and Badev (2017), who model peer effects among adolescents in extracurricular activities and smoking choice respectively, allowing agents to choose their action (activity/smoking decisions) simultaneously with their links. In both cases, the action and link decisions will generally be non-separable.

In Boucher (2016), agents get utility directly from links, from playing an action (activity choice) close to their type, and from conforming on action to the actions of the people they are linked to. He shows that, close to the optimum, utility is (locally) differentiable with respect to the action. Intuitively, since the action can be changed smoothly, while linking decisions are binary, utility should change smoothly with changes in the action around the optimum.[16] To also study link choice, Boucher shows that the game can be characterized by a potential function. He provides bounds on the maximum of this function, and assumes that this maximum is associated with the equilibrium that will be selected in practice. He then estimates (by quasi-maximum likelihood estimation) the equation determining the action combined with the network formation equation, for both of the bounds on the network. In practice when the network is sparse each bound will give similar answers: this is the case in his context.

In Badev (2017) link choice is strategic even in the absence of the action choice (smoking), since the value of a link to someone depends also on their links. Combining the individual utility functions with a random matching process between individuals and myopic decision-making, he shows that behaviour will converge to a k-player Nash stable state in finite time.[17] Adding Gumbel distributed preference shocks instead implies convergence to a stationary distribution over the set of possible network states, in particular one that is invariant to the choice of k. With these shocks the model maps to an Exponential Random Graph Model (see Section 4.2.4 for more details), for which an analytical characterization of the likelihood function is possible. However, as with the models discussed in the previous subsection, the large number of potential networks makes exact calculation of the denominator of the likelihood function computationally infeasible. Instead, as above, this is approximated using MCMC methods, and then maximum likelihood estimation can be used for this approximated likelihood function.

4. Measurement Error

The second challenge complicating identification of social effect parameters in network data is that of measurement error in the network. Measurement error can arise from a number of sources including: (1) missing data due to sampling method, (2) mis-specification of the network boundary, (3) top-coding of the number of edges, (4) mis-coding or mis-reporting and (5) non-response. We refer to the first three as sampling-induced error and the latter two as non-sampling-induced error. It is important to account

for these since, as we will show below, measurement error can induce important biases in measures of network statistics and in parameter estimates.

We focus on summarizing the consequences of sampling-induced measurement error, and outlining methods proposed in the literature to deal with these. Though a number of issues remain unresolved, this literature offers useful guidance to researchers planning to collect data to uncover social effects in terms of (i) how to construct a sample; and (ii) what data to collect and from whom. Note also that there is a large econometric and statistical literature on non-sampling induced measurement error, which could potentially apply or be extended to network contexts, for example, Chen *et al.* (2011) provide an overview of methods for dealing with misreporting in binary variables. However, these issues have been less studied in a networks context, and are thus not covered here.[18]

Measurement error issues arising from sampling are particularly problematic in the context of network data, since these data comprise of information on interrelated objects: nodes and edges. All sampling methods, other than a full census, sample at least one of these objects in a way that depends on the network structure: defining a random sampling process over one induces a particular process over the other.[19] To illustrate how this may happen, consider taking a random sample of nodes from a star network, which consists of a single central node directly connected to $N-1$ other peripheral nodes, with no other connections between them. If we were to randomly sample half the nodes in the network, we would sample the central node half the time. However, if we were to randomly sample half the links, we would always sample the central node, since every edge is connected to this node, and sample peripheral nodes roughly half the time only. Thus, random sampling of edges would lead to a higher chance of sampling nodes with many edges, giving a different sampling distribution for nodes compared to when directly sampling nodes. This means that methods for estimation and inference developed under classical sampling theory are often not applicable to network data.

In practice, censuses of networks that economists wish to study are rare, and feasible to collect only in a minority of cases (e.g. small classrooms or villages). Collection of data on the complete network is typically too expensive and cumbersome. Moreover, when data are collected from surveys, it is common to censor the number of edges that can be reported by nodes. Finally, to simplify data collection, one may erroneously limit the boundary of the network to a specified unit, for example, village or classroom, thereby missing edges connecting to nodes beyond this boundary. Section 4.1 outlines the consequences of missing data due to sampling on estimates of social effects and on network statistics. Until recently most research on these issues was done outside economics, so we draw also on research from other fields, including sociology, statistical physics and computer science. In Section 4.2, we then outline a number of methods developed to help deal with the consequences of measurement error.

Much of our discussion in the subsequent sections will consider two specific ways of constructing a network graph from sampled nodes. Given a sample of nodes, one could consider including only the edges among pairs of sampled nodes, generating an *induced subgraph*. Alternatively, one could include all edges of sampled nodes, including non-sampled nodes connected to sampled nodes within the network graph. This generates a *star subgraph*. These are displayed in Figure A1 in Appendix A. Panel (a) of the figure shows the network from which nodes are randomly sampled, while the shaded circles and dark lines in panels (b) and (c) display the network that emerges under star subgraph sampling and induced subgraph sampling, respectively.

4.1 *Measurement Error Due to Sampling*

4.1.1 *Local Network Models*

Missing data, for sampling or non-sampling reasons, can generate important biases in the estimates of social effects in the local average, local aggregate and hybrid local models. Identification strategies for

the social effect in these models exploit variation in network structure, typically using the exogenous characteristics of indirect neighbours as instruments for the outcomes of a node's direct neighbours ($\boldsymbol{w}_y(\boldsymbol{G},\boldsymbol{Y})$ in equation (1)). For example, in the local average model Bramoullé *et al.* (2009) suggest using the average exogenous characteristics of second- and third-degree neighbours, $\tilde{\boldsymbol{G}}^2\boldsymbol{X}$ and $\tilde{\boldsymbol{G}}^3\boldsymbol{X}$, as instruments for the endogenous $\tilde{\boldsymbol{G}}\boldsymbol{Y}$ ($\tilde{\boldsymbol{G}}^3\boldsymbol{X}$ is needed when we wish to account for network fixed effects). Critically, identification comes from knowledge of which edges are definitely *not* present. When data are missing or misclassified, one may not know definitively which nodes are only indirectly linked, complicating the use of this strategy.

Goldsmith-Pinkham and Imbens (2013) propose a test for measurement error in the network when more than one observation of the network is available. This will be the case, for example, in longitudinal network studies where the network is elicited on multiple occasions over time. The basic intuition underlying their test is that if measurement error is unconditionally random, and a link is absent in one observation of the network, there is a higher probability that it is missing spuriously (and hence was mismeasured) in the first observation if it is present in the second observation. If this is the case, we would expect these mismeasured links' characteristics and outcomes to also affect a node's outcome. To illustrate their method more formally, we introduce some additional notation: let $\boldsymbol{G}^A$ and $\boldsymbol{G}^{A'}$ denote the first and second measurements of the adjacency matrix related to the outcome of interest; and $\boldsymbol{G}^B$ denote a matrix that indicates which links are absent in $\boldsymbol{G}^A$ but present in $\boldsymbol{G}^{A'}$. The presence of unconditionally random measurement error can be tested by estimating the following equation for linear $\boldsymbol{w}_y$ and $\boldsymbol{w}_x$:

$$\begin{aligned}\boldsymbol{Y} = {} & \alpha\boldsymbol{\iota} + \boldsymbol{w}_y(\boldsymbol{G}^A,\boldsymbol{Y})\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{w}_x(\boldsymbol{G}^A,\boldsymbol{X})\boldsymbol{\delta} + \boldsymbol{w}_y(\boldsymbol{G}^B,\boldsymbol{Y})\boldsymbol{\beta}^B \\ & + \boldsymbol{w}_x(\boldsymbol{G}^B,\boldsymbol{X})\boldsymbol{\delta}^B + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{L}\boldsymbol{v} + \boldsymbol{\varepsilon} \end{aligned} \tag{4}$$

If $\boldsymbol{G}^A$ is well measured, links that are present in $\boldsymbol{G}^{A'}$ but not in $\boldsymbol{G}^A$ should not influence the outcome of interest, $\boldsymbol{Y}$. Hence, the coefficients on their outcomes and characteristics, $\boldsymbol{\beta}^B$ and $\boldsymbol{\delta}^B$, should be 0. Non-zero coefficients would be indicative of measurement error in the network. Note though that these coefficients could be non-zero even in the absence of measurement error if, for example, outcomes are correlated over time and the two measurements correspond to adjacency matrices collected at two points in time. Any such alternative explanations should be carefully considered when using this strategy to test for measurement error.

