Essays in Development Economics

Submitted by

Bansi Khimji Malde

for the degree of Doctor of Philosophy

of the

University College London

2016

Declaration

I, Bansi Khimji Malde, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed ..... (Bansi Khimji Malde)
Date:

To my spiritual master and guide, Param Pujya Bhaishree, without whose grace this would not have been possible. To my parents, Khimji and Chandan, and sister, Jigna.

Abstract

In this dissertation, I use household-level microdata from rural areas of developing countries, combined with simple theory and experimental and micro-econometric techniques, to study informal insurance arrangements within extended family networks, and the consequences of incorrect knowledge of the health production function on health and non-health choices.

The first chapter reviews methods for identifying the effects of social networks on outcomes (or social effects) using data with information on exact connections between agents, paying special attention to methods dealing with endogeneity of network formation, and measurement error in the network.

The second chapter studies the role of socially close and distant connections in informal risk sharing under imperfect enforcement. Socially close connections can better enforce informal arrangements, but may provide fewer risk sharing opportunities. A simple theoretical framework studies this trade-off and yields qualitative predictions for empirical testing with data from a large number of village-based extended family networks in rural Mexico.

In the third chapter, I study the relationship between risk sharing and group size in a setting with limited commitment and coalitional deviations. Building on Genicot and Ray (2003), the chapter shows that the relationship between risk sharing and group size is theoretically ambiguous. I then study the question empirically using data from rural Malawi and exploiting historical norms, which indicate that a woman's brothers play an important role in ensuring her household's wellbeing, to define the risk sharing group. I find that households where the wife has many brothers achieve worse risk sharing.

The final chapter studies the effects of a randomized intervention in rural Malawi which, over a six-month period, provided mothers of young infants with information on child nutrition only. Findings show that the intervention improved infant nutrition, household food consumption and child health. Male labour supply also increased, partially funding the increased consumption.

Acknowledgement

I am grateful to my advisors, Orazio Attanasio and Imran Rasul for their encouragement, guidance and support. Their supervision has made me a better economist. Many thanks also to Marcos Vera-Hernandez, who in addition to being a co-author on two chapters, was very generous with his time in discussing my research. These discussions and suggestions have improved the quality of this dissertation.

Thanks also to my co-authors Arun Advani, Emla Fitzsimons, Alice Mesnard and Marcos Vera-Hernandez for many useful discussions and insights. I'm grateful also to my colleagues and ex-colleagues at EDePo at the Institute for Fiscal Studies, specifically Laura Abramovsky, Alison Andrew, Britta Augsburg, Bet Caeyers, Elisa Cavatorta, Silvia Espinosa, Pamela Jervis, Sonya Krutikova, Dan Rogger and Marta Rubio-Codina, as well as my fellow PhD students at UCL for helping me keep perspective during the PhD.

I am indebted to my spiritual master, Param Pujya Bhaishree, and my constant and steadfast supports, P. Pradipbhai and Rashmiben. They have been at my side throughout, and kept me going through some tough periods in the PhD. I thank my parents and sister for their love, support and most of all for their patience as I embarked on yet more further study.

Financial support from the ESRC (Grants RES-576-25-0042, ES/K00123X/1, ESI03685/X, RES-183-25-008), the ERC and the IFS is gratefully acknowledged.

Contents

List of Figures

List of Tables

Contribution to chapters

The chapter 2 is coauthored with Arun Advani (UCL).
The chapter 3 is a solo document.
The chapter 4 is coauthored with Emla Fitzsimons (UCL-IOE and IFS) and Marcos Vera-Hernandez (UCL and IFS)
The chapter 5 is coauthored with Emla Fitzsimons (UCL-IOE and IFS), Alice Mesnard (City University) and Marcos Vera-Hernandez (UCL and IFS)

Chapter 1

Introduction

In this dissertation, I use household-level microdata from rural areas of developing countries, combined with simple theory and experimental and micro-econometric techniques, to study informal insurance arrangements within extended family networks, and the consequences of incorrect knowledge of the child health production function on health and non-health choices. Social networks, particularly the extended family, play a key role in helping households in developing countries cope with the risks they face. Identifying the features that make them effective is important for the design of good policies. However, identifying the effects of social networks on agents' outcomes is complicated by endogeneity of network formation – agents linking decisions are affected by variables that are unobserved to the econometrician – and measurement error in the network. A number of methods have been developed to deal with these issues. At the same time, child health is very poor in low income settings. Incorrect knowledge of the child health production function could drive this by distorting health and non-health choices.

