Credibly Identifying Social Effects: Accounting for Network Formation and Measurement Error

Arun Advani and Bansi Malde*

6th November 2016

Abstract

Understanding whether and how connections between agents (networks) such as declared friendships in classrooms, transactions between firms, and family connections in rural villages, 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 this 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 using observational data. 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.

Key Words: Networks, Social Effects, Econometrics, Endogeneity, Measurement Error, Sampling Design

JEL Classification: C31, C81, Z13

*Affiliations: Advani - University College London and Institute for Fiscal Studies (E-mail: arun.advani.10@ucl.ac.uk); Malde - University of Kent and Institute for Fiscal Studies (E-mail: b.k.malde@kent.ac.uk). 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, and Yves Zenou for their useful comments and suggestions. Financial support from the ESRC-NCRM Node 'Programme Evaluation for Policy Analysis', Grant reference RES-576-25-0042 is gratefully acknowledged.

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 relied on data with very limited information on networks, typically information on membership of mutually exclusive groups such as classrooms, neighbourhoods or villages. Estimating social effects using this type of information suffers from two key limitations: First, identifying the social effect is complicated by the reflection problem – a form of simultaneity where it isn't possible to identify who is influencing whom (Manski, 1993). Second, since more detail on interactions within a group is not available, studies impose (implicitly) the assumption that all agents within the same group interact with one another in the same way. However, the composition of the group on both observed and unobserved dimensions could influence within-group interactions, and through this the actual outcome. Ignoring 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 with information on exact interactions between agents (*networks data* hereon) to sidestep these issues. This growth has been spurred by the increasing availability of such detailed 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. These typically use information on network structure to break this 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) among others, 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 these assumptions are unlikely to hold in practice: for example, in a schooling context, personality traits which are rarely observed by a researcher might influence both, a child's choice of friends as well as 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: 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 for identification. Missing or mismeasured data on link status will affect the ability of these methods to yield unbiased and consistent social effect estimates.

The issue of endogenous link formation has long been recognised in the empirical literature, while that of measurement error has received increasing attention recently. A number of different methods

have been proposed to deal with both of these issues, in economics as well as other disciplines including sociology, mathematics and computer science.

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 with observational network data. A large swathe of empirical work on social effects uses linear models, motivating our focus on this class of model.[1] We draw on methods developed in a broad range of disciplines, 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.

We then provide a brief overview of strategies undertaken 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, five common approaches, using specific examples to illustrate ideas. The first approach exploits exogenous variation arising from randomly assigned links. Though this provides clean identification, random link 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, which may be challenging in many contexts. A fourth alternative relies on the observation that endogenous link formation induces a selection bias in social effect estimates. A natural solution is the control function, where one estimates the selection bias term, and 'controls' for it when estimating the social effect model. Finally, we consider the nascent literature modelling link and outcome choices simultaneously.

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 functionals of the network, and the parameter estimates of social effect models. This is because networks are comprised of two interrelated objects: agents (nodes) and links. A sampling strategy over one of these objects defines the (conditional) sampling process over the other, thereby biasing network measures and model parameters. We first discuss the implications of missing data for the estimation of network measures and regression parameters. Thereafter, we review the various 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 (2014), Topa and Zenou (2015) and Boucher and Fortin (2015) more than amply covers 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 (2016) and Chandrasekhar (2015). Finally, we do not survey findings on the size, magnitude and heterogeneity of social effects or peer effects found in applied economics: other reviews more than amply cover these (e.g. Epple and Romano (2011) and Sacerdote (2011) provide surveys of peer effects in

education, while Chuang and Schechter (2014) provide a survey of applied work on networks in developing countries).

The rest of the paper is organised 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; while 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 and 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, which nests a number of the key empirical specifications used in the literature, and elucidates the parameters of interest. We do this separately for individual- and network-level outcomes. Using this specification, we outline some of the common assumptions imposed to allow for identification of the parameters of interest, before illustrating the implications of endogenous network formation, and measurement error in the network.

Throughout, we will 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, co-working, economic transactions). We index networks by g, and with some abuse of notation, nodes within a network g by $i \in g$. 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 normalised to 0. We also define the *influence matrix,* $\tilde{\boldsymbol{G}}_g$, the row-stochastised[2] adjacency matrix. Elements of this matrix are generally defined as $\tilde{G}_{ij,g} = d_{i,g}^{-1} G_{ij,g}$ if two agents are linked and 0 otherwise.

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:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y})\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{w_x}(\boldsymbol{G}, \boldsymbol{X})\boldsymbol{\delta} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{L}\boldsymbol{\nu} + \boldsymbol{\varepsilon} \tag{2.1}$$

$\boldsymbol{Y}$ is an $\sum_{g=1}^{M} N_g \times 1$ vector stacking individual outcomes of nodes across all networks (indexed by $g = 1, ..., M$). $\boldsymbol{X} = (\boldsymbol{X}_1^{'}, ..., \boldsymbol{X}_M^{'})^{'}$ 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 with each column being 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. 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 and corresponds to $\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y}) = \boldsymbol{G}\boldsymbol{Y}$ in Equation 2.1. $\boldsymbol{w_x}(\boldsymbol{G}, \boldsymbol{X})$ 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. This allows the social effect to operate through both a desire for conformism and through strategic complementarities/substitutabilities. In the notation of Equation 2.1, it implies that $\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y}) = [\mathbf{G}\boldsymbol{Y}, \tilde{\boldsymbol{G}}\boldsymbol{Y}]$, 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 more general functionals or statistics of the network. 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, we can specialise the term $\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y})\boldsymbol{\beta}$ in Equation 2.1 for node i in network g as:

- $\sum_{r=1}^{R} \omega_{i,g}^r \beta_r$: R different network statistics; 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; 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.