Measurement error in the network due to sampling implies that the matrices $\boldsymbol{G}$ and $\tilde{\boldsymbol{G}}$ are misspecified. In particular, when some links are missing, any two nodes would appear to be, on average (weakly), further apart in the sampled network than they are in the true underlying network. This measurement error carries over to the endogenous covariate $\tilde{\boldsymbol{G}}\boldsymbol{Y}$ in the local average model, as well as the instruments $\tilde{\boldsymbol{G}}^2\boldsymbol{X}$ and $\tilde{\boldsymbol{G}}^3\boldsymbol{X}$. Further, since it is common to both the endogenous covariate and instrument, the instrument will be unable to purge the social effect parameter of bias (Chandrasekhar and Lewis, 2016). Simulations by Chandrasekhar and Lewis (2016) and Liu (2013) suggest (respectively) that these biases can be very large in local average and local aggregate models, with the magnitude falling as the proportion of the network sampled increases, and as the number of networks in the sample increases. Both papers also offer simple, direct solutions to this issue when data are available on a star subgraph: these are described in Section 4.2.1.

Patacchini *et al.* (2017) also use simulations to consider the robustness of social effect estimates in a model with heterogeneous social effects. Their data include a high proportion of missing nodes. Contrary to the simulations in Chandrasekhar and Lewis (2016) and Liu (2013), their simulations add new links to the observed network, some of which lead to mistakenly classifying neighbours as neighbours-of-neighbours. They show that their findings on peer effects hold qualitatively, though they over-estimate the magnitude of one type of peer effect. Such simulations offer one way for researchers to check the robustness of social effects estimates to missing data.

4.1.2 *Network Statistics*

Missing data arising from partial sampling can generate non-classical measurement error in measured network statistics, which in turn biases estimates of social effects. A number of studies, primarily in fields outside economics, have investigated the implications of sampled network data on measures of network statistics and model parameters. The following broad facts emerge from this literature:

1. *Network statistics computed from samples containing moderate (30–50%) and even relatively high (∼70%) proportions of nodes in a network can be highly biased. Sampling a higher proportion of nodes in the network generates more accurate network statistics.* Simulation evidence from studies including Galaskiewicz (1991), Costenbader and Valente (2003), Lee *et al.* (2006), Kim and Jeong (2007) and Chandrasekhar and Lewis (2016) indicates biases that are very large in magnitude, and which go in different directions, depending on the statistic being studied.[20] For example, the average path length – the average number of links one has to go through on the shortest path between any pair of nodes – was found to be over-estimated by 100% when constructed from an induced subgraph with 20% of nodes in the true network. Table A1 in Appendix A provides a more detailed summary of findings from these papers for some commonly used network statistics for data collected via random sampling of nodes as either a star subgraph or an induced subgraph.
2. *Measurement error due to sampling varies with the underlying network structure.* This is apparent from work by Frantz *et al.* (2009), who investigate the robustness of a variety of centrality measures to missing data when data are drawn from a range of underlying network structures: uniform random, small world, scale-free, core-periphery and cellular networks (see Appendix B for definitions). They find that the accuracy of centrality measures varies with the structure. Small world networks are especially vulnerable to missing data, since they have relatively high clustering and a few 'bridging' edges that reduce path lengths between nodes that would otherwise be distant. The estimated centrality statistics are therefore very sensitive to sampling the nodes that are part of a bridge. By contrast, centrality measures are less vulnerable to missing data when the underlying network is 'scale-free'.
3. *The magnitude of error in network statistics that is due to sampling varies with the sampling method.* Lee *et al.* (2006) compare the results of estimating network statistics using data collected via induced subgraph sampling, random sampling of nodes, random sampling of edges and snowball sampling (see Appendix C for more details on sampling strategies). They draw samples from networks with a power-law degree distribution, that is, where the fraction of nodes having k edges, $P(k)$, is asymptotically proportional to $k^{-\gamma}$, and usually $2 < \gamma < 3$. This distribution allows for 'fat tails', that is, the proportion of nodes with very high degrees constitutes a non-negligible proportion of all nodes. Lee *et al.* (2006) show that the sampling method impacts the magnitude and direction of bias in network statistics. For instance, random sampling of nodes and edges leads to over-estimation of the size of the exponent of the power-law degree distribution, which implies an over-estimation of the number of nodes with large degrees. Conversely, snowball sampling, which is less likely to find nodes with low degrees, underestimates this exponent.
4. *Parameters in economic models using mismeasured network statistics are subject to substantial bias.* Sampling induces non-classical measurement error in the measured statistic, that is, the measurement error is not independent of the true network statistic. Chandrasekhar and Lewis (2016) suggest that sampling-induced measurement error can generate upward bias, downward bias or even sign switching in parameter estimates. The bias is large in magnitude: for statistics such as degree, clustering and centrality measures, they find that the mean bias in parameters in network-level regressions ranges from over-estimation bias of 300% for some statistics to attenuation bias of 100% for others when a quarter of network nodes are sampled. As with network statistics, the bias becomes smaller in magnitude as the proportion of the network sampled increases. The

magnitude of bias is somewhat smaller, but nonetheless substantial, for node-level regressions. Table A2 summarizes the findings from the literature on the effects of random sampling of nodes on parameter estimates.

5. *Top-coding of edges or incorrectly specifying the boundary of the network biases network statistics.* Network data collected through surveys often place an upper limit on the number of edges that can be reported. Moreover, limiting the network boundary to an observed unit, for example, a village or classroom, will miss nodes and edges beyond the boundary. Kossinets (2006) investigates, via simulations, the implications of top-coding of reported edges and boundary misspecification. He considers a number of network statistics, including average degree, clustering and average path length. Both types of error cause average degree to be under-estimated, and average path length to be over-estimated. No bias arises in the estimated clustering parameter when only top-coding is present.

Overall, the literature indicates that even relatively little missing data (e.g. observing 75% of nodes) may generate severe non-classical measurement error in network statistics, as well as severely biased parameter estimates, highlighting the need for a census of the network. However, this can be very costly or infeasible to collect. Work in disciplines outside economics, as well as recent work in economics, has proposed a number of possible methods for dealing *ex post* with the consequences of missing data. We review this literature in the next subsection.

4.2 *Correcting for Measurement Error*

Having considered the problems posed by missing data on both the network and parameter estimates, we now discuss methods for dealing with measurement error *ex post*, that is, once data have been collected. These can be divided into four broad classes: (1) direct corrections, (2) design-based corrections, (3) likelihood-based corrections and (4) model-based corrections. We summarize the underlying ideas for each of these, and discuss their advantages and drawbacks.

4.2.1 *Direct Corrections*

As we saw earlier, missing data on network connections generate measurement error in both the endogenous regressor and the network-based instruments in local network models, thereby inducing bias in social effects. Chandrasekhar and Lewis (2016) suggest a simple, direct correction for this issue for the local average model when the network data available are a star subgraph collected from a random sample of nodes, and outcome data are available for all agents. In particular, they suggest restricting the estimation sample to include only the initially sampled nodes. For these nodes, data on all their neighbours (and the neighbours' outcomes) are observed, meaning that the regressor $\tilde{\boldsymbol{G}}\boldsymbol{Y}$ will not be subject to measurement error. The key instruments for identification, $\tilde{\boldsymbol{G}}^2\boldsymbol{X}$ and $\tilde{\boldsymbol{G}}^3\boldsymbol{X}$, can be constructed as usual using all the observed data. They will be mismeasured, but, crucially, the measurement error in the instruments will now not be correlated with the regressor, making them valid instruments. However, the measurement error in the instruments weakens the first-stage correlation with the endogenous regressors, particularly when the amount of missing data on the network is high, leading to a weak instrument problem. In this case, other methods, including model-based corrections, could be applied.