In chapter 2, I review the literature studying econometric methods for the analysis of linear models of social effects – the effects of social networks, such as declared friendships in classrooms, or extended family connections, on economic agents' outcomes. The class of linear social effect models includes the 'linear-in-means' local average model, the local aggregate model, and models where features of the network architecture (network statistics) affect outcomes. The chapter begins by providing a common empirical framework that nests these models, before summarising the underlying theoretical models that yield each empirical model. It then discuss conditions for identification of the social effects using observational and experimental/quasi-experimental data, before discussing methods to overcome endogeneity of network formation. These include models of network formation. The chapter provides a detailed overview of these, drawing on

methods developed within economics, as well as disciplines such as statistics and sociology. The final part of the review considers issues around collecting networks data and measurement error in the network. Constructing a network from a sample generates severe non-classical measurement error in the network structure, which in turn severely biases parameters estimated from sampled networks. Drawing on work in economics, computer science, statistical physics, statistics and sociology, I review the literature on the consequences of partial measurement of a network on measures of network structure as well as parameter estimates; before outlining methods developed in these literatures to deal with this issue.

The next part of the dissertation investigates how extended family networks, an important institution in developing countries, help households to cope with risk. Theory suggests that the structure of these networks are likely to influence their effectiveness in providing insurance (Bloch et al. 2008, Jackson et al. 2012, Ambrus et al. 2014). Variation in network structure will thus generate heterogeneity in informal insurance outcomes across households and networks. My research considers empirically this heterogeneity for two dimensions of network architecture – social distance and network size – in two different settings, and draws on theory linking these dimensions to channels for effective insurance provision to interpret the findings.

In chapter 3, I study the role of socially close (direct) and distant (indirect) connections in providing informal risk sharing in social networks. Socially close connections should be more effective in enforcing informal risk sharing arrangements, but may be more economically similar and less numerous than socially distant connections, and thereby provide fewer risk sharing opportunities. I begin by specifying a simple theoretical framework incorporating these features, and use it to conduct comparative statics on how the relationship between risk sharing and the number of socially close and distant connections changes as opportunities for risk sharing change. The analysis shows that the trade-off between enforcement and risk sharing opportunities yields a U-shaped (inverse U-shaped) relationship between risk sharing and the number of socially close (distant) connections. I then test the model predictions empirically using detailed data on a large number of village-based extended family networks in rural Mexico. I first document that socially distant connections provide more opportunities for risk sharing: they are more numerous, are less likely to be engaged in the same occupation and thus have less positively correlated incomes. Thereafter, I consider how risk sharing varies with the average number of socially close and distant connections in a household's extended family network. To measure risk sharing, I use a commonly used measure from the literature (Townsend 1994), which can also be motivated from the theoretical framework: the response of household consumption to income fluctuations,

net of aggregate network-level resources. My estimation accounts for time invariant household-level factors correlated with the network measures and risk sharing; as well as for common network-level variables. The findings indicate that risk sharing improves with more socially distant connections, while socially close connections have no effect. This suggests that opportunities for risk sharing are particularly important for the effective functioning of extended family based risk sharing in this context.

Chapter 4 seeks to understand and test empirically the relationship between group size and informal risk sharing. Models of risk sharing with limited commitment and grim-trigger punishments imply that larger groups provide better insurance. However, when subgroups of households can credibly deviate, so that arrangements ought to be coalition-proof, the relationship between group size and the amount of insurance is unclear. Building on Genicot & Ray (2003), the chapter shows that this relationship is theoretically ambiguous. I then investigate it empirically using data on the size of sibships of the household head and spouse in rural Malawi. To identify the potential risk sharing group, the chapter exploits a social norm among the main ethnic group in our sample – the Chewa – which indicates that the wife's brothers should play a key role in ensuring her household's wellbeing. I find that households where the wife has many brothers are poorly insured against crop loss events. I fail to uncover a similar relationship for the wife's sisters, ruling out that these findings are driven by wives with many siblings having poorer extended family networks. Finally, I calibrate the model to fit the empirical setting. The calibration indicates that the threat of coalitional deviations can explain the empirical findings.