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

The social effect parameter is $\boldsymbol{\beta}$ in Equation 2.1: 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 in many linear models, the presence of endogenous effects implies the presence of 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 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 2.1. These restrictions reflect assumptions on common unobserved shocks and on the network formation process. For example, $\mathbb{E}[\nu_g |\, \boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}_g] = 0 \quad \forall\, g \in \{1, ..., M\}$ implies nodes sort into networks exogenously, conditional on individual-level and network-level observables., while $\mathbb{E}[\varepsilon_{i,g} |\, \boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}_g] = 0 \quad \forall\, i \in g;\ g \in \{1, ..., M\}$ 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 affecting sorting into networks can be accounted for using network-level fixed effects, similar to 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 (2014), 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-, rather than node-level outcomes, in which case the following specification is typically estimated:

$$\bar{\boldsymbol{y}} = \phi_0 + \bar{\boldsymbol{w}}_{\bar{\boldsymbol{y}}}(\boldsymbol{G})\boldsymbol{\phi}_1 + \bar{\boldsymbol{X}}\boldsymbol{\phi}_2 + \bar{\boldsymbol{w}}_{\bar{\boldsymbol{X}}}(\boldsymbol{G}, \bar{\boldsymbol{X}})\boldsymbol{\phi}_3 + \boldsymbol{u} \tag{2.2}$$

where $\bar{\boldsymbol{y}}$ is an $(M \times 1)$ vector stacking the aggregate outcome of the M networks, $\bar{\boldsymbol{w}}_{\bar{\boldsymbol{y}}}(\boldsymbol{G})$ 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 (which could include network-averages of node characteristics) and $\bar{\boldsymbol{w}}_{\bar{\boldsymbol{X}}}(\boldsymbol{G}, \bar{\boldsymbol{X}})$ is a term interacting the network-level characteristics with the network statistics. $\boldsymbol{\phi}_1$ 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 on the outcome.

The key parameter of interest here is typically $\boldsymbol{\phi}_1$: the effect of some network statistic, such as network density or the average degree on the aggregate network outcome. The key identification assumption is that $\mathrm{E}[\boldsymbol{u}_g|\boldsymbol{G}_g, \bar{\boldsymbol{X}}_g] = 0$, which will not hold if there are unobserved variables in $\boldsymbol{u}$ that affect both the formation of the network and the outcome $\bar{\boldsymbol{y}}$; 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 agents do not take into account the influences of their connections on the outcome of interest when choosing their links; and second, that there are no unobserved (to the econometrician) agent-specific factors influencing both an agent's choice of connections and the outcome of interest. 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 understand what features influence take-up of a new, potentially risky technology. We might see that more connected farmers are 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 have great interest in learning about innovative practices, and may have thus chosen to have more connections for this reason! Both of these violate the condition that $\mathbb{E}[\varepsilon_{i,g}|\, \boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}_g] = 0 \;\; \forall\, i \in g;\; g \in \{1, ..., M\}$ in Equation 2.1. Section 3 discusses this endogeneity problem in more detail, and describes potential solutions.

Measurement error in $\boldsymbol{G}$ can also invalidate the assumption that $\mathbb{E}[\varepsilon_{i,g}|\, \boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}_g] = 0 \;\; \forall\, i \in g;\; g \in \{1, ..., M\}$, 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 2.1 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 2.1. Clearly, then $\mathbb{E}[\varepsilon_{i,g}|\, \boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}^*_g] \neq 0 \;\; \forall\, i \in g;\; g \in \{1, ..., M\}$, 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 measure.

A simple example illustrates this: surveys often place an upper limit, ψ, on the number of links a node can report. In the absence of other error, the number of misclassified links for node i can be expressed as $\sum_j \boldsymbol{\xi}(\boldsymbol{G})_{ij} = \min\left\{0, \psi - \sum_j \boldsymbol{G}_{ij}\right\}$. It is clear here that 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 methods that could be used to overcome this issue along with the conditions under which they apply.

3 Dealing with Endogeneity of Network Formation

In this section we discuss approaches taken to allow identification of social effects whilst relaxing the assumption that the network is exogenous. We now allow for the possibility that network links are chosen, and that these choices might be related to the unobservables determining individuals' outcomes.[3] We discuss five approaches to dealing with endogeneity that have been taken in the literature, providing examples of where they have been used, and discussing their limitations.

3.1 Random Assignment

The first method is random assignment of links. While this strategy has been widely applied in laboratory experiments of network effects, recent work has applied this to real-life contexts including classrooms (e.g. Carrell et al., 2009 among others), dorm rooms (e.g. Sacerdote, 2001), sport partners (e.g. Guryan et al., 2009), and among firm managers (e.g. Fafchamps and Quinn, 2016). Random assignment to a group such as a classroom, dorm room, or judging committee is likely to increase interactions between those assigned to the same group, and through this affect the social effects influencing the outcome of interest.

Though random assignment to a network alleviates biases associated with non-random network formation, researchers still need to account for unobserved network shocks, and where information on interactions within the network is not available, for the reflection problem, in order to obtain consistent estimates of the social effect. To account for these confounders, existing studies use pre-randomisation, rather than contemporaneous, values of outcomes and characteristics. As a result, the identified social effect parameter will not provide an estimate of the spillover of links' choices and outcomes on one's own outcome. It will also capture the influences of links' characteristics. More formally, specifications of the following type are estimated:

$$\boldsymbol{Y}_{post} = \alpha\tilde{\boldsymbol{\iota}} + \boldsymbol{w_y}(\boldsymbol{G},\, \boldsymbol{Y}_{pre})\tilde{\boldsymbol{\beta}} + \boldsymbol{X}_{pre}\tilde{\boldsymbol{\gamma}} + \boldsymbol{w_x}(\boldsymbol{G},\, \boldsymbol{X}_{pre})\tilde{\boldsymbol{\delta}} + \tilde{\boldsymbol{\varepsilon}} \tag{3.1}$$

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-randomisation outcome $\boldsymbol{Y}_{pre}$, will be uncorrelated with current unobserved shocks. However, it will not map cleanly to current choices which generate the endogenous social effect of interest, and could instead be considered to be an exogenous characteristic (and hence be part of $\boldsymbol{X}$). Thus, $\tilde{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$ in Equation 2.1.