For the local aggregate model, an alternative solution exists when network fixed effects are not necessary. In the absence of measurement error, the standard approach to identification uses node degree ($\boldsymbol{G}\boldsymbol{L}$), along with the network-based instruments, $\boldsymbol{G}^2\boldsymbol{X}$ and $\boldsymbol{G}\tilde{\boldsymbol{G}}\boldsymbol{X}$ as instruments for the mismeasured endogenous regressor $\boldsymbol{G}\boldsymbol{Y}$. This provides over-identification, since only one instrument is needed in the absence of network fixed effects. When data from a star subgraph are available, node out-degree is still

typically well-measured, meaning that it can be used as the only instrument for $\boldsymbol{GY}$, and the noisier mismeasured instruments using indirect neighbours can be ignored. This is supported by Monte Carlo simulation evidence in Liu (2013), which shows that estimates recovered using this strategy are very similar to the parameters from the pre-specified data generating process.

Liu *et al.* (forthcoming) suggest a solution for the case where there the network and covariates are perfectly observed, but outcome data are available for a sub-sample only. They note that the reduced form equation for the local average model, when restricted to the observations for whom complete outcome data are available, involves regressing the outcome on a non-linear transformation of $\boldsymbol{X}$ and $\tilde{\boldsymbol{G}}\boldsymbol{X}$. Such data are consistent with survey designs that collect network information and some key covariates from all nodes and detailed outcome data from a sample. Drawing on an argument in Wang and Lee (2013), they show that model parameters can be consistently estimated from the transformed reduced form equation using nonlinear least squares. Monte Carlo simulations suggest the method works well.

4.2.2 *Design-Based Corrections*

Design-based corrections rely on features of the sampling design to correct for sampling-induced measurement error. They are appropriate for correcting network-level statistics that can be expressed as totals or averages, such as average degree and clustering (Frank; 1978, 1980a, 1980b, 1981; Thompson, 2006).[21] Based on *Horvitz-Thompson* estimators, which use inverse probability-weighting to compute unbiased estimates of population totals and means from sampled data, they can be used to correct for the non-random sampling of either nodes or edges provided that the sample inclusion weights of the non-randomly sampled object can be calculated.

Formulae for node- and edge-inclusion probabilities are available for the random node and edge sampling schemes (see Kolaczyk, 2009). Recovering sample inclusion probabilities when using snowball sampling – where a sample is constructed by first collecting information on the neighbours of some (randomly) selected agents, then gathering information on the neighbours of these neighbours and so on (see Appendix C for more) – is typically not straightforward after the first step of sampling. This is because every possible sample path that can be taken in subsequent sampling steps must be considered when calculating the sample-inclusion probability, making this exercise very computationally intensive. However, Markov chain resampling methods make it feasible to estimate the sample inclusion probabilities (see Thompson, 2006, for more details). An application of this method in economics is given by Mastrobuoni and Patacchini (2012) and Mastrobuoni (2015), who use a Markov chain–based method to correct for non-random selection of nodes into a sample of mobsters followed by law enforcement officials in the United States. They model the sample construction of mobsters as a snowball sample, which can further be modelled as a Markov chain. The stationary distribution of the Markov chain of the sample inclusion probabilities provides the likelihood of a node i being found when following any randomly selected edge in the network.[22]

Frank (1978, 1980a, 1980b, 1981) derives unbiased estimators for a range of graph statistics. Chandrasekhar and Lewis (2016) characterize the biases in parameter estimates from linear univariate models for a range of network statistics, and provide guidance on how these biases may be corrected. They show that attenuation biases can be easily corrected by estimating the variance of the measurement error, and offer corrections for the scaling biases based on their characterisation. They further show that for four statistics – average degree, clustering coefficient, support and average graph span – estimators of social effect parameters are consistent when raw network statistics are replaced by their design-corrected counterparts. Numerical simulations suggest that this method reduces greatly the sampling-induced bias in parameter estimates.

A key drawback to this procedure is that it is not possible to compute Horvitz–Thompson estimators for network statistics that cannot be expressed as totals or averages. This includes node-level statistics, such

as eigenvector centrality, many of which are of interest to economists. Likelihood-based and model-based corrections offer alternative solutions that are more feasible in these cases.

4.2.3 *Likelihood-Based Corrections*

Likelihood-based corrections can also be applied to correct for measurement error. Such methods have been used to correct specific network-based statistics such as out-degree and in-degree. Conti *et al.* (2013) correct for sampling-induced measurement error in in-degree by adjusting the likelihood function. To do so, they first specify a process for outgoing and incoming edge nominations to obtain the outgoing and incoming edge probabilities. Specifically, they assume that outgoing (incoming) edge nominations from i to j are a function of i's (j's) observable preferences, the similarity between i and j's observable characteristics (capturing homophily), and a scalar unobservable for i and j. They allow for correlations between i's observable and j's unobservable characteristics (and vice versa). When edges are binary, the out-degree and in-degree have binomial distributions with the success probability given by the calculated outgoing and incoming edge probabilities. Random sampling of nodes to obtain a star subgraph generates measurement error in the in-degree, but not in the out-degree. However, since the true in-degree is binomially distributed, and nodes are randomly sampled, the observed in-degree has a hypergeometric distribution conditional on the true in-degree. Knowledge of these distributions allows the specification of the joint distribution of the true in-degree, the true out-degree, and the mismeasured in-degree. Pseudo-likelihood functions can be specified allowing for parameters to be consistently estimated via maximum likelihood methods.

4.2.4 *Model-Based Corrections*

Model-based corrections provide an alternative approach to correcting for measurement error. Such corrections involve specifying a model that maps the mismeasured network to the true network. Parameters of the model are estimated from the partially observed network data and the available data on the characteristics of nodes and edges. The estimated parameters are subsequently used to predict the value of non-sampled edges, essentially imputing the missing values. Network formation models usually recover the probability of a link, meaning that the predicted network is a matrix of probabilities. The predicted network can then be used in place of the mismeasured network to obtain an estimate of the social effect. To do this it is crucial to have information on individual characteristics (e.g. gender, ethnicity) that are predictive of link formation for *all* nodes in the network. It is also important that the network formation model estimated is sufficiently flexible to accurately capture the observed network(s).

When covariates on all nodes in the network are available, Chandrasekhar and Lewis (2016) derive conditions that must be satisfied for this approach to yield consistent estimates of the social effect parameter when allowing for the first stage network formation process to be heterogeneous across networks. In particular, the estimator of the network formation parameters must converge uniformly to the true parameters. This imposes restrictions on the available data – in particular, the number of networks must grow slower than the size of the networks – and on the first stage network formation model, assuming that data are missing at random.

Chandrasekhar and Lewis (2016) analyse three different classes of network formation models to derive the conditions under which they generate consistent social effect estimates. These models are known to have asymptotic frames which allow for consistent parameter estimation.[23]

The first model is the conditional edge independence model (Fafchamps and Gubert, 2007; Goldsmith-Pinkham and Imbens, 2013; among others), where links form independently, conditional on covariates. The probability of a link is typically modeled as a function of node- and link-level covariates. Chandrasekhar and Lewis (2016) show this model satisfies the conditions for uniform convergence as long as the level

of interdependence in covariates between a pair of nodes goes to 0 as the (social) distance between the two nodes increases to infinity. However, these models typically fail to generate clustering levels similar to those seen in real-life social networks.