The final part of the dissertation studies another policy relevant outcome in developing countries – child health and considers the implications of incorrect knowledge of the child health production function, on household health and non-health choices and child health outcomes in rural Malawi. Incorrect knowledge of the health production function may lead to inefficient household choices, and thereby to the production of suboptimal levels of health. Chapter 5 studies the effects of a randomized intervention in rural Malawi which, over a six-month period, provided mothers of young infants with information on child nutrition without supplying any monetary or in-kind resources. A simple model first investigates theoretically how nutrition and other household choices including labor supply may change in response to the improved nutrition knowledge observed in the intervention areas. The chapter then shows empirically that the intervention improved child nutrition, household food consumption and consequently health. It finds evidence that labor supply increased, which might have contributed to partially fund the increase in food consumption. Moreover, the chapter also pays careful attention to the important issue of inference in randomised experiments with few clusters,

using two leading methods proposed for this case – wild cluster bootstrap-t and randomisation inference – and evaluating their performance in the data.

The present thesis contributes to the literature on several fronts. First, chapter 2 provides a review of the fast-growing literature on methods for identifying social effects using detailed networks data, drawing on literatures both within economics and in other disciplines. Second, it provides a common framework for linear social effects, which nests many commonly used empirical specifications. Finally, this is one of the only reviews available that considers the issue of measurement error in the network in a detailed and comprehensive manner.

The third chapter adds to our understanding of how risk sharing in extended family networks varies with the number of connections at different social distances. In particular, it documents that socially close and distant connections offer varying opportunities for risk sharing, and these opportunities for risk sharing are very important for the effective functioning of extended family based insurance. This novel finding complements the much more widely accepted finding that spatially distant connections in agricultural settings are less likely to experience the same shock and thereby be able to provide insurance (Rosenzweig & Stark 1989). Most previous work on informal insurance in social networks doesn't consider this channel, (Bloch et al. 2008, Jackson et al. 2012, Ambrus et al. 2014). This chapter fills this gap. It also contributes to our understanding of how social distance affects household outcomes in developing country contexts.

Chapter 4 contributes to the literature on risk sharing with coalitional deviations Genicot & Ray (2003) by showing that the relationship between risk sharing and group size is theoretically ambiguous. It is also one of the first papers to estimate this relationship empirically when allowing for imperfect enforcement and coalitional deviations, using household micro-data rather than a laboratory experiment setting as in Chaudhuri et al. (2010).

The final chapter is one of the first to consider the empirical consequences of imperfect knowledge of the child health production function on non-health choices, specifically labour supply. Other studies had considered the effects of providing health information on specific health related behaviours, or health outcomes; finding mixed evidence (Madajewicz et al. 2007, Kremer & Miguel 2007, Jalan & Somanathan 2008, Dupas 2011*a*). This study considers a more multifaceted intervention, as well as assessing effects on non-health choices. The chapter also contributes to the literature investigating the causal effects of education on health by providing cleanly identified evidence of the importance of one of the key channels through which these effects are thought to operate - knowledge. Finally, in paying careful attention to the important issue of inference in randomised experiments with few clusters, it provides a detailed evaluation of the

performance of the leading inference methods in this case: the wild cluster bootstrap-t and randomisation inference.

In what follows I start by reviewing the literature on methods to identify social effects using networks data (chapter 2). Then I study how socially close and distant connections influence risk sharing in extended family networks (chapter 3) before analysing the relationship between group size and risk sharing in a setting with imperfect enforcement and coalitional deviations (chapter 4). Thereafter, I analyse how incorrect knowledge of the child health production function distorts health and non-health choices (Chapter 5). The final chapter 6 provides some concluding remarks and directions for future work.

Chapter 2

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

2.1 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.

[0]This chapter is co-authored with Arun Advani. We are grateful to Imran Rasul for his support and guidance. 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.

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 2.3 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 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

[1] A different presentation of some of the material in this part of Section 2.3 can be found in Topa & Zenou (2015). Of the models we discuss, their focus is on two of the more common specifications used. Topa & Zenou (2015) compare these models to each other, and also to neighbourhood effect models, and discuss the relationship between neighbourhood and network models.

understand how we can influence network structure and hence indirectly the distribution of outcomes.

In Section 2.4 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 2.5 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 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 & Durlauf (2001) and Brock & 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

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

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 & 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 2.7.1 then provides detailed definitions of the various network measures and topologies that are mentioned in the text below.

2.2 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 2.4 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, $\mathscr{E}_g = \{ij : \forall i, j \in \mathscr{N}_g, i \neq j\}$. Moreover, edges have *edge weights* $wei(i, j)$

[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 4.8 we provide further useful definitions.

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.

[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.

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 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.

2.3 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 2.4.

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

[9] We leave aside the important issues of inference, in order to keep the scope of this survey manageable.

