Institute for Fiscal Studies

Empirical Methods for Networks Data: Social Effects, Network Formation and Measurement Error

IFS Working Paper W14/34

Arun Advani
Bansi Malde

Empirical Methods for Networks Data: Social Effects, Network Formation and Measurement Error

Arun Advani and Bansi Malde*

19th December 2014

Abstract

In many contexts we may be interested in understanding whether direct connections between agents, such as declared friendships in a classroom or family links in a rural village, affect their outcomes. In this paper we review the literature studying econometric methods for the analysis of social networks. We begin by providing a common framework for models of *social effects*, a class that includes the 'linear-in-means' local average model, the local aggregate model, and models where network statistics affect outcomes. We discuss identification of these models using both observational and experimental/quasi-experimental data. We then discuss models of network formation, drawing on a range of literatures to cover purely predictive models, reduced form models, and structural models, including those with a strategic element. Finally we discuss how one might collect data on networks, and the measurement error issues caused by sampling of networks, as well as measurement error more broadly.

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

JEL Classification: C31, C81, Z13

*Both authors are affiliated with the Institute for Fiscal Studies and University College London. E-mail: arun_a@ifs.org.uk; bansi_m@ifs.org.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.

Contents

1 Introduction

Whilst anonymous markets have long been central to economic analysis, the role of networks as an alternative mode of interaction is increasingly being recognised. Networks might act as a substitute for markets, for example providing access to credit in the absence of a formal financial sector, or as a complement, for example transmitting information about the value of a product. Analysis that neglects the potential for such *social effects* when they are present is likely to mismeasure any effects of interest.

In this paper we provide an overview of econometric methods for working with network data – data on agents ('nodes') and the links between them – taking into account the peculiarities of the dependence structures present in this context. We draw on both the growing economic literature studying networks, and on research in other fields, including maths, computer science, and sociology. The discussion proceeds in three parts: (i) estimating social effects given a (conditionally) exogenous observed network; (ii) estimating the underlying network formation process, given only a single cross-section of data; and (iii) data issues, with a particular focus on accounting for measurement error, since in a network-context this can have particularly serious consequences.

The identification and estimation of social effects – direct spillovers from the characteristics or outcome of one agent to the outcome of others – are of central interest in empirical research on networks in economics. Whilst researchers have tended to focus on the effects from the average characteristics and outcomes of network 'neighbours', different theoretical models will imply different specifications for social effects. In Section 3 we begin by setting out a common framework for social effects, which has as a special case the common 'linear-in-means' specification, as well as a number of other commonly used specifications. Since the general model is not identified, we then go through some important special cases, first outlining the theoretical model which generates the specification, before discussing issues related to identification of parameters.[1] For most of our discussion we focus on identification of the parameters using only observational data, since this is typically what researchers have available to them. We then go on to consider the conditions under which experimental variation can help weaken the assumptions needed to identify the parameters of interest.

The key challenge for credible estimation of social effects comes from the likely endogeneity of the network. Thus far most of the empirical literature has simply noted this issue without tackling it head on, but more recently researchers have tried to tackle it directly. The main approach to doing this has been to search for instruments which change the probability of a link existing without directly affecting the outcome. Alternatively, where panel data are available, shocks to network structure – such as node death – have been used to provide exogenous variation. These approaches

[1] A different presentation of some of the material in this part of Section 3 can be found in Topa and Zenou (2015). Of the models we discuss, their focus is on two of the more common specifications used. Topa and Zenou (2015) compare these models to each other, and also to neighbourhood effect models, and discuss the relationship between neighbourhood and network models.

naturally have all the usual limitations: a convincing story must be provided to motivate the exclusion restriction, and where there is heterogeneity they identify only a local effect. Additionally, they rely on the underlying network formation model having a unique equilibrium. Without uniqueness we do not have a complete model, as we have not specified how an equilibrium is chosen. Hence a particular realisation of the instrument for some group of nodes is consistent with multiple resulting network structures, and a standard IV approach cannot be used.

This provides one natural motivation for the study of network formation models: being able to characterise and estimate a model of network formation would, in the presence of exclusion restrictions (or functional form assumptions motivated by theory) allow us to identify social effects using the predicted network. Formation models can also be useful for tackling measurement error, by imputing unobserved links. Finally, in some circumstances we might be interested in these models *per se*, for example to understand how we can influence network structure and hence indirectly the distribution of outcomes.

In Section 4 we consider a range of network formation models, drawing from literatures outside economics as well as recent work by economists, and show how these methods relate to each other. We first consider purely descriptive models that make use of only data on the observed links, and can be used to make in-sample predictions about unobserved links given the observed network structure. Next we turn to reduced form economic models, which make use of node characteristics in predicting links, but which do not allow for dependencies in linking decisions. Lastly we discuss the growing body of work estimating games of strategic network formation, which allow for such dependencies and so at least, in principle, can have multiple equilibria.[2]

The methods discussed until now have all assumed access to data on a population of nodes and all the relevant interconnections between them. However, defining and measuring the appropriate network is often not straightforward. In Section 5 we begin by discussing issues in network definition and measurement. We then discuss different sampling approaches: these are important because networks are comprised of interrelated nodes and links, meaning that a sampling strategy over one of these objects will define a non-random sampling process over the other. For example if we sample edges randomly, and compute the mean number of neighbours for the nodes to whom those edges belong, this estimated average will be higher than if the average were computed across all nodes, since nodes with many edges are more likely to have been included in the sample by construction. Next we discuss different sources of measurement error, and their implications for the estimation of network statistics and regression parameters. We end with an explanation of 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. In the context of social effects estimation, we omit entirely any discussion of *peer effects* where all that is known about agents' links are the groups to which they belong. A recent survey by Blume et al. (2010) more than amply covers this ground, and

[2]Another review of the material on strategic network formation is provided by Graham (2015).

we direct the interested reader to their work. We also restrict our focus to linear models, which are appropriate for continuous outcomes but may be less suited to discrete choice settings such as those considered by Brock and Durlauf (2001) and Brock and Durlauf (2007). Similarly in our discussion of network formation, we do not consider in any detail the literature on the estimation of games. Although strategic models of network formation can be considered in this framework, the high dimension of these models typically makes it difficult to employ the same methods as are used in the game context. For readers who wish to know more about these methods, the survey paper by de Paula (2013) is a natural starting point. Finally, for a survey of applied work on networks in developing countries, see the review by Chuang and Schechter (2014).

We round off the paper with some concluding remarks, drawing together the various areas discussed, noting the limits of what we currently know about the econometrics of networks, and considering the potential directions for future research. Appendix A then provides detailed definitions of the various network measures and topologies that are mentioned in the text below.

2 Notation

Before we proceed, we first outline the notation we use throughout the paper. We define a *network* or *graph* $g = (\mathscr{N}_g, \mathscr{E}_g)$[3] as a set of nodes, $\mathscr{N}_g$, and edges or links, $\mathscr{E}_g$.[4] The nodes represent individual agents, and the edges represent the links between pairs of nodes. In economic applications, nodes are usually individuals, households, firms or countries. Edges could be social ties such as friendship, kinship, or co-working, or economic ties such as purchases, loans, or employment relationships. The number of nodes present in g is $N_g = |\mathscr{N}_g|$, and the number of edges is $E_g = |\mathscr{E}_g|$. We define $\mathcal{G}_N = \{g : |\mathscr{N}_g| = N\}$ as the set of all possible networks on N nodes.

In the simplest case – the *binary network* – any (ordered) pair of nodes $i, j \in \mathscr{N}_g$ is either linked, $ij \in \mathscr{E}_g$, or not linked, $ij \notin \mathscr{E}_g$. If $ij \in \mathscr{E}_g$ then j is often described as being a *neighbour* of i. We denote by $nei_{i,g} = \{j : ij \in \mathscr{E}_g\}$ the *neighbourhood* of node i, which contains all nodes with whom i is linked. Nodes that are neighbours of neighbours will often be referred to as *'second degree neighbour'*. Typically it is convenient to assume that $ii \notin \mathscr{E}_g \ \forall i \in \mathscr{N}_g$. Edges may be directed, so that a link from node i to node j is not the same as a link from node j to node i; in this case the network is a *directed graph* (or *digraph*). In Section 4 we will at times find it useful to explicitly enumerate the edges; we denote by Λ this set of enumerated edges, with typical element l. Unlike $\mathscr{E}_g$, Λ is an ordered set, with order 12, 13, ...$N(N-1)$, so that we may use $(l-1)$ to denote the element in the set one position before l.

A more general case than the binary graph is that of a *weighted graph*, in which the edge set contains all possible combinations of nodes, other than to the node itself. That is, $\mathscr{E}_g = \{ij : \forall i, j \in \mathscr{N}_g,\ i \neq$

[3] In a slight abuse of notation, we will also use g to index individual networks when data from multiple networks is available.

[4] In Appendix A we provide further useful definitions.

$j\}$. Moreover, edges have *edge weights* $wei(i,j)$ which measure some metric of distance or link strength. Care is needed in interpreting the value of weights, as these differ by context. 'Distance' weighted graphs, which arise for example when weights represent transaction costs between two nodes, would typically have $wei^d(i,j) \in [0,\infty)$, with $wei^d(i,j) = \infty$ being equivalent to i and j being unconnected in the binary graph case. Conversely, 'strength' weighted graphs, where weights capture for example the frequency of interaction between agents, typically have $wei^s(i,j) \in [0,\bar{w}]$, with $wei^s(i,j) = 0$ being equivalent to i and j being unconnected in the binary graph case and $\bar{w} < \infty$.[5] Which definition is used depends on the context and application, but similar methods can be used for analysis in either case.[6]

Network graphs, whether directed or not, can also be represented by an *adjacency matrix*, $\boldsymbol{G}_g$, with typical element $G_{ij,g}$. This is an $N_g \times N_g$ matrix with the leading diagonal normalised to 0. When the network is binary, $G_{ij,g}$= 1 if $ij \in \mathscr{E}_g$, and 0 otherwise, while for weighted graphs, $G_{ij,g} = wei(i,j)$. We will use the notation $\boldsymbol{G}_{i,g}$ to denote the i^{th} row of the adjacency matrix $\boldsymbol{G}_g$, and $\boldsymbol{G}'_{i,g}$ to denote its i^{th} column.[7] Many models defined for binary networks make use of the row-stochastic[8] adjacency matrix or *influence matrix,* $\tilde{\boldsymbol{G}}_g$. Elements of this matrix are generally defined as $\tilde{G}_{ij,g} = {}^{G_{ij,g}}/_{\sum_j G_{ij,g}}$ if two agents are linked and 0 otherwise.

When we describe empirical methods for identifying and estimating social effects, we will frequently work with data from a number of network graphs. Graphs for different networks will be indexed, in a slight abuse of notation, by $g = 1, ..., M$, where M is the total number of networks in the data. Node-level variables will be indexed with $i = 1, ..., N_g$, where N_g is the number of nodes in graph g. Node-level outcomes will be denoted by $y_{i,g}$, while exogenous covariates will be denoted by the $1 \times K$ vector $\boldsymbol{x}_{i,g}$ and common network-level variables will be collected in the $1 \times Q$ vector, $\boldsymbol{z}_g$.

The node-level outcomes, covariates and network-level variables can be stacked for each node in a network. In this case, we will denote the stacked $N_g \times 1$ outcome vector as $\boldsymbol{y}_g$ and the $N_g \times K$ matrix stacking node-level vectors of covariates for graph g as $\boldsymbol{X}_g$. Common network-level variables for graph g will be gathered in the matrix $\boldsymbol{Z}_g = \boldsymbol{\iota}_g \boldsymbol{z}_g$ where $\boldsymbol{\iota}_g$ denotes an $N_g \times 1$ vector of ones. The adjacency and influence matrices for network g will be denoted by $\boldsymbol{G}_g$ and $\tilde{\boldsymbol{G}}_g$. At times we will also make use of the $N_g \times N_g$ identity matrix, $\boldsymbol{I}_g$, consisting of ones on the leading diagonal, and zeros elsewhere.

Finally, we introduce notation for vectors and matrices stacking together the network-level outcome vectors, covariate matrices and adjacency matrices for all networks in the data. $\boldsymbol{Y} = (\boldsymbol{y}_1', ..., \boldsymbol{y}_M')'$ is an $\sum_{g=1}^{M} N_g \times 1$ vector that stacks together the outcome vectors; $\boldsymbol{G} = diag\{\boldsymbol{G}_g\}_{g=1}^{g=M}$ denotes the $\sum_{g=1}^{M} N_g \times \sum_{g=1}^{M} N_g$ block-diagonal matrix with network-level adjacency matrices along the leading

[5] In both of these examples, $wei(i,j) = wei(j,i)$. More generally this need not be true. For example, in some settings one might use 'flow weights' where $wei^f(i,j)$ represents the net flow of, say, resources from i to j. Then by definition $wei^f(i,j) = -wei^f(j,i)$, and the weighted adjacency matrix, defined shortly, is skew-symmetric.

[6] With distance weighted graphs, one must be careful in dealing with edges where $wei^d(i,j) = \infty$. A good approximation can usually be made by replacing infinity with an arbitrarily high finite value.

[7] $\boldsymbol{G}'_{i,g}$ is the i^{th} row of $\boldsymbol{G}'_g$, which is the i^{th} column of $\boldsymbol{G}_g$.

[8] A row stochastic (also called 'right stochastic' matrix) is one whose rows are normalised so they each sum to one.

diagonal and zeros off the diagonal, and analogously $\tilde{\boldsymbol{G}} = diag\{\tilde{\boldsymbol{G}}_g\}_{g=1}^{g=M}$ (with similar dimensions as $\boldsymbol{G}$) for the influence matrices; and $\boldsymbol{X} = (\boldsymbol{X}_1', ..., \boldsymbol{X}_M')'$ and $\boldsymbol{Z} = (\boldsymbol{Z}_1', ..., \boldsymbol{Z}_M')'$ are respectively, $\sum_{g=1}^{M} N_g \times K$ and $\sum_{g=1}^{M} N_g \times Q$ matrices, that stack together the covariate matrices across networks. Finally, we define the vector $\boldsymbol{\iota}$ as a $\sum_{g=1}^{M} N_g \times 1$ vector of ones and the matrix $\boldsymbol{L} = diag\{\boldsymbol{\iota}_g\}_{g=1}^{g=M}$, as an $\sum_{g=1}^{M} N_g \times M$ matrix with each column being an indicator for being in a particular network.

3 Social Effects

Researchers are typically interested in understanding how the behaviour, choices and outcomes of agents are influenced by the agents that they interact with, *i.e.* by their neighbours. This section reviews methods that have been used to identify and estimate these social effects.[9] We consider a number of restrictions that would allow parameters of interest to be recovered, and place them into a broader framework. We focus on linear estimation models, which cover the bulk of methods used in practice.

We begin by providing a common organisational framework for the different empirical specifications that have been applied in the literature. Thereafter, we discuss in turn a series of commonly used specifications, the underlying theoretical models that generate them, and outline conditions for the causal identification of parameters with observational cross-sectional data. We then briefly discuss how experimental and quasi-experimental variation could be used to uncover social effects. Finally, we discuss some methods that can be applied to overcome confounding due to endogenous formation of edges, and discuss their limitations. A comprehensive overview of models of network formation is provided in Section 4.

We will use a specific example throughout this section to better illustrate the restrictions imposed by each of the different models and empirical specifications. Specifically, we will consider how we can use these methods to answer the following question: How is a teenager's schooling performance influenced by his friends? This is a widely studied question in the education and labour economics literatures, and is of great policy interest.[10]

We take as given throughout this section that the researcher knows the network(s) for which he is trying to estimate social effects and that he observes the entirety of this network without error. In Section 5 we will discuss how these data might be collected, and the consequences of having only a partial sample of the network and/or imperfectly measured networks.

3.1 Organising Framework

Almost all (linear) economic models of social effects can be written as a special case of the following equation (written in matrix terms using the notation specified in Section 2):

[9] We leave aside the important issues of inference, in order to keep the scope of this survey manageable.
[10] See Sacerdote (2011) for an overview of this literature.

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \boldsymbol{w_y}(\boldsymbol{G},\, \boldsymbol{Y})\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{w_x}(\boldsymbol{G},\, \boldsymbol{X})\boldsymbol{\delta} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{L}\boldsymbol{\nu} + \boldsymbol{\varepsilon} \tag{3.1}$$

$\boldsymbol{Y}$ is a vector stacking individual outcomes of nodes across all networks.[11] $\boldsymbol{X}$ is a matrix of observable background characteristics that influence a node's own outcome and potentially that of others in the network. $\boldsymbol{G}$ is a block-diagonal matrix with the adjacency matrices of each network along its leading diagonal, and zeros on the off-diagonal. $\boldsymbol{w_y}(\boldsymbol{G},\, \boldsymbol{Y})$ 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 (possibly all) nodes in the network, influence the outcome, $\boldsymbol{Y}$. 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. $\boldsymbol{Z}$ is a matrix of observed network-specific variables; $\boldsymbol{\nu} = \{\nu_g\}_{g=1}^{g=M}$ is the associated vector of network-specific mean effects, unobserved by the econometrician but known to agents; and $\boldsymbol{\varepsilon}$ is a vector stacking the (unobservable) error terms for all nodes across all networks.

We make the following assumptions on the $\boldsymbol{\varepsilon}$ term:

$$\mathbb{E}[\varepsilon_{i,g}|\, \boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}_g] = 0 \quad \forall\, i \in g;\ g \in \{1, ..., M\} \tag{3.2}$$

$$Cov[\varepsilon_{i,g}\varepsilon_{k,h}|\, \boldsymbol{X}_g, \boldsymbol{X}_h, \boldsymbol{Z}_g, \boldsymbol{Z}_h, \boldsymbol{G}_g, \boldsymbol{G}_h] = 0 \quad \forall\, i \in g;\ k \in h;\ g, h \in \{1, ..., M\};\ g \neq h \tag{3.3}$$

Equation 3.2 says that the error term for individual nodes in a network is mean independent of observed node-level characteristics of all network members, of network-level characteristics and of the network structure, as embodied in the adjacency matrix $\boldsymbol{G}_g$. The network, is in this sense assumed to be exogenous, conditional on individual-level observable characteristics and network-level observable characteristics. Later in Subsection 3.7 below, we will review some approaches taken to relax this assumption. In addition, Equation 3.3 implies that the error terms of all nodes, i and k in different networks, g and h, are uncorrelated conditional on observable characteristics of the nodes, the observable characteristics of the networks, and the structure of the network. Finally, note that no assumptions are imposed on the covariance of node-level error terms within the same network.

In some cases, the following assumption is made on $\boldsymbol{\nu}$:

$$\mathbb{E}[\nu_g|\, \boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}_g] = 0 \quad \forall\, g \in \{1, ..., M\} \tag{3.4}$$

That is, the network-level unobservable is mean independent of observable node- and network-level characteristics, and of the network. Many of the models that we consider below relax this

[11] We allow $\boldsymbol{Y}$ to be univariate, so individuals have only a single outcome. A recent paper by Cohen-Cole et al. (forthcoming) discusses how to relax this assumption, and provides some initial evidence that restricting outcomes to only a single dimension might be important in empirical settings.

assumption and allow for correlation between $\boldsymbol{\nu}$ and the other right hand side variables in Equation 3.1.

The social effect parameter that is most often of interest to researchers is $\boldsymbol{\beta}$ - the effect of a function of a node's neighbours' outcomes (*e.g.* an individual's friends' schooling performance) and the network. This is also known as the *endogenous effect*, to use the term coined by Manski (1993). This parameter is often of policy interest, since 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 parameters $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ are known as the *exogenous or contextual effect* while $\boldsymbol{\nu}$ captures a *correlated effect*.

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

1. *Local average models*: This model corresponds with $\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y}) = \tilde{\boldsymbol{G}}\boldsymbol{Y}$ and $\boldsymbol{w_x}(\boldsymbol{G}, \boldsymbol{X}) = \tilde{\boldsymbol{G}}\boldsymbol{X}$, which arises when node outcomes are influenced by the average behaviour and characteristics of his direct neighbours. In our schooling example, this model implies that an individual's schooling performance is a function of the average schooling performance of his friends, his own characteristics, the average characteristics of his friends and some background network characteristics. This can apply, for example, when social effects operate through a desire for a node to conform to the behaviour of its neighbours. The identifiability of the parameters β, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$ from the data available to a researcher depends on the structure of the network and the level of detail available about the network:[12]

   (a) With data containing information only on the broad peer group that a node belongs to and where a node can belong to a single group only (*e.g.* a classroom), it is common to assume that the node is directly linked with all other nodes in the same group and that there are no links between nodes in different groups. In this case, the peer group corresponds to the network. All elements of the influence matrix of a network g, $\tilde{\boldsymbol{G}}_g$, (including the diagonal) are set to $\frac{1}{N_g}$ where N_g is the number of agents within the network.[13] This generates the linear-in-means peer group model studied by Manski (1993) among others. Manski (1993) shows that identification of the parameter β is hampered by a simultaneity problem that he labels the *reflection problem*: it is not possible to differentiate whether the choices of a node i in the network influence the choices of node j, or vice versa. An alternative definition for $\tilde{\boldsymbol{G}}$ sets all diagonal terms of the network-level influence matrices, $\tilde{\boldsymbol{G}}_g$, to 0 and off-diagonal terms to $\frac{1}{N_g - 1}$, which implies using the leave-self-out mean outcome as the regressor generating social effects. With this definition, identification of the parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$ is possible in some circumstances as shown by Lee (2007).[14] Identification issues related to this model with

[12] The parameter $\boldsymbol{\eta}$ can also be identified under the assumption that $\mathbb{E}[\boldsymbol{\nu}|\, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{G}] = 0$.

[13] Note that in this case, since all nodes are linked to all others (including themselves), the total number of i's edges (or *degree*), $d_{i,g} = \sum_j G_{ij,g} = N_g \,\forall\, i \in g$. Hence by definition, all elements of $\tilde{\boldsymbol{G}}_g$ are set to $\frac{1}{N_g}$.

[14] Other solutions to the reflection problem have also been proposed, such as those by Glaeser et al. (1996), Moffitt

single peer groups have been surveyed in detail elsewhere, and thus will not be considered here. The interested reader should consult the comprehensive review by Blume et al. (2010).

(b) If instead detailed network data (*i.e.* information on nodes and the edges between them) are available, or if nodes belong to multiple partially overlapping peer groups, it may be possible to separately identify the parameters β, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$ from a single cross-section of data. In this case, elements of the network-level influence matrices, $\tilde{\boldsymbol{G}}_g$ are defined as $\tilde{G}_{ij,g} = \frac{1}{d_{i,g}}$ when a link between i and j exists, where $d_{i,g}$ is the total number of i's links (or degree); and 0 otherwise. Identification results for observational network data have been obtained by Bramoullé et al. (2009). These are explored in more detail in Subsection 3.2 below.

2. *Local aggregate models*: 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 our schooling example, it may be more productive for an individual to put in more effort in studying if his friends also put in more effort, consequently leading to better schooling outcomes. In this case a node's outcome depends on the aggregate outcome of its neighbours. In the context of Equation 3.1, this implies that $\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y}) = \boldsymbol{GY}$ and $\boldsymbol{w_x}(\boldsymbol{G}, \boldsymbol{X})$ is typically defined to be $\tilde{\boldsymbol{G}}\boldsymbol{X}$, implying that the outcome of interest is influenced by the *average* exogenous characteristics of a node's neighbours.[15] Identification and estimation of this model in observational networks data has been studied by Calvó-Armengol et al. (2009), Lee and Liu (2010) and Liu et al. (2014b). More details are provided in Subsection 3.3 below.

3. *Hybrid local models:* 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 schooling example, the model implies that individuals may want to 'fit-in' and thus put in similar amounts of effort in studying as their friends, but their studying efforts may also be more productive if their friends also put in effort. Both of these channels then influence their schooling performance. In the notation of Equation 3.1, it implies that $\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y}) = \mathbf{G}\boldsymbol{Y} + \tilde{\boldsymbol{G}}\boldsymbol{Y}$. As in the local average and aggregate models above, $\boldsymbol{w_x}(\boldsymbol{G}, \boldsymbol{X})$ is typically defined to be $\tilde{\boldsymbol{G}}\boldsymbol{X}$. Identification and estimation of this model with observational data is studied by Liu et al. (2014a). See Subsection 3.4 for more details.

4. Networks may influence node outcomes (and consequently aggregate network outcomes) through more general features or functionals of the network. For instance, the DeGroot (1974)

(2001), and Graham (2008). Kwok, 2013 provides a general study of the conditions under which identification of parameters can be achieved. He finds that network *diameter* – the length of the longest geodesic – is the key parameter in determining identification.

[15] This choice of definition for $\boldsymbol{w_x}(\boldsymbol{G}, \boldsymbol{X})$ is, to our understanding, not based on any explicit theoretical justification. It does, however, ease identification as $\boldsymbol{w_x}(.)$ and $\boldsymbol{w_y}(.)$ are now different functions of $\boldsymbol{G}$.

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.[16] In the schooling context, if an individual's friends are also friends of each other (a phenomenon captured by clustering), he may have to spend less time maintaining these friendships due to scale economies, allowing him more time for school work thereby leading to better schooling performance.
Denoting a specific network statistic (such as eigenvector centrality in the social learning model above) by ω^r, where r indexes the statistic, we can specialise the term $\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y})\boldsymbol{\beta}$ in Equation 3.1 for node i in network g in a model with node-level outcomes 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 are used for $\boldsymbol{w_x}(\boldsymbol{G}, \boldsymbol{X})\boldsymbol{\delta}$. Models of this type have been estimated by Jackson et al. (2012) and Alatas et al. (2012).
When researchers are interested in *aggregate* network outcomes, rather than node level outcomes, the following specification is typically estimated:

$$\bar{\boldsymbol{y}} = \phi_0 + \bar{\boldsymbol{w}}_{\bar{\boldsymbol{y}}}(\boldsymbol{G})\boldsymbol{\phi}_1 + \bar{\boldsymbol{X}}\boldsymbol{\phi}_2 + \bar{\boldsymbol{w}}_{\bar{\boldsymbol{X}}}(\boldsymbol{G}, \bar{\boldsymbol{X}})\boldsymbol{\phi}_3 + \boldsymbol{u} \tag{3.5}$$

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 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. Models of this type have been estimated by among others, Banerjee et al. (2013), and are discussed further in Subsection 3.5.

In Subsections 3.2 to 3.5 below, we review methods relating to identification of the parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$, $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_3$ in these models,[17] under the assumption that the network is exogenous

[16] Eigenvector centrality is a more general function of the network than those considered above, since it relies on the whole structure of the network.