A second class of models are the *subgraph generated models* of Chandrasekhar and Jackson (2016), which model the network to be the union of different network features (pairs, triangles, etc.) that each form with a certain probability. Chandrasekhar and Lewis (2016) shows that this model satisfies the conditions for uniform convergence given an assumption on convergence rates is satisfied. This class of models does not require information on node-level covariates.

A final class of models considered is the *group* or *block model*, where the link formation probability is a function of group-specific parameters. A group is defined based on the values of a combination of (bounded) characteristics (e.g. high educated females aged < 40 years). In other words, the model can be thought of one with group-fixed effects and a growing number of groups, which allows for substantial flexibility in characterizing the underlying network formation process. However, since the number of parameters to be estimated can grow with the network size, Chandrasekhar and Lewis (2016) show that sufficiently fast convergence can only be achieved for network-level analysis, not for node-level analysis.

It should be noted that misspecification of the first stage model could undermine the ability of this method to correct for measurement error. In particular, conditional edge independence models may not be well suited to correcting measurement errors in network clustering, but may be sufficient in correcting measurement error in average degree. Thus, the characteristic one is trying to correct should be taken into consideration when choosing the first stage model. Simulations in Chandrasekhar and Lewis (2016) show that model based corrections work well in greatly reducing and almost eliminating biases in social effect parameters arising from missing data for a number of social effect models including the local average model.

5. Conclusion

Networks are thought to play an important role in shaping the preferences, behaviour and outcomes of agents. Uncovering empirical evidence in support of this has proven to be difficult, particularly when using information on membership of mutually exclusive groups as the key measure for social interactions. A burgeoning literature in economics has turned instead to using network data – data with detailed information on agents and the links between them – to uncover this evidence. However, there exist important challenges that are not present in other contexts. In this paper we outline econometric methods for working with network data to identify social effects: the influence of a node's neighbours on its choices. We focus particularly on methods for dealing with the endogenous formation of links, and solutions to account for measurement error.

There have been a number of approaches taken to account for network endogeneity, including random assignment of interventions or links, use of local network shocks, instrumental variables and jointly modelling the choice of links and outcomes (either sequentially or simultaneously). The first three do not require explicit specification of the process of network formation. Where they are feasible, they can provide credible identification. However, randomly assigning interventions or links is frequently infeasible; and exogenous local network shocks and suitable instruments might not be available in many contexts. Explicit specification of the network formation model, as is required by the last method, provides an alternative approach. This uses knowledge (or assumptions) about the payoffs from forming links to provide a different route to identification. The challenges to this solution are not only in determining what assumptions about payoffs are reasonable, but also technical. Such models are typically difficult to estimate: they are slow to compute, and estimated parameters are frequently unstable. There is much scope for future work in advancing these methods.

Finally, the paper discussed the issue of measurement error, focusing particularly on sampling-induced measurement error. Since networks comprise of interrelated nodes and edges, a particular sampling scheme over one of these objects will imply a structure for sampling over the other. Hence, one must think carefully in this context about how data are collected, and not simply rely on the usual intuitions that random sampling will allow us to treat the sample as the population. When collecting census data is not feasible, it will in general be necessary to make corrections for the induced measurement error, in order to get unbiased parameter estimates. Whilst there are methods for correcting some network statistics for some forms of sampling, again there are few general results, and consequently much scope for research.

Much work has been done to develop methods for working with network data, both in economics and in other fields. Applied researchers can therefore take some comfort in knowing that many of the challenges they face using these data are ones that have been considered before, and for which there are typically at least partial solutions already available. Whilst the limitations of currently available techniques mean that empirical results should be interpreted with some caution, attempting to account for social effects is likely to be less restrictive than simply imposing that they cannot exist.

Acknowledgements

We are grateful to Imran Rasul for his support and guidance on this project. We also thank Richard Blundell, Andreas Dzemski, Toru Kitagawa, Aureo de Paula, Ian Preston, Karen Macours and Yves Zenou for their useful comments and suggestions. Financial support from the ESRC-NCRM Node 'Programme Evaluation for Policy Analysis', Grant reference ES/I03685X/1 is gratefully acknowledged. Malde acknowledges support from the ESRC Future Research Leaders grant ES/K00123X/1.

1. These methods are less suited to discrete choice settings, such as those considered by Brock and Durlauf (2001) and Brock and Durlauf (2007).
2. A row stochastic, or 'right stochastic', matrix is one whose rows are normalized so they each sum to one.
3. The name 'local average' is used here to denote that only *local* (direct) connections affect an individual directly, and the way in which they matter is only through the *average* outcome of these agents to whom the individual directly connects.
4. For a survey of the main network statistics used and the contexts in which they are relevant, see Jackson *et al.* (2017).
5. A discussion of the key network-level statistics used is provided by Jackson *et al.* (2017), where they are described as 'macro characteristics' of the network.
6. It is important to note that this implies that individuals already have some information about the unobservables. If these unobservables are identically distributed, are realized after the network formation decisions are taken, and do not themselves depend on the network structure, then network formation does not create an endogeneity problem. Goldsmith-Pinkham and Imbens (2013) suggest a method to test for endogeneity.
7. A social effect can also be identified by comparing the outcomes of treated nodes with different levels of exposure to other treated nodes. However, such an effect would have a different interpretation.
8. Researchers will also need to account for the reflection problem when information on interactions within the network is not available.
9. Here we think of 'local' effects in terms of a local treatment effect, rather than in the sense of local interactions.
10. Booij *et al.* (2017) and Tincani (2017) provide different interpretations of this result. The former suggests that the problem with the assignment based on the results of Carrell *et al.* (2009) is that the peer groups constructed fall far outside the support of the data used. Hence, predictions about student performance come from extrapolation based on the functional form assumptions used, which

should have been viewed with caution. Tincani (2017) suggests that the findings can be explained by an education production function allowing for competition between students.

11. As with random assignment approaches, quasi-random assignment of interventions on a pre-specified network have also been used to identify social effects. Examples of papers taking such an approach include Banerjee *et al.* (2013).
12. One also needs access to panel data for the network, which may often not be available. Moreover, measurement error in either round of network data will reduce the power of this strategy.
13. To give a sense of scale, for a network of more than seven agents the support of this space is larger than the number of neurons in the human brain (estimated to be around 8.5×10^{10}); with 13 agents it is larger than the number of board configurations in chess (around $10^{46.25}$); and with 17 agents it is larger than the number of atoms in the observed universe (around 10^{80}).
14. They also develop a fixed effects approach, which can only be applied in contexts where the social effect is heterogeneous.
15. In a dyadic model, the link choice is modelled to be a function of characteristics of each node (the sum and/or difference), as well as characteristics of the link. Some models allow for node-specific unobserved heterogeneity.
16. This requires that agents are not indifferent about any of their linking decisions, so are not at kink points of their utility function, which they will not be generically.
17. A network is k-player Nash stable if any subset of k players is in a Nash equilibrium of the game between them when only the links between the k players are decided together with their action choices. This equilibrium concept is well suited in modelling myopic behaviour, but less so for networks formed with the intention of influencing behaviour with a long horizon.
18. Comola and Fafchamps (2017) develop and implement a correction for this class of measurement error in a networks context, while Patacchini *et al.* (2017) use simulations to assess the robustness of estimated peer effects to misspecification of links and link types.
19. We consider a random sample to consist of independently and identically distributed units.
20. With the exception of average degree in a star subgraph, the evidence on the direction and magnitude of biases described here come from simulation studies with specific designs. These may not always hold for all network structures and sampling techniques, as explained below.
21. Chapter 5 of Kolaczyk (2009) provides useful background on these methods.
22. Mastrobuoni (2015) observes less than 20% of nodes in the whole network, which creates further biases. He corrects for these by first taking logarithmic values of the network statistics (and his outcome of interest), and then using instrumental variables. The logarithmic transformation accounts for a scaling bias related to the proportion of the network sampled.
23. Importantly, they do not consider the properties of the so-called p^*-models (Wasserman and Pattison, 2013) or *exponential random graph models* (ERGMs), which model the probability of a link to depend on the links around it, since these models do not have a suitable asymptotic frame.