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. Moreover, the formation of more complex network structures such as transitive or intransitive triads is not currently well understood, making it difficult to use encourage these through this method.

Second, the identified parameter will capture a local, rather than average, effect. 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 sub-set of group members, and make this choice non-randomly (e.g. choosing those that provide them with the highest net value), these estimates will not be so informative about the likely social effect when the group is constructed in some other 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 maximising 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 not accounting for the choice of links formed by individuals within squadrons.[4]

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.[5] Examples include unanticipated deaths of individuals (e.g. 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 (e.g. Waldinger (2010, 2012)), the 2011 Great East Japan earthquake (Carvalho et al. (2016)), among others. This method recovers a social effect parameter by comparing outcomes of agents affected by a local shock to their (immediate) network with those of agents with similar pre-shock characteristics (including network position and characteristics) who do not face a local shock to their network. The key underlying assumption is that agents with similar pre-shock observed characteristics and network position would have faced a similar trend in their outcomes in the absence of the shock.

However, in order to be a valid identification strategy, this method also requires that agents choose not to directly respond to the natural or quasi-experiment.[6] Importantly, non-response in this case includes both, not adjusting edges in response to the changes that occur, as well as not *ex-ante* choosing edges strategically to insure against the probabilistic exogenous edge 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 not considering the board member's life expectancy when hiring. Finally, if there is heterogeneity in the social effect, then this approach provides only a local effect, based on an average over the links that change. This may not be representative of the average social effect if, for example, older board members have larger effects and are more likely to die.

3.3 Instrumental Variables

An alternative approach is to use instrumental variables – variable(s) that are correlared with the endogenous network covariate ($\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y})$ in Equation 2.1) but excluded from the equation itself.

Some examples of applications of this approach include Munshi and Myaux (2006), Mihaly (2009), Cohen-Cole et al. (forthcoming), König et al. (2014) and Acemoglu 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, Munshi and Myaux (2006) exploit strong social norms that prevent the formation of cross-religion edges even where these might otherwise be very profitable, when studying fertility in rural Bangladesh. The restrictions on cross-religion connections mean that having different religions is a strong predictor that two women are not linked.

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 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) makes use of instruments based on the network adjacency matrix predicted from a dyadic network formation model, where link formation is a function of variables that influence link formation decisions, but don't otherwise affect the outcome. They study spillovers from R&D collaborations between firms connected by a web of collaboration agreements (and who also might compete with one another), and model link formation as a function of having a common R&D collaborator in the past, or having collaborated on R&D in the past, as well as time-lagged measures of technological proximities of firms.

It is important to note that this type of solution can only be employed when the underlying network formation model has a unique equilibrium, so that there is only one network structure consistent with the (observed and unobserved) characteristics of the agents and environment. However, when multiple equilibria are possible, which will generally be the case if 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 formation equation in these circumstances is 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 effect. In particular, the estimated $\hat{\beta}_{IV}$ will be a weighted average of the individual-specific βs, with more weight given to 'compliers' – those for whom the network covariate of interest is induced to change by the instrument. Hence, the estimated social effect might be larger than the average social effect if the compliers 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

A fourth method that has been proposed (e.g. by Blume et al., 2013) and implemented in recent work is that of the control function. Endogenous linking decisions yield selectivity bias in social effect estimates. Control function methods propose to correct this by including an estimated

selectivity bias term as an additional regressor in the main equation of interest (e.g. Heckman, 1979; Lee, 1983; Heckman and Robb, 1985). Recent work by Goldsmith-Pinkham and Imbens (2013), Horrace et al. (2016), Arduini et al. (2015), and Hsieh and Lee (forthcoming) extends control function methods to a networks context. Specifically, the selectivity bias term is estimated from a first stage network formation model, and then included as an additional regressor in the social effects estimation. The selection correction term is a non-linear function of the predicted network (and thus of variables determining link choice).[7] Identification of the social effect parameter can thus be achieved even in the absence of variables influencing the outcome only through link choices (i.e. exclusion restrictions) by relying on functional form assumptions. The presence of an exclusion restriction, however, makes identification more credible.

The key challenge in operationalising this method lies in specifying a sufficiently tractable first-stage model of link formation. This is a result of the size of the joint distribution for edges, which for a directed binary network, is a $N(N-1)$-dimensional simplex, which has $2^{N(N-1)}$ points of support (potential networks).[8] Recent advances in specifying and estimating network formation models are provided in Advani and Malde (2014), Graham (2015), Chandrasekhar (2015), and de Paula (2016).

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 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.[9] As explained above, model parameters are identified through functional form assumptions, though the presence of a variable that affects the outcome only through link choice would aid identification and make it more credible.

Recent studies including Goldsmith-Pinkham and Imbens (2013), Hsieh and Lee (forthcoming) and Arduini et al. (2015) use dyadic models of link formation.[10] 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 characterise 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. 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 (forthcoming) consider linking decisions in directed networks, allowing for decisions to be affected by multiple unobserved variables, in a framework similar to Goldsmith-Pinkham and

Imbens (2013).[11] Linking decisions are assumed to be homophilous, and are influenced by dyad-specific characteristics, $\boldsymbol{C}$, as well as individual characteristics, $\boldsymbol{X}$. Functionals of the network, particularly transitivity, are treated as an unobserved term. Assuming that the unobservable terms in the social effects equation and the network formation equation are joint normal, Hsieh and Lee (forthcoming) are able to characterise 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.[12] As with Goldsmith-Pinkham and Imbens (2013), likelihood estimation using maximum likelihood methods is computationally difficult, necessitating the use of a Bayesian MCMC approach.

Arduini et al. (2015) consider two potential ways of modelling the first stage: (i) a dyadic link formation model of Graham (2014), which assumes homophilous link formation and agent-specific unobserved heterogeneity, and (ii) 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 simultanbeously chosen. This approach is taken by Badev (2013), who models peer effects in smoking for adolescents, allowing agents to choose their smoking decisions simultaneously with their links. His model also incorporates a 'strategic' element to network formation, and uses an equilibrium concept of k-player Nash stability, where 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.[13] At least one equilibrium to the game exists, and equilibria can be probabilistically ranked thereby providing a systematic equilibrium selection rule. He shows that the game can be specified in terms of a potential function, which allows for an analytical characterisation of the likelihood function. However, estimating the model through maximum likelihood methods is computationally infeasible since the set of possible networks is extremely large. Instead, Badev (2013) proposes approximating the likelihood using MCMC methods.