[17] $\boldsymbol{\eta}$ can also be identified in some cases, particularly when the assumption $\mathbb{E}[\boldsymbol{\nu}|\, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{G}] = 0$ is imposed.

conditional on observable individual and network-level variables. For each case discussed, we start by outlining a theoretical model that generates underlying the resulting empirical specification, and outline identification conditions using observational data.

Thereafter, in Subsection 3.6, we outline how experimental and quasi-experimental variation has been used to uncover social effects, and highlight some of the challenges faced in using such variation to uncover parameters of the structural models outlined in Subsections 3.2 to 3.4 below.

Subsection 3.7 outlines methods used by researchers to relax the assumption made in equation 3.2: that the individual error term is mean independent of the network and observed individual and network-level characteristics. Dealing with endogenous formation of social links is quite challenging, and so most of the methods outlined in this section fail to satisfactorily deal with the identification challenges posed by endogenous network formation. Moreover, none of these methods deal with the issue of measurement error in the network. These issues are considered in Sections 4 and 5 respectively.

3.2 Local Average Models

In local average models, a node's outcome (or choice) is influenced by the average outcome of its neighbours. Thus, an individual's schooling performance is influenced by the average schooling performance of his friends. The outcome for node i in network g, $y_{i,g}$, is typically modelled as being influenced by its own observed characteristics, $\boldsymbol{x}_{i,g}$, scalar unobserved heterogeneity $\varepsilon_{i,g}$, observed network characteristics $\boldsymbol{z}_g$, unobserved network characteristic ν_g, and also the average outcomes and characteristics of neighbours. Below, we consider identification conditions when data are available from multiple networks, though some results apply to data from a single network.[18]

Stacking together data from multiple networks yields the following empirical specification, expressed in matrix terms:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \beta\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\gamma} + \tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\delta} + \boldsymbol{Z}\boldsymbol{\eta} + \boldsymbol{L}\boldsymbol{\nu} + \boldsymbol{\varepsilon} \tag{3.6}$$

where $\boldsymbol{Y}$, $\boldsymbol{\iota}$, $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{L}$ and $\boldsymbol{\nu}$ are as defined previously; and $\tilde{\boldsymbol{G}}$ is a block diagonal matrix stacking network-level influence matrices along its leading diagonal, with all off-diagonal terms set to 0. The social effect, β, is a scalar in this model.

Given the simple empirical form of this model, it has been widely applied in the economics literature. Examples include:

- Understanding how the average schooling performance of an individual's peers influences the individual's own performance in a setting where students share a number of different classes

[18] When data on only a single network are available, the empirical specification is as follows: $\boldsymbol{y}_g = \boldsymbol{a} + \beta\tilde{\boldsymbol{G}}_g\boldsymbol{y}_g + \boldsymbol{X}_g\boldsymbol{\gamma} + \tilde{\boldsymbol{G}}_g\boldsymbol{X}_g\boldsymbol{\delta} + \boldsymbol{\varepsilon}_g$, where $\boldsymbol{a} = \alpha\boldsymbol{\iota}_g + \boldsymbol{Z}_g\boldsymbol{\eta} + \boldsymbol{\iota}_g\nu_g$ in our earlier notation, capturing all of the network-level characteristics.

(*e.g.* De Giorgi et al., 2010), or where students have some (but not all) common friends (*e.g.* Bramoullé et al., 2009).

- Understanding how non-market links between firms arising from company directors being members of multiple company boards influence firm choices on investment and executive pay (*e.g.* Patnam, 2013).

Although this specification is widely used in the empirical literature, few studies consider or acknowledge the form of its underlying economic model, even though parameter estimates are subsequently used to evaluate alternative policies and to make policy recommendations. Indeed, parameters are typically interpreted as in the econometric model of Manski (1993), whose parameters do not map back to 'deep' structural (*i.e.* policy invariant) parameters without an economic model.

An economic model that leads to this specification is one where nodes have a desire to conform to the average behaviour and characteristics of their neighbours (Patacchini and Zenou, 2012). In our schooling example, conformism implies that individuals would want to exert as much effort in their school work as their friends so as to 'fit in'. Thus, if one's friends may want to exert no effort in their school work, the individual would also not want to exert any effort in his school work.

Below we show how this model leads to Equation 3.6. However, this is not the only economic model that leads to an empirical specification of this form: a similar specification arises from, for example, models of perfect risk sharing, where a well-known result is that under homogeneous preferences, when risk is perfectly shared, the consumption of risk-averse households will move with average household consumption in the risk sharing group or network (Townsend, 1994).

Conformism is commonly modelled by node payoffs that are decreasing in the distance between own outcome and network neighbours' average outcomes. Payoffs are also allowed to vary with an individual heterogeneity parameter, $\pi_{i,g}$, which captures the individual's ability or productivity associated with the outcome:[19]

$$U_i(y_{i,g}; \boldsymbol{y}_{-i,g}, \boldsymbol{X}_g, \tilde{\boldsymbol{G}}_{i,g}) = \left(\pi_{i,g} - \frac{1}{2} \left(y_{i,g} - 2\beta \sum_{j=1}^{N_g} \tilde{G}_{ij,g} y_{j,g} \right) \right) y_{i,g} \tag{3.7}$$

β in Equation 3.7 can be thought of as a taste for conformism. Although we write this model as though nodes are perfectly able to observe each others' actions, this assumption can be relaxed. In particular, an econometric specification similar to Equation 3.6 can be obtained from a static model with imperfect information (see Blume et al., 2013).

The best response function derived from the first order condition with respect to $y_{i,g}$ is thus:

$$y_{i,g} = \pi_{i,g} + \beta \sum_{j=1}^{N_g} \tilde{G}_{ij,g} y_{j,g} \tag{3.8}$$

[19]Notice that in Equation 3.7, $\sum_{j=1}^{N_g} \tilde{G}_{ij,g} y_{j,g}$ is identical to the i^{th} row of $\tilde{\boldsymbol{G}}_g \boldsymbol{y}_g$, which appears in Equation 3.6.

Patacchini and Zenou (2012) derive the conditions under which a Nash equilibrium exists, and characterise properties of this equilibrium.

The individual heterogeneity parameter, $\pi_{i,g}$, can be decomposed into a linear function of individual and network characteristics (both observed and unobserved):

$$\pi_{i,g} = \boldsymbol{x}_{i,g}\boldsymbol{\gamma} + \sum_{j=1}^{N_g} \tilde{G}_{ij,g}\boldsymbol{x}_{j,g}\boldsymbol{\delta} + \boldsymbol{z}_g\boldsymbol{\eta} + \nu_g + \varepsilon_{i,g} \tag{3.9}$$

Substituting for this in Equation 3.8, we obtain the following best response function for individual outcomes:

$$y_{i,g} = \beta \sum_{j=1}^{N_g} \tilde{G}_{ij,g} y_{j,g} + \boldsymbol{x}_{i,g}\boldsymbol{\gamma} + \sum_{j=1}^{N_g} \tilde{G}_{ij,g}\boldsymbol{x}_{j,g}\boldsymbol{\delta} + \boldsymbol{z}_g\boldsymbol{\eta} + \nu_g + \varepsilon_{i,g} \tag{3.10}$$

Then, stacking observations for all nodes in multiple networks, we obtain Equation 3.6, which can be taken to the data.

Bramoullé et al. (2009) study the identification and estimation of Equation 3.6 in observational data with detailed network information or data from partially overlapping peer groups.[20] To proceed further, one needs to make some assumptions on the relationship between the unobserved variables – $\boldsymbol{\nu}$ and $\boldsymbol{\varepsilon}$ – and the other right hand side variables in Equation 3.6.

One specific assumption is that $\mathbb{E}[\boldsymbol{\varepsilon}|\boldsymbol{X}, \boldsymbol{Z}, \tilde{\boldsymbol{G}}] = 0$, *i.e.* the individual level error term, $\boldsymbol{\varepsilon}$, is assumed to be mean independent of the observed individual and network-level characteristics and of the network. The network level unobservable is also initially assumed to be mean independent of the right hand side variables, *i.e.* $\mathbb{E}[\boldsymbol{\nu}|\ \boldsymbol{X}, \boldsymbol{Z}, \tilde{\boldsymbol{G}}] = 0$; though this assumption will be relaxed further on.

Under these assumptions, the parameters $\{\alpha, \beta, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\eta}\}$ are identified if $\{\boldsymbol{I}, \tilde{\boldsymbol{G}}, \tilde{\boldsymbol{G}}^2\}$ are linearly independent. Identification thus relies on the network structure. In particular, the condition would not hold in networks composed only of cliques – subnetworks comprising of completely connected components – of the same size, and where the diagonal terms in the influence matrix, $\tilde{\boldsymbol{G}}$ are not set to 0. In this case, $\tilde{\boldsymbol{G}}^2$ can be expressed as a linear function of $\boldsymbol{I}$ and $\tilde{\boldsymbol{G}}$. Moreover, the model is then similar to the single peer group case of Manski (1993), and the methods outlined in Blume et al. (2010) apply.

In an undirected network (such as the in the left panel in Figure 1 below), this identification condition holds when there exists a triple of nodes (i, j, k) such that i is connected to j but not k, and j is connected to k. The exogenous characteristics of k, $\boldsymbol{x}_{k,g}$, directly affect j's outcome, but not (directly) that of i, hence forming valid instruments for the outcome of i's neighbours

[20] Similar identification results have been independently described by De Giorgi et al. (2010), who have data with overlapping peer groups of students who share a number of classes.

(*i.e.* j's outcome) in the equation for node i. Intuitively this method uses the characteristics of second-degree neighbours who are not direct neighbours as instruments for outcomes of direct neighbours.

(a) Intransitive triad in undirected network

(b) Intransitive triad in directed network

Figure 1: Intransitive triad in a undirected network (left panel) and a directed network (right panel)

It is thus immediately apparent why identification fails in networks composed only of cliques: in such networks, there is no triple of nodes (i, j, k) such that i is connected to j, and j is connected to k, but i is not connected to k.

In the directed network case, the condition is somewhat weaker, requiring only the presence of an intransitive triad: that is, a triple such that $ij \in \mathscr{E}$, $jk \in \mathscr{E}$ and $ik \notin \mathscr{E}$ (as in the right panel of Figure 1 above).[21] This is weaker than in undirected networks, which would also require that $ki \notin \mathscr{E}$.

As an example, consider using this method to identify the influence of the average schooling performance of an individual's friends on the individual, controlling for the individual's age and gender, the average age and gender of his friends, and some observed school characteristics (such as expenditure per pupil). Assume first that the underlying friendship network in this school is undirected as in the left panel of Figure 1, so that if i considers j to be his friend, j also considers i to be his friend. j also has a friend k who is not friends with i. We could then use the age and gender of k as instruments for the schooling performance of j in the equation for i. If instead, the network were directed as in the right panel of Figure 1, where the arrows indicate who is affected by whom (*i.e.* i is affected by j in the Figure, and so on), we can still use the age and gender of k as instruments for the school performance of j in the equation for i even though k is connected with i. This is possible since the direction of the relationship is such that k's school performance is affected by i's performance, but the converse is not true.

The identification result above requires that the network-level unobservable term be mean independent of the observed covariates, $\boldsymbol{X}$ and $\boldsymbol{Z}$, and of the network, $\tilde{\boldsymbol{G}}$. However, in many circumstances one might be concerned that unobservable characteristics of the network might be correlated with $\boldsymbol{X}$, so that $\mathbb{E}[\boldsymbol{\nu}|\boldsymbol{X}, \boldsymbol{Z}, \tilde{\boldsymbol{G}}] \neq 0$. For example, in our schooling context, when we take the network of interest to be constrained to be within the school, it is plausible that children with higher parental

[21] Equivalently, a triple such $ji \in \mathscr{E}$, $kj \in \mathscr{E}$ and $ki \notin \mathscr{E}$ forms an intransitive triad.

income will be in schools with teachers who have better unobserved teaching abilities, since wealthier parents may choose to live in areas with schools with good teachers. In this case, a natural solution when data on more than one network is available, is to include network fixed effects, $\boldsymbol{L}\tilde{\boldsymbol{\nu}}$ in place of the network-level observables, $\boldsymbol{Z}$, and the network-level unobservable, $\boldsymbol{L}\boldsymbol{\nu}$; where $\tilde{\boldsymbol{\nu}}$ is an $M \times 1$ vector that captures the network fixed effects.

Since the fixed effects themselves are generally not of interest, to ease estimation they are removed using a *within transformation*. This is done by pre-multiplying Equation 3.6 by $\boldsymbol{J}^{glob}$, a block diagonal matrix that stacks the network-level transformation matrices $\boldsymbol{J}_g^{glob} = \boldsymbol{I}_g - \frac{1}{N_g}(\boldsymbol{\iota}_g\boldsymbol{\iota}_g')$ along the leading diagonal, and off-diagonal terms are set to 0.[22] The resulting model, suppressing the superscript on $\boldsymbol{J}^{glob}$ for legibility, is of the following form:

$$\boldsymbol{JY} = \beta\boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{JX}\boldsymbol{\gamma} + \boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\delta} + \boldsymbol{J}\boldsymbol{\varepsilon} \tag{3.11}$$

In this case, the identification condition imposes a stronger requirement on network structure. In particular, the matrices $\{\boldsymbol{I}, \tilde{\boldsymbol{G}}, \tilde{\boldsymbol{G}}^2, \tilde{\boldsymbol{G}}^3\}$ should be linearly independent. This requires that there exists a pair of agents (i, j) such that the shortest path between them is of length 3, that is, i would need to go through at least two other nodes to get to j (as in Figure 2 below). The presence of at least two intermediate agents allows researchers to use the characteristics of third-degree neighbours (neighbours-of-neighbours-of-neighbours who are not direct neighbours or neighbours-of-neighbours) as an additional instrument to account for the network fixed effect.

Figure 2: Identification with network fixed effects. The picture on the left panel shows an undirected network with an agent l who is at least 3 steps away from i, while the picture on the right panel shows the same for a directed network.

A concern that arises when applying this method is that of instrument strength. Bramoullé et al. (2009) find that this varies with graph *density*, *i.e.*, the proportion of node pairs that are linked; and the level of *clustering*, *i.e.* the proportion of node triples such that precisely two of the possible three edges are connected.[23] Instrument strength is declining in density, since the number of intransitive triads tends to zero. The results for clustering are non-monotone, and depend on density.

The discussion thus far has assumed that the network through which the endogenous social effect operates is the same as the network through which the contextual effect operates. It is possible to

[22] This is a *global* within transformation, which subtracts the average across the entire network from the individual's value. Alternatively, a *local* within transformation, $\boldsymbol{J}_g^{loc} = \boldsymbol{I}_g - \tilde{\boldsymbol{G}}_g$, can be used, which would subtract only the average of the individual's peers rather than the average for the whole network. The latter transformation has slightly stricter identification conditions than the former, since it does not make use of the fact that the network fixed effect is common across all network members, and not just among directly linked nodes.

[23] This definition is also referred to as the clustering coefficient.

allow for these two networks to be distinct. This could be useful in a school setting, for instance, where contextual effects could be driven by the average characteristics of all students in the school, while endogenous effects by the outcomes of a subset of students who are friends. This might occur if the contextual effect operates through the level of resources the school has, which depends on the parental income of all students, whilst the peer learning might come only from friends.

Let $\boldsymbol{G}_{\boldsymbol{X},g}$ and $\boldsymbol{G}_{\boldsymbol{y},g}$ denote the network-level adjacency matrices through which, respectively, the contextual and endogenous effects operate. As before we define the block diagonal matrices $\boldsymbol{G}_{\boldsymbol{X}} = diag\{\boldsymbol{G}_{\boldsymbol{X},g}\}_{g=1}^{g=M}$ and $\boldsymbol{G}_{\boldsymbol{y}} = diag\{\boldsymbol{G}_{\boldsymbol{y},g}\}_{g=1}^{g=M}$. Blume et al. (2013) study identification of this model assuming that the two networks are (conditionally) exogenous and show that when the matrices $\boldsymbol{G}_{\boldsymbol{y}}$ and $\boldsymbol{G}_{\boldsymbol{X}}$ are observed by the econometrician, and at least one of δ and γ is non-zero, then the necessary and sufficient conditions for the parameters of Equation 3.6 to be identified are that the matrices $\boldsymbol{I}$, $\boldsymbol{G}_{\boldsymbol{y}}$, $\boldsymbol{G}_{\boldsymbol{X}}$ and $\boldsymbol{G}_{\boldsymbol{y}}\boldsymbol{G}_{\boldsymbol{X}}$ are linearly independent.

Although all parameters of interest can be identified by this method, the assumption that the network structure is conditionally exogenous is highly problematic. Though endogeneity caused by selection into a network can be overcome by allowing for group fixed effects which can be differenced out, endogenous formation of links within the network remains problematic and is substantially more difficult to overcome. Formally, the problem arises from the fact that agents' choices of with whom to link are correlated with unobservable (at least to the researcher) characteristics of both agents, so $\Pr(G_{ij,g} = 1|\varepsilon_{i,g}) \neq \Pr(G_{ij,g})$.

This means that the absence of a link between two nodes i and k may be correlated with $\varepsilon_{i,g}$ and $\varepsilon_{k,g}$, meaning that $\mathbb{E}[\varepsilon_{i,g}|\boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}_g] \neq 0$.[24] Consequently the condition in Equation 3.2 no longer holds. This is problematic for the method of Bramoullé et al. (2009), where the absence of a link is used to identify the social effect, and this absence could be for reasons related to the outcome of interest, thereby invalidating the exclusion restriction. For instance, more motivated pupils in a school may choose to link with other motivated pupils; or individuals may choose to become friends with other individuals who share a common interest (such as an interest in reading, or mathematics) that is unobserved in the data available to the researcher. In such examples, the absence of a link is due to the unobserved terms of the two agents being correlated in a specific way rather than the absence of correlation between these terms. Solutions to this problem are considered in Subsection 3.7.

3.3 Local Aggregate Model

The local aggregate class of models considers settings where agents' utilities are a function of the aggregate outcomes (or choices) of their neighbours. Such a model applies to situations where there are strategic complementarities or strategic substitutabilities. For example:

[24] Similarly, $\mathbb{E}[\varepsilon_{k,g}|\boldsymbol{G}_g] \neq 0$.

- An individual's costs of engaging in crime may be lower when his neighbours also engage in crime (*e.g.* Bramoullé et al., 2014[25]).
- An agent is more likely to learn about a new product and how it works if more of his neighbours know about it and have used it.

The local aggregate model corresponds empirically to Equation 3.1 with $\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y})=\boldsymbol{GY}$ and $\boldsymbol{w_x}(\boldsymbol{G}, \boldsymbol{X})=\tilde{\boldsymbol{G}}\boldsymbol{X}$, and a scalar social effect parameter, β. This specification can be motivated by the best responses of a game in which nodes have linear-quadratic utility and there are strategic complementarities or substitutabilities between the actions of a node and those of its neighbours. A model of this type has studied by Ballester et al. (2006).[26] In particular, the utility function for node i in network g takes the following form:

$$U_i(y_{i,g}; \boldsymbol{y}_{-i,g}, \boldsymbol{X}_g, \boldsymbol{G}_g) = \left(\pi_{i,g} - \frac{1}{2} y_{i,g} + \beta \sum_{j=1}^{N_g} G_{ij,g} y_{j,g} \right) y_{i,g} \tag{3.12}$$

where $y_{i,g}$ is i's action or choice, and $\pi_{i,g}$ is, as before, an individual heterogeneity parameter.[27] $\pi_{i,g}$ is parameterised as

$$\pi_{i,g} = \boldsymbol{x}_{i,g}\boldsymbol{\delta} + \sum_{j=1}^{n} \tilde{G}_{ij,g} \boldsymbol{x}_{j,g} \boldsymbol{\gamma} + \boldsymbol{z}_g \boldsymbol{\eta} + \nu_g + \varepsilon_{i,g}$$

so that individual heterogeneity is a function of a node's own characteristics, the *average* characteristics of its neighbours, network-level observed characteristics, and some unobserved network- and individual-level terms.

The quadratic cost of own actions means that in the absence of any network, there would be a unique optimal amount of effort the node would exert. $\beta > 0$ implies that neighbours' actions are complementary to a node's own actions, so that the node increases his actions in response to those of his neighbours. If $\beta < 0$, then nodes' actions are substitutes, and the reverse is true. Nodes choose $y_{i,g}$ so as to maximise their utility.

The best response function is:

$$y^*_{i,g}(\boldsymbol{G}_g) = \beta \sum_{j=1}^{n} G_{ij,g} y_{j,g} + \boldsymbol{x}_{i,g}\boldsymbol{\delta} + \sum_{j=1}^{n} \tilde{G}_{ij,g} \boldsymbol{x}_{j,g} \boldsymbol{\gamma} + \boldsymbol{z}_g \boldsymbol{\eta} + \nu_g + \varepsilon_{i,g} \tag{3.13}$$

Ballester et al. (2006) solve for the Nash equilibrium of this game when $\beta > 0$ and show that when $|\beta \omega_{max}(\boldsymbol{G}_g)| < 1$, where $\omega_{max}(\boldsymbol{G}_g)$ is the largest eigenvalue of the matrix $\boldsymbol{G}_g$, the equilibrium is

[25] The games considered in both Bramoulle and Kranton (2007) and Bramoullé et al. (2014) are not strictly linear models, since there are corner solutions at zero.

[26] Ballester et al. (2006) focus on the case where there are strategic complementarities. Bramoullé et al. (2014) study the case where there are strategic substitutabilities and characterise all equilibria of this game.

[27] Notice that $\sum_{j=1}^{N_g} G_{ij,g} y_{j,g} = \boldsymbol{G}_{i,g}\boldsymbol{y}_g$.

unique and the equilibrium outcome relates to a node's Katz-Bonacich centrality, which is defined as $\boldsymbol{b}(\boldsymbol{G}_g, \beta) = (\boldsymbol{I}_g - \beta \boldsymbol{G}_g)^{-1}(\boldsymbol{\iota}_g)$.[28]

Bramoullé et al. (2014) study the game with strategic substitutabilities between the action of a node and those of its neighbours. They characterise the set of Nash equilibria of the game and show that, in general, multiple equilibria will arise. A unique equilibrium exists only when $\beta|\omega_{min}(\boldsymbol{G}_g)| < 1$, where $\omega_{min}(\boldsymbol{G}_g)$ is the lowest eigenvalue of the matrix $\boldsymbol{G}_g$. When there are multiple equilibria possible, they must be accounted for in any empirical analysis. Methods developed in the literature on the econometrics of games may be applied here (Bisin et al., 2011). See de Paula (2013) for an overview.

When a unique equilibrium exists, this theoretical set-up implies the following empirical model (stacking data from multiple networks):

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \beta\boldsymbol{G}\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\gamma} + \tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\delta} + \boldsymbol{Z}\boldsymbol{\eta} + \boldsymbol{L}\boldsymbol{\nu} + \boldsymbol{\varepsilon} \tag{3.14}$$

which corresponds to Equation 3.1 with $\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y})=\boldsymbol{G}\boldsymbol{Y}$ and $\boldsymbol{w_x}(\boldsymbol{G}, \boldsymbol{X})=\tilde{\boldsymbol{G}}\boldsymbol{X}$, and where all other variables and parameters are as defined above in Subsection 3.1.

Identification of Equation 3.14 using observational data has been studied by Calvó-Armengol et al. (2009), Lee and Liu (2010) and Liu et al. (2014b). They proceed under the assumption that $\mathbb{E}[\boldsymbol{\varepsilon}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{G}, \tilde{\boldsymbol{G}}] = 0$ and $\mathbb{E}[\boldsymbol{\nu}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{G}, \tilde{\boldsymbol{G}}] \neq 0$. That is, the node-varying error component is conditionally independent of node- and network-level observables and of the network, while the network-level unobservable could be correlated with node- and network-level characteristics and/or the network itself.

These assumptions imply a two-stage network formation process. First agents select into a network based on a set of observed individual- and network-level characteristics and some common network-level unobservables. Then in a second stage they form links with other nodes. There are no network-level unobservable factors that determine link formation once the network has been selected by the node. Moreover, there are no node-level unobservable factors that determine the choice of network or link formation within the chosen network.

To proceed, we assume that data is available for multiple networks. Then, as in Subsection 3.2, we replace the network-level observables, $\boldsymbol{Z}$, and the network-level unobservable, $\boldsymbol{L}\boldsymbol{\nu}$ in Equation 3.14 with network fixed effects, $\boldsymbol{L}\tilde{\boldsymbol{\nu}}$, where $\tilde{\boldsymbol{\nu}}$ is a $M \times 1$ vector that captures the network fixed effects.

To account for the fixed effect, a global within-transformation is applied, as in Subsection 3.2. This transformation is represented by the block diagonal matrix $\boldsymbol{J}^{glob}$ that stacks the following network-level transformation matrices – $\boldsymbol{J}_g^{glob} = \boldsymbol{I}_g - \frac{1}{N_g}(\boldsymbol{\iota}_g\boldsymbol{\iota}_g')$ – along the leading diagonal, with off-diagonal terms set to 0. Again we suppress the superscript on $\boldsymbol{J}^{glob}$ in the rest of this subsection. The resulting model, analogous to Equation 3.11, is:

[28] A more general definition for Katz-Bonacich centrality is $\boldsymbol{b}(\boldsymbol{G}_g, \beta, a) = (\boldsymbol{I}_g - \beta\boldsymbol{G}_g)^{-1}(a\boldsymbol{G}_g\boldsymbol{\iota}_g)$, where $a > 0$ is a constant (Jackson, 2008).

$$\boldsymbol{JY} = \beta \boldsymbol{JGY} + \boldsymbol{JX}\boldsymbol{\gamma} + \boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\delta} + \boldsymbol{J}\boldsymbol{\varepsilon} \tag{3.15}$$

The model above suffers from the reflection problem, since $\boldsymbol{Y}$ appears on both sides of the equation. However, the parameters of Equation 3.15 can be identified using linear IV if the deterministic part of the right hand side, $[\mathbb{E}(\boldsymbol{JGY}), \boldsymbol{JX}, \boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{X}]$, has full column rank. To see the conditions under which this is satisfied, we examine the term with the endogenous variable, $\mathbb{E}(\boldsymbol{JGY})$. Under the assumption that $|\beta\omega_{max}(\boldsymbol{G}_g)| < 1$, we obtain the following from the reduced form equation of Equation 3.14:

$$\begin{aligned}\mathbb{E}(\boldsymbol{JGY}) = \boldsymbol{J}(\boldsymbol{GX} + \beta\boldsymbol{G}^2\boldsymbol{X} + ...)\boldsymbol{\gamma} + \boldsymbol{J}(\boldsymbol{G}\tilde{\boldsymbol{G}}\boldsymbol{X} + \beta\boldsymbol{G}^2\tilde{\boldsymbol{G}}\boldsymbol{X} + ...)\boldsymbol{\delta} \\ + \boldsymbol{J}(\boldsymbol{GL} + \beta\boldsymbol{G}^2\boldsymbol{L} + ...)\tilde{\boldsymbol{\nu}}\end{aligned} \tag{3.16}$$

We can thus see that if there is variation in node degree within at least one network g (which means that $\boldsymbol{G}_g$ and $\tilde{\boldsymbol{G}}_g$ are linearly independent), and the matrices $\{\boldsymbol{I}, \mathbf{G}, \tilde{\mathbf{G}}, \mathbf{G}\tilde{\mathbf{G}}\}$ are linearly independent with $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$, and $\tilde{\boldsymbol{\nu}}$ each having non-zero terms, the parameters of Equation 3.14 are identified.[29] This is a special case of the Blume et al. (2013) result discussed earlier. Node degree ($\boldsymbol{GL}$), along with the total and average exogenous characteristics of the node's direct neighbours (*i.e.* $\boldsymbol{GX}$ and $\tilde{\boldsymbol{G}}\boldsymbol{X}$) and sum of the average exogenous characteristics of its second-degree neighbours (*i.e.* $\boldsymbol{G}\tilde{\boldsymbol{G}}\boldsymbol{X}$) can be used as instruments for the total outcome of the node's neighbours (i.e. $\boldsymbol{GY}$). The availability of node degree as an instrument can allow one to identify parameters without using the exogenous characteristics, $\boldsymbol{X}$, of second- or higher-degree network neighbours, which could be advantageous in some situations as we will see in Section 5 below.