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 2.5 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.

2.3.1 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 2.2):

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

$\boldsymbol{Y}$ 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:

$$\mathbb{E}[\varepsilon_{i,g} |\, \boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}_g] = 0 \quad \forall\, i \in g;\ g \in \{1, ..., M\} \tag{2.2}$$

$$Cov[\varepsilon_{i,g}\varepsilon_{k,h} | \boldsymbol{X}_g, \boldsymbol{X}_h, \boldsymbol{Z}_g, \boldsymbol{Z}_h, \boldsymbol{G}_g, \boldsymbol{G}_h] = 0 \quad \forall i \in g; k \in h; g, h \in \{1, ..., M\}; g \neq h \tag{2.3}$$

[10] See Sacerdote (2011) for an overview of this literature.

[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.

Equation 2.2 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 2.3.7 below, we will review some approaches taken to relax this assumption. In addition, Equation 2.3 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 $\boldsymbol{\nu}$:

$$\mathbb{E}[\nu_g |\, \boldsymbol{X}_g, \boldsymbol{Z}_g, \boldsymbol{G}_g] = 0 \quad \forall\, g \in \{1, ..., M\} \tag{2.4}$$

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 assumption and allow for correlation between $\boldsymbol{\nu}$ and the other right hand side variables in Equation 2.1.

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 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 2.3.2 below.

[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 (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.

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 2.1, 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 & Liu (2010) and Liu, Patacchini, Zenou & Lee (2014). More details are provided in Subsection 2.3.3 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 2.1, 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, Patacchini & Zenou (2014). See Subsection 2.3.4 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) 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.

[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 $\boldsymbol{G}$.

[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.

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 2.1 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. (2014).
When researchers are interested in *aggregate* network outcomes, rather than node level outcomes, the following specification is typically estimated:

$$\bar{\boldsymbol{y}} = \phi_0 + \bar{\boldsymbol{w}}_{\bar{\boldsymbol{y}}}(\boldsymbol{G})\boldsymbol{\phi}_1 + \bar{\boldsymbol{X}}\boldsymbol{\phi}_2 + \bar{\boldsymbol{w}}_{\bar{\boldsymbol{X}}}(\boldsymbol{G}, \bar{\boldsymbol{X}})\boldsymbol{\phi}_3 + \boldsymbol{u} \tag{2.5}$$

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 2.3.5.

In Subsections 2.3.2 to 2.3.5 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 conditional on observable individual and network-level variables. For each case discussed, we start by outlining a theoretical model that generates under-

[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.

lying the resulting empirical specification, and outline identification conditions using observational data.

Thereafter, in Subsection 2.3.6, 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 2.3.2 to 2.3.4 below.

Subsection 2.3.7 outlines methods used by researchers to relax the assumption made in equation 2.2: 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 2.4 and 2.5 respectively.

2.3.2 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:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \beta\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\gamma} + \tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\delta} + \boldsymbol{Z}\boldsymbol{\eta} + \boldsymbol{L}\boldsymbol{\nu} + \boldsymbol{\varepsilon} \tag{2.6}$$

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:

[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.

- 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 (*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 & 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 2.6. 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]

$$U_i(y_{i,g}; \boldsymbol{y}_{-i,g}, \boldsymbol{X}_g, \tilde{\boldsymbol{G}}_{i,g}) = \left(\pi_{i,g} - \frac{1}{2} \left(y_{i,g} - 2\beta \sum_{j=1}^{N_g} \tilde{G}_{ij,g} y_{j,g} \right) \right) y_{i,g} \tag{2.7}$$

β in Equation 2.7 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 as-

[19]Notice that in Equation 2.7, $\sum_{j=1}^{N_g} \tilde{G}_{ij,g} y_{j,g}$ is identical to the i^{th} row of $\tilde{\boldsymbol{G}}_g \boldsymbol{y}_g$, which appears in Equation 2.6.

sumption can be relaxed. In particular, an econometric specification similar to Equation 2.6 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:

$$y_{i,g} = \pi_{i,g} + \beta \sum_{j=1}^{N_g} \tilde{G}_{ij,g} y_{j,g} \tag{2.8}$$

Patacchini & 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):

$$\pi_{i,g} = \boldsymbol{x}_{i,g}\boldsymbol{\gamma} + \sum_{j=1}^{N_g} \tilde{G}_{ij,g}\boldsymbol{x}_{j,g}\boldsymbol{\delta} + \boldsymbol{z}_g\boldsymbol{\eta} + \nu_g + \varepsilon_{i,g} \tag{2.9}$$