References

Acemoglu, D., Garcia-Jimeno, C. and Robinson, J. (2015) State capacity and economic development: a network approach. *American Economic Review* 105: 2364–2409.

Advani, A. and Malde, B. (2014) Empirical methods for networks data: Social effects, network formation and measurement error. IFS Working Paper W14/34.

Advani, A. and Malde, B. (forthcoming) Methods to identify linear network models: a review. *Swiss Journal of Economics and Statistics*.

Angelucci, M., De Giorgi, G., Rangel, M.A. and Rasul, I. (2010) Family networks and school enrolment: Evidence from a randomized social experiment. *Journal of Public Economics* 94: 197–221.

Angelucci, M., De Giorgi, G. and Rasul, I. (forthcoming) Consumption and investment in resource pooling family networks. *Economic Journal*.
Aral, S. and Walker, D. (2012) Identifying influential and susceptible members of social networks. *Science* 337: 337–341.
Arduini, T., Patacchini, E. and Rainone, E. (2015) Parametric and semiparametric IV estimation of network models with selectivity. EIEF Working Paper.
Babcock, P.S. and Hartman, J.L. (2010) Networks and workouts: treatment size and status specific peer effects in a randomized field experiment. NBER Working Paper, WP 16581.
Babcock, P.S., Bedard, K., Charness, G., Hartman, J.L. and Royer, H. (2015) Letting down the team? Social effects of team incentives. *Journal of the European Economic Association* 13: 841–870.
Badev, A.I. (2017) Discrete games in endogenous networks: theory and policy. Arxiv preprint arXiv:1705.03137.
Banerjee, A., Chandrasekhar, A.G., Duflo, E. and Jackson, M. (2013) The diffusion of microfinance. *Science* 341: 1236498.
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: 25–53.
Blume, L.E., Brock, W.A., Durlauf, S.N. and Ioannides, Y.M. (2010) Identification of social interactions. In J. Benhabib, A. Bisin, and M. Jackson (eds.), *Handbook of Social Economics*, Volume 1B. Amsterdam, The Netherlands: North Holland.
Blume, L.E., Brock, W.A., Durlauf, S.N. and Jayaraman, R. (2015) Linear social interaction models. *Journal of Political Economy* 123: 444–496.
Booij, A.S., Leuven, E. and Oosterbeek, H. (2017) Ability peer effects in university: evidence from a randomized experiment. *Review of Economic Studies* 84: 547–578.
Boucher, V. (2016) Conformism and self-selection in social networks. *Journal of Public Economics* 136: 30–44.
Boucher, V. and Fortin, B. (2015) Some challenges in the empirics of the effects of networks. In Y. Bramoulle, A. Galeotti, and B. Rogers (eds.), *The Oxford Handbook of the Economics of Networks* (pp. 277–302). New York, USA: Oxford University Press.
Bramoullé, Y., Djebbari, H. and Fortin, B. (2009) Identification of peer effects through social networks. *Journal of Econometrics* 150: 41–55.
Bramoullé, Y., Kranton, R. and D'Amours, M. (2014) Strategic interaction and networks. *American Economic Review* 104: 898–930.
Breza, E. and Chandrasekhar, A.G. (2015) Social networks, reputation and commitment: evidence from a Savings Monitors Experiment. NBER Working Paper, WP 21169.
Brock, W.A. and Durlauf, S.N. (2001) Discrete choice with social interactions. *Review of Economic Studies* 68: 235–260.
Brock, W.A. and Durlauf, S.N. (2007) Identification of binary choice models with social interactions. *Journal of Econometrics* 140: 52–75.
Cai, J., De Janvry, A. and Sadoulet, E. (2015) Social networks and the decision to insure. *American Economic Journal: Applied Economics* 7: 81–108.
Calvó-Armengol, A., Patacchini, E. and Zenou, Y. (2009) Peer effects and social networks in education. *Review of Economic Studies* 76: 1239–1267.
Carrell, S., Fullerton, R. and West, J. (2009) Does your cohort matter? estimating peer effects in college achievement. *Journal of Labor Economics* 27: 439–464.
Carrell, S., Sacerdote, B. and West, J. (2013) From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica* 81: 855–882.
Carvalho, V., Nirei, M., Saito, Y. and Tahbaz-Salehi, A. (2016) Supply chain disruptions: Evidence from the great east Japan earthquake. Mimeo, University of Cambridge.
Chandrasekhar, A.G. (2015) Econometrics of network formation. In Y. Bramoulle, A. Galeotti, and B. Rogers (eds.), *The Oxford Handbook of the Economics of Networks* (pp. 303–357). New York, USA: Oxford University Press.
Chandrasekhar, A.G. and Jackson, M. (2016) A network formation model based on subgraphs. Mimeo, Stanford.
Chandrasekhar, A.G. and Lewis, R. (2016) Econometrics of sampled networks. Mimeo, Massachusetts Institute of Technology.

Chen, X., Hong, H. and Nekipelov, D. (2011) Nonlinear models of measurement errors. *Journal of Economic Literature* 49: 901–937.
Chuang, Y. and Schechter, L. (2015) Social networks in developing countries. *Annual Review of Resource Economics* 7: 451–472.
Cohen-Cole, E., Liu, X. and Zenou, Y. (2017) Multivariate choices and identification of social interactions. *Journal of Applied Econometrics*. 1–14. https://doi.org/10.1002/jae.2590
Comola, M. and Fafchamps, M. (2017) The missing transfers: Estimating mis-reporting in dyadic data. *Economic Development and Cultural Change* 65: 549–582.
Comola, M. and Prina, S. (2017) Treatment effects accounting for network changes. Mimeo, Case Western Reserve University.
Conti, G., Galeotti, A., Mueller, G. and Pudney, S. (2013) Popularity. *Journal of Human Resources* 48: 1072–1094.
Costenbader, E. and Valente, T.W. (2003) The stability of centrality measures when networks are sampled. *Social Networks* 25: 283–307.
Cruz, C., Labonne, J. and Querubin, P. (forthcoming) Politician family networks and electoral outcomes: evidence from the Philippines. *American Economic Review*. 107(10): 3006–3037.
Dahl, G. (2002) Mobility and the return to education: testing a Roy model with multiple markets. *Econometrica* 70: 2367–2420.
De Giorgi, G., Pellizzari, M. and Redaelli, S. (2010) Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics* 2: 241–275.
de Paula, A. (2017) Econometrics of network models. In B. Honore, A. Pakes, M. Piazzesi and L. Samuelson (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Eleventh World Congress*. (Econometric Society Monographs, pp. 268–323). Cambridge: Cambridge University Press.
DeGroot, M. (1974) Reaching a consensus. *Journal of the American Statistical Association* 69: 118–121.
Delavallade, C., Griffith, A. and Thornton, R. (2016) Network partitioning and social exclusion under different selection regimes. Mimeo, University of Illinois at Urbana-Champaign.
Dupas, P., Keats, A. and Robinson, J. (forthcoming) The effect of savings accounts on interpersonal financial relationships: evidence from a field experiment in rural kenya. *Economic Journal*.
Epple, D. and Romano, R.E. (2011) Peer effects in education: A survey of the theory and evidence. In Alberto Bisin, Jess Benhabib and Matthew O. Jackson (eds.), *Handbook of Social Economics*, Vol. 1 (pp. 1053–1163). Amsterdam, The Netherlands: North-Holland.
Fafchamps, M. and Gubert, F. (2007) The formation of risk sharing networks. *Journal of Development Economics* 83: 326–350.
Fafchamps, M. and Quinn, S. (2016) Networks and manufacturing firms in Africa: results from a randomized field experiment. *World Bank Economic Review*. https://doi.org/10.1093/wber/lhw057
Frank, O. (1978) Sampling and estimation in large social networks. *Social Networks* 1: 91–101.
Frank, O. (1980a) Estimation of the number of vertices of different degrees in a graph. *Journal of Statistical Planning and Inference* 4: 45–50.
Frank, O. (1980b) Sampling and inference in a population graph. *International Statistical Review/Revue Internationale de Statistique* 48: 33–41.
Frank, O. (1981) A survey of statistical methods for graph analysis. *Sociological Methodology* 23: 110–155.
Frantz, T.L., Cataldo, M. and Carley, K.M. (2009) Robustness of centrality measures under uncertainty: examining the role of network topology. *Computational and Mathematical Organization Theory* 15: 303–328.
Galaskiewicz, J. (1991) Estimating point centrality using different network sampling techniques. *Social Networks* 13: 347–386.
Godlonton, S. and Thornton, R. (2012) Peer effects in learning HIV results. *Journal of Development Economics* 97: 118–129.
Goldsmith-Pinkham, P. and Imbens, G.W. (2013) Social networks and the identification of peer effects. *Journal of Business and Economic Statistics* 31: 253–264.
Graham, B.S. (2015) Methods of identification in social networks. *Annual Review of Economics* 7: 465–485.
Graham, B.S. (2017) An econometric model of link formation with degree heterogeneity. *Econometrica* 85: 1033–1063.