In terms of practical application, consider using this method to identify whether there are complementarities between the schooling performance of an individual and that of his friends, conditional on how own characteristics (age and gender), the composition of his friends (average age and gender), and some school characteristics. Then, if there are individuals in the same network with different numbers of friends, and the matrices $\{\boldsymbol{I}, \mathbf{G}, \tilde{\mathbf{G}}, \mathbf{G}\tilde{\mathbf{G}}\}$ are linearly independent, the individual's degree, along with the total and average characteristics of his friends (*i.e.* total and average age and gender) and the sum of the average age and gender of the individual's friends of friends can be used as instruments for the sum of the individual's friends' schooling performance.

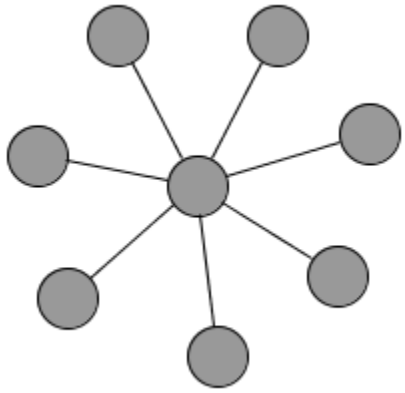
Parameters can still be identified if there no variation in node degree within a network for all networks in the data, but there is variation in degree across networks. In this case, $\boldsymbol{G}_g = \bar{d}_g\tilde{\boldsymbol{G}}_g$ and $[E(\boldsymbol{JGY}), \boldsymbol{JX}, \boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{X}]$ has full column rank if the matrices $\{\boldsymbol{I}, \boldsymbol{G}, \tilde{\boldsymbol{G}}, \boldsymbol{G}\tilde{\boldsymbol{G}}, \tilde{\boldsymbol{G}}^2, \boldsymbol{G}\tilde{\boldsymbol{G}}^2\}$ are linearly

[29] See Liu et al. (2014b) for a different identification condition that allows for some linear dependence among these matrices under additional restrictions.

independent and $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ each have non-zero terms.[30] Finally, when there is no variation in node degree within and across all networks in the data, parameters can be identified using a similar condition as encountered in Subsection 3.3 above: the matrices $\{\boldsymbol{I}, \tilde{\boldsymbol{G}}, \tilde{\boldsymbol{G}}^2, \tilde{\boldsymbol{G}}^3\}$ should be linearly independent.

It is possible to identify model parameters in the local aggregate model in networks where the local average model parameters cannot be identified. For example, in a star network (see Figure 3) there is no pair of agents that has a geodesic distance (*i.e.* shortest path) of 3 or more, so this fails the identification condition for the local average model (see Subsection 3.2 above). However, there is variation in node degree within the network and the matrices $\boldsymbol{I}_g, \mathbf{G}_g, \tilde{\mathbf{G}}_g, \mathbf{G}_g\tilde{\mathbf{G}}_g$ can be shown to be linearly independent, thus satisfying the identification conditions for the local aggregate model.

Figure 3: Star Network



3.4 Hybrid Local Models

The local average and local aggregate models embody distinct mechanisms through which social effects arise. One may be interested in jointly testing these mechanisms, and empirically identifying the most relevant one for a particular context. Liu et al. (2014a) present a framework nesting both the local aggregate and local average models, allowing for this.

The utility function for node i in network g that nests both the (linear) local aggregate and local average models has the following form:

$$U_i(y_{i,g}; \boldsymbol{y}_{-i,g}, \boldsymbol{X}_g, \tilde{\boldsymbol{G}}_{i,g}, \boldsymbol{G}_{i,g}) = \left(\pi_{i,g} + \beta_1 \sum_{j=1}^{N_g} G_{ij,g} y_{j,g} - \frac{1}{2} \left(y_{i,g} - 2\beta_2 \sum_{j=1}^{N_g} \tilde{G}_{ij,g} y_{j,g} \right) \right) y_{i,g} \quad (3.17)$$

[30] See Liu et al. (2014b) for a different identification condition that allows for some linear dependence among these matrices under additional restrictions.

where $\pi_{i,g}$ is node-specific observed heterogeneity, which affects the node's marginal return from the chosen outcome level $y_{i,g}$. A node's utility is thus affected by the choices of its neighbours through changing the marginal returns of its own choice (*e.g.* in a schooling context, an individual's studying effort is more productive if his friends also study), as in the local aggregate model, and by a cost of deviating from the average choice of its neighbours (*i.e.* individuals face a utility cost if they study when their friends don't study), as in the local average model.

The best reply function for a node i nests both the local average and local aggregate terms. Liu et al. (2014a) prove that under the condition that $\beta_1 \geq 0$, $\beta_2 \geq 0$ and $d_g^{max}\beta_1 + \beta_2 < 1$, where d_g^{max} is the largest degree in network g, the simultaneous move game has a unique interior Nash equilibrium in pure strategies.

The econometric model, assuming that the node-specific observed heterogeneity parameter takes the form $\pi_{i,g} = \boldsymbol{x}_{i,g}\boldsymbol{\gamma} + \sum_{j=1}^{N_g} \tilde{G}_{ij,g}\boldsymbol{x}_{j,g}\boldsymbol{\delta} + \boldsymbol{z}_g\eta_g + \nu_g + \varepsilon_{i,g}$, is as follows:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \beta_1\boldsymbol{GY} + \beta_2\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{X\gamma} + \tilde{\boldsymbol{G}}\boldsymbol{X\delta} + \boldsymbol{Z\eta} + \boldsymbol{L\nu} + \boldsymbol{\varepsilon} \tag{3.18}$$

using the same notation as before (see *e.g.* Subsection 3.1).

With data from only a single network it is not possible to separately identify β_1 and β_2 and hence test between the local aggregate and local average models (or indeed find that the truth is a hybrid of the two effects). Identification of parameters is considered when data from multiple networks are available under the assumption that $\mathbb{E}[\varepsilon_{i,g}|\boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}_g, \tilde{\boldsymbol{G}}_g] = 0$ and $\mathbb{E}[\boldsymbol{\nu}_g|\boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}_g, \tilde{\boldsymbol{G}}_g] \neq 0$. Thus, as in Subsections 3.2 and 3.3 above, the individual error term, $\varepsilon_{i,g}$ is assumed to be mean independent of node- and network-level observable characteristics and the network. The network-level unobservable, $\boldsymbol{\nu}_g$, by contrast is allowed to be correlated with node- and network-level characteristics and/or the network.

To proceed, as in the local average and local aggregate model, $\boldsymbol{Z\eta}$ and $\boldsymbol{L\nu}$ are replaced by a network-level fixed effect, $\boldsymbol{L\tilde{\nu}}$, which is then removed using the global within-transformation, $\boldsymbol{J}^{glob}$. Again, we suppress the superscript on $\boldsymbol{J}^{glob}$. The resulting transformed network model is:

$$\boldsymbol{JY} = \beta_1\boldsymbol{JGY} + \beta_2\boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{JX\gamma} + \boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{X\delta} + \boldsymbol{J\varepsilon} \tag{3.19}$$

When there is variation in the degree within a network g, then the reduced form equation of Equation 3.19 implies that $\boldsymbol{JG}(\boldsymbol{I} - \beta_1\boldsymbol{G} - \beta_2\tilde{\boldsymbol{G}})^{-1}\boldsymbol{L}$ can be used as an instrument for the local aggregate term $\boldsymbol{JGY}$ and $\boldsymbol{J}\tilde{\boldsymbol{G}}(\boldsymbol{I} - \beta_1\boldsymbol{G} - \beta_2\tilde{\boldsymbol{G}})^{-1}\boldsymbol{L}$ can be used as an instrument for the local average term $\boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{Y}$. The model parameters may thus be identified even if there are no node-level exogenous characteristics, $\boldsymbol{X}$, in the model. Caution must be taken though when the model contains no exogenous characteristics, $\boldsymbol{X}$, since, in this case, the model may be only tautologically identified if $\beta_1 = 0$ (Angrist, 2013). The availability of such characteristics offers more possible IVs: in particular, the total and average exogenous characteristics of direct and indirect neighbours can

be used as instruments. These are necessary for identification when all nodes within a network have the same degree, though average degree may vary across networks. In this case, parameters can be identified if the matrices $\{\boldsymbol{I}, \mathbf{G}, \tilde{\mathbf{G}}, \mathbf{G}\tilde{\mathbf{G}}, \tilde{\boldsymbol{G}}^2, \boldsymbol{G}\tilde{\boldsymbol{G}}^2, \tilde{\boldsymbol{G}}^3\}$ are linearly independent. If, however, all nodes in all networks have the same degree, it is not possible to identify separately the parameters β_1 and β_2.

This specification nests both the local average and local aggregate models, so a J-test for non-nested regression models can be applied to uncover the relevance of each mechanism. The intuition underlying the J-test is as follows: if a model is correctly specified (in terms of the set of regressors), then the fitted value of an alternative model should have no additional explanatory power in the original model, *i.e.* its coefficient should not be significantly different from zero. Thus, to identify which of the local average or local aggregate mechanisms is more relevant for a specific outcome, one could first estimate one of the models (*e.g.* the local average model), and obtain the predicted outcome value under this mechanism. In a second step, estimate the other model (in our example, the local aggregate model), and include as a regressor the predicted value from the other (*i.e.* local average) model. If the mechanism underlying the local average model is also relevant for the outcome, the coefficient on the predicted value will be statistically different from 0. The converse can also be done to test the relevance of the second model (the local aggregate model in our case). See Liu et al. (2014a) for more details.

3.5 Models with Network Characteristics

The models considered thus far allow for a node's outcomes to be influenced only by outcomes of its neighbours. However, the broader network structure may affect node- and aggregate network-outcomes through more general functionals or features of the network. Depending on the theoretical model used, there are different predictions on which network features relate to different outcomes of interest. For example, the DeGroot (1974) model of social learning implies that a node's eigenvector centrality, which measures its 'importance' in the network by how important its neighbours are, determines how influential it is in affecting the beliefs of other nodes.

Empirical testing and verification of the predictions of these theoretical models has greatly lagged the theoretical literature due to a lack of datasets with both information on network structure and socio-economic outcomes of interest. The recent availability of detailed network data from many contexts has begun to relax this constraint.

The following types of specification are typically estimated when assessing how outcomes vary with network structure, for node-level outcomes:

$$\boldsymbol{Y} = \boldsymbol{f_y}(\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y}), \boldsymbol{X}, \boldsymbol{w_x}(\boldsymbol{G}, \boldsymbol{X}), \boldsymbol{Z}) + \boldsymbol{\varepsilon} \tag{3.20}$$

and network-level outcomes:

$$\bar{\boldsymbol{y}} = \boldsymbol{f}_{\bar{\boldsymbol{y}}}(\bar{\boldsymbol{w}}_{\bar{\boldsymbol{y}}}(\boldsymbol{G}), \bar{\boldsymbol{X}}, \bar{\boldsymbol{w}}_{\bar{\boldsymbol{x}}}(\boldsymbol{G}, \bar{\boldsymbol{X}})) + \boldsymbol{u} \tag{3.21}$$

$\boldsymbol{f}_{\boldsymbol{y}}(.)$ and $\boldsymbol{f}_{\bar{\boldsymbol{y}}}(.)$ are functions that specify the shape of the relationship between the network statistics and the node- and network-level outcomes. When $\boldsymbol{f}_{\boldsymbol{y}}(.)$ is simply a linear index in its argument, Equation 3.22 remains nested in Equation 3.1. Though, in principle, the shape of $\boldsymbol{f}_{\boldsymbol{y}}(.)$ should be guided by theory (where possible), through the rest of this Subsection, we take $\boldsymbol{f}_{\boldsymbol{y}}(.)$ to be a linear index in its argument. $\boldsymbol{w}_{\boldsymbol{y}}(\boldsymbol{G}, \boldsymbol{Y})$ includes R network statistics that vary at the node- or network-level and that may be interacted with $\boldsymbol{Y}$[31] while $\bar{\boldsymbol{w}}_{\bar{\boldsymbol{y}}}(\boldsymbol{G})$ contains the $\bar{R}$ network statistics in the network-level regression. $\boldsymbol{X}$ is a matrix of observable characteristics of nodes, $\boldsymbol{w}_{\boldsymbol{x}}(\boldsymbol{G}, \boldsymbol{X})$ interacts network statistics with exogenous characteristics of nodes, and $\boldsymbol{Z}$ and $\bar{\boldsymbol{X}}$ are network-level observable characteristics. $\bar{\boldsymbol{w}}_{\bar{\boldsymbol{X}}}(\boldsymbol{G}, \bar{\boldsymbol{X}})$ interacts network statistics with network-level observable characteristics.

The complexity of networks poses an important challenge in understanding how outcomes vary with network structure. In particular, there are no sufficient statistics that fully describe the structure of a network. For example, networks with the same average degree may vary greatly on dimensions such as density, clustering and average path length among others. Moreover, the adjacency matrix, $\boldsymbol{G}$, which describes fully the structure of a network, is too high-dimensional an object to include directly in tests of the influence of broader features of network structure. Theory can provide guidance on which statistics are likely to be relevant, and also on the shape of the relationship between the network statistic and the outcome of interest. A limitation though is that theoretical results may not be available (given currently known techniques) for outcomes one is interested in studying. This is a challenge faced by, for instance Alatas et al. (2012) who study how network structure affects information aggregation.

Below we outline methods that have been applied to analyse the effects of features of network structure on socio-economic outcomes. We do so separately for node-level specifications and network-level specifications. This literature is very much in its infancy and few methods have been developed to allow for identification of causal parameters.

3.5.1 Node-Level Specifications

Many theoretical models predict how node-level outcomes vary with the 'position' of a node in the network, captured by node varying network statistics such as centrality; or with features of the node's local neighbourhood such as node clustering; or with the 'connectivity' of the network, represented by statistics that vary at the network-level such as network density.

A common type of empirical specification used in the literature correlates network statistics with some relevant socio-economic outcome of interest. This approach is taken by, for example, Jackson

[31] The term $\boldsymbol{w}_{\boldsymbol{y}}(\boldsymbol{G}, \boldsymbol{Y})$ will be endogenous when network statistics are interacted with $\boldsymbol{Y}$.

et al. (2012) who test whether informal favours take place across edges that are supported (*i.e.* that nodes exchanging a favour have a common neighbour), which is the prediction of their theoretical model.

This corresponds with $\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y})$ in Equation 3.20 above being defined as $\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y}) = \boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is an $(\sum_{g=1}^{M} N_g \times R)$ matrix stacking $\boldsymbol{\omega}_{i,g}$, the $(1 \times R)$ node-level vector of network statistics of interest for all nodes in all networks, and $\boldsymbol{w_x}(.)$ being defined as $\boldsymbol{\iota}$. Here, $\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y})$ is defined to be a function of the network only.

When $\boldsymbol{f_y}(.)$ is linear, the specification is as follows:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \boldsymbol{\omega}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\gamma} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{3.22}$$

where the variables and parameters are as defined above and the parameter of interest is $\boldsymbol{\beta}$. Defining $\boldsymbol{W} = (\boldsymbol{\omega}, \boldsymbol{X}, \boldsymbol{Z})$, the key identification assumption is that $\mathrm{E}[\boldsymbol{\varepsilon}'\boldsymbol{W}] = 0$, that is that the right hand side terms are uncorrelated with the error term. This may not be satisfied if there are unobserved factors that affect both the network statistic (through affecting network formation decisions) and the outcome, $\boldsymbol{Y}$ or if the network statistic is mismeasured. Both of these are important concerns that we cover in detail in Sections 4 and 5 below.

In some cases, one may also be interested in estimating a model where an agent's outcome is affected by the outcomes of his neighbours, weighted by a measure of their network position. For example, in the context of learning about a new product or technology, the DeGroot (1974) model of social learning implies that nodes' eigenvector centrality determines how influential they are in influencing others' behaviour. Thus, conditional on the node's eigenvector centrality, its choices may be influenced more by the choices of his neighbours with high eigenvector centrality. Thus, one may want to weight the influence of neighbours' outcomes on own outcomes by their eigenvector centrality, conditional on own eigenvector centrality. This implies a model of the following form:

$$\begin{aligned} \boldsymbol{Y} = \alpha\boldsymbol{\iota} + \boldsymbol{w_y}(\boldsymbol{G},\, \boldsymbol{Y})\boldsymbol{\beta} + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\gamma}} + \boldsymbol{w_x}(\boldsymbol{G},\, \tilde{\boldsymbol{X}})\tilde{\boldsymbol{\delta}} \\ + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{L}\boldsymbol{\nu} + \boldsymbol{\varepsilon} \end{aligned} \tag{3.23}$$

$\boldsymbol{w_y}(\boldsymbol{G},\, \boldsymbol{Y})$ is an $\sum_g N_g \times R$ matrix, with the $(i,r)^{th}$ element being the weighted sum of i's neighbours' outcomes, $\sum_{j \neq i} G_{ij,g} y_{j,g} \omega^r_{j,g}$ or $\sum_{j \neq i} \tilde{G}_{ij,g} y_{j,g} \omega^r_{j,g}$, with weights $\omega^r_{j,g}$ being the neighbour's r^{th} network statistic. $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{X}}'_1, \tilde{\boldsymbol{X}}'_2, ..., \tilde{\boldsymbol{X}}'_M)'$, where $\tilde{\boldsymbol{X}}_g = (\boldsymbol{X}_g, \boldsymbol{\omega}_g)$ is a matrix stacking together the network-level matrices of exogenous explanatory variables and network statistics of interest. $\boldsymbol{w_x}(\boldsymbol{G},\, \tilde{\boldsymbol{X}})$ could be defined as $\boldsymbol{G}\tilde{\boldsymbol{X}}$ or $\tilde{\boldsymbol{G}}\tilde{\boldsymbol{X}}$. Identification of parameters in this case is complicated by the fact that $\boldsymbol{w_y}(\boldsymbol{G},\, \boldsymbol{Y})$ is a (possibly non-linear) function of $\boldsymbol{Y}$, and thus endogenous. It may be possible to achieve identification using network-based instrumental variables, as done in

Subsections 3.2, 3.3 and 3.4 above, though it is not immediately obvious how such an IV could be constructed. Future research is needed to shed light on these issues.

3.5.2 Network-level Specifications

Aggregate network-level outcomes, such as the degree of risk sharing or the aggregate penetration of a new product, may also be affected by how 'connected' the network is, or the 'position' of nodes that experience a shock or who first hear about a new product.

Empirical tests of the relationship between aggregate network-level outcomes and network statistics involves estimating specifications such as Equation 3.21, where the shape of the function $\boldsymbol{f}_{\bar{\boldsymbol{y}}}(.)$ and the choice of statistics in $\bar{\boldsymbol{w}}_{\bar{\boldsymbol{y}}}(\boldsymbol{G}) = \bar{\boldsymbol{\omega}}$, where $\bar{\boldsymbol{\omega}}$ is an $(M \times \bar{R})$ matrix of network statistics, are, ideally, motivated by theory. With linear $\boldsymbol{f}_{\bar{\boldsymbol{y}}}(.)$, this implies the following equation:

$$\bar{\boldsymbol{y}} = \phi_0 + \bar{\boldsymbol{\omega}}\boldsymbol{\phi}_1 + \bar{\boldsymbol{X}}\boldsymbol{\phi}_2 + \bar{\boldsymbol{w}}_{\bar{\boldsymbol{X}}}(\boldsymbol{G}, \bar{\boldsymbol{X}})\boldsymbol{\phi}_3 + \boldsymbol{u} \tag{3.24}$$

where the variables are as defined after Equation 3.21. The parameter of interest is typically $\boldsymbol{\phi}_1$. Defining $\bar{\boldsymbol{W}} = (\boldsymbol{\omega}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{w}}_{\bar{\boldsymbol{X}}}(\boldsymbol{G}, \bar{\boldsymbol{X}}))$, the key identification assumption is that $\mathrm{E}[\boldsymbol{u}\bar{\boldsymbol{W}}] = 0$, 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. Recent empirical work, such as that by Banerjee et al. (2013), has used quasi-experimental variation to try and alleviate some of the challenges posed by the former issue in identifying the parameter $\boldsymbol{\phi}_1$.

Since this specification uses data at the network-level, estimation will require a large sample of networks in order to recover precise estimates of the parameters, even in the absence of endogeneity from network formation and mismeasurement of the network. This is a problem in practice, since as we will see below in Section 5.3, the difficulties and costs involved in collecting network data often mean that in practice researchers have data for a small number of networks only.

3.6 Experimental Variation

Subsections 3.2 to 3.5 above considered the identification of the social effect parameters using observational data. In this section, we consider identification of these parameters using experimental data. We focus on the case where a policy is assigned randomly to a sub-set of nodes in a network. Throughout we assume that the network is pre-determined and unchanged by the exogenously assigned policy.[32]

We focus the discussion on identifying parameters of the local average model specified in Subsection 3.2 above. The issues related to using experimental variation to uncover the parameters of the local

[32]This assumption is not innocuous. Comola and Prina (2014) provide an example where the policy intervention does change the network.

aggregate model are similar. As outlined above, this model implies that a node's outcome is affected by the average outcome of its network neighbours, its own and network-level exogenous characteristics (which may be subsumed into a network fixed effect), and the average characteristics of its network neighbours. We are typically interested in parameters β, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ in the following equation:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \beta\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\gamma} + \tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\delta} + \boldsymbol{L}\tilde{\boldsymbol{\nu}} + \boldsymbol{\varepsilon} \tag{3.25}$$

where the variables are as defined above.

Throughout this section, we assume that the policy shifts outcomes for the nodes that directly receive the policy.[33] To proceed further, we first assume that a node that does not receive the policy (*i.e.* is untreated, to use the terminology from the policy evaluation literature), is only affected by the policy through its effects on the outcomes of the node's network neighbours. This implies the following model for the outcome $\boldsymbol{Y}$:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \beta\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\gamma} + \tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\delta} + \rho\boldsymbol{t} + \boldsymbol{L}\tilde{\boldsymbol{\nu}} + \boldsymbol{\varepsilon} \tag{3.26}$$

where $\boldsymbol{t}$ is the treatment vector, and ρ is the direct effect of treatment. We assume that $\mathbb{E}[\boldsymbol{\varepsilon}|\boldsymbol{X}, \boldsymbol{Z}, \tilde{\boldsymbol{G}}, \boldsymbol{t}] = 0$. Moreover, random allocation of the treatment implies that $\boldsymbol{t} \perp\!\!\!\perp \boldsymbol{X}, \boldsymbol{Z}, \tilde{\boldsymbol{G}}, \boldsymbol{\varepsilon}$.

Applying the same within-transformation as in Subsection 3.2 above to account for the network-level fixed effect leads to the following specification:

$$\boldsymbol{J}\boldsymbol{Y} = \alpha\boldsymbol{J}\boldsymbol{\iota} + \beta\boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{J}\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\delta} + \rho\boldsymbol{J}\boldsymbol{t} + \boldsymbol{J}\boldsymbol{\varepsilon} \tag{3.27}$$

We can use instrumental variables to identify β as long as the deterministic part of the right hand side of Equation 3.27, $[\mathrm{E}(\boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{Y}), \boldsymbol{J}\boldsymbol{X}, \boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{X}]$ has full column rank. $\boldsymbol{J}\boldsymbol{X}$ and $\boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{X}$ can be used as instruments for themselves. We thus need an instrument for $\mathbb{E}[\boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{Y}]$. We use the following expression for $\boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{Y}$, derived from the reduced form of Equation 3.26 under the assumption that $|\beta| < 1$, to construct instruments:

$$\begin{aligned}\mathbb{E}[\boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{Y}] = \boldsymbol{J}\tilde{\boldsymbol{G}}\sum_{s=0}^{\infty}\beta^s\tilde{\boldsymbol{G}}^s\alpha\boldsymbol{\iota} + \boldsymbol{J}(\tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\gamma} + \beta\tilde{\boldsymbol{G}}^2\boldsymbol{X}\boldsymbol{\gamma} + ...) + \boldsymbol{J}(\tilde{\boldsymbol{G}}^2\boldsymbol{X}\boldsymbol{\delta} + \beta\tilde{\boldsymbol{G}}^3\boldsymbol{X}\boldsymbol{\delta} + ...) \\ + \boldsymbol{J}(\rho\tilde{\boldsymbol{G}}\boldsymbol{t} + \beta\rho\tilde{\boldsymbol{G}}^2 t + ...)\end{aligned} \tag{3.28}$$

From this equation, we can see that $\tilde{\boldsymbol{G}}\boldsymbol{t}$, the average treatment status of a node's network neighbours, does not appear in Equation 3.26. It can thus be used as an instrument for $\tilde{\boldsymbol{G}}\boldsymbol{Y}$, either

[33] Below, we will consider identification conditions in the case where a node may be affected by the treatment status of his network neighbours even if their outcomes do not shift in response to the treatment.

in addition to, or as an alternative to $\tilde{\boldsymbol{G}}^2\boldsymbol{X}$ and $\tilde{\boldsymbol{G}}^3\boldsymbol{X}$, the average characteristics of the node's second- and third-degree neighbours. Thus, the policy could be used to identify the model parameters, albeit under a strong assumption on who it affects.[34]

In many cases, however, the assumption that the policy affects a node's outcome only if it is directly treated may be too strong. The treatment status of a node's neighbours could affect its outcome even when the neighbours' outcomes do not shift in response to receiving the policy. An example of such a case, studied by Banerjee et al. (2013), is when the treatment involves providing individuals with information on a new product, and the outcome of interest is the take-up of the product. Then neighbours' treatment status could affect the individual's own adoption decision by (1) shifting his neighbours' decision (endorsement effects), and also (2) through neighbours passing on information about the product and letting the individual know of its existence (diffusion effect).[35] In this case, a more appropriate model would be as follows:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \beta\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\gamma} + \tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\delta} + \rho\boldsymbol{t} + \tilde{\boldsymbol{G}}\boldsymbol{t}\mu + \boldsymbol{\varepsilon} \tag{3.29}$$

where ρ captures the direct treatment effect, *i.e.* the effect of a node itself being treated, and μ is the direct effect of the average treatment status of social contacts. This highlights the limits to using exogenous variation from randomised experiments to identify social effect parameters. We might want to use the exogenous variation in the average treatment allocation of a node's neighbours, $\tilde{\boldsymbol{G}}\boldsymbol{t}$, as an instrument for neighbours' outcomes, $\tilde{\boldsymbol{G}}\boldsymbol{Y}$. However, this will identify β only under the assumption that $\mu = 0$, *i.e.* there is no direct effect of neighbours' treatment status. This rules out economic effects such as the diffusion effect.