4 Measurement Error

The second challenge complicating the 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 summarising 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, first, whether and how to construct a sample; and second, what data to collect and for whom. Note also that there is a large econometric and statistical literature on non-sampling induced measurement error: For example, Chen et al. (2011) provide a summary of methods for dealing with misreporting in binary variables, which could potentially apply to network contexts. However, these issues have been less studied in a networks context, and are thus not covered here.[14]

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 undertaking a full census – generate a (conditionally) non-random sample of at least one of these objects, since any random sampling process over one will induce a particular (non-random) structure over the other.[15] This means that econometric and statistical 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 ease logistics of data collection, one may erroneously limit the boundary of the network to a specified unit, *e.g.* village or classroom, thereby missing nodes and edges lying beyond this boundary. Subsection 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 also draw on research from a range of fields, including sociology, statistical physics, and computer science. Thereafter, in Subsection 4.2 we outline a number of methods developed to help deal with the consequences of measurement error.

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 the structure of the network, typically using the exogenous characteristics of indirect neighbours as instruments for the outcomes of one's direct neighbours ($\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y})$ in Equation 2.1). 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 occassions over time. To illustrate their method, we introduce some additional notation: Let $\boldsymbol{G}^A$ denote the adjacency matrix related to the outcome of interest; and $\boldsymbol{G}^B$ denote a matrix indicating which links are absent in $\boldsymbol{G}^A$ but are present in the alternative measurement, $\boldsymbol{G}^{A'}$. If the measurement error is unconditionally random, for any link $l(i,j)$ that is reported to not exist in $\boldsymbol{G}^A$, so that $G^A_{ij} = 0$, there will be a higher probability that it is missing if it is reported to be present in $\boldsymbol{G}^{A'}$(so that $G^B_{ij} = 1$). The presence of this type of measurement error can be tested by estimating the following generalisation of Equation 2.1:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota}+\boldsymbol{w_y}(\boldsymbol{G}^A, \boldsymbol{Y})\boldsymbol{\beta}+\boldsymbol{X}\boldsymbol{\gamma}+\boldsymbol{w_x}(\boldsymbol{G}^A, \boldsymbol{X})\boldsymbol{\delta}+\boldsymbol{w_y}(\boldsymbol{G}^B, \boldsymbol{Y})\boldsymbol{\beta}^B+\boldsymbol{w_x}(\boldsymbol{G}^B, \boldsymbol{X})\boldsymbol{\delta}^B+\mathbf{Z}\boldsymbol{\eta}+\boldsymbol{L}\boldsymbol{\nu}+\boldsymbol{\varepsilon} \quad (4.1)$$

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 discounted 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 on average to be (weakly) further apart in the sampled network than they are in the true underlying network. This measurement error carries over 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 not be able to purge the social effect parameter of bias (Chandrasekhar and Lewis, 2011). Simulations by Chandrasekhar and Lewis (2011) 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. However, both papers also offer simple, direct solutions to this issue when data are available on a star subgraph (where nodes are randomly sampled, and all edges are included regardless of whether the incident nodes are sampled, *i.e.* if i is randomly sampled, the edge ij will be included regardless of whether or not j is sampled): these are described in Subsection 4.2.1.

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 (2011) indicates biases that are very large in magnitude, and which go in different directions, depending on the statistic being studied. 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 a graph constructed from a sample of nodes and informaton on edges among the sampled nodes only (also known as an induced subgraph) with 20% of nodes in the true network. Table 1 provides a more detailed summary of findings from these papers for some commonly used network statistics for data collected via random sampling of nodes, which could be done in two ways: (i) as a star subgraph (defined above); or (ii) as an induced subgraph where nodes are randomly sampled, and information on edges is collected for connections among the sampled nodes only. Figure 1 in Appendix A provides a graphical example of a star and 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 A 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 B for more details on sampling strategies). They draw samples from networks with a power-law degree distribution *i.e.* where the fraction of nodes having k edges, $P(k)$ is asymptotically proportional to $k^{-\gamma}$, and usually $2 < \gamma < 3$. Such a distribution allows for fat tails, *i.e.* 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, *i.e.* the measurement error is not independent of the true network statistic. Chandrasekhar and Lewis (2011) 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 2 summarises 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, *e.g.*, 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 underestimated, 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 very 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 might be very costly or infeasible to collect. Work in disciplines outside economics (e.g. sociology) 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 sub-section.

Table 1: Findings from literature on sampling-induced bias in measures of network statistics

Statistic	Measurement error in statistic	
Network-Level Statistics	Star Subgraph	Induced Subgraph
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]

Notes: Little bias refers to |bias| of $< 20\%$; large bias to |bias| of 20%; and very large bias to |bias| $> 50\%$.

Source: [a]Chandrasekhar and Lewis (2011); [b]Lee et al. (2006).

Table 1 contd.

Statistic	Measurement error in statistic	
Node - Level Statistics	Star Subgraph	Induced Subgraph
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| $> 50\%$.

Source: [c]Costenbader and Valente (2003); [d]Lee et al. (2006); [e]Kim and Jeong (2007)

Table 2: Findings from literature on sampling-induced bias in parameter estimates

Statistic	Bias in Parameter Estimates	
Network Level Statistics	Star Subgraph	Induced Subgraph
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.
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.

Notes: Little bias refers to |bias| of $< 20\%$; large bias to |bias| of 20%; and very large bias to |bias| $> 50\%$.

Source: Chandrasekhar and Lewis (2011)

Table 2 contd.

Statistic	Bias in Parameter Estimates	
Node - Level Statistics	Star Subgraph	Induced Subgraph
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| $> 50\%$.