Guryan, J., Kroft, K. and Notowidigdo, M.J. (2009) Peer effects in the workplace: evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics* 1: 34–68.
Heckman, J. (1979) Sample selection bias as a specification error. *Econometrica* 47: 153–61.
Heckman, J.J. and Robb, R. (1985) Alternative methods for evaluating the impacts of interventions: an overview. *Journal of Econometrics* 30: 239–267.
Horrace, W., Liu, X. and Patacchini, E. (2016) Endogenous network production functions with selectivity. *Journal of Econometrics* 190: 222–232.
Hoxby, C. (2000) Peer effects in the classroom: Learning from gender and race variation. NBER Working Paper, WP 7867.
Hoxby, C. and Weingarth, G. (2005) Taking race out of the equation: School reassignment and the structure of peer effects. Mimeo, Stanford University.
Hsieh, C.-S. and Lee, L-F. (2016) A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics* 31: 301–319.
Jackson, M.O., Rogers, B.W. and Zenou, Y. (2017) The economic consequences of social-network structure. *Journal of Economic Literature* 55: 49–95.
Jackson, M.O., Rodriguez-Barraquer, T. and Tan, X. (2012) Social capital and social quilts: network patterns of favor exchange. *American Economic Review* 102: 1857–97.
Kim, P. and Jeong, H. (2007) Reliability of rank order in sampled networks. *The European Physical Journal B* 55: 109–114.
Kolaczyk, E. (2009) *Statistical Analysis of Network Data: Methods and Models*. New York: Springer-Verlag.
König, M., Liu, X. and Zenou, Y. (2014) R&D networks: theory, empirics, and policy implications. Technical report, CEPR Discussion Paper 9872.
Kossinets, G. (2006) Effects of missing data in social networks. *Social Networks* 28: 247–268.
Lee, L-F. (1983) Generalized econometric models with selectivity. *Econometrica* 51: 507–12.
Lee, L-F. and Liu, X. (2010) Identification and GMM estimation of social interactions models with centrality. *Journal of Econometrics* 159: 99–115.
Lee, S.H., Kim, P. and Jeong, H. (2006) Statistical properties of sampled networks. *Physical Review E* 73(1): 016102.
Liu, X. (2013) Estimation of a local-aggregate network model with sampled networks. *Economics Letters* 118: 243–246.
Liu, X., Patacchini, E. and Zenou, Y. (2014a) Endogenous peer effects: local aggregate or local average? *Journal of Economic Behavior and Organization* 103: 39–59.
Liu, X., Patacchini, E., Zenou, Y. and Lee, L-F. (2014b) Criminal networks: who is the key player? Mimeo.
Liu, X., Patacchini, E. and Rainone, E. (2017) Peer effects in bedtime decisions among adolescents: a social network model with sampled data. *Econometrics Journal* 20(3): S103–S125.
Manski, C. (1993) Identification of endogenous social effects: the reflection problem. *Review of Economic Studies* 60: 531–542.
Mastrobuoni, G. (2015) The value of connections: Evidence from the Italian-American Mafia. *Economic Journal* 125: F256–F288.
Mastrobuoni, G. and Patacchini, E. (2012) Organized crime networks: an application of network analysis techniques to the American Mafia. *Review of Network Economics* 11(3): 1–43.
Méango, R. (2014) International student migration: A partial identification analysis. Mimeo, Munich Center for the Economics of Aging.
Mihaly, K. (2009) Do more friends mean better grades? Student popularity and academic achievement. RAND Working Papers, WR-678.
Moffitt, R. (2001) Policy interventions, low-level equilibria, and social interactions. In S. Durlauf and , H.P. Young (eds.), *Social Dynamics* (pp. 45–82). Cambridge: MIT Press.
Mohnen, M. (2016) Stars and brokers: peer effects among medical scientists. Mimeo, University College London.
Munshi, K. and Myaux, J. (2006) Social norms and the fertility transition. *Journal of Development Economics* 80: 1–38.
Oster, E. and Thornton, R. (2012) Determinants of technology adoption: peer effects in menstrual cup take-up. *Journal of the European Economic Association* 10: 1263–1293.

Patacchini, E. and Zenou, Y. (2016) Social networks and parental behavior in the intergenerational transmission of religion. *Quantitative Economics* 7: 969–995.
Patacchini, E., Rainone, E. and Zenou, Y. (2017) Heterogeneous peer effects in education. *Journal of Economic Behavior & Organization* 134: 190–227.
Patnam, M. (2013) Corporate networks and peer effects in firm policies. Mimeo, ENSAE-CREST.
Sacerdote, B. (2001) Peer effects with random assignment: results for dartmouth roommates. *Quarterly Journal of Economics* 116: 681–704.
Sacerdote, B. (2011) Peer effects in education: how might they work, how big are they and how much do we know thus far? In E. Hanushek, S. Machin, and L. Woessman (eds.), *Handbook of the Economics of Education*, Vol. 3. Amsterdam, The Netherlands: Elsevier.
Thompson, S.K. (2006) Adaptive web sampling. *Biometrics* 62: 1224–1234.
Tincani, M. (2017) Heterogeneous peer effects and rank concerns: theory and evidence. HCEO Working Paper 2017-006.
Topa, G. and Zenou, Y. (2015) Neighborhood and network effects. In G. Duranton, V. Henderson, and W. Strange (eds.), *Handbook of Regional and Urban Economics*, Vol. 5A. Amsterdam, The Netherlands: Elsevier.
Waldinger, F. (2010) Quality matters: The expulsion of professors and the consequences for PhD students outcomes in nazi germany. *Journal of Political Economy* 118: 787–831.
Waldinger, F. (2012) Peer effects in science: evidence from the dismissal of scientists in Nazi Germany. *Review of Economic Studies* 79: 838–861.
Wang, W. and Lee, L-F. (2013) Estimation of spatial autoregressive models with randomly missing data in the dependent variable. *Econometrics Journal* 16: 73–102.
Wasserman, S. and Pattison, P. (2013) Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*. *Psychometrika* 61: 401–425.