We can still make use of the treatment effect for identification, by using the average treatment status of a node's second-degree (and higher-degree) neighbours, $\tilde{\boldsymbol{G}}^2\boldsymbol{t}$, as instruments for the average outcome of his neighbours ($\tilde{\boldsymbol{G}}\boldsymbol{Y}$). This is the same identification result as discussed earlier, from Bramoullé et al. (2009), and simply treats $\tilde{\boldsymbol{G}}^2\boldsymbol{t}$ in the same way the other covariates of second-degree neighbours, $\tilde{\boldsymbol{G}}^2\boldsymbol{X}$. Such instruments rely not only on variation in treatment status, but also on the network structure, with identification not possible for certain network structures as we saw in Subsection 3.2.[36]

Thus far, we have discussed how exogenous variation arising from the random assignment of a policy can be used to identify the social effect associated with a specific model – the local average model – which, as we saw, arises from an economic model where agents conform to their peers. In empirical work, though, it is common for researchers to directly include the average treatment

[34]Similar results can be shown for the local aggregate model when $|\beta\omega_{max}(\boldsymbol{G})| < 1$. However, as shown above, node degree can also be used as an additional instrument in this model.

[35]The study of how to use these effects to maximise the number of people who adopt relates closely to study of the 'key player' in work by Ballester et al. (2006) and Liu et al. (2014b).

[36]Note that instruments based on random treatment allocation and network structure (*e.g.* $\tilde{\boldsymbol{G}}\boldsymbol{t}$ and $\tilde{\boldsymbol{G}}^2\boldsymbol{t}$) may be more plausible than those based on the exogenous characteristics, $\boldsymbol{X}$, and the network structure (*e.g.* $\tilde{\boldsymbol{G}}^2\boldsymbol{X}$), since $\boldsymbol{t}$ has been randomly allocated, whereas $\boldsymbol{X}$ need not be.

status of network neighbours, rather than their average outcome, as a regressor in the model. In other words, the following type of specification is usually estimated:

$$\boldsymbol{Y} = b_1\boldsymbol{\iota} + b_2\tilde{\boldsymbol{G}}\boldsymbol{t} + \boldsymbol{X}\boldsymbol{b}_3 + \tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{b}_4 + b_5\boldsymbol{t} + \boldsymbol{u} \tag{3.30}$$

A non-zero value for b_2 is taken to indicate the presence of some social effect. However, without further modelling, it is not possible to shed light on the exact mechanism underlying this social effect, or the value of some 'deep' structural parameter.

3.7 Identification of Social Effects with Endogenous Links

In the previous subsections we focused on the identification of social effects under the assumption that the edges along which the effects are transmitted are exogenous. By exogenous we mean that the probability that agent i forms an edge with agent j is mean independent of any unobservables that might influence the outcome of interest for any individual in our social effects model. Formally, we assumed $\mathbb{E}[\boldsymbol{\varepsilon}|\boldsymbol{X}, \boldsymbol{Z}, \tilde{\boldsymbol{G}}] = 0$.[37]

However, in many contexts this may not be hold. 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 practice. We might see that more connected farmers are more likely to take up the practice. However, without further analysis we cannot necessarily interpret this as being *caused* by the network.

One possibility is that there is some underlying correlation in the unobservables of the outcome and connection equations. More risk-loving people, who might be more likely to take up new farming practices, may also be more sociable, and thus have more connections. The endogeneity problem here comes from not being able to hold constant risk-preferences. Hence the coefficient on the network measures is not independent of this unobserved variable. This problem could be solved if we could find an instrument: something correlated with network connections that is unrelated to risk-preferences.

Another possibility is that connections were formed explicitly because of their relationship with the outcome. If agents care about their outcome $y_{i,g}$, and if the network has some impact on $y_{i,g}$, then they have incentives to be strategic in choosing the links in which they are involved. Suppose agents' utility (or profit) varies with $y_{i,g}$, but that some agents have a higher marginal utility from increases in $y_{i,g}$. Agents have incentives to manipulate the parts of the network they are involved in *i.e.* the elements of the i^{th} row and i^{th} columns of $\boldsymbol{G}_g$ – $\{\boldsymbol{G}_{i,g}, \boldsymbol{G}'_{i,g}\}$ – to try to maximise $y_{i,g}$. Moreover, if links are costly, but there is heterogeneity in the agents' valuations of $y_{i,g}$, then agents who value $y_{i,g}$ most should form more costly links, and have higher $y_{i,g}$, but the network is a consequence and not a cause of the individual value for $y_{i,g}$.

[37] Goldsmith-Pinkham and Imbens (2013) suggest a test for endogeneity.

Returning to the farming example, some agents may have a greater preference for taking up new technologies. If talking to others is costly, but can help in understanding the new techniques, these farmers will form more connections. Now the unobservable factors which influence the outcome – preference for take up – will be be correlated with the number of connections. Unlike the previous case, this time we cannot find an 'instrumental' solution: it is the same unobservable driving both y_i and $\boldsymbol{G}_i$.

To overcome this issue experimentally one would need to be able to assign links in the network. However, with the exception of rare examples (including one below), this is difficult to achieve in practice. Additionally there can be external validity issues, as knowing the effect that randomly assigned networks have may not be informative about what effect non-randomly assigned networks have. Alternatively, one can randomly assign treatment status, as discussed in Section 3.6. [38]

Carrell et al. (2013) provide a cautionary example of the importance of considering network formation when using estimated social effects to inform policy reform. Carrell et al. (2009) use data from the US Air Force Academy, where students are randomly assigned to classrooms. They estimate a non-linear model of peer effects, implicitly assuming that conditional on classroom assignment friendship formation is exogenous. They find large and significant peer effects in maths and English test scores, and some non-linearity in these effects. Carrell et al. (2013) use these estimated effects to 'optimally assign' a random sample of students to classrooms, with the intention of maximising the achievement of lower ability students. However, test performance in the 'optimally assigned' classrooms is worse than in the randomly assigned classrooms. They suggest that this finding comes from not taking into account the structure of the linkages between individuals within classrooms.[39]

3.7.1 Instrumental Variables

In the first example above, the outcome y was determined by an equation of the form of Equation 3.1, where the network $\boldsymbol{G}$ was determined potentially by some of the observables already in Equation 3.1 and also the unobservables $\boldsymbol{u}$, and $\mathbb{E}[\boldsymbol{\varepsilon}|\boldsymbol{X}, \boldsymbol{Z}, \tilde{\boldsymbol{G}}] \neq 0$. The failure of the mean independence assumption prevents us from identifying the parameters of Equation 3.1 in the ways suggested previously.

If our interest is in identifying only those parameters, one (potential) solution to the problem is to randomly assign the network structure. However, this is typically prohibitively difficult to enforce

[38] However, when the network is allowed to be endogenous, one needs to make (implicit) assumptions on the network formation process in order to obtain causal estimates. For example, if we assume that the network formation process is such that nodes with similar observed and unobserved characteristics hold similar positions in the resulting network, we can obtain causal estimates if we compare outcomes of nodes with similar network characteristics and different levels of indirect treatment exposure – i.e. exposure to the treatment through their neighbours. See Manski (2013) for more discussion on these issues.

[39] **?** have a different interpretation of this result. They suggest 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.

in real world settings. It is also unlikely to be representative of the edges people actually choose (see for example Carrell et al., 2013).[40]

Alternatively we can attempt to overcome the endogeneity of the network by taking an instrumental variables (IV) approach and finding an exclusion restriction. Here one needs to have a covariate that affects the structure of the network in a way relevant to the outcome equation – something which changes $\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y})$ – but is excluded from the outcome equation itself. For example, if the outcome equation has only in-degree as a network covariate, then one needs to find a covariate that is correlated with in-degree but not the outcome. If instead the outcome equation included some other network covariate, for example Bonacich centrality, a different variable might be appropriate as an instrument.

Mihaly (2009) takes this approach. In trying to uncover the effect of popularity – measured in various ways[41] – on the educational outcomes of adolescents in the US, she uses an interaction between individual and school characteristics as an instrument for popularity. This is a valid instrument if the composition of the school has no direct effect on educational attainment (something which the education literature suggests is unlikely), but does affect all of the measures of popularity.

As ever with instrumental variables, the effectiveness of this approach relies on having a good instrument: something which has strong predictive power for the network covariate but does not enter the outcome equation directly. As noted earlier, if individuals care about the outcome of interest, they will have incentives to manipulate the network covariate. Hence such a variable 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) consider the role of 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 means 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 provide an independent source of variation in edge formation probabilities.[42]

It is important to note that this type of solution can only be employed when the underlying network formation model has a unique equilibrium. Uniqueness requires 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, IV solutions

[40] In the models discussed this means we might observe outcomes that wouldn't be seen without manipulation, because we have changed the support of $\boldsymbol{G}$. In interpreting these results in the context of unmanipulated data we need to be cautious, since we are relying heavily on the functional form assumptions as extrapolate outside the support of what we observe.

[41] She uses four definitions of popularity: in-degree, network density (which only varies between networks), eigenvector centrality, and Bonacich centrality.

[42] An application of this idea is provided by 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.

cannot be used. We discuss further in Section 4 issues of uniqueness in network formation models, and how one might estimate the formation equation in these circumstances.

One should also be aware, when interpreting the results, that if there is heterogeneity in β then this approach delivers a local average treatment effect (LATE). This is a particular weighted average of the individual-specific β's, putting more weight on those for whom the instrument (in our example, school composition) creates most variation in the network characteristic. Hence if the people whose friendship decisions are most affected by school characteristics are also those who, perhaps, are most affected by their friends' outcomes, then the estimated social effect will be higher than the average social effect across all individuals.

3.7.2 Jointly model formation and social effects

In our second example at the beginning of Subsection 3.7 we considered the case where the outcome y was determined by an equation of the form of Equation 3.1, and the network $\boldsymbol{G}$ was strategically chosen to maximise the (unobserved) individual return from this outcome, subject to unobserved costs of forming links. Here the endogeneity comes from $\boldsymbol{G}$ being a function of u. If there is heterogeneity in the costs of forming links, these costs might be useful as instruments, if observed.[43] Without this we must take an alternative approach.

Rather than treating the endogeneity of the network as a problem, jointly modelling $\boldsymbol{G}$ and y uses the observed choices over links to provide additional information about the unobservables which enter the outcome equation. Rather than looking for a variable that can help explain the endogenous covariate but is excluded from the outcome, we now model an explicit economic relationship, and rely on the imposed model to provide identification. Such an approach is taken, for example, by Badev (2013), Blume et al. (2013), Hsieh and Lee (forthcoming), and Goldsmith-Pinkham and Imbens (2013).

Typically the process is modelled as a two-stage game,[44] where agents first form a network and then make outcome decisions. Agents are foresighted enough to see the effect of their network decisions on their later outcome decisions. Consequently they solve the decision process by backward induction, first determining actions for each possible network, and then choosing network links with knowledge of what this implies for outcomes. For this approach to work one needs to be able to characterise the payoff of each possible network, so as to account for agents' network formation incentives in a tractable way.

There are two main limitations for this approach. First, by avoiding the use of exclusion restrictions, the role of functional form assumptions in providing identification becomes critical. Since theory

[43] However, even this will depend on the timing of decisions. See Blume et al. (2013) for details on when such an argument might not hold.

[44] Of the papers mentioned above, Badev (2013) models the choice of friendships and actions simultaneously, whilst the others assume a two-stage process.

rarely specifies precise functional forms, it is not unreasonable to worry about the robustness of results based on assumptions that are often due more to convenience than conviction.

Second, we typically need to impose limits on the form of the network formation model that mean the model is unable to generate many of the features of observed networks, such as the relatively high degree of clustering and low diameter. Particularly restrictive, and discussed further in Section 4, is the restriction that links are formed conditionally independently.

3.7.3 Changes in network structure

An alternative approach to those suggested above relies on *changes* in network structure to provide exogenous variation. In some circumstances one might believe that particular nodes or edges are removed from the network for exogenous reasons (this is sometimes described as 'node/edge failure'). For example, Patnam (2013) considers a network of interlocking company board memberships in India. A pair of firms is considered to be linked if the firms have a common board member. Occasionally edges between companies are severed due to the death of a board member, and to the extent that this is unpredictable, it provides plausibly exogenous variation in the network structure. One can then see how outcomes change as the network changes, and this gives a local estimate of the effect of the network on the outcome of interest. A similar idea is used by Waldinger (2010, 2012) using the Nazi expulsion of Jewish scientists to provide exogenous changes in academic department membership.

The difficulty with this approach in general is finding something that exogenously changes the network, but to which agents do not choose to respond.[45] Non-response includes both not adjusting edges in response to the changes that occur, and not *ex ante* choosing edges strategically to insure against the probabilistic exogenous edge destruction process. In the examples above these relate to not taking into account a board member's probability of death when hiring (*e.g.* not considering age when recruiting), and not hiring new scientists to replace the expelled ones.

4 Network Formation

Network formation is commonly defined as the process of edge formation between a fixed set of nodes. Although, in principle, one could also consider varying the nodes, in most applications the set of nodes will be well-defined and fixed. The empirical study and analysis of this process is important for three reasons.

First, the analysis in most of the previous section described how one might estimate social effects under the critical assumption that the networks of connections were themselves exogenous, or exogenous conditional on observed variables. In many circumstances, such as those described in

[45] It is important to note that one also needs access to a panel of data for the network, which is not often available.

Subsection 3.7, one might think that economic agents are able to make some choice over the connections they form, and that if their connections influence their outcomes they might be somewhat strategic in which edges they choose to form. In this case the social effects estimated earlier will be contaminated by correlations between an individual's observed covariates and the unobserved covariates of his friends. This is in addition to the problems of correlations in group-level unobservables that is well-known in the peer effects literature. For example, someone with a pre-disposition towards smoking is likely to choose to form friendships with others who might also enjoy smoking. An observed correlation in smoking decision, even once environmental characteristics are controlled for, might then come from the choice of friends, rather than any social influence. One solution to this problem, is to use a two-step procedure, in which a predicted network is estimated as a first stage. This predicted network is then used in place of the observed network in the second stage. This approach is taken by König et al. (2014).[46] Again the first stage will require estimation of a network formation process.

Second, an important issue when working with network data is that of measurement error. We return to this more fully in the next section, but where networks are incompletely observed, direct construction of network statistics using the sampled data typically introduces non-classical measurement error in these network statistics. If these statistics are used as covariates in models such as those in Section 3, we will obtain biased parameter estimates. One potential solution to this problem – proposed in different contexts by Goldberg and Roth (2003), Popescul and Ungar (2003), Hoff (2009), and Chandrasekhar and Lewis (2011) – is to use the available data and any knowledge of the sampling scheme to predict the missing data. This can be used to recover the (predicted) structure of the entire network, which can then be used for calculating any network covariates. Such procedures require estimation of network formation models on the available data.

Finally, we saw in Section 3 that social contacts can be important for a variety of outcomes, including education outcomes (Duflo et al., 2011; De Giorgi et al., 2010), risk-sharing (Ambrus et al., 2014; Angelucci et al., 2012; Jackson et al., 2012), and agricultural practices (Conley and Udry, 2010). Hence one might want to understand where social connections come from *per se* and how they can be influenced, in order to create more desirable outcomes. For example, there is substantial evidence of homophily (Currarini et al., 2010). Homophily might in some circumstances limit the benefits of connections, since there may be bigger potential gains from interaction by agents who are more different, *e.g. ceteris paribus* the benefits of mutual insurance are decreasing in the correlation of income. We might then want to consider what the barriers are to the creation of such links, and what interventions might support such potentially profitable edges.

The key challenge to dealing with network formation models is the size of the joint distribution for edges. For a directed binary network, this is a $N(N-1)$-dimensional simplex, which has $2^{N(N-1)}$ points of support (potential networks).[47] To give a sense of scale, for a network of more than 7

[46] The same idea is used by Kelejian and Piras (2014) in the context of spatial regression.

[47] Through Section 4 we will be concerned with the identification and estimation of network formation models using data on a single network only. Throughout this section we therefore suppress the subscript g.

agents the support of this space is larger than the number of neurons in the human brain,[48] with 13 agents it is larger than the number of board configurations in chess,[49] and with 17 agents it is larger than the number of atoms in the observed universe.[50] Yet networks with so few agents are clearly much smaller than one would like to work with in practice. Hence simplifications will typically need to be made to limit the complexity of the probability distribution defined on this space, in order to make work with these distributions computationally tractable.

We begin in Subsection 4.1 by considering methods which allow us to use data on a subset of observed nodes to predict the status of unsampled nodes. Here the focus is purely on in-sample prediction of link probabilities, not causal estimates of model parameters, so econometric concerns about endogeneity can be neglected. Such methods allow us to impute the missing network edges, providing one method for dealing with measurement error.

In Subsection 4.2, we then discuss conditions for estimating a network formation model, when the ultimate objective is controlling for network endogeneity in the estimation of a social effects model, as discussed in Subsection 3.7. Now we may have data on some or all of the edges of the network, and methods used for estimation will in many cases be similar to those for in-sample prediction. The key difference is that only exogenous predictors/covariates may be used. Additionally, in order to be useful as a first-stage for a social effects model, there must be at least one covariate which is a valid instrument *i.e.* it must have explanatory power for edge status, and not directly affect the outcome in the social effects model.

Next in Subsection 4.3, we consider economic models of network formation. Here we think about individual nodes as being economic agents, who make choices to maximise some objective *e.g.* students maximising their utility by choosing who to form friendships with. We first consider non-strategic models of formation, where the formation of one edge does not generate externalities, so that $\Pr(G_{ij} = 1|G_{kl}) = \Pr(G_{ij} = 1) \, \forall ij \neq kl$. Estimation of these models is relatively straightforward, and again relates closely to the discussion in the first two subsections.

Finally, we end with a discussion of more recent work on network formation, which has begun allowing for strategic interactions. Here the value to i of forming edges with j might depend on the status of other edges in the network. For example, when trying to gather information about jobs, individuals might find it more profitable to form edges with highly linked individuals who are more likely to obtain information, rather than those with few contacts. This dependence of edges on the status of other edges introduces important challenges, particularly when only a single cross-section of data are observed, as will typically be the case in applications. Since this work is at the frontier of research in network formation, we will focus on describing the assumptions and methods that have so far been used to estimate these models, without being able to provide any general guidance on how practitioners should use these methods.

[48] Estimated to be around 8.5×10^{10} (Azevedo et al., 2009).
[49] Around $10^{46.25}$ (Chinchalkar, 1996).
[50] Around 10^{80} (Schutz, 2003).

4.1 In-sample prediction

Network formation models have long been studied in maths, computer science, statistical physics, and sociology. These models are characterised by a focus on the probability distribution $\Pr(\boldsymbol{G})$ as the direct object of interest.[51] For economists the main use for such models is likely to be for imputation/in-sample prediction when all nodes, and only a subset of edges in a network are observed.

The data available are typically a single realisation for a particular network, although occasionally multiple networks are observed and/or the network(s) is (are) observed over time. We focus on the case of one observation for a single network, since even when multiple networks are observed their total number is still small.[52] If multiple networks are available one could clearly at a minimum use the procedures described below, treating each separately, although one could also impose some restrictions on how parameters vary across networks if there is a good justification for doing so in a particular context. For example, suppose one observed edges between children in multiple classrooms in a school, with no cross-edges existing between children in different classes. If one believed that the parameters affecting edge formation were common across classrooms then one could improve the efficiency of estimation by combining the data. It could also provide additional identifying power, as network-level variables could also be incorporated into the model.

Identifying any non-trivial features of the probability distribution over the set of possible (directed) networks, $\Pr(\boldsymbol{G})$, is not possible from a single observation without making further restrictive assumptions. It is useful to note that $\Pr(\boldsymbol{G})$ is by definition equal to the joint distribution over all of the individual edges, $\Pr\left(G_{12}, ..., G_{N(N-1)}\right)$. Hence a single network containing N agents can be seen instead as $N(N-1)$, potentially dependent, observations of directed edge statuses.[53] This joint distribution can be decomposed into the product of a series of conditionals. For notational ease, let $l \in \Lambda$ index edges, so $\Lambda = \{12, 13, ..., 1N, 21, 23, ..., N(N-1)\}$. Then we can write $\Pr(\boldsymbol{G}) = \prod_{l \in \Lambda} \Pr(G_l | G_{l-1}, ..., G_1)$, so that each conditional distribution in the product is the distribution for a particular edge conditional on all previous edges. This conditioning encodes any dependencies which may exist between particular edges.

We begin with the simplest model of network formation, which assumes away both heterogeneity and dependence in edge propensities, and then reintroduce these features, describing the costs and benefits associated with doing so.

[51] Economists, in contrast, are often interested in microfoundations, so the focus is typically instead on understanding the preferences, constraints, and/or beliefs of the agents involved in forming $\boldsymbol{G}$. We consider models of this form in Subsection 4.3.

[52] As noted in footnote 47, we therefore suppress the subscript g throughout this section to avoid unnecessarily cluttered notation.

[53] If the network is undirected there are only half that many edges.

4.1.1 Independent edge formation

The *Bernoulli random graph* model is the simplest model of network formation. It imposes a common edge probability for each edge, and that probabilities are independent across edges. Independence ensures that the joint distribution $\Pr\left(G_{12}, ..., G_{N(N-1)}\right)$ is just the product of the marginals, $\prod_{l \in \Lambda} \Pr(G_l)$. A common probability for each edge means that $\Pr(G_l) = p \; \forall \, l \in \Lambda$, so all information about the distribution $\Pr(\boldsymbol{G})$ is condensed into a single parameter, p, the probability an edge exists.[54] This can be straightforwardly estimated by maximum likelihood, with the resulting estimate of the edge probability $\hat{p} = \frac{|E|}{N(N-1)}$,[55] equal to the proportion of potential edges that are present.

A natural extension of this model allows the probability $\Pr(G_{ij} = 1)$ to depend on characteristics of the nodes involved, $(\boldsymbol{x}_i, \boldsymbol{x}_j)$, but conditional on these characteristics independence across edges is maintained. This type of model can be motivated either by pairs of individuals with particular characteristics $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ being more likely to meet each other and hence form edges, or by the benefits of forming an edge depending on these characteristics, or some combination of these. In general one cannot separate meeting probabilities from the utility of an edge without either parametric restrictions or an exclusion restriction, so additional assumptions will be needed if one wants to interpret the parameters structurally. We discuss this further in Subsection 4.3.1.

The key restriction here is the assumption of independence across edge decisions. In many cases this is unlikely to be reasonable. For example, in a model of directed network formation, there might well be correlation in edges G_{ij} and G_{il} driven by some unobservable node-specific fixed effect for node i *e.g.* i might be very friendly, so be relatively likely to form edges. Use of the estimated model to generate predicted networks will be problematic, as it will fail to generate some of the key features typically observed, such as the high degree of clustering.

4.1.2 Allowing for fixed effects

The simplest form of dependencies that one might want to allow for are individual-specific propensities to form edges with others, and to be linked to by others. Such models were developed by Holland and Leinhardt (1977, 1981) and are known as p_1*-models*. They parameterise the log probability an edge exists, $\log(p_{ij})$, as a linear index in a (network-specific) constant θ_0, a fixed effect for the edge 'sender' $\theta_{1,i}$, and a fixed effect for the edge 'receiver' $\theta_{2,j}$, so $\log(p_{ij}) = \theta_0 + \theta_{1,i} + \theta_{2,j}$. The fixed effects are interpreted as individual heterogeneity in propensity to make or receive edges. Additional restrictions $\sum_i \theta_{1,i} = \sum_j \theta_{2,j} = 0$ provide a normalisation that deals with the perfect collinearity that would otherwise be present.

The use of such fixed effects creates inferential problems, since increasing the size of the network also

[54] Theoretical work on this type of model was done by Gilbert (1959), and it relates closely to the model of Erdös and Rényi (1959).

[55] Or twice that probability if edges are undirected, so that there are only $\frac{1}{2}N(N-1)$ potential edges.

increases the number of parameters,[56] sometimes described as an *incidental parameters problem*. One natural solution to the latter problem is to impose homogeneity of the θ_1 and θ_2 parameters within certain groups, such as gender and race.[57] If there are C groups, then the number of parameters is now $2C + 1$ and this remains fixed as N goes to infinity. This removes the inference problem and also allows agents' characteristics to be used in predicting edge formation.[58]

Alternatively, if node-specific effects are uncorrelated with node characteristics, then variations in edge formation propensity 'only' create a problem for inference. This comes from the unobserved node-specific effects inducing a correlation in the residuals, analogous to random effects. Fafchamps and Gubert (2007) show how clustering can be used to adjust standard errors appropriately.