Substituting for this in Equation 2.8, we obtain the following best response function for individual outcomes:

$$y_{i,g} = \beta \sum_{j=1}^{N_g} \tilde{G}_{ij,g} y_{j,g} + \boldsymbol{x}_{i,g}\boldsymbol{\gamma} + \sum_{j=1}^{N_g} \tilde{G}_{ij,g}\boldsymbol{x}_{j,g}\boldsymbol{\delta} + \boldsymbol{z}_g\boldsymbol{\eta} + \nu_g + \varepsilon_{i,g} \tag{2.10}$$

Then, stacking observations for all nodes in multiple networks, we obtain Equation 2.6, which can be taken to the data.

Bramoullé et al. (2009) study the identification and estimation of Equation 2.6 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 2.6.

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 partic-

[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.

ular, 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 2.1 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 (*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 2.1: 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 2.1, 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

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

for the schooling performance of j in the equation for i. If instead, the network were directed as in the right panel of Figure 2.1, 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 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 2.6 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:

$$\boldsymbol{JY} = \beta \boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{JX}\boldsymbol{\gamma} + \boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\delta} + \boldsymbol{J}\boldsymbol{\varepsilon} \tag{2.11}$$

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 2.2 below). The presence of at least two intermediate agents allows researchers to use the characteristics of third-degree neighbours (neighbours-of-

[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.

neighbours-of-neighbours who are not direct neighbours or neighbours-of-neighbours) as an additional instrument to account for the network fixed effect.

Figure 2.2: 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 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 $\boldsymbol{G_{X,g}}$ and $\boldsymbol{G_{y,g}}$ denote the network-level adjacency matrices through which, respectively, the contextual and endogenous effects operate. As before we define the block diagonal matrices $\boldsymbol{G_X} = diag\{\boldsymbol{G_{X,g}}\}_{g=1}^{g=M}$ and $\boldsymbol{G_y} = diag\{\boldsymbol{G_{y,g}}\}_{g=1}^{g=M}$. Blume et al. (2013) study identification of this model assuming that the two networks are (conditionally) exogenous and show that when the matrices $\boldsymbol{G_y}$ and $\boldsymbol{G_X}$ 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 2.6 to be identified are that the matrices $\boldsymbol{I}$, $\boldsymbol{G_y}$, $\boldsymbol{G_X}$ and $\boldsymbol{G_y G_X}$ 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,

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

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 2.2 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 2.3.7.

2.3.3 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:

- 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 2.1 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

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

[25] The games considered in both Bramoullé & 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.

form:

$$U_i(y_{i,g}; \boldsymbol{y}_{-i,g}, \boldsymbol{X}_g, \boldsymbol{G}_g) = \left(\pi_{i,g} - \frac{1}{2} y_{i,g} + \beta \sum_{j=1}^{N_g} G_{ij,g} y_{j,g} \right) y_{i,g} \tag{2.12}$$

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:

$$y^*_{i,g}(\boldsymbol{G}_g) = \beta \sum_{j=1}^{n} G_{ij,g} y_{j,g} + \boldsymbol{x}_{i,g}\boldsymbol{\delta} + \sum_{j=1}^{n} \tilde{G}_{ij,g} \boldsymbol{x}_{j,g} \boldsymbol{\gamma} + \boldsymbol{z}_g \boldsymbol{\eta} + \nu_g + \varepsilon_{i,g} \tag{2.13}$$

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 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

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

[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).

an overview.

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

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \beta\boldsymbol{GY} + \boldsymbol{X\gamma} + \tilde{\boldsymbol{G}}\boldsymbol{X\delta} + \boldsymbol{Z\eta} + \boldsymbol{L\nu} + \boldsymbol{\varepsilon} \tag{2.14}$$

which corresponds to Equation 2.1 with $\boldsymbol{w_y}(\boldsymbol{G}, \boldsymbol{Y})=\boldsymbol{GY}$ and $\boldsymbol{w_x}(\boldsymbol{G}, \boldsymbol{X})=\tilde{\boldsymbol{G}}\boldsymbol{X}$, and where all other variables and parameters are as defined above in Subsection 2.3.1.