Source: Chandrasekhar and Lewis (2011)

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 i.e.* 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 summarise 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 network-based instruments in local network models, thereby inducing bias in social effects. Chandrasekhar and Lewis (2011) 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 is 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 ($\boldsymbol{GL}$), 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 $\boldsymbol{GY}$. This provides overidentification, 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.

When there is no missing data in the network or covariates, but outcome data is available for a sub-sample only,[16] In the case of the local average model, Liu et al. (2013) show that the reduced form equation restricted to the observations for whom complete outcome data is available involves regressing the outcome on a non-linear transformation of $\boldsymbol{X}$ and $\tilde{\boldsymbol{G}}\boldsymbol{X}$. 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 show 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).[17] 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 B 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. Estimators based on Markov chain resampling methods, however, make it feasible to estimate the sample inclusion probabilities. See Thompson (2006) for more details.

Frank (1978, 1980a, 1980b, 1981) derives unbiased estimators for a range of graph statistics. Chandrasekhar and Lewis (2011) show for three statistics – average degree, clustering coefficient, and average graph span – that 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, and eliminates at sufficiently high sampling rates, 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 for the specification of the joint distribution of the true in-degree, the true out-degree and the mismeasured in-degree. Pseudolikelihood functions can therefore 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 for *all* nodes in the network on individual characteristics (*e.g.* gender, ethnicity) that are predictive of link formation.

As yet, no specific guidelines are available on how to best specify the model used for link prediction.[18] However, it is clear that we want a model that delivers the maximum possible variation in link probabilities, so that we have some power when using the imputed link probabilities in estimating social effects. For economists, one possible approach is the *dyadic regression* model of Fafchamps and Gubert (2007), where the probability of a link between a pair of agents is modelled to depend on the characteristics of each agent, the (absolute) difference between these characteristics (capturing homophily), and idiosyncratic link-specific errors.

Where covariates are not observed, it might be possible to impute missing links using statistical models of link formation. The simplest of these is the Bernoulli random graph model (Gilbert, 1959, Erdős and Rényi, 1959), which assumes that each link forms with common probability, independent of all else. However, this model is unsuitable in this context since it yields, by definition, no variation in link probabilities. Nonetheless, more general models, which model the probability of a link existing as depending in some way on the other links around it, can be used. These models are known as p^*-models (Wasserman and Pattison, 1996) or *exponential random graph models* (ERGMs). Whilst they can easily be extended to handle individual covariates, we can now already generate variation in the predicted link probability, using only network information. Of course, these models are only useful if the probability a link exists does actually vary with the status of other links in the network. Evidence on such models has been provided by Mele (2013), who shows that these models can arise as the result of utility maximising behaviour by individual agents. Another drawback of these models is that they are more difficult to estimate. For more details on the specification and estimation of such models, see Section 4 of Advani and Malde (2014).

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 their links – 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 effects of the influence of one's neighbours on one's choices and decisions. 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 link assignment, use of local network shocks, instrumental variables, control function and modelling the choice of links and outcomes to be simultaneous. 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 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 latter two methods, provides an alternative approach. Knowledge (or assumptions) about the payoff from forming links provides a different route to identification. The challenges to this solution are not only in determining what assumptions about payoffs are reasonable, but also technical, since 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 (which is not even well-defined until we specify whether it is nodes or edges over which we define the 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 networks 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.

Notes

[1]These 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 normalised so they each sum to one.

[3]It is important to note that this implies that individuals already have some information about the unobservables. If these unobservables are identically distributed, realised 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.

[4]Booij et al. (2015) and Tincani (2015) 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 (2015) suggests that the findings can be explained by an education production function allowing for competition between students.

[5]A related recent literature shows that randomly assigned policy interventions can change the likelihood as well as strength of links between households (e.g. Feigenberg et al., 2013; Comola and Prina, 2014), but as yet such variation has not been used to overcome endogenous link formation.

[6]One also needs access to panel data for the network, which may not often be available. Moreover, measurement error in either round of network data will reduce the power of this strategy.

[7]The network formation models specified by the studies covered model links as limited dependent variables, and are thus also nonlinear models.

[8]To give a sense of scale, for a network of more than 7 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}).

[9]They also develop a fixed effects approach, which can only be applied in contexts where the social effect is heterogeneous.

[10]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.

[11]In addition to homophily, they also allow for more general functionals of the network to influence linking decisions.

[12]Note that as before, since the network formation model is nonlinear, model parameters could also be identified through functional form assumptions.

[13]Such an 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.

[14]A paper that develops and implements a correction for this class of measurement error in a networks context is Comola and Fafchamps (2014).

[15]We consider a random sample to consist of independentally and identically distributed units.

[16]This is consistent with survey designs collecting networks information and some key covariates from all nodes and detailed outcome data from a sample

[17]Chapter 5 of Kolaczyk (2009) provides useful background on these methods.

[18]Sections 3.4.1 and 3.4.2 considered some models of link formation

References

D. Acemoglu, C. Garcia-Jimeno, and J. Robinson. "State Capacity and Economic Development: A Network Approach". *American Economic Review*, forthcoming.

A. Advani and B. Malde. "Empirical Methods for Networks Data: Social Effects, Network Formation and Measurement Error". *IFS Working Paper W14/34*, 2014.

T. Arduini, E. Patacchini, and E. Rainone. "Parametric and Semiparametric IV Estimation of Network Models with Selectivity". *EIEF Working Paper*, 2015.

A. I. Badev. "Discrete Games in Endogenous Networks: Theory and Policy", 2013.

A. Banerjee, A. G. Chandrasekhar, E. Duflo, and M. Jackson. "The Diffusion of Microfinance". *Science*, 341:1236498, 2013.

L. E. Blume, W. A. Brock, S. N. Durlauf, and Y. M. Ioannides. "Identification of Social Interactions". In J. Benhabib, A. Bisin, and M. Jackson, editors, *Handbook of Social Economics*, volume 1B. North Holland, 2010.

L. E. Blume, W. A. Brock, S. N. Durlauf, and R. Jayaraman. "Linear Social Interaction Models". *NBER Working Paper*, WP 19212, 2013.