However, in both cases the maintenance of the conditional independence assumption across edges continues to present a problem for the credibility of this method. In particular it rules out cases where the *status* of other edges, rather than just their probability of existence, affects the probability of a given edge being present. This would be inappropriate if for example i's decision on whether to form an edge with j depends on how many friends j actually has, not just on how friendly j is.

4.1.3 Allowing for more general dependencies

As discussed earlier in this section, identification of features of $\Pr(\boldsymbol{G})$ whilst allowing for completely general dependencies in edge probabilities is not possible. However, it is possible to allow the probability of an edge to depend on a subset of the network, where this subset is specified *ex ante* by the researcher. Such models are called p^*-models (Wasserman and Pattison, 1996) or *exponential random graph models* (ERGMs). These have already been used in economics by, for example, Mele (2013), who shows how such models can arise as the result of utility maximising decisions by individual agents, and Jackson et al. (2012) studying favour exchange among villagers in rural India.

Frank and Strauss (1986) showed how estimation could be performed in the absence of edge independence under the assumption that the structure of any dependence is known. For example, one might want to assume that edge ij depends not on all other edges, but only on the other edges that involve either i or j. This dependency structure, $\Pr_{\boldsymbol{\theta}}(G_{ij}|\boldsymbol{G}_{-ij}) = \Pr_{\boldsymbol{\theta}}(G_{ij}|G_{rs}\ \forall\ r \in \{i,j\}\text{or } s \in \{i,j\} \text{ but } rs \neq ij)$ where $\boldsymbol{\theta}$ is a vector of parameters and $\boldsymbol{G}_{-ij} = \boldsymbol{G} \backslash G_{ij}$, is called the *pairwise Markovian* structure.

Drawing from the spatial statistics literature, where this is a more natural assumption, Frank and Strauss show how an application of the Hammersley-Clifford theorem[59] can be used to account for *any* arbitrary form of dependency. The key result is that if the probability of the observed network

[56] Every new node adds two new parameters to be estimated.

[57] This is sometimes described as *block modelling*, since we allow the parameters, and hence edge probability, to vary across 'blocks'/groups.

[58] A related approach to solving this problem is suggested by Dzemski (2014).

[59] Originally due to Hammersley and Clifford (1971) in an unpublished manuscript, and later proved independently by Grimmett (1973); Preston (1973); Sherman (1973); and Besag (1974).

is modelled as an exponential function of a linear index of network statistics, appropriately defined, any dependency can be allowed for.

To construct the appropriate network statistics, they first construct a *dependency graph,* g^{dep}. This graph contains $N(N-1)$ nodes, with each node here representing one of the $N(N-1)$ edges in the original graph.[60] Then an edge between a pair of nodes ij and rs in the dependency graph denotes that the conditional probability that edge ij exists is not independent of the status of edge rs *i.e.* $\Pr_{\boldsymbol{\theta}}(G_{ij}=1|G_{rs}) \neq \Pr_{\boldsymbol{\theta}}(G_{ij}=1)$. Further, conditional on the set of neighbours of node ij in the dependency graph, nei_{ij}^{dep}, $\Pr(G_{ij}=1)$ is independent of all other edges in the original graph. So $\Pr_{\boldsymbol{\theta}}(G_{ij}=1|\boldsymbol{G}_{-ij}) = \Pr_{\boldsymbol{\theta}}(G_{ij}=1|G_{rs} \in nei_{ij}^{dep})$. For example, the p_1 graph, with independent edges, has a dependency graph containing no edges. By contrast, a 5-node graph with a pairwise Markovian dependency structure would have, for example, edge 12 dependent on edges $(13, 14, 15, 23, 24, 25, 31, 32, 41, 42, 51, 52)$, *i.e.* all edges which have one end at either 1 or 2.

We let $\mathcal{A}$ be the set of cliques[61] of the dependency graph, where isolates are considered to be cliques of size one. For example, if G_{ij} is independent of all other edges conditional on G_{ji} then $\mathcal{A} = \{(ij),\, (ij, ji)\}_{i \neq j}$.[62] Then we define A as representing the different architectures or *motifs* in $\mathcal{A}$. In the previous example these would be 'edges', (ij), and 'reciprocated edges' (ij, ji). This imposes a homogeneity assumption: that the probability a particular graph g is selected from $\mathcal{G}_N$ depends only on the number of edges and reciprocated edges, rather than to whom those edges belong, so all networks with the same overall architecture (called 'isomorphic networks'[63]) are equally likely. If instead we allow dependence between any edges that share a common node, then $\mathcal{A}$ is the set of all edges (ij), reciprocated edges (ij, ji), triads (ij, ir, rj),[64] and *k-stars* $(ij_1, ij_2, ..., ij_k)$. Now A represents 'edges', 'reciprocated edges', 'triads', and 'k-stars'.

Invoking the Hammersley-Clifford theorem, Frank and Strauss (1986) note that the probability distribution over the set of graphs $\mathcal{G}_N$ allows for the imposed dependencies if it takes the form

$$\Pr_{\boldsymbol{\theta}}(\boldsymbol{G}) = \frac{1}{\kappa(\boldsymbol{\theta})} \exp\left\{\sum_A \theta_A S_A(\boldsymbol{G})\right\} \tag{4.1}$$

where $S_A(\boldsymbol{G})$ is a summary statistic for motif A calculated from $\boldsymbol{G}$, θ_A is the parameter associated with that statistic, and $\kappa(\boldsymbol{\theta})$ is a normalising constant, sometimes described as the *partition function*, such that $\sum_{\boldsymbol{G} \in \mathcal{G}_N} \Pr_{\boldsymbol{\theta}}(\boldsymbol{G}) = 1$.[65] In particular, $S_A(\boldsymbol{G})$ must be a positive function of

[60] Nodes in this graph will be referred to by the name of the edge they represent in the original graph.

[61] A clique is any group of nodes such that every node in the group is connected to every other node in the group.

[62] (i, j) is always a member of $\mathcal{A}$, since we defined isolates as cliques of size one. Dependence of ij on ji means that we can also define (ij, ji) as a clique, since in the dependency graph these nodes are connected to each other.

[63] Formally, two networks are isomorphic iff we can move from one to the other only by permuting the node labels. For example, all six directed networks composed of three nodes and one edge are isomorphic. Isomorphism implies that all network statistics are also identical, since these statistics are measured at a network level so are not affected by node labels.

[64] This represents all triads in an undirected network, but in a directed network there are six possible edges between three nodes, since $ij \neq ji$, so we may define a number of different triads.

[65] In a slight abuse of notation we write $\sum_{\boldsymbol{G} \in \mathcal{G}_N} \Pr_{\boldsymbol{\theta}}(\boldsymbol{G})$ to mean $\sum_{g \in \mathcal{G}_N} \Pr_{\boldsymbol{\theta}}(\boldsymbol{G}_g)$.

the number of occurrences of motif A in $\boldsymbol{G}$. Since we are working with binary edges, without loss of generality we can define $S_A(\boldsymbol{G})$ as simply a count of the number of occurrences of motif A in the graph represented by $\boldsymbol{G}$. For example, defining $\mathbf{S}(\boldsymbol{G})$ as the vector containing the $S_A(\boldsymbol{G})$, if $\mathcal{A} = \{(ij),\, (ij, ji)\}_{i \neq j}$ then $\mathbf{S}(\boldsymbol{G})$ is a 2×1 vector containing a count of the number of edges and a count of the number of reciprocated edges.

Estimation of the ERGM model is made difficult by the presence of the partition function, $\kappa(\boldsymbol{\theta})$. Since this function normalises the probability of each graph so that the probabilities across all potential graphs sum to unity, it is calculated as $\sum_{\boldsymbol{G} \in \mathcal{G}_N} \exp\{\sum_A \theta_A S_A(\boldsymbol{G})\}$. The outer summation is a sum over the $2^{N(N-1)}$ possible graphs. As noted earlier, even for moderate N this is a large number, so computing the sum analytically is rarely possible.

Three approaches to estimation have been taken to overcome this difficulty: (1) the *coding method*; (2) the *pseudolikelihood* approach; and (3) the *Markov Chain Monte Carlo* approach. The first two are based on the maximising the conditional likelihoods of edges, rather than the joint likelihood, thus obviating the need for calculating the normalising constant, whilst the third instead calculates an approximation to this constant.

Coding Method The coding method (Besag, 1974) writes the joint distribution of the edge probabilities as the product of conditional distributions $\Pr_{\boldsymbol{\theta}}(\boldsymbol{G}) = \prod_{l \in \Lambda} \Pr_{\boldsymbol{\theta}}(G_l | G_{l-1}, ..., G_1)$, where as before Λ is the set of all $N(N-1)$ potential edges. Under the assumption that edge G_l depends only on a subset of other edges $G_{l'} \in nei_l^{dep}$ one could 'colour' each edge, such that each edge depends only on edges of a different colour.[66, 67] All edges of the original graph that have the same colour are therefore independent of each other by construction. Let Λ_c be the set of all edges of a particular colour. One could then estimate the parameter vector of interest, $\boldsymbol{\theta}$, by maximum likelihood, using only $\Pr_{\boldsymbol{\theta}}(G_l | G_{l'} \in nei_l^{dep})\, \forall l \in \Lambda_c$, which treats only edges of the same colour as containing any independent information.

We define the 'change statistic' $D_A(\boldsymbol{G}; l) := S_A(G_l = 1, \boldsymbol{G}_{-l}) - S_A(G_l = 0, \boldsymbol{G}_{-l})$ as the change in statistic S_A from edge G_l being present, compared with it not being present, given all the other edges $\boldsymbol{G}_{-l}$. Then, given the log-linear functional form assumption that we have made (see Equation 4.1), the conditional probability of an edge l can be estimated from the logit regression $\log\left\{\frac{\Pr(G_l=1|\boldsymbol{G}_{-l})}{\Pr(G_l=0|\boldsymbol{G}_{-l})}\right\} = \sum_A \theta_A D_A(\boldsymbol{G}; l)$. This can be implemented in most standard statistical packages. Hence we can estimate $\boldsymbol{\theta}$ using maximum likelihood under the assumption that the edge probability takes a logit form and treating the edges $l \in \Lambda_c$ as independent, conditional on the edges not in Λ_c. Since all the conditioning edges which go into S_A are of different colours, they are not included in the maximisation, so $\hat{\boldsymbol{\theta}}_c$ will be consistent.

[66] This is equivalent to saying that no two adjacent (*i.e.* linked) nodes of the dependency graph should have the same colour.

[67] Note that this colouring will not be unique. For example, one could trivially always colour every edge a different colour. However, for estimation it is optimal to try to minimise the number of colours used, as this makes the most of any information available about independence.

By performing this maximisation separately for each colour, a number of different estimates can be recovered. Researchers may choose to then report the range of estimates produced, or to create a single estimate from these many results, for example taking a mean or median.

The main disadvantage of this approach is that the resulting estimates will each be inefficient, since they treat the edges $l \notin \Lambda_c$ as if they contain no information about the parameters. In practice the proportion of edges in even the largest colour set Λ_c is likely to be small. For example, if any edges that share a node are allowed to be dependent, then the number of independent observations will only be $\frac{1}{2}N$[68]. Hence efficiency is far from a purely theoretical concern in the environment.

Pseudolikelihood approach The pseudolikelihood approach[69] attempts to overcome the inefficiency problem, by finding $\boldsymbol{\theta}$ which jointly maximises *all* the conditional distributions, not just those of the same colour. We write the log likelihood based on edges of colour c as $L_c = \sum_{l \in \Lambda_c} \log \Pr_{\boldsymbol{\theta}}(G_l = 1 | G_{l'} \in nei_l^{dep})$, with $\hat{\boldsymbol{\theta}}_c$ as the maximiser of this. Besag (1975) notes that the log (pseudo)likelihood $PL = \sum_c L_c = \sum_c \sum_{l \in \Lambda_c} \log \Pr_{\boldsymbol{\theta}}(G_l = 1 | G_{l'} \in nei_l^{dep})$, constructed by simply combining all the data as if there were no dependencies, is equivalent to a particular weighting of the individual, 'coloured' log likelihoods. This likelihood is misspecified,[70] since the correct log likelihood using all the data should be $L = \sum_l \log \Pr_{\boldsymbol{\theta}}(G_l = 1 | G_{l-1}, ..., G_1)$, whilst here we have instead $L = \sum_l \log \Pr_{\boldsymbol{\theta}}(G_l = 1 | \boldsymbol{G}_{-l}) = \sum_l \log \Pr_{\boldsymbol{\theta}}(G_l = 1 | G_L, ...G_{l+1}, G_{l-1}, ..., G_1)$. Nevertheless, under a particular form of asymptotics it may still yield consistent estimates.

We have already noted that for any given colour, the standard maximum likelihood consistency result applies, as the observations included are independent. If the number of colours are held fixed as the number of potential edges is increased,[71] then under some basic regularity conditions (Besag, 1975), maximising the log pseudolikelihood function $PL(\boldsymbol{\theta})$ as though there were no dependencies will also give a consistent estimate of $\boldsymbol{\theta}$.

Unfortunately, in practice this approach suffers from a number of problems. First, although it makes use of more information in the data, so is potentially more efficient, the standard errors that are produced by standard statistical packages such as Stata will clearly be incorrect as they will not take into account the dependence in the data. Little is known about how to provide correct standard errors, but in some cases inference can proceed using an alternative, non-parametric procedure: *multiple regression quadratic assignment procedure* (MRQAP). This method can provide a test as to whether particular edge characteristics or features of the local network, such as a common friend, are important for predicting the probability that a pair of individuals is linked. It is based

[68] Or $\frac{1}{2}(N-1)$ if N is odd.

[69] Introduced to the social networks literature by Strauss and Ikeda (1990).

[70] A likelihood based on $\Pr_{\boldsymbol{\theta}}(G_l | \boldsymbol{G}_{-l})$ without any correction suffers from simultaneity, since the probability of each edge is being estimated conditional on all others remaining unchanged. In a two node directed network, as a simple example, we effectively have two simultaneous equations, one for $\Pr_{\boldsymbol{\theta}}(G_{12} | G_{21})$ and $\Pr_{\boldsymbol{\theta}}(G_{21} | G_{12})$. It is well-known that such systems will not generally yield consistent parameter estimates if the dependence between the equations is not considered, and that strong restrictions will typically be needed even to achieve identification.

[71] In the language of spatial statistics, this is described as 'domain increasing asymptotics'.

on the quadratic assignment procedure (QAP): a type of permutation test for correlation between variables. For more details see Appendix B.

A second issue is that in network applications we need to impose some structure on the way in which new nodes are added to the network when we do asymptotics (Boucher and Mourifié, 2013; Goldsmith-Pinkham and Imbens, 2013). If, as we increase the sample size, new nodes added could be linked to all the existing nodes, then there is no reduction in dependence between links. In the spatial context for which the theory was developed, the key idea is that increasing sample size creates new geographic locations that are added at the 'edge' of the data. If correlations reduce with distance, then as new, further away, locations are added, they will be essentially independent from most existing locations. Such asymptotics are called *domain-increasing* asymptotics. The analogy in a networks context, proposed by Boucher and Mourifié (2013) and Goldsmith-Pinkham and Imbens (2013), is that new nodes are further away in the support of the covariates. If there is homophily, so that nodes which are far apart in covariates never link, then the decisions of these nodes are almost independent. Asymptotics results from the spatial case can then be used.

Third, Kolaczyk (2009) suggests that in practice this method only works well when the extent of dependence in the data is small. In general there is no reason to assume dependence will be small in network data; indeed it is precisely because we did not wish to assume this that we considered ERGMs at all.

Markov Chain Monte Carlo Maximum Likelihood An alternative approach, not based on the *ad-hoc* weighting provided by the pseudolikelihood approach, is to use Markov Chain Monte Carlo (MCMC) maximum likelihood (Geyer and Thompson, 1992, Snijders, 2002, Handcock, 2003). As noted earlier, the key difficulty with direct maximum likelihood estimation of Equation 4.1 is the presence of the partition function $\kappa(\boldsymbol{\theta}) = \sum_{\boldsymbol{G}\in\mathcal{G}} \exp\{\sum_A \theta_A S_A(\boldsymbol{G})\}$. This normalising constant is an intractable function of the parameter vector $\boldsymbol{\theta}$. In this estimation approach, MCMC techniques can be used to create an estimate of $\kappa(\boldsymbol{\theta})$ based on a sample of graphs drawn from $\mathcal{G}_N$.

The original log likelihood can be written as $L(\boldsymbol{\theta}) = \sum_A \theta_A S_A(\boldsymbol{G}) - \kappa(\boldsymbol{\theta})$. Maximising this is equivalent to maximising the likelihood ratio $LR = L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^{(0)})$ since the latter is just a constant for some arbitrary initial $\boldsymbol{\theta}^{(0)}$. Writing this out in full we get $LR = \sum_A \left[\theta_A - \theta_A^{(0)}\right] S_A(\boldsymbol{G}) - \left[\kappa(\boldsymbol{\theta}) - \kappa(\boldsymbol{\theta}^{(0)})\right]$. The second component can be approximated by drawing a sequence of W graphs, $(\boldsymbol{G}_1, ..., \boldsymbol{G}_W)$, from the ERGM under $\boldsymbol{\theta}^{(0)}$, and computing $\log \sum_{w\in W} \exp\left\{\sum_A (\theta_A - \theta_A^{(0)}) S_A(\boldsymbol{G}^{(w)})\right\}$ (see Kolaczyk (2009) pp185-187 for details). Under this procedure the maximiser of the approximated log likelihood will converge to its true value $\boldsymbol{\theta}$ as the number of sampled graphs W goes to infinity.

This approach has two major disadvantages. The first is that implementation of this method is very computationally intensive. Second, although this approach avoids the approximation of the likelihood by directly evaluating the normalising constant, its effectiveness depends significantly on the quality of the estimate of $\left[\kappa(\boldsymbol{\theta}) - \kappa(\boldsymbol{\theta}^{(0)})\right]$. If this cannot be approximated well then it is not clear that this approach, although more principled, should be preferred in practical applications.

Recent work by Bhamidi et al. (2008) and Chatterjee et al. (2010) suggests that in practice the mixing time – time taken for the Markov chain to reach its steady state distribution – of such MCMC processes is very slow (exponential time). This means that as the space of possible networks grows, the number of replications in the MCMC process that must be performed in order to achieve a reasonable approximation to $\left[\kappa(\boldsymbol{\theta}) - \kappa(\boldsymbol{\theta}^{(0)})\right]$ rises rapidly, making this approach difficult to justify in practice.

Statistical ERGMs Chandrasekhar and Jackson (2014) also note that practitioners often report obtaining wildly different estimates from repeated uses of ERGM techniques on the same set of data with the same model, with variation far exceeding that expected given the claimed standard errors. They propose a technique which they call *Statistical ERGM* (SERGM), which is easier to estimate, as an alternative to the usual ERGM. With this they are not able to recover the probability that we observe a particular network, but instead focus on the probability of observing a given realisation, $\boldsymbol{s}$, of the network statistics, $\boldsymbol{S}$.[72]

In an ERGM the sample space consists of the set of possible distinct networks on the N nodes. This set has $2^{N(N-1)}$ elements (in the case of a directed network), and we treat each isomorphic element as being equally likely. Our *reference distribution* is a uniform distribution across these $2^{N(N-1)}$ elements *i.e.* this is the null distribution against which we are comparing the observed network.

If our interest is only in the realisations of the network statistics, we can reduce the size of the sample space we are working with. Chandrasekhar and Jackson (2014) define SERGMs as ERGMs on the space of possible network statistics, $\mathcal{S}$. This sample space will typically contain vastly fewer elements than the space of possible networks.

We can then rewrite Equation 4.1 using the space of network statistics as sample space. In this case the probability of observing statistics $\boldsymbol{S}(\boldsymbol{G})$ taking value $\boldsymbol{s}$ is $\Pr_{\boldsymbol{\theta}}(\boldsymbol{S}(\boldsymbol{G}) = \boldsymbol{s}) = \frac{\#_{\boldsymbol{S}}(\boldsymbol{s})\exp(\boldsymbol{\theta s})}{\sum_{\boldsymbol{s}'}\#_{\boldsymbol{S}}(\boldsymbol{s}')\exp(\boldsymbol{\theta s}')}$, where $\#_{\boldsymbol{S}}(\boldsymbol{s}) = |\{\boldsymbol{G} \in \mathcal{G} : \boldsymbol{S}(\boldsymbol{G}) = \boldsymbol{s}\}|$ is the number of potential networks which have $\boldsymbol{S} = \boldsymbol{s}$.

So far we have only rewritten our originally ERGM by defining it over a new space. We defined our reference distribution in the ERGM to put equal weight on each possible *network*. To maintain this distribution when the sample space is the space of statistics, we must weight the usual (unnormalised) probability of observing network $\boldsymbol{G}$, $\exp(\boldsymbol{\theta s})$, by the number of networks which exhibit this configuration of statistics, $\#_{\boldsymbol{S}}(\boldsymbol{s}')$.

Much of the difficulty in estimating ERGM models comes from use of these weights, since we are required to know in how many networks a particular combination of statistics exists. Since this is typically not possible to calculate analytically, we discussed how MCMC approaches might be used to sample from the distribution of networks.

[72] $\boldsymbol{S}$ is a $|\mathcal{A}| \times 1$ dimensional vector stacking the network statistics S_A, and $\boldsymbol{\theta}$ a $1 \times |\mathcal{A}|$ dimensional vector of parameters.

Chandrasekhar and Jackson (2014) complete their definition of SERGMs as a generalisation of ERGMs by allowing any reference distribution, $K_{\boldsymbol{S}}(\boldsymbol{s})$ to be used in the place of $\#_{\boldsymbol{S}}(\boldsymbol{s}')$. However, to ease estimation relative to ERGMs, they then define the 'count SERGM', which imposes $K_{\boldsymbol{S}}(\boldsymbol{s}) = \frac{1}{|\mathcal{S}|}$.[73] The key here is not that these weights are constant, but that they no longer depend on the space of networks. Since $K_{\boldsymbol{S}}(\boldsymbol{s})$ is now known, unlike $\#_{\boldsymbol{S}}(\boldsymbol{s}')$ which needed to be calculated, if $|\mathcal{S}|$ is sufficiently small, exact evaluation of the partition function $\tilde{\kappa}(\boldsymbol{\theta}) = \sum_{\boldsymbol{s}'} K_{\boldsymbol{S}}(\boldsymbol{s}') \exp\{\boldsymbol{\theta}\boldsymbol{s}'\}$ is now possible.

Since count SERGMs – and any other SERGMs with known $K_{\boldsymbol{S}}(\boldsymbol{s}')$ – can be estimated directly and without approximation, they are easier to implement than standard ERGMs. Chandrasekhar and Jackson (2014) also provide assumptions under which the parameters of the SERGM, $\boldsymbol{\theta}_{SERGM}$, can be estimated consistently.

The key drawback to this method is in interpretation. The estimated parameters, $\boldsymbol{\theta}_{SERGM}$, are not the same as the parameters $\boldsymbol{\theta}$ in Equation 4.1, and the predicted probabilities are now the probability of a particular configuration of statistics, rather than of a particular network. Nevertheless, for a researcher interested in which network motifs are more likely to be observed than one would expect under independent edge formation, SERGMs offer an appropriate alternative.

4.2 Reduced form models of network formation

The methods discussed in the previous subsection focused on in-sample prediction of network edges. However, since they (mostly) predict these probabilities based on the structure of the networks, without use of other characteristics, they both fail to make use of all the information typically available to researchers, and also do not contain the necessary independent variation needed for use as the first stage of a social effects model with an endogenous network (of the sort discussed in Subsection 3.7). When our ultimate aim is to estimate a social effects model but we are concerned about the network being endogenous, one solution discussed in Subsection 3.7 is to estimate the edge probability using individual characteristics, including at least one covariate that is not included in the outcome equation (an exclusion restriction), as in a standard two-stage least squares setting. In this subsection we describe estimation of models that include individual (node) characteristics. As long as at least one of these is a valid instrument, then this approach to overcoming the endogeneity of network formation is possible.

A well-recognised feature of many kinds of interaction networks is the prevalence of homophily: a propensity to be linked to relatively similar individuals.[74] This observation may arise from a preference for interacting with agents who are similar to you (preference homophily), a lower cost of interacting with such agents (cost homophily), or a higher probability of meeting such agents

[73] Count SERGMs also restrict the set $\mathcal{A}$ to include only network motifs such as triangles and nodes of particular degree, which can be counted. This rules out, for example, statistics such as density.

[74] Homophily may be casually described as the tendency of 'birds of a feather to flock together'.

(meeting homophily). However, they all have the reduced form implication that more similar agents are more likely to be linked.[75]

Fafchamps and Gubert (2007) provide a discussion of the conditions that must be fulfilled by a model used for *dyadic regression*, *i.e.* a regression model of edge formation when edges are being treated as observations and node characteristics are included in the regressors. They note the regressors must enter the model symmetrically, so that the effect of individual characteristics ($\boldsymbol{x}_i$, $\boldsymbol{x}_j$) on edge G_{ij} is the same as that of ($\boldsymbol{x}_j$, $\boldsymbol{x}_i$) on G_{ji}. Additionally the model may contain some edge-specific covariates, such as the distance between agents, which must by definition be symmetric $\boldsymbol{w}_{ij} = \boldsymbol{w}_{ji}$. If edges are modelled as directed, then the model takes the general form

$$G_{ij} = f\left(\lambda_0 + (\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j})\boldsymbol{\lambda}_1 + \boldsymbol{x}_{2i}\boldsymbol{\lambda}_2 + \boldsymbol{x}_{3j}\boldsymbol{\lambda}_3 + \boldsymbol{w}_{ij}\boldsymbol{\lambda}_4 + u_{ij}\right) \tag{4.2}$$

This specification allows a term that varies with the difference between i and j in some characteristics, $(\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j})$; terms varying in the characteristics of both the sender and the receiver of the edge, $\boldsymbol{x}_{2i}$ and $\boldsymbol{x}_{3j}$ respectively; some edge-specific characteristics, $\boldsymbol{w}_{ij}$; and an edge-specific unobservable, u_{ij}. There may be partial or even complete overlap between any of $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$. Since G_{ij} is typically binary, the function $f(.)$ and the distribution of u are usually chosen to make the equation amenable to probit or logit estimation. However, in some cases other functional forms are chosen. For example, Marmaros and Sacerdote (2006) model $f(.)$ as exp(.) since they are working with email data, measuring edges by the number of emails between the individuals, which takes only non-negative values and varies (almost) continuously.