Appendix A: Additional Figures and Tables

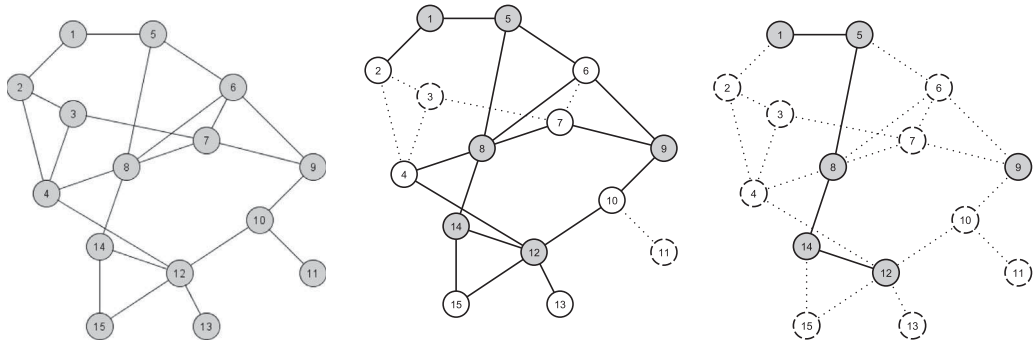

Figure A1. Star and Induced Subgraph.

Notes: Panel (a) of the figure shows the network from which nodes are randomly sampled. The shaded circles and dark lines in panels (b) and (c) display the network that emerges under star subgraph sampling and induced subgraph sampling, respectively, when 40% of nodes are sampled. The white circles and dotted lines represent missing nodes and edges under the two different sampling schemes. The average degree is the average number of edges that a researcher would get from dividing the number of sampled edges by the number of sampled nodes.

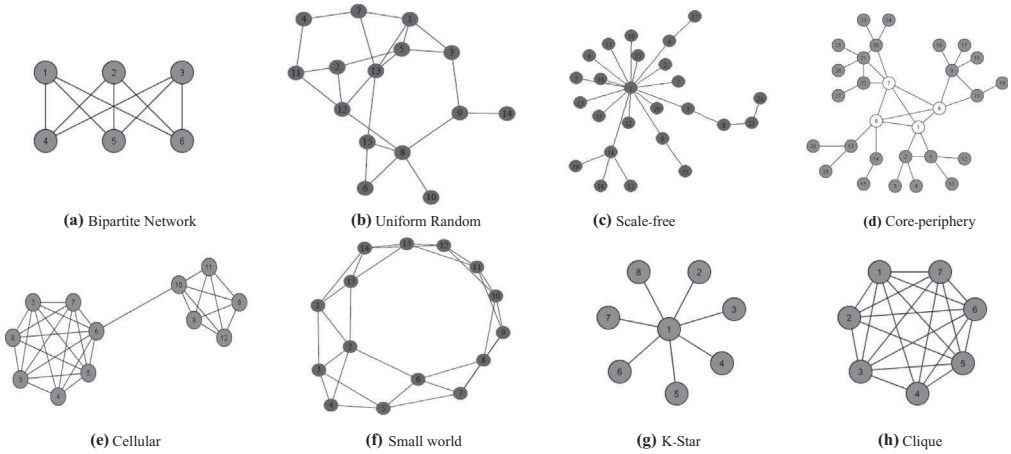

Figure A2. Network Topologies.

Table A1. Findings from Literature on Sampling-Induced Bias in Measures of Network Statistics

Statistic	Measurement error in statistic	
	Star subgraph	Induced subgraph
Network-level Statistics		
Average degree	Underestimated (–) if non-sampled nodes are included in the calculation. Otherwise sampled data provide an accurate measure.[a]	Underestimated (–).[a]
Average path length	Not known.	Over-estimated (+); network appears less connected; magnitude of bias very large at low sampling rates, and falls with sampling rate.[b]
Clustering coefficient	Attenuation (–) since triangle edges appear to be missing.[a]	Little or no bias; random sampling yields same share of connected edges between possible triangles.[a,b]
Average graph span	Overestimation (+) of the graph span: sampled network is less connected than the true network. At low sampling rates, graph span may appear to be small, depending on how nodes not in the giant component are treated.[a]	Overestimation (+) of the graph span: sampled network is less connected than the true network. At low sampling rates, graph span may appear to be small, depending on how nodes not in the giant component are treated.[a]

(*Continued*)

Table A1. *Continued*

Statistic	Measurement error in statistic	
	Star subgraph	Induced subgraph
Node-level Statistics		
Degree (in- and out- in directed graphs)	In-degree and out-degree both underestimated (–) if all nodes in sample included in calculation. If only sampled nodes included, out-degree is accurately estimated. In undirected graphs, underestimation (–) of degree for non-sampled nodes.[c]	Degree (in undirected graphs) of highly connected nodes is underestimated (–).[d]
Degree centrality (degree distribution)	Not known.	Overestimation (+) of exponent in scale-free networks ⇒ degree of highly connected nodes is underestimated. Rank order of nodes across distribution considerably mismatched as sampling rate decreases.[d]
Betweenness centrality	Distance between true betweenness centrality distribution and that from sampled graph decreases with the sampling rate. At low sampling rates (*e.g.* 20%), correlations can be as low as 20%.[c]	Shape of the distribution relatively well estimated. Ranking in distribution much worse, *i.e.* nodes with high betweenness centrality can appear to have low centrality.[e]
Eigenvector centrality	Very low correlation between vector of true node eigenvector centralities and that from sampled graph.[c]	Not known.

Notes: Little bias refers to |bias| of $< 20\%$; large bias to |bias| of 20%; and very large bias to |bias| $> 50\%$. With the exception of average degree in the star subgraph, the evidence on the direction and magnitude of biases comes from simulation studies with specific designs, which need not hold for all types of network structure.
Source: [a]Chandrasekhar and Lewis (2016); [b]Lee *et al.* (2006). [c]Costenbader and Valente (2003); [d]Lee *et al.* (2006); [e]Kim and Jeong (2007).

Table A2. Findings from Literature on Sampling-Induced Bias in Parameter Estimates

Statistic	Bias in parameter estimates	
	Star subgraph	Induced subgraph
Network level statistics		
Average degree	Scaling (+) and attenuation (–), both of which fall with sampling rate when all nodes in sample included in calculation; |scaling| > |attenuation|. No bias if only sampled nodes included.	Scaling (+) and attenuation (–), both of which fall with sampling rate; |scaling| > |attenuation|. Magnitude of bias higher than for star subgraphs.
Average path length	Attenuated (–). Magnitude of bias large and falls with sampling rate.	Attenuated (–) more than star subgraphs. Magnitude of bias is very large at low sampling rates, and falls with sampling rate.

(*Continued*)

Table A2. *Continued*

Statistic	Bias in parameter estimates	
	Star subgraph	Induced subgraph
Clustering coefficient	Scaling (+) and attenuation (–); |scaling| > |attenuation|. Very large biases, which fall with sampling rate.	Attenuation (–), falls with sampling rate. Little bias even at node sampling rates of <40%.
Average graph span	Estimates have same sign as true parameter if node sampling rate is sufficiently large. Can have wrong sign if sampling rate is too low, depending on how nodes not connected to the giant component are treated in the calculation.	Estimates have same sign as true parameter if node sampling rate is sufficiently large. Can have wrong sign if sampling rate is too low, depending on how nodes not connected to the giant component are treated in the calculation.
Node-level statistics		
Degree (in- and out- in directed graphs)	Attenuation (–), with the magnitude of bias falling with the sampling rate. The magnitude of bias is large even when 50% of nodes are sampled.	Scaling (+), with the bias falling with the node sampling rate. Bias is very large in magnitude.
Degree centrality (degree distribution)	Not known.	Not known.
Betweenness centrality	Not known.	Not known.
Eigenvector centrality	Attenuation (–), with magnitude of bias falling with the sampling rate. Magnitude of bias large even when 50% of nodes are sampled.	Attenuation (–), with magnitude of bias falling with the sampling rate. Magnitude of bias very large.

Notes: Little bias refers to |bias| of $<20\%$; large bias to |bias| of 20%; and very large bias to |bias| $>$ 50%. The evidence on the direction and magnitude of biases comes mostly from simulation studies with specific (univariate) designs, which need not hold for all types of network structure.
Source: Chandrasekhar and Lewis (2016).