Adam S. Booij, Edwin Leuven, and Hessel Oosterbeek. Ability Peer Effects in University: Evidence from a Randomized Experiment. IZA Discussion Papers 8769, Institute for the Study of Labor (IZA), January 2015. URL `https://ideas.repec.org/p/iza/izadps/dp8769.html`.

V. Boucher and B. Fortin. "Some Challenges in the Empirics of the Effects of Networks". In Y. Bramoulle, A. Galeotti, and B. Rogers, editors, *The Oxford Handbook of the Economics of Networks*, pages 277–302. Oxford University Press, 2015.

Y. Bramoullé, H. Djebbari, and B. Fortin. "Identification of Peer Effects through Social Networks". *Journal of Econometrics*, 150:41–55, 2009.

Y. Bramoullé, R. Kranton, and M. D'Amours. "Strategic Interaction and Networks". *American Economic Review*, 104(3):898–930, 2014.

W. A. Brock and S. N. Durlauf. "Discrete Choice with Social Interactions". *Review of Economic Studies*, 68:235–260, 2001.

W. A. Brock and S. N. Durlauf. "Identification of Binary Choice Models with Social Interactions". *Journal of Econometrics*, 140:52–75, 2007.

A. Calvó-Armengol, E. Patacchini, and Y. Zenou. "Peer Effects and Social Networks in Education". *Review of Economic Studies*, 76:1239–1267, 2009.

S. Carrell, R. Fullerton, and J. West. "Does Your Cohort Matter? Estimating Peer Effects in College Achievement". *Journal of Labor Economics*, 27(3):439–464, 2009.

S. Carrell, B. Sacerdote, and J. West. "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation". *Econometrica*, 81(3):855–882, 2013.

V. Carvalho, M. Nirei, Y. Saito, and A. Tahbaz-Salehi. "Supply Chain Disruptions: Evidence from the Great East Japan Earthquake". 2016.

A. G. Chandrasekhar. "Econometrics of Network Formation". In Y. Bramoulle, A. Galeotti, and B. Rogers, editors, *The Oxford Handbook of the Economics of Networks*, pages 303–357. Oxford University Press, 2015.

A. G. Chandrasekhar and R. Lewis. "Econometrics of Sampled Networks". *mimeo, Massachusetts Institute of Technology*, 2011.

X. Chen, H. Hong, and D. Nekipelov. "Nonlinear Models of Measurement Errors". *Journal of Economic Literature*, 49(4):901–937, 2011.

Y. Chuang and L. Schechter. "Social Networks In Developing Countries", 2014.

E. Cohen-Cole, X. Liu, and Y. Zenou. "Multivariate Choice and Identification of Social Interactions". *Journal of Econometrics*, forthcoming.

M. Comola and M. Fafchamps. "Estimating Mis-reporting in Dyadic Data: Are Transfers Mutually Beneficial?". *mimeo, Paris School of Economics*, 2014.

M. Comola and S. Prina. "Do Interventions Change the Network? A Dynamic Peer Effect Model Accounting for Network Changes". *SSRN Working Paper No. 2250748*, 2014.

G. Conti, A. Galeotti, G. Mueller, and S. Pudney. "Popularity". *Journal of Human Resources*, 48(4):1072–1094, 2013.

E. Costenbader and T. W. Valente. "The stability of centrality measures when networks are sampled". *Social Networks*, 25:283–307, 2003.

G. Dahl. "Mobility and the Return to Education: Testing a Roy Model with Multiple Markets". *Econometrica*, 70(6):2367–2420, 2002.

G. De Giorgi, M. Pellizzari, and S. Redaelli. "Identification of Social Interactions through Partially Overlapping Peer Groups". *American Economic Journal: Applied Economics*, 2(2):241–275, 2010.

A. de Paula. "Econometrics of Network Models". *CeMMAP working paper CWP06/16*, 2016.

M. DeGroot. "Reaching a Consensus". *Journal of the American Statistical Association*, 69:118–121, 1974.

Dennis Epple and Richard E. Romano. Chapter 20 - peer effects in education: A survey of the theory and evidence. volume 1 of *Handbook of Social Economics*, pages 1053 – 1163. North-Holland, 2011. doi: http://dx.doi.org/10.1016/B978-0-444-53707-2.00003-7. URL `http://www.sciencedirect.com/science/article/pii/B9780444537072000037`.

P. Erdős and A. Rényi. "On Random Graphs". *Publicationes Mathematicae*, 6:290–297, 1959.

M. Fafchamps and F. Gubert. "The Formation of Risk Sharing Networks". *Journal of Development Economics*, 83:326–350, 2007.

M. Fafchamps and S. Quinn. "Networks and Manufacturing Firms in Africa: Results from a Randomized Field Experiment". 2016.

B. Feigenberg, E. Field, and R. Pande. "The Economic Returns to Social Interaction: Experimental Evidence from Microfinance". 80(4):1459–1483, 2013.

O. Frank. "Sampling and Estimation in Large Social Networks". *Social Networks*, 1:91–101, 1978.

O. Frank. "Estimation of the Number of Vertices of Different Degrees in a Graph". *Journal of Statistical Planning and Inference*, 4:45–50, 1980a.

O. Frank. "Sampling and Inference in a Population Graph". *International Statistical Review/Revue*

Internationale de Statistique, 48(1):33–41, 1980b.

O. Frank. "A Survey of Statistical Methods for Graph Analysis". *Sociological Methodology*, 23: 110–155, 1981.

T. L. Frantz, M. Cataldo, and K.M. Carley. "Robustness of centrality measures under uncertainty: Examining the role of network topology". *Computational and Mathematical Organization Theory*, 15:303–328, 2009.

J. Galaskiewicz. "Estimating Point Centrality Using Different Network Sampling Techniques". *Social Networks*, 13:347–386, 1991.

E. N. Gilbert. "Random Graphs". *Annals of Mathematical Statistics*, 30:1141–1144, 1959.

P. Goldsmith-Pinkham and G. W. Imbens. "Social Networks and the Identification of Peer Effects". *Journal of Business and Economic Statistics*, 31:253–264, 2013.