If edges are undirected, then $(\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j})$ must be replaced with $|\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j}|$;[76] $\boldsymbol{x}_2 = \boldsymbol{x}_3$ and $\boldsymbol{\lambda}_2 = \boldsymbol{\lambda}_3$; and $u_{ij} = u_{ji}$, so that G_{ij} necessarily equals G_{ji}. The identification of parameters $\boldsymbol{\lambda}_2$ and $\boldsymbol{\lambda}_3$ requires variation in degree. As Fafchamps and Gubert (2007) note, if all individuals in the data have the same number of edges, such as a dataset of only married couples, then it is possible to ask whether people are more likely to form edges with people of the same race, captured by $\boldsymbol{\lambda}_1$, but not possible to ask whether some races are more likely to have edges.

Careful attention needs to be paid to inference in this model, since there is dependence across multiple dyads for any individual, similar to the Markov random graph assumption discussed in the previous subsection. Fafchamps and Gubert (2007) show that standard errors can be constructed analytically using a 'four-way error components model'. This is a type of clustering, allowing for correlation between u_{ij} and u_{rs} if either of i or j is equal to either of r and s. The analytic correction they propose provides an alternative to using MRQAP, described in Subsection 4.1.3, which may also be used in this circumstance.

[75] In Subsection 4.3.1 below, we consider homophily in more detail, and structural models that try to separate these causes of observed homophily.

[76] Or $(\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j})^2$ may also be used.

4.3 Structural models of network formation

Economic models of network formation consider nodes as motivated agents, endowed with preferences, constraints, and beliefs, choosing which edges to form. The focus for applied researchers is to estimate parameters of the agents' objective functions. For example, to understand what factors are important for students in deciding which other students to form friendships with.

These models allow us to think about counterfactual policy scenarios. For example, if friendships affect academic outcomes, then there might be a role for policy in considering how best to organise students into classrooms, given knowledge of their endogenous friendship formation response. If students tend to form homophilous friendships *i.e.* with others who have similar predetermined characteristics, but not to form friendships across classrooms, there may be a case for not streaming students into classes of similar academic abilities. This would create more heterogeneity in the characteristics of friends than if streaming were used, which might improve the amount of peer learning that takes place.[77] We begin by discussing non-strategic models, in which these decisions depend only on the characteristics of the agents involved in the edge. We then discuss strategic network formation, which occurs when network features directly enter into the costs or benefits of forming particular edges.[78]

4.3.1 Structural Homophily

As noted above, a key empirical regularity which holds across a range of network types is the presence of *homophily*. This is related to the more familiar (in economics) concept of positive assortative matching, *i.e.* that people with similar characteristics form edges with one another. As we have already seen, many reduced form models include homophilic terms – captured by $\boldsymbol{\lambda}_1$ in Equation 4.2 – to allow the probability a tie exists to vary with similarity on various node characteristics.[79] In this subsection, we consider the economic models of network formation that are based on homophily.

We define homophily formally as follows. Let the individuals in a particular environment be members of one of H groups, with typical group h. Groups might be defined according to sex, race, height, or any other characteristics. Continuous characteristics will typically need to be discretised. We denote individual i's membership of group h as $i \in h$. Relationships for individuals in group h exhibit homophily if $\Pr(G_{ij} = 1 | i \in h,\, j \in h) > \Pr(G_{ij} = 1 | i \in h,\, j \notin h)$. In words, a group h exhibits homophily if its members are more likely to form edges with other members of the same group than one would expect if edges were formed uniformly at random among the population of

[77] Clearly this is just an example, and there are many other factors to consider, such as the effectiveness of teachers when faced with more heterogeneous classrooms, the ability to tailor lessons to challenge high ability students, and other outcomes that might be influenced by changing friendships.

[78] See also a recent survey by Graham (2015), which became available after work on this manuscript.

[79] In principle this probability could be falling in similarity, known as *heterophily*. This may be relevant, for example in models of risk sharing with heterogeneous risk preferences and complete commitment.

nodes. In general there will be multiple characteristics $\{H^1, ..., H^K\}$ according to which individuals can be classified, and relationships may exhibit homophily on any number of these characteristics.

As noted earlier there are (at least) three possible sources of homophily: *preference homophily*, *cost homophily*, and *meeting homophily*.

Preference homophily implies that, conditional on meeting, people in a group are more likely to form edges with other members of the same group as they value these edges more. For example, within a classroom boys and girls might have equal opportunities to interact, but boys may choose to form more friendships with other boys (and *mutatis mutandis* for girls) if they have more similar interests.

Cost homophily occurs when the cost of maintaining an edge to a dissimilar agent is greater than the cost of maintaining an edge to a more similar agent. For example, one might have an equal preference for all potential friends, but find it 'cheaper' to maintain a friendship with individuals who live relatively nearer. Unlike preferences, which are in some sense fundamental to the individual, costs might be manipulable by policy. To the extent that they are environmental these can also change the value of an edge over time, *e.g.* a friend moving further away may lead to the friendship being broken.

Meeting homophily occurs when people of a particular group are more likely to meet other members of the same group. For example, if we thought of all students in a school year as being part of a single network, then there is likely to be meeting homophily within class groups, since students in the same class have more opportunities to interact. Again this is amenable to manipulation by policy, for example changing seating arrangements across desks in a classroom. However, unlike cost homophily, once individuals have met, changes in the environment should not change the value of a friendship.

These three sources of homophily all have the reduced form implication that the coefficient on the *absolute* difference in characteristics, $\boldsymbol{\lambda}_1$ in Equation 4.2, should be negative for any characteristics on which individuals exhibit homophily. However, since they may have different policy implications, there is a case for trying to distinguish which of these channels are operating to cause the observed homophily.

Currarini et al. (2009) suggest how one can distinguish between preference and meeting homophily under the assumption that cost homophily does not exist. They note that if group size varies across groups, then preference homophily should lead to more friendships among the larger group, whereas meeting homophily should not. Intuitively this is because under preference homophily, a larger own-group means there are more people with whom one might potentially form a profitable friendship. One could then use regression analysis to test for the presence of preference homophily by interacting group size with absolute difference in characteristics, and testing whether the estimated parameter is significantly different from zero.

Alternatively one might want to estimate the magnitude of the effect of changing particular features of the environment, such as the classrooms to which individuals are assigned. In this case one could

parameterise an economic model of behaviour, and then directly estimate the parameters of the model. Currarini et al. (2009) do this using a model of network formation that incorporates a biased meeting process, so individuals can meet their own-type more frequently than other types, and differences in the value of a friendship depending on whether agents are the same type.[80] They simulate the model with a number of different parameters for meeting probabilities and relative values of friendships, and use a minimum distance procedure to choose the parameters that best explain the data.

As ever with structural models, whilst this approach allows one to perform counterfactual policy experiments, the main cost is that the reasonableness and interpretation of results depend on the accuracy with which the imposed model fits reality. Also, without time series variation in friendships, one cannot also allow for cost heterogeneity, which might show up either in preferences by changing the value of forming an edge, or in meeting probabilities since those with lower meeting probabilities will typically have a greater cost to maintaining a friendship. Finally, it is important to note that estimation of such models requires the unobserved component of preferences to be independent of the factors influencing meeting. If the unobserved preference for partying is correlated with choosing to live in a particular dormitory, and hence meeting other people living here, then this will bias the parameter estimate of the probability of meeting in this environment.

Mayer and Puller (2008) develop an enriched version of this model which allows again for meeting and preference homophily, but they allow the bias in the meeting process to depend not only on exogenous characteristics, but also on sharing a mutual friend. Formally, $\Pr(meet_{ij} = 1 | G_{ir} = G_{jr} = 1) > \Pr(meet_{ij} = 1)$, where $\Pr(meet_{ij})$ denotes the probability that nodes i and j meet (and hence have the opportunity to form an edge). This allows for the stylised fact that individuals who are friends often also share mutual friends, which helps the model match the observed clustering in the data.

However, although the model fit is improved, their model cannot distinguish whether this clustering is in fact generated by a greater probability of meeting such individuals, a greater benefit to being friends with someone you share a friend with already, or a lower cost of maintaining that friendship. They show how one can estimate their model using a simulated method of moments procedure. However, this method suffers from the same constraints as those in the model suggested by Currarini et al. (2009): the utility of the model for counterfactuals depends on how closely it matches reality; cost homophily is neglected; and it is important the unobserved component of preferences is independent of the meeting process.

In the next subsection we consider extensions to these models that allow network statistics, such as sharing a common friend, to enter into individuals' utility functions. These create strategic interactions which can complicate estimation.

[80] Again they do not allow for cost homophily.

4.3.2 Strategic network formation

Much of the theoretical literature on networks has emphasised the strategic nature of interactions, setting up games of network formation as well as games to be played on existing networks (as seen in Section 3 above). The empirical literature has recently begun to take a similar approach, trying to estimate games of network formation. The key extension of such models, beyond those already considered, is to include network covariates into the objective function of agents. This creates two complications: first such models may have zero, one, or many equilibria, and this must be accounted for in estimation; and second, as with ERGM models, the presence of network covariates necessitates the calculation of intractable functions of the unknown parameters.

Before considering estimation in more detail, we discuss the modelling choices that one needs to make. First, as with all structural modelling one must explicitly determine the nature of the objective function that agents are trying to maximise. For example one might have individuals with utility functions that depend on some feature of the network,[81,82] who are trying to maximise this utility. Second, the 'rules of the game': are decisions made simultaneously or sequentially? Unilaterally or bilaterally? What do agents know, and how do they form beliefs? Given that we typically only observe a single cross-section of data, additional assumptions about the nature of any meeting process are necessary. Similarly, data may be reported as directed or undirected, but whether we treat unreciprocated directed edges as measurement error or evidence of unilateral linking is an important consideration, particularly given the consequences of such measurement error (see Section 5.3). Finally, one needs to take a stand on the appropriate concept of equilibrium and the strategies being played. At the weakest, one could impose only that strategies must be rationalisable, and hence many strategy profiles are likely to be equilibria. On the other hand, depending on the information available to agents one could impose Nash equilibrium, or Bayes-Nash equilibrium where individuals have incomplete information and need to form beliefs. Alternatively one could use a partly cooperative notion of equilibrium such as pairwise stability (Jackson and Wolinsky, 1996), which models link formation as requiring agreement from both parties involved, although dissolution remains one-sided.[83]

Since these models are at the frontier of research on network formation, few general results are currently available. We therefore instead briefly discuss the approaches that have been taken so far to write estimable models, and estimate the parameters of these models. Our aim is to highlight some of the choices that need to be made, and their relative advantages and costs.

Christakis et al. (2010) and Mele (2013) both model network formation as a sequential game: there is some initial network, and then a sequential process by which edge statuses may be adjusted.

[81] For example their centrality, or the number of edges they have subject to some cost of forming edges.

[82] It is important to note that although it is the *realised* network feature that typically enters an agent's objective function, their strategy will depend on their *beliefs* about how others will act.

[83] As in the literature on coalition formation, the issue of whether utility is transferable or not is also critical. Typically this issue is not discussed in networks papers (Sheng (2012) is an exception to this), and it is implicitly assumed that utility is not transferable.

Crucial, also, to their models, is that at each meeting agents only weigh the static benefits of updating the edge status (*i.e.* play a *myopic best response*), rather than taking into account the effect this decision will have on both their own and others' future decisions. Allowing for such forward-looking behaviour has so far proved insolvable from an economic theory perspective, and hence they rule this out.

Christakis et al. (2010) assume the initial network is empty, and allow each pair to meet precisely once, uniformly at random, in some unknown order. Mele (2013) also allows uniform at random meeting, but pairs may meet many times until no individual wants to change any edge. In both cases these assumptions about the meeting process – the number of meetings, order in which pairs meet, and probability with which each pair meets – will influence the set of possible networks that may result. However, in the latter case, the resulting network will be an equilibrium network, something which is not true in Christakis et al. (2010).

A different approach, taken by Sheng (2012), avoids making assumptions about the meeting order. Instead she uses only an assumption about the relevant equilibrium concept (pairwise stability). For the network to be pairwise stable, the utility an agent gets from each link that is present must be greater than the utility he would get if the link were not present, and conversely for a link which is not present at least one of the agents it would involve must not prefer it. Sheng uses the moment inequalities this implies for estimation, but is only able to find bounds on the probability of observing particular networks.[84] Hence assumptions about meeting order seem important for the point identification of the parameter of interest (we discuss this further below).

de Paula et al. (2014) also avoid assumptions on the meeting order. Rather than using individual-level data, they identify utility parameters by aggregating individuals into 'types', and looking at the share of each type that is observed in equilibrium. This can be seen as an extension of the work of Currarini et al. (2009). Individuals' characteristics are discretised, so that each individual can be defined as a single type. Agent characteristics might, for example, be sex and age. Typically age is measured to the nearest month or year, so is already discretised. However, if the number of elements in the support is large, broader discretisation might be desirable (*e.g.* in the age example, measure age in ten-year bands). Then we might define one type as (male, 25-35years) and another as (female, 15-25). de Paula et al. (2014) assume that agents have preferences only over the types they connect to both directly and indirectly, not who the individuals are, and that preference shocks are also defined in terms of type rather than individuals. They further assume that there is some maximum distance such that there is no value to a having connections beyond this distance, and there is a maximum number of direct connections that would be desired. Under these restrictions they can set identify the set of parameters for which the observed outcome – distribution of network types – is an equilibrium, without making any assumptions on equilibrium selection. They are even

[84] Sheng (2012) is actually only able to estimate an 'outer region' in which these probabilities lie, rather than a sharp set. More information is, in principle, available in the data, but making use of it would increase the computational burden.

able to allow for non-existence of equilibrium, in which case the identified set is empty. Estimation can be performed using a quadratic program.

Recent work by Leung (2014) takes a fourth approach, and is able to achieve point identification without assumptions on the meeting order. Instead the game is modelled as being simultaneous (so there is no meeting order to consider), but there is also incomplete information. Specifically, the unobserved (by the econometrician) link-specific component of utility is assumed to also be unobserved by other agents. Hence agents make their decisions with only partial knowledge about what network will form. Estimation proceeds using a so-called 'two-step' estimator, analogous to that used by Bisin et al. (2011) in a different context. First agents' beliefs about the expected state of the network are estimated non-parametrically. The observed conditional probability of a link in the network is used as an estimate for agents' belief about the probability such a link should form. This estimated network is used to replace the endogenous observed network variables that enter the utility function. Then the parameters of the utility function can be estimated directly in a second step. One advantage of this approach is that only a single network is needed to be able to estimate the utility parameters, although the network must be large.

Whether edges should be modelled as directed has consequences for identification and estimation, as well as the interpretation of the results, and will depend on features of the data used. Both Christakis et al. (2010) and Mele (2013) use data on school students from the *National Longitudinal Study of Adolescent Health* (*Add Health*), but Christakis et al. (2010) assume friendship formation is a bilateral decision whilst Mele (2013) assumes it is unilateral. The data show some edges that are not reciprocated, and it is an issue for researchers how this should be interpreted.[85] Theoretically, networks based on unilateral linking are typically modelled as being Nash equilibria of the network formation game, whilst those based on bilateral edges use *pairwise stability* (Jackson and Wolinsky, 1996) as their equilibrium concept.[86]

Both Christakis et al. (2010) and Mele (2013) assume utility functions such that the marginal utility of an edge depends on characteristics of the individuals involved, the difference in their characteristics (homophily), and some network statistics. This has two crucial implications.

First, since they assume network formation occurs sequentially, they need to assume a meeting process to 'complete' their models. This process acts as an equilibrium selection mechanism. Although they do not discuss equilibrium, Christakis et al. (2010) use the meeting process to determine what network should be realised for a given set of covariates and parameters. Mele (2013) makes assumptions on the structure of the utility function to ensure that at least one Nash equilibrium exists, but potentially there are multiple equilibria. The meeting process is then used to provide

[85] It is sometimes argued when data contain edges that are not reciprocated that the underlying relationships are reciprocal, but that some agents failed to state all their edges. The union of the edges is then used to form an undirected graph, so $g_{ij}^{undir} = max(g_{ij}, g_{ji})$.

[86] Loosely, an undirected network is pairwise stable if (i) $G_{ij} = 1$ implies that neither i nor j would prefer to break the edge, and (ii) $G_{ij} = 0$ implies that if i would like to edge with j then j must *strictly* not want to edge with i.

an ergodic distribution over these equilibria. In both cases functional form assumptions and use of a meeting order are critical to identification.[87]

Second, both papers assume that the relevant network statistics are based on purely 'local' network features. By this we mean that the marginal utility to i of forming an edge with j depends only on edges that involve either i or j. This is equivalent to the *pairwise Markovian* assumption discussed in Subsection 4.1. Estimation of these models can therefore be performed using the MCMC techniques described there. It also suffers from the same difficulties, *viz.* that estimation is time-consuming, and often the parameter estimates are highly unstable between runs of the estimation procedure because of the difficulty in approximating the partition function.

Hence, although in principle, it has recently become possible to estimate economic models of strategic network formation, there is still significant scope for further work to generalise these results and relax some of the assumptions that are used.

5 Empirical Issues

The discussion thus far has taken as given some, possibly multiple, networks $g = \{1, .., M\}$ of nodes and edges. In this section we consider where this network comes from. We begin by outlining the issues involved in defining the network of interest. We then discuss the different methods that may be used to collect data on the network, focusing on practical considerations for direct data collection and sampling methods. Our discussion thereafter examines in detail the issue of measurement error in networks data. We divide issues into those where measurement error depends on the sampling procedure, and those from other sources. Since networks are composed of interrelated nodes and edges, random (*i.e.* i.i.d.) sampling of either nodes or edges imposes some (conditionally) non-random process on the other, which depends on the structure of the underlying network, thereby generating non-classical measurement error. We discuss the implications of measurement error arising from both these sources – sampling and other – on network statistics, and on parameter estimates of models that draw on these data. Researchers working in a number of disciplines including economics, statistics, sociology and statistical physics have suggested methods for dealing with measurement error in networks data, which are described in detail thereafter.

5.1 Defining the network

A first step in network data collection is to define, based on the research question of interest, the interaction that one would like to measure. For example, suppose one were studying the role of social learning in the adoption of a new technology, such as a new variety of seeds. In this situation, information sharing with other farmers cultivating the new variety could be considered to be the

[87] Without a meeting order, both Sheng (2012) and de Paula et al. (2014) only achieve partial identification. Leung (2014) achieves point identification by assuming agents move simultaneously and have incomplete information.

most relevant interaction. The researcher would then aim to capture interactions of this type in a network of nodes and edges. It should be noted that different behaviours and choices will be influenced by different interactions. For example, amongst households in a village, fertiliser use might be affected by the actions of other farmers, whilst fertility decisions may be influenced by social norms of what the whole village chooses. Similarly, (extended) family members are more likely to lend one money, while friends and acquaintances are often better sources of information on new opportunities.[88]

Moreover, even when the interaction of interest is well-defined, *e.g.* risk-sharing between households, there is an additional question of whether *potential* network neighbours – that is households who are willing to make a transfer or lend to one's own household – or *realised* network neighbours – the households that one's household actually received transfers or loans from – are of interest. Hence the research question of interest and the context matter, and having detailed network data is not a panacea: one must still justify why the measured network is the most relevant one for the research question being considered.

In addition, researchers are typically also forced to define a *boundary* for the network, within which all interactions are assumed to take place. Geographic boundary conditions are very common in social networks – for instance, edges may only be considered if both nodes are in the same village, neighbourhood or town – supported by the implicit assumption that a majority of interactions takes place among geographically close individuals, households and firms. Such an assumption is questionable,[89] but greatly eases the logistics and costs of collecting primary network data, and is often considered to be the most reasonable when no further information is available on the likely reach of the network being studied.

Network data collection involves collecting information on two interrelated objects – nodes and edges between nodes – within the pre-defined boundary. Data used in most economic applications are typically collected as a set of observations on nodes (individuals, households, or firms), with information on the network (or group(s)) they belong to, and perhaps with information on other nodes within the network (or group) that they are linked to. As an example, in a development context, we may have a dataset with socio-economic information on households (nodes), the village or ethnic group they belong to (group), and potentially which other households within the village its members talk to about specific issues (edges). Our focus, as elsewhere in this paper, continues to be cases where detailed information on network neighbours (*i.e.* edges) is available, although where multiple group memberships are known these may also be used to implicitly define a set of neighbours, as in De Giorgi et al. (2010).

[88] The classic example of this issue comes from Granovetter (1973), who shows the importance of 'weak ties' in providing job vacancy information.

[89] For example, a household's risk sharing might depend more on its edges to other households outside the village, since the geographic separation is likely to reduce the correlation between the original household's shocks and the shocks of these out-of-village neighbours.

5.2 Methods for Data Collection

In practical terms, a range of methods can be and have been used to collect the information needed to construct network graphs. In order to construct undirected network graphs, researchers need information on the nodes in the network, and on the edges between nodes.[90] Depending on the interaction or relationship being studied, it may furthermore be possible to obtain information on the directionality of edges between nodes, and on the strength of edges, allowing for the construction of *directed* and *weighted* graphs. The methods include:

1. Direct Elicitation from nodes:

   (a) Asking nodes to report all the other nodes they interact with in a specific dimension within the specified network boundary, *e.g.* all individuals within the same village that one lends money to. In this case, nodes are free to list whomever they want. Information on the strength of edges can similarly be collected.[91]

   (b) Asking nodes to report for every other node in the network whether they interacted with that node (and potentially the strength of these interactions). In contrast to (a), nodes are provided with a list of all other nodes in the network. Though this method has the advantage of reducing recall errors, it may generate errors from respondent fatigue in networks with a large number of nodes.

   (c) Asking nodes to report their own network neighbours and their perception of edges between other nodes in the network. This method would presumably work reasonably well in settings where, and in interactions for which, private information issues are not very important (*e.g.* kinship relations in small villages in developing countries). Alatas et al. (2012) use this method to collect information on networks in Indonesian hamlets.

   (d) Asking nodes to report their participation in various groups or activities, and then imposing assumptions on interactions within the groups and activities, *e.g.* two nodes are linked if they are members of the same group. The presence of multiple groups can generate a partially-overlapping peer group structure.

2. Collection from Existing Data Sources: Edges between nodes can be constructed from information in available databases *e.g.* citation databases (Ductor et al., 2014), corporate board memberships (Patnam, 2013), online social networks (*e.g.* LinkedIn, Twitter, Facebook).

   The resulting networks often have a partially-overlapping peer group structure, with agents that share a common environment (such as a university) belonging to multiple subgroups (*e.g.* classes within the university). Network structure is then imposed by assuming that

[90] Some features of network graphs can be obtained without detailed information on all nodes and the edges between nodes. Degree, for instance, can be captured by asking nodes directly about the number of edges they have, without enquiring further about who these neighbours are.

[91] In practice, edge strength is usually proxied by the frequency of interaction, or the amount of time spent together, or in the case of family relationships, by the amount of shared genetic material between individuals.

an edge exists between nodes that share a subgroup. Examples include students in a school sharing different classes (*e.g.* De Giorgi et al., 2010) or company directors belonging to the same board of directors (*e.g.* Patnam, 2013) or households which, through marriage ties of members, belong to multiple families (*e.g.* Angelucci et al., 2010).

Moreover, the directionality of the edge can sometimes, though not always, be inferred from available data, *e.g.* data from Twitter includes information on the direction of the edge, while the existence of an edge in LinkedIn requires both nodes to confirm the edge. However, it is not possible to infer directionality among, for instance, students in a school belonging to multiple classes, since we don't even know if they actually have any relationship.

In order to generate the full network graph, researchers would need to collect data on all nodes and edges, *i.e.* they need to collect a census. This is typically very expensive, particularly since a number of methods described above in Section 3 exploit cross-network variation to identify parameters, meaning that many networks would need to be fully sampled.

In general, it is very rare to have data available from a census of all nodes and edges. Even when a census of nodes is available, it is very common to observe only a subset of edges because of censoring in the number of edges that can be reported.[92] In practice, given the high costs of direct elicitation of networks, and the potentially large size of networks from existing data sources,[93] researchers usually collect data on a sample of the network only, rather than on all nodes and edges. Various sampling methods have been used, of which the most common are:

1. Random Sampling: Random samples can be drawn for either nodes or edges. This is a popular sampling strategy due to its low cost relative to censuses. 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.[94] At times, information may also be available on the identities, and in some rare cases, on some socio-economic variables of all nodes in the network. Data on outcomes and socio-economic characteristics of non-sampled nodes are crucial in order to be able to implement many of the identification strategies discussed in Section 3 above. Moreover, as we will see below, this information is also useful for correcting for measurement error in the network. Recent analyses with networks data in the economics literature have featured datasets with edges collected from random samples of nodes. Examples include data on social networks and the diffusion of microfinance used by both Banerjee et al. (2013) and Jackson et al. (2012); and

[92] This is a feature of some commonly used datasets, including the popular National Longitudinal Study of Adolescent Health (AddHealth) dataset.

[93] For instance, Facebook has over 1 billion monthly users, while Twitter reports having around 200 million regular users.

[94] 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. The network constructed from data where nodes are randomly sampled and all their 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), is called a star subgraph.

data on voting and social networks used in Fafchamps and Vicente (2013).
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.

2. 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, so that one would get an insufficiently large sample through random sampling from the population *e.g.* sex workers. Link tracing is usually used to collect data from vast online social networks. 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 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, which is needed to apply the methods outlined in Section 3 above. Examples in economics of datasets collected by snowball sampling include that of student migrants used in Méango (2014).

The sampling method used has important implications for how accurately the network graph and its features are measured. In the next subsection we will discuss some of the common measurement errors arising from the above methods (as well as measurement error from non-sampling sources), their implications for model parameters, and methods for overcoming these often substantial biases.

5.3 Sources of Measurement Error

An important challenge that complicates identification of parameters using overlapping peer groups and detailed network data is the issue of measurement error. 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) miscoding and misreporting errors, (5) spurious nodes and (6) non-response. We refer to the first three of these as sampling-induced error, and the latter three as non-sampling error. It is important to account for this, since as we will show in this Subsection, measurement error can induce important biases in measures of network statistics and in parameter estimates.

Measurement error issues arising from sampling are very important in the context of networks data, since these data comprise 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 a particular sampling distribution over one will induce a particular (non-random) structure for sampling over the other.[95] This means that econometric and statistical methods for estimation and inference developed under classical sampling theory are often not applicable to networks data, since many of the underlying assumptions fail to hold. Consequently the use of standard techniques, without adjustments for the specific features of network data, leads to errors in measures of the network, and hence biases model parameters.