Appendix B: Definitions

- **Adjacency Matrix**, $\boldsymbol{G}$: An $N \times N$ matrix, $\boldsymbol{G}$, whose ijth element, G_{ij}, represents the relationship between i and j. In a binary network, $G_{ij} = 1$ if i and j are linked, and 0 otherwise.
- **Influence Matrix**, $\tilde{\boldsymbol{G}}$: A row-stochastic adjacency matrix, $\tilde{\boldsymbol{G}}$ with $\tilde{G}_{ij} = {}^{G_{ij}}/_{\sum_j G_{ij}}$ if two agents are linked and 0 otherwise.
- **Degree**, d_i: Number of edges of a node in an undirected graph, $d_i = \sum_j G_{ij}$ in a binary graph (more generally, $d_i = \sum_j 1(G_{ij} > 0)$). In a directed graph, a node's in-degree is the number of edges from other nodes to that node, and its out-degree is the number of edges from that node to other nodes.
- **Average degree**, $\bar{d}$: Average number of links per node in the network, $\bar{d} = N^{-1} \sum_i d_i$.
- **Density:** Fraction of possible edges that are present in a network, $\frac{\bar{d}}{N-1}$.
- **Path:** A path in a network g between nodes i and j is a sequence of edges, $i_1 i_2, i_2 i_3, \ldots, i_{R-1} i_R$, such that $i_r i_{r+1} \in g$, for each $r \in \{1, \ldots, R\}$ with $i_1 = i$ and $i_R = j$ and such that each node in the sequence $i_1, \ldots, i_R$ is distinct.

- **Shortest path length (geodesic):** The shortest path length between i and j is minimum number of edges that must be traversed on a path from i and j.
- **Average path length:** The average geodesic for every pair of nodes in the network. For pairs of nodes for which no path exists, it is common to either exclude them from the calculation or to define the geodesic for these nodes to be some large number ($\geq$largest observed geodesic).
- **Induced subgraph:** A subset of nodes from the network, and all the edges in the network for which both nodes involved in that edge are in the subset. See the right panel of Figure A1 for an example.
- **Star subgraph:** A subset of nodes from the network, and all the edges in the network for which at least one of the nodes involved in that edge is in the subset. The middle panel of Figure A1 illustrates an example of a star subgraph.
- **Component:** In an undirected network, this is a subgraph of a network such that every pair of nodes in the subgraph is connected via some path, and there exists no edge from the subgraph to the rest of the network.
- **Bridge:** The edge ij is a bridge in network g if removing it results in an increase in the number of components in g.
- **Degree centrality:** A measure of centrality based on the number of direct neighbours a node has. For node i this is given by $\frac{d_i}{N-1}$.
- **Betweenness centrality**: A measure of centrality based on how well situated a node is in terms of the paths it lies on. The importance of node i in connecting nodes j and k is the ratio of the no. of geodesics between j and k that i lies on to the total no. of geodesics between j and k. Averaging this ratio across all pairs of nodes (excluding i) yields the betweenness centrality of node i.
- **Eigenvector centrality**: A relative measure of centrality, the centrality of node i is proportional to the sum of the centrality of its neighbours. It is given by $[C^e(\boldsymbol{G})]_i$, the ith element of vector $C^e(\boldsymbol{G})$, where $C^e(\boldsymbol{G})$ is the eigenvector associated with the largest eigenvalue of $\boldsymbol{G}$, $\lambda_{max}(\boldsymbol{G})$. This is calculated as a solution to $\lambda_{max}(\boldsymbol{G})C^e(\boldsymbol{G}) = \boldsymbol{G}C^e(\boldsymbol{G})$.
- **Clustering coefficient:** For an undirected network, this is the proportion of fully connected triples of nodes out of all potential triples for which at least two edges are present.
- **Graph span:** A measure that is closely related to the average path length. It is defined as $span = \frac{log(N)-log(\bar{d})}{log(\tilde{d})-log(\bar{d})} + 1$ where N is the size (number of nodes) in the network, $\bar{d}$ is the average degree, and $\tilde{d}$ is the average number of second-degree neighbours.
- **Cliques:** Subgraph of a network where every node is directly connected to every other node in the subgraph.
- **Uniform random network:** Network where the *ex ante* probability of an edge between any pair of nodes is constant across all edges in the network.
- **Bipartite network:** A network whose set of nodes can be divided into two sets, U and V, such that every edge connects a node in U to one in V.
- **Scale-free network:** Network whose degree distribution follows a power law, *i.e.* where the fraction of nodes having k edges, $P(k)$, is asymptotically proportional to $k^{-\gamma}$. Such a distribution allows for fat tails.
- C**ore-periphery network:** Network that can be partitioned into a set of nodes that is completely connected ('core'), and another set ('periphery') who are linked only with nodes in the 'core'.
- **Cellular network:** Networks containing many cliques, with few edges connecting the different cliques.
- **Small world network:** Network where most nodes are not directly linked to one another, but where geodesics between nodes are small.
- ***k*-star:** Component with k nodes and $k-1$ links such that there is one 'hub' node who has a direct link to each of the $(k-1)$ other ('periphery') nodes.

Appendix C: Collecting Network Data: Sampling Methods

In order to construct the full network, researchers would need to collect data on all nodes and edges, that is, collect a census. This is typically very expensive, as well as logistically challenging. Instead researchers usually collect data on a sample of the network. A number of sampling methods have been used to do this, of which the most common are as follows:

Random Sampling

Random samples can be drawn for either nodes or edges. Data collected from a random sample of nodes typically contain information on socio-economic variables of interest and some (or all) edges of the sampled nodes, although data on edges are usually censored. The network graph constructed from data where nodes are randomly sampled and where edges are included only if both nodes are randomly sampled is known as an induced subgraph.

Information may also be available on some socio-economic variables of all nodes in the network. Recent analyses with network data in the economics literature have featured datasets with edges collected from random samples of nodes, where covariate information was available for all nodes. Examples include data on social networks and the diffusion of microfinance used by both Banerjee *et al.* (2013) and Jackson *et al.* (2012).

Datasets constructed through the random sampling of edges include a node only if any one of its edges is randomly selected. Examples of such datasets include those constructed from random samples of email communications, telephone calls or messages. In these cases researchers often have access to the full universe of all e-mail communication, but are obliged to work with a random sample due to computational constraints.

Snowball Sampling and Link Tracing

Snowball sampling is popularly used in collecting data on 'hard to reach' populations, that is, those for whom there is a relatively small proportion in the population. For these groups one would get a very small sample through random sampling from the whole population. Link tracing is a related method that is usually used to collect data from vast online social networks, where the average degree is relatively large.

Under both these methods, a dataset is constructed through the following process. Starting with an initial, possibly non-random, sample of nodes from the population of interest, information is obtained on either all, or a random sample of their edges. Snowball sampling collects information on all edges of the initially sampled nodes, while link tracing collects information on only a random sample of these edges. In the subsequent step, data on edges and outcomes are collected from any node that is reported to be linked to the initial sample of nodes. This process is then repeated for the new nodes, and in turn for nodes linked to these nodes (i.e. second-degree neighbours of the initially drawn nodes) and so on, until some specified node sample size is reached or up to a certain social distance from the initial 'source' nodes.

It is hoped that, after k steps of this process, the generated dataset is representative of the population, that is, the distribution of sampled nodes no longer depends on the initial 'convenience' sample. However, this typically happens only when k is large. Moreover, the rate at which the dependence on the original sample declines is closely related to the extent of homophily, both on observed and unobserved characteristics, in the network. In particular, stronger homophily is associated with lower rates of decline of this dependence. Nonetheless, this method can collect, at reasonable costs, complete information on local neighbourhoods. Examples in economics of datasets collected by snowball sampling include that of student migrants used in Méango (2014).