B. S. Graham. "An Econometric Model of Link Formation with Degree Heterogeneity". *NBER Working Paper 20341*, 2014.

B. S. Graham. "Methods of Identification in Social Networks". *Annual Review of Economics*, 7, 2015.

Jonathan Guryan, Kory Kroft, and Matthew J. Notowidigdo. Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. 1(4):34–68, October 2009. doi: 10.1257/app.1.4.34. URL `http://www.aeaweb.org/articles?id=10.1257/app.1.4.34`.

J. J. Heckman and R. Robb. "Alternative Methods for Evaluating the Impacts of Interventions: An Overview". *Journal of Econometrics*, 30(1):239–267, 1985.

James Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–61, 1979.

W. Horrace, X. Liu, and E. Patacchini. "Endogenous Network Production Functions with Selectivity". *Journal of Econometrics*, 190:222–232, 2016.

C. Hoxby and G. Weingarth. "Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects". *mimeo, Stanford University*, 2005.

C.-S. Hsieh and L-F. Lee. "A Social Interactions Model with Endogenous Friendship Formation and Selectivity". *Journal of Applied Econometrics*, forthcoming.

M.O. Jackson, T. Rodriguez-Barraquer, and X. Tan. "Social Capital and Social Quilts: Network Patterns of Favor Exchange". *American Economic Review*, 102(5):1857–97, 2012.

P. Kim and H. Jeong. "Reliability of rank order in sampled networks". *The European Physical Journal B*, 55:109–114, 2007.

E. Kolaczyk. *"Statistical Analysis of Network Data"*. Springer, 2009.

M. König, X. Liu, and Y. Zenou. "R&D Networks: Theory, Empirics, and Policy Implications". Technical report, CEPR Discussion Paper 9872, 2014.

G. Kossinets. "Effects of missing data in social networks". *Social Networks*, 28:247–268, 2006.

L-F. Lee and X. Liu. "Identification and GMM Estimation of Social Interactions Models with Centrality". *Journal of Econometrics*, 159:99–115, 2010.

Lung-Fei Lee. Generalized Econometric Models with Selectivity. *Econometrica*, 51(2):507–12, March 1983.

S. H. Lee, P. Kim, and H. Jeong. "Statistical properties of sampled networks". *Physical Review E*, 73(1), 2006.

X. Liu. "Estimation of a local-aggregate network model with sampled networks". *Economics Letters*, 118:243–246, 2013.

X. Liu, E. Patacchini, and E. Rainone. "The Allocation of Time in Sleep: a Social Network Model with Sampled Data". *CPR Working Paper No. 162*, 2013.

X. Liu, E. Patacchini, and Y. Zenou. "Endogenous Peer Effects: Local Aggregate or Local Average?". *Journal of Economic Behavior and Organization*, 103:39–59, 2014a.

X. Liu, E. Patacchini, Y. Zenou, and L-F. Lee. "Criminal Networks: Who is the Key Player?". *mimeo*, 2014b.

C. Manski. "Identification of Endogenous Social Effects: The Reflection Problem". *Review of Economic Studies*, 60:531–542, 1993.

R. Méango. "International Student Migration: A Partial Identification Analysis". *mimeo*, 2014.

A. Mele. "A Structural Model of Segregation in Social Networks". *mimeo*, 2013.

K. Mihaly. "Do More Friends Mean Better Grades? Student Popularity and Academic Achievement". *RAND Working Papers*, WR-678, 2009.

M. Mohnen. "Stars and Brokers: Peer Effects among Medical Scientists". *mimeo, University College London*, 2016.

K. Munshi and J. Myaux. "Social Norms and the Fertility Transition". *Journal of Development Economics*, 80 (1):1–38, 2006.

M. Patnam. "Corporate Networks And Peer Effects In Firm Policies". *mimeo, ENSAE-CREST*, 2013.

B. Sacerdote. "Peer Effects With Random Assignment: Results For Dartmouth Roommates". *Quarterly Journal of Economics*, 116:681–704, 2001.

B. Sacerdote. "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, editors, *Handbook of the Economics of Education*, volume 3. Elsevier, 2011.

S. K. Thompson. "Adaptive Web Sampling". *Biometrics*, 62(4):1224–1234, 2006.

M. Tincani. "Heterogeneous Peer Effects and Rank Concerns: Theory and Evidence". 2015.

G. Topa and Y. Zenou. "Neighborhood and Network Effects". In G. Duranton, V. Henderson, and W. Strange, editors, *Handbook of Regional and Urban Economics*, volume 5A, chapter 9. Elsevier, 2015.

F. Waldinger. "Quality Matters: The Expulsion of Professors and the Consequences for PhD Students Outcomes in Nazi Germany". *Journal of Political Economy*, 118 (4):787–831, 2010.

F. Waldinger. "Peer Effects in Science - Evidence from the Dismissal of Scientists in Nazi Germany". *Review of Economic Studies*, 79 (2):838–861, 2012.

Wei Wang and Lung-Fei Lee. Estimation of spatial autoregressive models with randomly missing data in the dependent variable. *Econometrics Journal*, 16(1):73–102, 02 2013.

S. Wasserman and P. Pattison. "Logit models and logistic regressions for Social Networks: I. An Introduction to Markov Graphs and p*". *Psychometrika*, 61:401–425, 1996.

A Definitions

- **Adjacency Matrix, $\boldsymbol{G}$:** An $N \times N$ matrix, $\boldsymbol{G}$, whose ij^{th} 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, $i_1 i_2, i_2 i_3, ..., i_{R-1} i_R$, such that $i_r i_{r+1} \in g$, for each $r \in \{1, ..., R\}$ with $i_1 = i$ and $i_R = j$ and such that each node in the sequence $i_1, ..., i_R$ 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 1 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 1 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 i^{th} 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 $span = \dfrac{log(N) - log(\bar{d})}{log(\tilde{d}) - log(\bar{d})} + 1$ 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.