In practice, however, 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). Frequently, it is too expensive and cumbersome to collect data on the whole network. 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 exercises, 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 5.3.1 outlines the consequences of missing data due to sampling on estimates of social effects arising from outcomes of network neighbours (such as those considered in Subsections 3.2, 3.3 and 3.4) and network statistics (as in Subsection 3.5). Until recently most research into these issues was done outside economics, so we draw on research from a range of fields, including sociology, statistical physics, and computer science.

Measurement error arising from the other three sources – misreporting or miscoding errors, spurious nodes, and non-response – which we label as non-sampling measurement error, can also generate large biases in network statistics and parameters in network models. Though there is a large literature on these types of measurement error in the econometrics and statistics (see, for example, Chen et al. (2011) for a summary of methods for dealing with misreporting errors in binary variables, also known as misclassification errors), these issues has been less studied in a networks context. Subsection 5.3.2 below summarises findings from this literature.

Finally, a number of methods have been suggested to help deal with the consequences of measurement error, whether due to sampling or otherwise. Subsection 5.4 outlines the various methods that have been developed for this purpose.

[95] We consider a random sample to consist of units that are independent and identically distributed.

5.3.1 Measurement Error Due to Sampling

Node-Specific Neighbourhoods Collecting only a sample of data, rather than a complete census, can lead to biased and inconsistent parameter estimates in social effect models. This is because sampling of the network leads to misspecification of nodes' neighbours. In particular, a pair of nodes in the sampled network may appear to be further away than they actually are. Recall from Section 3 that with observational data, methods for identifying the social effects parameters in the local average, local aggregate and hybrid local model use the exogenous characteristics of direct, second- and, in some cases, third-degree neighbours as instrumental variables for the outcomes of a node's neighbours. Critically, these methods require us to know which edges are definitely *not* present to give us the desired exclusion restrictions. Misspecification of nodes' direct and indirect (*i.e.* second- and third-degree) neighbours may consequently result in mismeasured and invalid instruments.

Chandrasekhar and Lewis (2011) show that this is indeed the case for the local average model, where the instruments are the average characteristics of nodes' second- and third-degree neighbours. The measurement error in the instruments is correlated with the measurement error in the endogenous regressors, leading to bias in the social effect estimates. Simulations in their paper suggest that these biases can be very large, with the magnitude falling as the proportion of the network sampled increases, and as the number of networks in the sample increases.[96] Chandrasekhar and Lewis (2011) offer a simple solution to this problem when (i) network information is collected via a star subgraph – *i.e.* where a subset of nodes is randomly sampled ('sampled nodes') and all their edges are included in constructing the network graph; and (ii) data on the outcome and exogenous characteristics are available for all nodes in the network, or at least for the direct and second- and potentially third-degree neighbours of the 'sampled' nodes. In this case, all variables in the second stage regression (i.e. Equation 3.6) are correctly measured for the 'sampled' nodes, since for any node, the regressors, $\tilde{\boldsymbol{G}}_{i,g}\boldsymbol{Y}_g = \sum_{j \in nei_{i,g}} \tilde{G}_{ij,g} y_{j,g}$ and $\tilde{\boldsymbol{G}}_{i,g}\boldsymbol{X}_g = \sum_{j \in nei_{i,g}} \tilde{G}_{ij,g} \boldsymbol{x}_{j,g}$, are fully observed. Including only sampled nodes in the second stage thus avoids issues of erroneously assuming that nodes in the observed network are further away from one another than they actually are. The influence matrix constructed with the sampled network is, however still mismeasured, leading to measurement error in the instruments (which use powers of this matrix), and thus in the first stage. However, this measurement error is uncorrelated with the second stage residual, thus satisfying the IV exclusion restriction. Note though that the measurement error in the instruments reduces their informativeness (strength), particularly when the sampling rate is low. This is because this strategy requires the existence of nodes that have a (finite) geodesic of at least 2 or 3 between them. At low sampling rates there will be very few such pairs of nodes, since many sampled nodes will seem completely unconnected as the nodes that connect them will be missing from the data.

A similar issue applies to local aggregate and hybrid models. Simulations in Liu (2013) show that

[96] A limitation of these simulations is that the authors only considered simulations with either 1 or 20 networks. It is unclear how large such biases may be when a large number (*e.g.* 50) of networks is available.

parameters of local aggregate models are severely biased and unstable when estimated with partial samples of the true network. In this model, however, as shown in Subsection 3.3, a node's degree can be used as an instrument for neighbours' outcomes. When the sampled data take the form of a star subgraph, the complications arising from random sampling of nodes can be circumvented by using the out-degree, which is not mismeasured, as an instrument for the total outcome of edges. This allows for the consistent estimation of model parameters. This is supported by simulation evidence in Liu (2013), which shows that estimates of the local aggregate model computed using out-degrees as an additional instrument are very close to the parameters of a pre-specified data generating process. Other possible ways around this problem include the model-based and likelihood-based corrections outlined in Subsection 5.4.

Network Statistics Missing data arising from partial sampling generate non-classical measurement error in measured network statistics. This is an important issue in estimating the effects of network statistics on outcomes using regressions of the form seen in Subsection 3.5, because measurement error leads to substantial bias in model parameter estimates. A number of studies, primarily in fields outside economics, have investigated the consequences and 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.* We illustrate the severity of this issue using a stylised example. Consider the network in panel (a) of Figure 4, which contains 15 nodes and has an average degree of 3.067. We sample 60%, 40% and 20% of nodes and elicit information on all their edges (*i.e.* we elicit a star subgraph). The resulting network graphs are plotted in panels (b), (c) and (d), with the unshaded nodes being those that were not sampled. Average degree is calculated based on all nodes and edges in the star subgraph, *i.e.* including all sampled nodes, the edges they report, and nodes they are linked with.[97] When only 20% of nodes are sampled, the average degree of the sampled graph is 2, which is around 35% lower than the true average degree.[98] However, when a higher proportion of nodes are sampled, average degree of the sampled graph becomes closer to that of the true graph. More generally, simulation evidence[99] from studies including Galaskiewicz (1991), Costenbader and Valente (2003), Lee et al. (2006), Kim and Jeong (2007) and Chandrasekhar and Lewis (2011) have estimated the magnitude of sampling induced bias in statistics such as

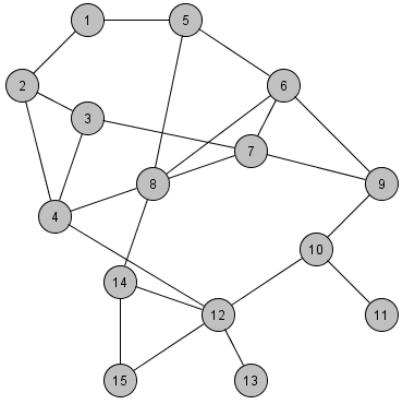
[97] This is equivalent to taking an average of the row-sums of the (undirected) adjacency matrix constructed from the sampled data, in which two nodes are considered to be connected if one reports an edge. This is a common way of constructing the adjacency matrix in empirical applications. However, for data collected through star subgraph sampling, an accurate estimate of average degree can be obtained by including only the sampled nodes in the calculation.

[98] We will discuss methods that allow one to correct for this bias in Subsection 5.4.

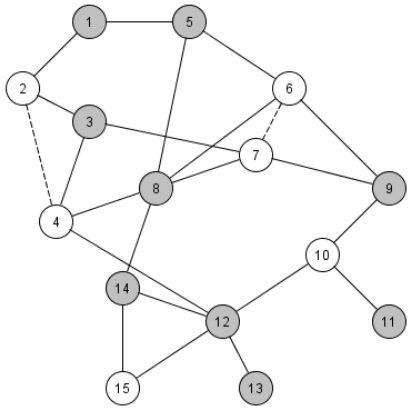
[99] Simulations are typically conducted by taking the observed network to be the true network, and constructing 'sampled' networks by drawing samples of different sizes using various sampling methods.

degree (in-degree and out-degree in the directed network case), degree centrality, betweenness centrality, eigenvector centrality, transitivity (also known as local clustering), and average path length. They find biases that are very large in magnitude, and the direction of the bias varies depending on the statistic. For example, the average path length may be over-estimated by 100% when constructed from an induced subgraph with 20% of nodes in the true network. This concern is particularly relevant for work in the economics literature: a literature review of studies in economics by Chandrasekhar and Lewis (2011) reports a median sampling rate of 25% of nodes in a network. Table 1 below summarises findings from these papers for various commonly used network statistics.

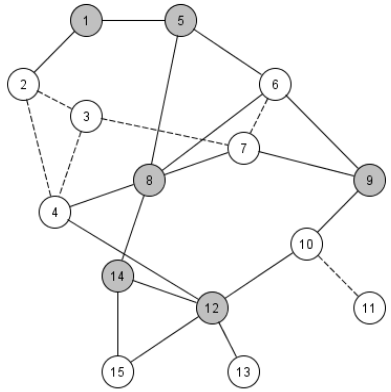
Figure 4: Sampled networks with different sampling rates



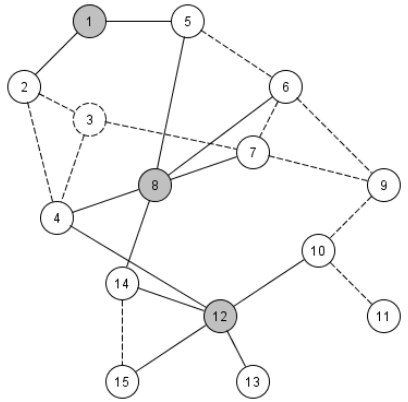
Nodes in sample = 15
Edges in sample = 46
Average degree = 3.067
(a) Full Graph



Nodes in sample = 15
Edges in sample = 42
Average degree = 2.800
(b) 60% of nodes sampled



Nodes in sample = 13
Edges in sample = 34
Average degree = 2.615
(c) 40% of nodes sampled

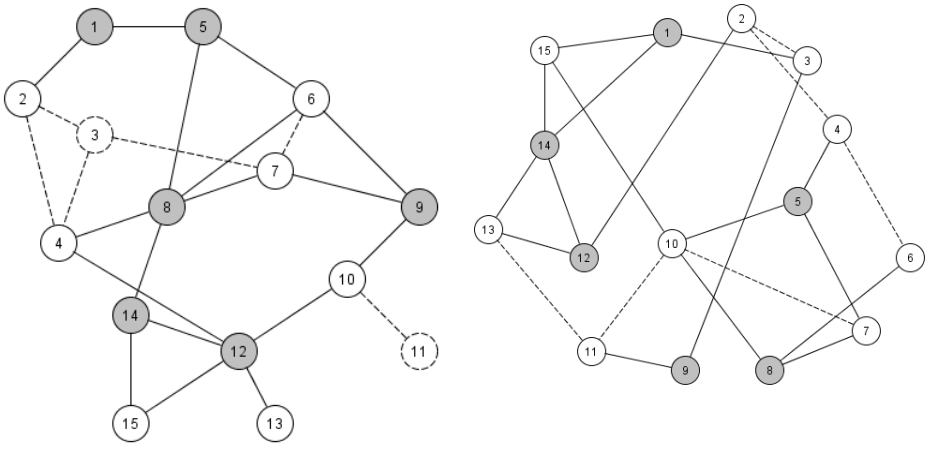


Nodes in sample = 12
Edges in sample = 24
Average degree = 2
(d) 20% of nodes sampled

Notes to Figure: This figure displays the full graph (panel (a)), and the star subgraphs obtained from sampling 60% (panel (b)), 40% (panel (c)) and 20% (panel (d)) of nodes. The unshaded nodes in panels (b), (c) and (d) represent nodes that were not sampled, and the dotted lines represent nodes and edges on which no data were collected. Though the average degree in the original graph is 3.067, that in the sampled graphs ranges from 2.8 to 2. The # Nodes, and # Edges indicated in the figure refer to the numbers included in the calculation of the displayed average degree.

2. *Measurement error due to sampling varies with the underlying network topology (i.e. 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 topologies: 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 topology: small world networks, which have relatively high clustering and 'bridging' edges that reduce path lengths between nodes that would otherwise be far away from one another, are especially vulnerable to missing data. This is not surprising since key nodes that are part of a bridge could be missed in the sample and hence give a picture of a less connected network. By contrast, scale-free networks are less vulnerable to missing data. Such effects are evident even in the simple stylised example in Figure 5 below, where we sample the same nodes from networks with different topologies – uniform random, and small world. Though each network has the same average degree,[100] and the same number of nodes is sampled in both cases, the average degree in the graph sampled from the uniform random network is closer to the true value than that sampled from the small world network.

Figure 5: Sampling from uniform random and small world networks



Nodes in sample = 13
Edges in sample = 34
True average degree = 3.067;
Sampled avg. Degree = 2.615
(a) Uniform Random Graph

Nodes in sample = 15
Edges in sample = 34
True average degree = 3.067
Sampled avg. Degree = 2.267
(b) Small world graph

Notes to Figure: This figure displays the star subgraphs obtained from sampling 40% of nodes in a network with a uniform random topology (panel (a)) and a small world topology (panel(b)). The unshaded nodes represent nodes that were not sampled, and the dotted lines represent nodes and edges on which no data were collected.

[100] As in (1) above, average degree is calculated from the adjacency matrix with all nodes and edges in the sample (*i.e.* all the nodes and edges with firm lines).

3. *The magnitude of error in network statistics due to sampling varies with the sampling method.* Different sampling methods result in varying magnitudes of errors in network statistics. Lee et al. (2006) compare data sampled via induced subgraph sampling, random sampling of nodes, random sampling of edges, and snowball sampling, from networks with a power-law degree distribution.[101] They 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 an over-estimation of the size of the exponent of the power-law degree distribution.[102] Conversely, snowball sampling, which is less likely to find nodes with low degrees, underestimates this exponent. We illustrate this fact further using a simple example that compares two node sampling methods common in data used by economists – *induced subgraph*, where only edges between sampled nodes are retained; and *star subgraph*, in which all edges of sampled nodes are retained regardless of whether or not the nodes involved in the edges were sampled. Consider again the network graph considered in panel (a) of Figure 4 above, and displayed again in panel (a) of Figure 6 below. We sample the same set of nodes – 1, 5, 8, 9, 12, and 14 – from the full network graph. Panels (b) and (c) of Figure 6 display the resulting network graphs under star and induced subgraph sampling respectively. Though the proportion of the network sampled is the same under both types of sampling, the resulting network structure is very different. This is reflected in the estimated network statistics as well: the average degree for the induced subgraph is just over a half of that for the star subgraph, which is not too different from the average degree of the full graph.[103]

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.[104] 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.

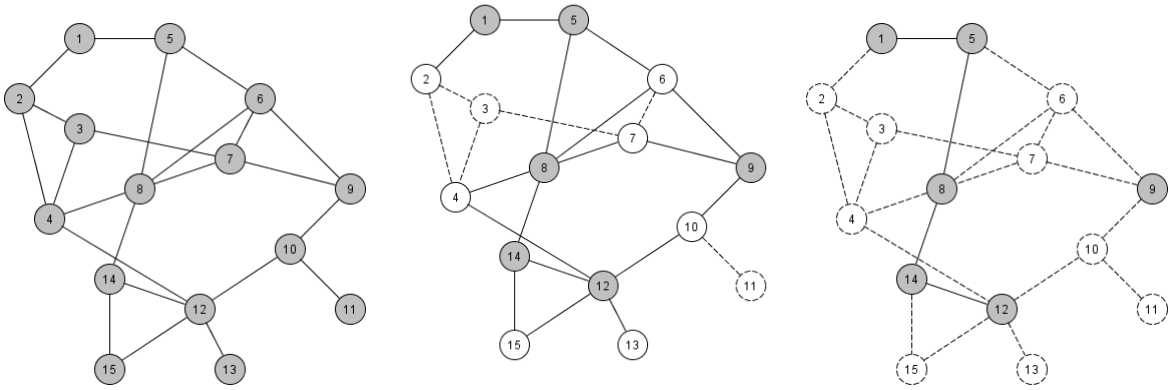
[101] Power law degree distributions are those where the fraction of nodes having k edges, $P(k)$ is asymptotically proportional to $k^{-\gamma}$, where usually $2 < \gamma < 3$. Such a distribution allows for fat tails, *i.e.* the proportion of nodes with very high degrees constitutes a non-negligible proportion of all nodes.

[102] A larger exponent on the power law degree distribution indicates a greater number of nodes with large degrees.

[103] Average degree is calculated as above, including all nodes and edges in the sample, *i.e.* those with firm lines in Figure 6.

[104] Simulations typically report bias in parameters from models where the outcome variable is a linear function of the network statistic.

Figure 6: Sampling with star and induced subgraphs



Nodes in sample = 15
Edges in sample = 46
Average degree = 3.067
(a) Full Graph

Nodes in sample = 13
Edges in sample = 34
Average degree = 2.615
(b) Star Subgraph

Nodes in sample = 6
Edges in sample = 8
Average degree = 1.333
(c) Induced Subgraph

Notes to Figure: Panel (a) of the figure displays the true network graph and panels (b) and (c) display the star and induced subgraph obtained when the darker-shaded nodes are sampled. The unshaded nodes in panels (b) and (c) represent nodes that were not sampled, and the dotted lines represent nodes and edges on which no data were collected. In the star subgraph, an edge is present as long as one of the two nodes involved in the edge is sampled. This is not the case in the induced subgraph, where an edge is present only if both nodes involved in the edge are sampled.

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 in reported edges and boundary specification on network statistics such as average degree, clustering and average path length. Both types of error cause average degree to be under-estimated, while average path length is over-estimated. No bias arises in the estimated clustering parameter if the consequence of the error is to simply limit the number of edges of each node.

Tables 1 and 2 below summarises findings on the consequences of missing data for both estimates of network statistics and parameter estimates when using data on networks collected through random sampling of nodes. We consider two types of graph induced by data collected via random node sampling: induced subgraph, and star subgraph, which are as shown in Figure 6 above.

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]
Spectral gap	Direction of bias ambiguous (±); depends on the relative magnitudes of bias in the first and second eigenvalues, both of which are attenuated.[a]	Direction of bias ambiguous (±): depends on the relative magnitudes of bias in the first and second eigenvalues, both of which are attenuated.[a]
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: Non-negligible, or little bias refers to |bias| of 0-20%, large bias to |bias| of 20%-50% and very large bias to |bias| > 50%. [a] Source: Chandrasekhar and Lewis (2011); [b] Source: 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.[a]	Degree (in undirected graphs) of highly connected nodes is underestimated (–).[b]
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.[b]
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%.[a]	Shape of the distribution relatively well estimated. Ranking in distribution much worse, *i.e.* nodes with high betweenness centrality appear to have low centrality.[d]
Eigenvector Centrality	Very low correlation between vector of true node eigenvector centralities and that from sampled graph.[a]	Not known.

Notes: Source: [a]Costenbader and Valente (2003);[b]Source: Lee et al. (2006); [c]Source: 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.
Spectral gap	Attenuated (–), with bias falling with sampling rate. Magnitude of bias large even when 50% of nodes are sampled.	Attenuated (–) (more than star subgraphs). Magnitude of bias very large and falls with the sampling rate.
Clustering Coefficient	Scaling (+) and attenuation (–); |scaling| > |attenuation|. Very large biases, which fall with sampling rate.	Attenuation (–), falls with sampling rate. Magnitude of bias non-negligible 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: Non-negligible bias refers to |bias| of 0-20%, large bias to |bias| of 20%-50% 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: Large bias refers to |bias| of 20%-50% and very large bias to |bias| $>$ 50%. Source: Chandrasekhar and Lewis (2011)

5.3.2 Other Types of Measurement Error

Beyond sampling-induced measurement error, networks could be mismeasured for a variety of other reasons including:

1. Miscoding and misreporting errors: Edges could be miscoded, either because of respondent or interviewer error: respondents may forget nodes or interview fatigue may lead them to misreport edges. In some cases, there may be strategic reporting of edges, *e.g.*, respondents may report desired rather than actual edges, as in Comola and Fafchamps (2014).

2. Spurious Nodes: Spelling mistakes in node names or multiple names for the same nodes can lead to the presence of spurious nodes. This is a concern when edges are inferred from existing data.

3. Non-response: Edges are missing as a result of non-response from nodes.

Wang et al. (2012) consider, in a simulation study, the consequences of these types of measurement error on network statistics including degree centrality, the clustering coefficient and eigenvector centrality. They find that degree centrality and eigenvector centrality are relatively robust to measurement error arising from spurious nodes and miscoded edges, while clustering coefficient is biased by mismeasured data. Though there is a large literature on these types of measurement error in the econometrics and statistics (see, for example, Chen et al. (2011) for a summary of methods for dealing with misreporting errors in binary variables, also known as misclassification errors), these issues has been less studied in a networks context. An exception is Comola and Fafchamps (2014), who propose a method for identifying and correcting misreported edges.

5.4 Correcting for Measurement Error

Ex-post (*i.e.* once data have been collected) methods of dealing with measurement error can be divided into three broad classes: (1) design-based corrections, (2) model-based corrections, and (3) likelihood-based corrections. Design-based corrections apply primarily to correcting sampling-induced measurement error, while model-based and likelihood-based corrections can apply to both sampling-induced and non-sampling-induced measurement error. We briefly summarise the underlying ideas behind each of these, discussing some advantages and drawbacks of each.

5.4.1 Design-Based Corrections

Design-based corrections rely on features of the sampling design to correct for sampling-induced measurement error (Frank 1978, 1980a, 1980b, 1981; Thompson, 2006).[105] They are based on

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

Horvitz-Thompson estimators, which use inverse probability-weighting to compute unbiased estimates of population totals and means from sampled data. This method can be applied to correct mismeasured network statistics that can be expressed as totals, such as average degree and clustering. We illustrate how Horvitz-Thompson estimators work using a simple example.

A researcher has data on an outcome y for a sample of n units drawn from the population. Under the particular sampling scheme used to draw this sample, each unit i in the population $U = \{1, ..., N\}$ has a probability p_i of being in the sample. The researcher wants to use the sample to compute an estimate of the sum of y in the population, $\tau = \sum_{i \in U} y_i$. The Horvitz-Thompson estimator for this total can be computed by summing the y's for the sampled units, weighted by their probability of being in the sample. That is, $\hat{\tau}_p = \sum_{i \in U} \frac{y_i}{p_i}$. Essentially, the estimator computes an inverse probability-weighted estimate to correct for bias arising from unequal probability sampling. In the case of network statistics, this thus corrects for the non-random sampling of either nodes or edges induced by the particular sampling scheme. The key to this approach is the construction of the sample inclusion weights, p_i.

Formulae for node- and edge-inclusion probabilities are available for the random node and edge sampling schemes (see Kolaczyk (2009) for more details). Recovering sample inclusion probabilities when using snowball sampling 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 graph parameters such as dyad and triad counts, degree distribution, average degree, and clustering under random sampling of nodes. Chandrasekhar and Lewis (2011) show that parameter estimates in network regressions using design-based corrected network statistics as regressors are consistent for three statistics: average degree, clustering coefficient, and average graph span. Their results show that the Horvitz-Thompson estimators can correct for sampling-induced measurement error. Numerical simulations suggest that this method reduces greatly, and indeed eliminates at sufficiently high sampling rates, the sampling induced bias in parameter estimates.

There are two drawbacks of this procedure. First, 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 statistics of interest for economists. Second, they can't be used to correct for measurement error arising from reasons other than sampling (unless the probability of correct reporting is known). Model-based and likelihood-based corrections can, by placing more structure on the measurement error problem, offer alternative ways of dealing with measurement error in these cases.

5.4.2 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 and have primarily been used to correct for measurement error arising from sampling related reasons. Thus the model is typically a network formation model of the type seen in Subsection 4.1 above. Parameters of the network formation model are estimated from the partially observed network, and available data on the identities and characteristics of nodes and edges; with the estimated parameters subsequently used to predict missing edges (in-sample edge prediction). Note that it is crucial to have information on the identities and, if possible, the characteristics (*e.g.* gender, ethnicity, *etc.*) of all nodes in the network. This is important from a data requirements perspective. Without this information, it is not possible to use this method to correct for measurement error.

In most economics applications, researchers would typically want to use the predicted networks to subsequently identify social effect parameters using models similar to those in Section 3 above. Chandrasekhar and Lewis (2011) show that the network formation model must satisfy certain conditions in order to allow for consistent estimation of the parameters of social effects models such as those discussed in Section 3.

They study a setting where data on the network is assumed to be missing at random, and where the identities and some characteristics of all nodes are observed. Data are assumed to be available for multiple, possibly large networks. This is necessary since in their results the rate of convergence of the estimated parameter to the true parameter depends on both the number of nodes within a network, and the number of networks in the data. Their analysis shows that consistent estimation of social effect parameters is possible with network formation models similar to those outlined in Section 4.1 above, as long as the interdependence between the covariates of pairs of nodes decays sufficiently fast with network distance between the nodes. This may not be satisfied for instance, in a model where a network statistic (such as degree distribution) is a sufficient statistic for the network formation process. In this case, Chandrasekhar and Lewis (2011) show that parameters of the network formation process do not converge sufficiently fast to allow for consistent estimation of the social effect parameters in models at the node-level (*e.g.* Equation 3.1), though parameters of network-level models, such as Equation 3.5 can be consistently estimated. Their analysis also shows that network formation processes that allow for specific network effects in edge formation (*i.e.* some strategic models of network formation such as the model of Christakis et al., 2010) also satisfy conditions under which the social effect parameter can be consistently estimated.

5.4.3 Likelihood-Based Corrections

Likelihood-based corrections can be applied to correct for measurement error when only a sub-sample of nodes in a network are observed. Such methods have, however, been used to correct specific network-based statistics such as out-degree and in-degree, but may not apply to other

statistics. Here, we discuss two likelihood-based methods to correct for measurement error: the first method from Conti et al. (2013), corrects for sampling related measurement error when data is available only for sampled nodes; while the second has been proposed and applied by Comola and Fafchamps (2014) to correct for misreporting.

Conti et al. (2013) correct for non-classical measurement error in in-degree arising from random sampling of nodes by adjusting the likelihood function to account for the measurement error. The method involves first, specifying the process for outgoing and incoming edge nominations, and as a result obtaining the outgoing and incoming edge probabilities. Specifically, Conti et al. (2013) 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 (to capture homophily) and a scalar unobservable for i and j. Moreover, the process allows 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.[106]

Comola and Fafchamps (2014) propose a maximum likelihood based framework to correct for measurement error arising from misreporting by nodes of their neighbours and/or flows across the edges. To illustrate this method, we take the case of binary edges. In survey data, where nodes are asked to declare the presence or not of an edge with other nodes, misreporting could mean that one of two nodes in any edge omits to report the edge; or both forget to report the edge even if it exists, or both report an edge when it doesn't exist or, one of the two nodes erroneously reports an edge when it doesn't exist. Misreporting in this case is a form of misclassification error. Assuming that the misreporting process is such that either nodes forget to declare neighbours, or they spuriously report neighbours, it is possible to use a maximum likelihood framework to correct for this misreporting bias. By assuming a statistical process for edges (*e.g.* Comola and Fafchamps (2014) assume that edges follow a logistic process, and are a function of observed characteristics), and given that the mismeasured variable is binary, it is possible to write down a likelihood function that incorporates the measurement error. Maximising this function provides the correct parameter estimates for the edge formation process, which can then be used to correct for misreporting.