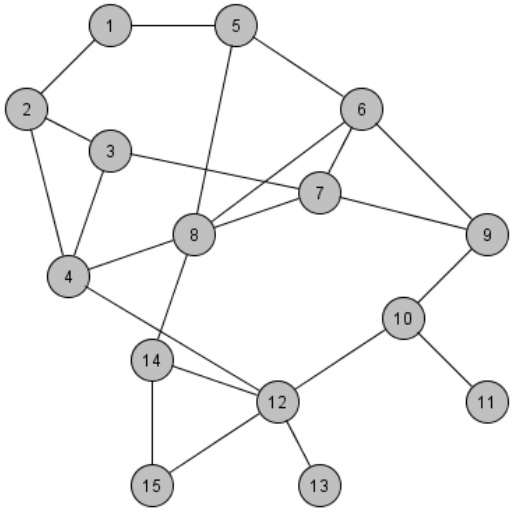
- C**ore-periphery network:** Network that can be partitioned into a set of nodes that is completely connected ('core'), and another set ('periphery') who are linked primarily with nodes in the 'core'.

- **Cellular network:** Networks containing many cliques, with few edges connecting the different cliques.

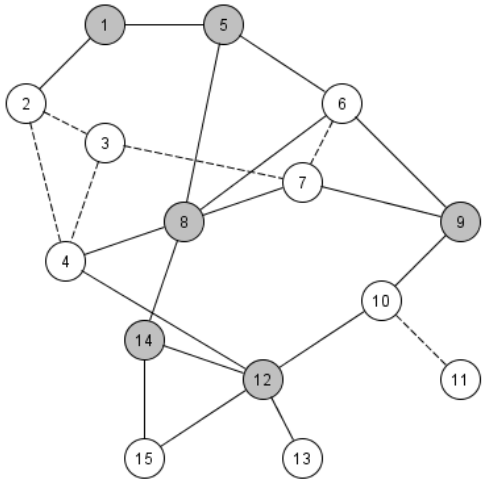
- **Small world network:** Network where most nodes are not directly linked to one another, but where geodesics between nodes are small.

- **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.

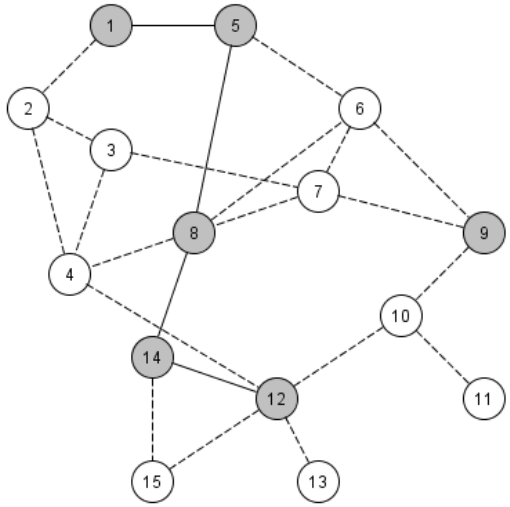
Figure 1: Star and Induced Subgraph



Average degree = 3.067
(a) Full Graph

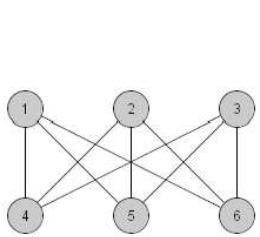


Average degree = 2.615
(b) Star Subgraph

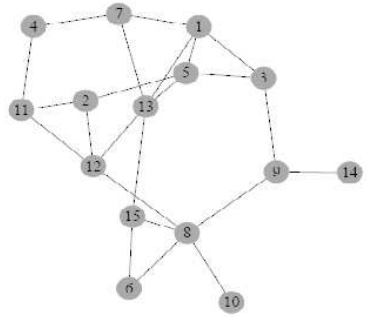


Average degree = 1.333
(c) Induced Subgraph

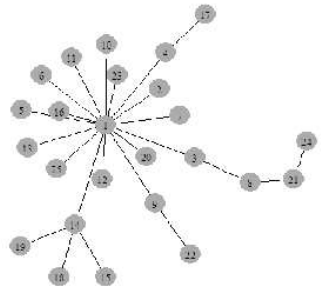
Figure 2: Network Topologies



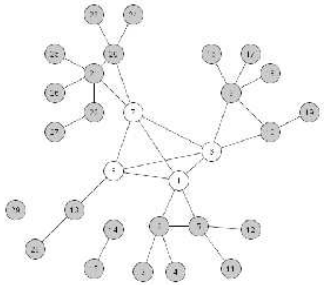
(a) Bipartite Network



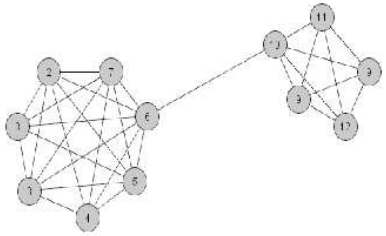
(b) Uniform Random



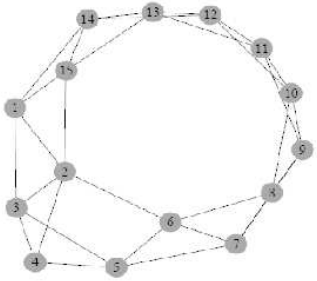
(c) Scale-free



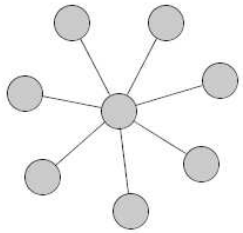
(d) Core-periphery



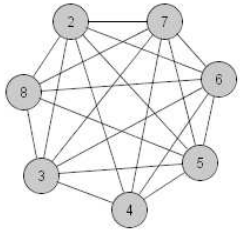
(e) Cellular



(f) Small world



(g) K-Star



(h) Clique

B Collecting Network Data: Sampling Methods

In order to construct the full network, researchers would need to collect data on all nodes and edges, *i.e.* 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:

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 networks 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 *i.e.* those for whom there is a relatively small proportion in the population. For these groups one would get an insufficiently large 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 *i.e.* 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).