[106] Conti et al. (2013) also account for censoring by using a truncated distribution in the likelihood function.

6 Conclusion

Networks can play an important role both as a substitute for incomplete or missing markets and a complement to markets, for example, by transmitting information, or even preferences. Whether such effects exist in practice is an important empirical question, and recent work across a range of fields in economics has tried to provide some evidence about this. However, working with networks data creates important challenges that are not present in other contexts.

In this paper we outline econometric methods for working with network data that take account of the peculiarities of the dependence structures present in this context. It divides the issues into three parts: (i) estimating social effects given a conditionally exogenous observed network; (ii) estimating the underlying network formation process, given only a single cross-section of data; and (iii) accounting for measurement error, which in a network context can have particularly serious consequences.

When data are available on only agents and the reference groups to which they belong, researchers have for some time worried about how social effects might be identified. However, when detailed data on nodes and their individual links are present, identification of social effects (taking the network as conditionally exogenous) is generic, and estimation is relatively straightforward. Two broader conceptual issues exist in this case: First, theory is often silent on the precise form that peer effects should take when they exist. Since Manski (1993), many people have focused on the 'local average' framework, often without discussion of the implications for economic behaviour, but social effects might instead take a local aggregate, or indeed local maximum/minimum form where the best child in a classroom provides a good example to all others, or the worst disrupts the lesson. Until a non-parametric way of allowing for social effects is developed, researchers need to use theory to guide the empirical specification they use. Second, researchers typically treat the observed network as the network which mediates the social effect, and where many networks are observed the union of these is taken. Given what we know about measurement error in networks, this behaviour will generally create important biases in results, if the relevant network is a network defined by a different kind of relationship, or is actually some subset of the union taken. Here again it is important that some justification is given for why the network used should be the appropriate one.

In addition to these conceptual issue, the key econometric challenge in identifying social effects is allowing for network endogeneity. In recent years there have been attempts to account directly for network endogeneity. A natural first direction for this work has been to use exclusion restrictions to provide an instrument for the network structure. As ever, this requires us to be able to credibly argue that there is some variable that indirectly affects the outcome of interest, through its effect on the network structure, but has no direct effect. Whether this seems reasonable will depend on the circumstance, but an important issue here is that the network formation process must have a unique equilibrium for these methods to be valid.

This leads naturally to a discussion of network formation models that can allow for dependence between links. Drawing from work in a number of fields, this paper brings together the main estimation methods and assumptions, describing them in a common language. Although other fields have modelled network formation for some time, and developed methods to estimate parameters, they are often unsuitable when we treat the data as observations of decisions made by optimising agents. There is still much scope in this area to develop more general methods and results which do not rely on strong assumptions about the structure of utility functions or meeting processes in order to achieve identification.

Finally, the paper discussed data collection and 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.

References

V. Alatas, A. Banerjee, A. G. Chandrasekhar, R. Hanna, and B. A. Olken. "Network Structure and the Aggregation of Information: Theory and Evidence from Indonesia". *NBER Working Paper*, WP 18351, 2012.

A. Ambrus, M. Mobius, and A. Sziedl. "Consumption Risk-Sharing in Social Networks". *American Economic Review*, 140:149–182, 2014.

M. Angelucci, G. De Giorgi, M. A. Rangel, and I. Rasul. "Family Networks and School Enrolment: Evidence from a Randomized Social Experiment". *Journal of Public Economics*, 94:197–221, 2010.

M. Angelucci, G. De Giorgi, and I Rasul. "Resource Pooling Within Family Networks: Insurance and Investment". *mimeo, University College London*, 2012.

J. Angrist. "The Perils of Peer Effects". *NBER Working Paper*, WP 19774, 2013.

F. A. Azevedo, L. R. Carvalho, L. T. Grinberg, J. M. Farfel, R. E. Ferretti, R. E. Leite, W. Jacob Filho, R. Lent, and S. Herculano-Houzel. "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain". *Journal of Computational Neurology*, 513:532–41, 2009.

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

C. Ballester, A. Calvó-Armengol, and Y. Zenou. "Who's who in networks. Wanted: the key player". *Econometrica*, 74:1403–1417, 2006.

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

J. Besag. "Spatial Interaction and the Statistical Analysis of Lattice Systems". *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.

J. Besag. "Statistical Analysis of Non-Lattice Data". *The Statistician*, 24:179–195, 1975.

S. Bhamidi, G. Bresler, and A. Sly. "Mixing Time of Exponential Random Graphs". Arxiv preprint arXiv:0812.2265, 2008.

A. Bisin, A. Moro, and G. Topa. "The Empirical Content of Models with Multiple Equilibria in Economies with Social Interactions". *NBER Working Paper*, WP 17196, 2011.

F. Bloch, G. Genicot, and D. Ray. "Informal Insurance in Social Networks". *Journal of Economic Theory*, 143(1):36–58, 2008.

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.

V. Boucher and I. Mourifié. "My Friend Far Far Away: Asymptotic Properties of Pairwise Stable Networks". Technical report, Working Papers, University of Toronto Dept. of Economics, 2013.

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.

Y. Bramoulle and R. Kranton. "Public Goods in Networks". *Journal of Economic Theory*, 135(1): 478–494, 2007.

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.

A. G. Chandrasekhar and M. O. Jackson. "Tractable and Consistent Exponential Random Graph Models"'. *NBER working paper*, WP 20276, 2014.

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

S. Chatterjee, P. Diaconis, and A. Sly. "Random Graphs with a Given Degree Sequence". Arxiv preprint arXiv:1005.1136, 2010.

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

S. Chinchalkar. "An Upper Bound for the Number of Reachable Positions". *ICCA Journal*, 19: 181–183, 1996.

N. A. Christakis, J. H. Fowler, G. W. Imbens, and K. Kalyanaraman. "An Empirical Model for Network Formation". *NBER Working Paper*, WP 16039, 2010.

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.

T. G. Conley and C. R. Udry. "Learning About a New Technology: Pineapple in Ghana". *American Economic Review*, 100:35–69, 2010.

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.

S. Currarini, M. O. Jackson, and P. Pin. "An Economic Model of Friendship: Homophily, Minorities, and Segregation". *Econometrica*, 77:1003–1045, 2009.

S. Currarini, M. O. Jackson, and P. Pin. "Identifying the Roles of Race-based Choice and Chance in High School Friendship Network Formation". *Proceedings of the National Academy of Sciences of the USA*, 107:4857–4861, 2010.

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, April 2010. URL `http://ideas.repec.org/a/aea/aejapp/v2y2010i2p241-75.html`.

A. de Paula. "Econometric Analysis of Games with Multiple Equilibria". *Annual Review of Economics*, 5:107–131, 2013.

A. de Paula, S. Richards-Shubik, and E. Tamer. "Identification of Preferences in Network Formation Games". 2014.

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

L. Ductor, M. Fafchamps, S. Goyal, and M. van der Leij. "Social Networks and Research Output". *Review of Economics and Statistics*, Forthcoming, 2014.

E. Duflo, P. Dupas, and M. Kremer. "Peer Effects and the Impacts of Tracking: Evidence from a Randomized Evaluation in Kenya". *American Economic Review*, 101:1739–1774, 2011.

A. Dzemski. "An empirical model of dyadic link formation in a network with unobserved heterogeneity". 2014.

P. Erdös 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 P. Vicente. "Political Violence and Social Networks: Experimental Evidence". *Journal of Development Economics*, 101(C):27–48, 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.

O. Frank and D. Strauss. "Markov Graphs". *Journal of the American Statistical Association*, 81: 832–842, 1986.

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.

C. Geyer and E. Thompson. "Constrained Monte Carlo maximum likelihood for dependent data". *Journal of the Royal Statistical Society, Series B*, 54 (3):657–699, 1992.

E. N. Gilbert. "Random Graphs". *Annals of Mathematical Statistics*, 30:1141–1144, 1959. doi: doi:10.1214/aoms/1177706098.

E. Glaeser, B. Sacerdote, and J. Scheinkman. Crime and social interactions. *Quarterly Journal of Economics*, 115:811–846, 1996.

D. Goldberg and F. Roth. "Assessing experimentally derived interactions in a small world". *Proceedings of the National Academy of Sciences of the USA*, 100 (8):4372–4376, 2003.

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. "Identifying Social Interactions through Conditional Variance Restrictions". *Econometrica*, 76:643–660, 2008.

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

M. S. Granovetter. "The Strength of Weak Ties". *American Journal of Sociology*, 78:1360–1380, 1973.

G. R. Grimmett. "A theorem about random fields". *Bulletin of the London Mathematical Society*, 5:81–84, 1973.

J. Hammersley and P. Clifford. "Markov fields on finite graphs and lattices". 1971.

M. S. Handcock. "Assessing Degeneracy in Statistical Models of Social Networks". Technical report, CSSS Working Paper no. 39, 2003. URL `http://www.csss.washington.edu/Papers/wp39.ps`.

P. Hoff. "Multiplicative latent factor models for description and prediction of social networks". *Computational and Mathematical Organization Theory*, 15 (4):261–272, 2009.

P. W. Holland and S. Leinhardt. "Notes on the Statistical Analysis of Network Data". 1977.

P. W. Holland and S. Leinhardt. "An Exponential Family of Probability Distributions for Directed Graphs". *Journal of the American Statistical Association*, 76:33–50, 1981.

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

L. J. Hubert. *"Assignment Methods in Combinatorial Data Analysis"*. Marcel Dekker, 1987.

L. J. Hubert and J. Schultz. "Quadratic Assignment as a General Data Analysis Strategy". *British Journal of Mathematical and Statistical Psychology*, 29:190–241, 1976.

M. O. Jackson and A. Wolinsky. "A Strategic Model of Social and Economic Networks". *Journal of Economic Theory*, 71:44–74, 1996.

M.O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.

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.

H. H. Kelejian and G. Piras. "Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes". *Regional Science and Urban Economics*, 46:140–149, 2014.

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.

D. Krackhardt. "Predicting with Networks: A Multiple Regression Approach to Analyzing Dyadic Data". *Social Networks*, 10:359–381, 1988.

H. H. Kwok. "Identification problems in linear social interaction models: a general analysis based on matrix spectral decompositions". 2013.

L-F. Lee. "Identification and Estimation of Econometric Models with Group Interactions, Contextual Factors and Fixed Effects". *Journal of Econometrics*, 140:333–374, 2007.

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

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

M. Leung. "Two-Step Estimation of Network-Formation Models with Incomplete Information". *mimeo, Stanford University*, 2014.

X. Liu. "Estimation of a local-aggregate network model with sampled networks". *Economics Letters*, 118:243–246, 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?". *Unpublished Manuscript*, 2014b.

R. Méango. "International Student Migration: A Partial Identification Analysis". *Available at SSRN: http://ssrn.com/abstract=2392732 or http://dx.doi.org/10.2139/ssrn.2392732*, 2014.

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

C. Manski. "Identification of Treatment Response with Social Interactions". *Econometrics Journal*, 16:S1–S23, 2013.

N. Mantel. "The Detection of Disease Clustering and a Generalised Regression Approach". *Cancer Research*, 27:209–220, 1967.

D. Marmaros and B. Sacerdote. "How do Friendships Form". *Quarterly Journal of Economics*, 121(1):79–119, 2006.

A. Mayer and S. L. Puller. "The Old Boy (and Girl) Network: Social network formation on university Campuses". *Journal of Public Economics*, 92:329–347, 2008.

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

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

R. Moffitt. "Policy interventions, low-level equilibria, and social interactions". In S. Durlauf and H. P. Young, editors, *Social Dynamics*, pages 45–82. MIT Press, Cambridge, 2001.

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

E. Patacchini and Y. Zenou. "Juvenile Delinquency and Conformism". *Journal of Law, Economics and Organization*, 1:1–31, 2012.

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

A. Popescul and L. Ungar. "Statistical relational learning for link prediction". *Proceedings of the Workshop on Learning Statistical Models from Relational Data at IJCAI-2003*, 2003.

C. J. Preston. "Generalised Gibbs states and Markov random fields". *Advances in Applied Probability*, 5:242–261, 1973.

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.

B. Schutz. *"Gravity from the ground up"*. Cambridge Univ Press, 2003.

S. Sheng. "Identification and Estimation of Network Formation Games". Job Market Paper, 2012.

S. Sherman. "Markov random fields and Gibbs random fields". *Israel Journal of Mathematics*, 14: 92–103, 1973.

T. A. B. Snijders. "Markov Chain Monte Carlo Estimation of Exponential Random Graph Models". *Journal of Social Structure*, 3:1–40, 2002.

D. Strauss and M. Ikeda. "Pseudolikelihood Estimation for Social Networks". *Journal of the American Statistical Association*, 85:204–212, 1990.

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

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.

R. Townsend. "Risk and Insurance in Village India". *Econometrica*, 62(3):539–591, 1994.

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.

D. J. Wang, X. Shi, D. McFarland, and J. Leskovec. "Measurement error in network data: A re-classification". *Social Networks*, 34(4):396–409, 2012.

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

Here we provide an index of definitions for the different network representations and summary statistics used.

- **Adjacency Matrix:** This is an $N \times N$ matrix, $\boldsymbol{G}$, whose ij^{th} element, G_{ij}, represents the relationship between node i and node j in the network. In the case of a binary network, the elements G_{ij} take the value 1 if i and j are linked, and 0 if they are not linked; while in a weighted network, $G_{ij} = w(i,j)$, where $w(i,j)$ is some measure of the strength of the relationship between i and j. Typically, the leading diagonal of $\boldsymbol{G}$ is normalised to 0.

- **Influence Matrix:** This is a row-stochastic (or 'right stochastic') adjacency matrix, $\tilde{\boldsymbol{G}}$ whose elements are generally defined as $\tilde{G}_{ij} = {}^{G_{ij}}/_{\sum_j G_{ij}}$ if two agents are linked and 0 otherwise.

- **Degree:** A node's degree, d_i, is the number of edges of the node in an undirected graph. The degree of node i in the network with a binary adjacency matrix, $\boldsymbol{G}$, can be calculated by summing the elements of the i^{th} row of this matrix.[107] In a directed graph, a node's **in-degree** is the number of edges from other nodes to that node, and it's **out-degree** is the number of edges from that node to other nodes in the network. For node i, the former can be calculated by summing the elements of the i^{th} column of the binary adjacency matrix for the network, while the latter is obtained by summing the i^{th} row of this matrix.

- **Average degree:** The average degree for a network graph is the average number of edges that nodes in the network have.

- **Density:** The relative fraction of edges that are present in a network. It is calculated as the average degree divided by $N-1$, where N is the number of nodes in the network.

- **Shortest path length (geodesic):** A path in a network g between nodes i and j is a sequence of edges, $i_1i_2, i_2i_3, ..., i_{R-1}i_R$, such that $i_ri_{r+1} \in g$, 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. The shortest path length or geodesic between i and j is the path between i and j that contains the fewest edges. The average geodesic of a network is the average geodesic for every pair of nodes in the network. For nodes for whom no path exists, it is common to either exclude them from the calculation of the average geodesic (i.e. to calculate the average geodesic from the connected part of the network) or to define the geodesic for these nodes to be some large number (usually greater than the largest geodesic in the network).

- **Diameter:** The diameter of a graph is the largest geodesic in the connected part of the network, where by connected, we refer to nodes for whom a path exists to get from one node to the other.

[107] Similarly, for a weighted graph, summing the elements for row i in the adjacency matrix yields the weighted degree.

- **Component:** A connected component, or component, in an undirected network 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 considered to be a bridge in the network g if removing the edge ij results in an increase in the number of components in g.
- **Complete Network:** A network in which all possible edges are present.
- **Degree Centrality:** This is the node's degree divided by $N-1$, where N is total number of nodes in the network. It measures how well a node is connected in terms of direct neighbours. Nodes with a large degree have a high degree centrality.
- **Betweenness centrality**: This is 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 can be calculated as the ratio of the number of geodesics between j and k that i lies on to the total number of geodesics between j and k. Averaging this ratio across all pairs of nodes yields the betweenness centrality of node i.
- **Eigenvector centrality**: A relative measure of centrality, the centrality of node i is the sum of the centrality of its neighbours. It can be calculated by solving the following equation in matrix terms, $\lambda C^e(\boldsymbol{G}) = \boldsymbol{G}C^e(\boldsymbol{G})$, where $C^e(\boldsymbol{G})$ is an eigenvector of $\boldsymbol{G}$, and λ is the corresponding eigenvalue.
- **Bonacich Centrality:** Another measure of centrality that defines a node's centrality as a function of their neighbours' centrality. It is defined as $\boldsymbol{b}(\boldsymbol{G}_g, \beta) = (\boldsymbol{I}_g - \beta\boldsymbol{G}_g)^{-1}.(\alpha\boldsymbol{G}_g\boldsymbol{\iota})$.
- **Dyad count:** A dyad is a pair of nodes. In an undirected network, the dyad count is the number of edges in the network.
- **Triad count:** A triad is a triple of nodes such that a path connecting all 3 nodes exists. The triad count of an undirected network is the number of such triples in the network.
- **Clustering coefficient:** For an undirected network, this measures the proportion of fully connected triples of nodes out of all potential triples in which at least two edges are present.
- **Support:** An edge $ij \in \mathscr{E}_g$ is supported if there exists an agent $k \neq i, j$ such that $ik \in \mathscr{E}_g$ and $jk \in \mathscr{E}_g$.
- **Expansiveness:** For subsets of connected nodes in the network, the ratio of the number of edges connecting the subset to the rest of the network to the number of nodes in the subset.
- **Sparseness:** A property of the network related with the length of all minimal cycles connecting triples of nodes in the network. For any integer, $q \geq 0$, a network is q-sparse if all minimal cycles connecting any triples of nodes (i, j, k) such that $ij \in \mathscr{E}_g$ and $jk \in \mathscr{E}_g$ have length $\leq q + 2$. See Bloch et al. (2008) for more details.

- **Graph span:** The graph span is a measure that mimics the average path length. It is defined as

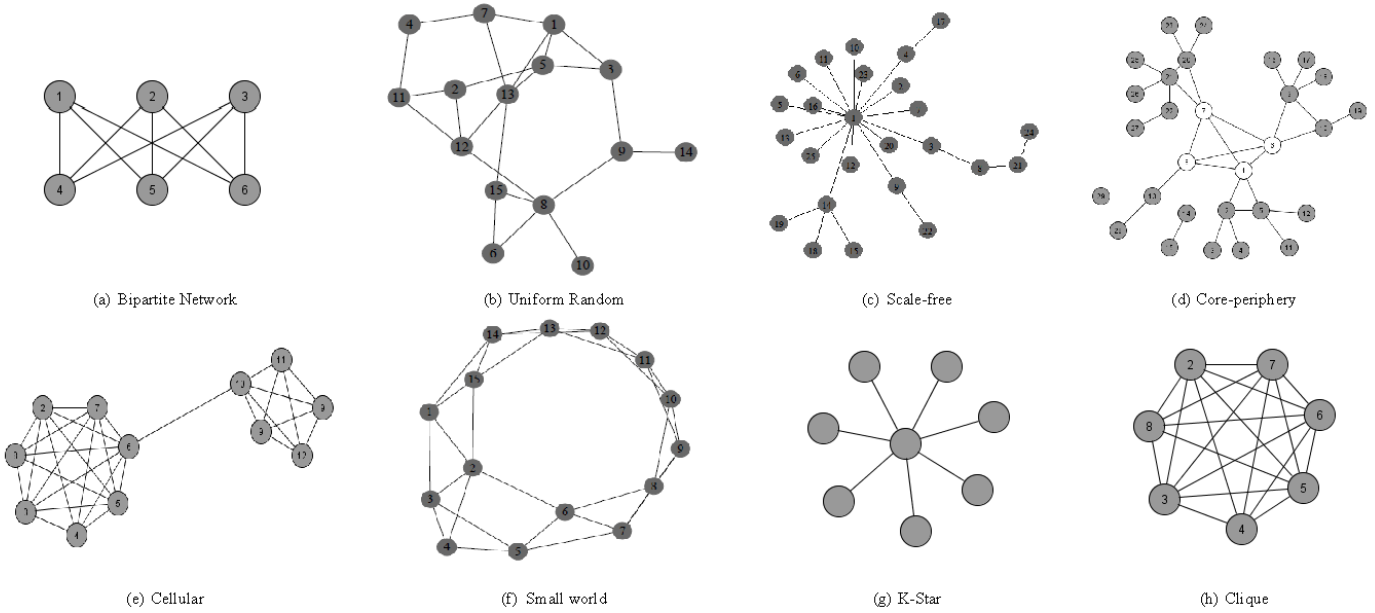
$$span_g = \frac{log(N_g) - log(d_g)}{log(\tilde{d}_g) - log(d_g)} + 1$$

where N_g is the number of nodes in network g, d_g is the average degree of network g and $\tilde{d}_g$ is the average number of second-degree neighbours in the network.

Network Topologies

- **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.
- **Uniform random network:** A graph where edges between nodes form randomly.
- **Scale-free network:** A 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, *i.e.* the proportion of nodes with very high degrees constitutes a non-negligible proportion of all nodes.
- C**ore-periphery network:** A network that can be partitioned into a set of nodes that is completely connected ('core'), and another set of agents ('periphery') who are linked primarily with nodes in the 'core'.
- **Cellular network:** Networks containing many sets of completely connected nodes (or 'cliques'), with few edges connecting the different cliques.
- **Small world network:** A network where most nodes are not directly linked to one another, but where geodesics between nodes are small, i.e. a node can reach every other node in the network by passing through a small number of nodes.
- **k-star:** A 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.
- **Cliques:** A clique is any induced subgraph of a network (*i.e.* subset of nodes and all edges between them) such that every node in the subgraph is directly connected to every other node in the subgraph.

Figure 7: Network Topologies



(a) Bipartite Network (b) Uniform Random (c) Scale-free (d) Core-periphery

(e) Cellular (f) Small world (g) K-Star (h) Clique

- **Induced Subgraph:** The network graph constructed from data where nodes are randomly sampled and where edges are included only if both nodes are randomly sampled are known as induced subgraph.

- **Star Subgraph:** The network constructed from data where nodes are randomly sampled and all their 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), is called a star subgraph.

- **Network Motif:** Any subgraph of the network which has a particular structure. For example, the reciprocated link motif is defined as any pair of nodes, $\{i, j\}$, such that both of the possible directed links between them, $\{ij, ji\}$, are present in the subgraph. Another example is the k-star motif, which is defined as any k nodes such that one of the nodes is linked to all (k-1) other nodes, and the other nodes are not linked to each other.

- **Isomorphic Networks:** Two networks are isomorphic iff we can move from one to the other only by permuting the node labels. For example, all six directed networks composed of three nodes and one edge are isomorphic. Isomorphism implies that all network statistics are also identical, since these statistics are measured at a network level so are not affected by node labels.

B Quadratic Assignment Procedure

The Quadratic Assignment Procedure (QAP) was developed originally by Mantel (1967) and Hubert and Schultz (1976).[108] It tests for correlation between a pair of network variables by calculating the correlation in the data, and comparing this to the range of estimates computed from the same calculation after permutation of the rows and columns of the adjacency matrix $\boldsymbol{G}$. For example, suppose we have two vectors $\mathbf{y}(\boldsymbol{G}) = \{y_i(\boldsymbol{G}_g)\}_{i \in \mathscr{N}_g}$ and $\mathbf{x}(\boldsymbol{G}) = \{x_i(\boldsymbol{G}_g)\}_{i \in \mathscr{N}_g}$ which are functions of the network. We first calculate $\hat{\rho}_{0,YX}$, the correlation between $\mathbf{y}$ and $\mathbf{x}$ observed in the data. In order to respect the dependencies between edges that involve the same node, we then jointly permute the rows and columns of the argument of $\mathbf{y}$. This amounts to effectively relabelling the nodes, so that we calculate a new estimate $\hat{\rho}_{w,YX}$: the correlation between $\mathbf{y}(\boldsymbol{G}_w)$ and $\mathbf{x}(\boldsymbol{G})$, where $\boldsymbol{G}_w$ is the permuted adjacency matrix. It is generally *not* the same as permuting the elements of the vectors $\mathbf{y}$. This is repeated W times, to give a range of estimates $\{\hat{\rho}_{w,YX}\}_{w=1,...,W}$. Under the null hypothesis of no correlation, we can perform, for example, a two-sided test at the 10% level, by considering whether $\hat{\rho}_{0,YX}$ lies between the 5th and 95th percentiles of $\{\hat{\rho}_{w,YX}\}_{w=1,...,W}$. If it does not, we can reject the null at the 10% level.

Ideally one would like to use all the possible permutations available, but typically this number is too large. Hence a random sample of permutations is typically used. This is done by drawing the from the set of nodes of the network, $\{1, ..., N\}$, without replacement. The order in which the indices are drawn is defined as the new, permuted ordering, for calculating $\mathbf{y}(\boldsymbol{G}_w)$.

Krackhardt (1988) extended QAP to a multivariate setting. Now we have variables $\{\mathbf{y}(\boldsymbol{G}), \mathbf{x}_1(\boldsymbol{G}), ..., \mathbf{x}_K(\boldsymbol{G})\}$ and are interested in testing whether there is a statistically significant correlation between $\mathbf{y}$ and the K other variables. To test for a relationship between $\mathbf{y}$ and $\mathbf{x}_1$, Krackhardt suggests we first regress $\mathbf{y}$ and $\mathbf{x}_1$, separately, on $(\mathbf{x}_2 ... \mathbf{x}_K)$ to give residuals $\mathbf{y}_1^*$ and $\mathbf{x}_1^*$. Then one can perform QAP on $\mathbf{y}_1^*$ and $\mathbf{x}_1^*$, as in the bivariate setting, where $\hat{\rho}_{0,Y^*X_1^*}$ is an estimate of the partial correlation between $\mathbf{y}$ and $\mathbf{x}_1$ conditioning on the other $(\mathbf{x}_2 ... \mathbf{x}_K)$. This process can be repeated for all K covariates.

[108] See Hubert (1987) for a review of developments of this method.