Identification of Equation 2.14 using observational data has been studied by Calvó-Armengol et al. (2009), Lee & Liu (2010) and Liu, Patacchini, Zenou & Lee (2014). 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 2.3.2, we replace the network-level observables, $\boldsymbol{Z}$, and the network-level unobservable, $\boldsymbol{L\nu}$ in Equation 2.14 with network fixed effects, $\boldsymbol{L\tilde{\nu}}$, 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 2.3.2. 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 2.11, is:

$$\boldsymbol{JY} = \beta\boldsymbol{JGY} + \boldsymbol{JX\gamma} + \boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{X\delta} + \boldsymbol{J\varepsilon} \tag{2.15}$$

The model above suffers from the reflection problem, since $\boldsymbol{Y}$ appears on both

sides of the equation. However, the parameters of Equation 2.15 can be identified using linear IV if the deterministic part of the right hand side, $[\mathbb{E}(\boldsymbol{JGY}), \boldsymbol{JX}, \boldsymbol{J\tilde{G}X}]$, 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 2.14:

$$\begin{aligned}\mathbb{E}(\boldsymbol{JGY}) = \boldsymbol{J}(\boldsymbol{GX} + \beta\boldsymbol{G}^2\boldsymbol{X} + ...)\boldsymbol{\gamma} + \boldsymbol{J}(\boldsymbol{G\tilde{G}X} + \beta\boldsymbol{G}^2\boldsymbol{\tilde{G}X} + ...)\boldsymbol{\delta} \\ + \boldsymbol{J}(\boldsymbol{GL} + \beta\boldsymbol{G}^2\boldsymbol{L} + ...)\tilde{\boldsymbol{\nu}} \end{aligned} \tag{2.16}$$

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 2.14 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.* $\boldsymbol{G\tilde{G}X}$) 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 2.5 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.

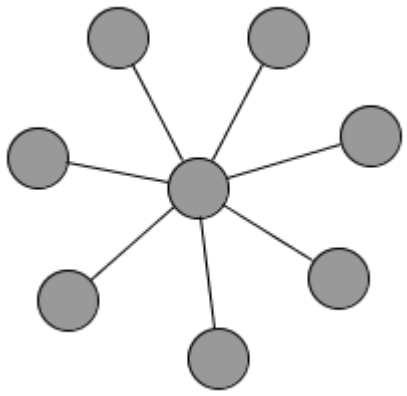
Parameters can still be identified if there no variation in node degree within a network for all networks in the data, but there is variation in degree across networks. In this case, $\boldsymbol{G}_g = \bar{d}_g\tilde{\boldsymbol{G}}_g$ and $[E(\boldsymbol{JGY}), \boldsymbol{JX}, \boldsymbol{J\tilde{G}X}]$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

[29] See Liu, Patacchini, Zenou & Lee (2014) for a different identification condition that allows for some linear dependence among these matrices under additional restrictions.

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 2.3.3 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 2.3) 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 2.3.2 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 2.3: Star Network



2.3.4 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, Patacchini & Zenou (2014) 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:

[30] See Liu, Patacchini, Zenou & Lee (2014) for a different identification condition that allows for some linear dependence among these matrices under additional restrictions.

$$U_i(y_{i,g}; \boldsymbol{y}_{-i,g}, \boldsymbol{X}_g, \tilde{\boldsymbol{G}}_{i,g}, \boldsymbol{G}_{i,g}) = \left(\pi_{i,g} + \beta_1 \sum_{j=1}^{N_g} G_{ij,g} y_{j,g} - \frac{1}{2} \left(y_{i,g} - 2\beta_2 \sum_{j=1}^{N_g} \tilde{G}_{ij,g} y_{j,g} \right) \right) y_{i,g} \tag{2.17}$$

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, Patacchini & Zenou (2014) 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:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \beta_1\boldsymbol{GY} + \beta_2\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{X\gamma} + \tilde{\boldsymbol{G}}\boldsymbol{X\delta} + \boldsymbol{Z\eta} + \boldsymbol{L\nu} + \boldsymbol{\varepsilon} \tag{2.18}$$

using the same notation as before (see *e.g.* Subsection 2.3.1).

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 2.3.2 and 2.3.3 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:

$$\boldsymbol{JY} = \beta_1\boldsymbol{JGY} + \beta_2\boldsymbol{J\tilde{G}Y} + \boldsymbol{JX\gamma} + \boldsymbol{J\tilde{G}X\delta} + \boldsymbol{J\varepsilon} \tag{2.19}$$

When there is variation in the degree within a network g, then the reduced form equation of Equation 2.19 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 $\boldsymbol{J\tilde{G}}(\boldsymbol{I}-\beta_1\boldsymbol{G}-\beta_2\tilde{\boldsymbol{G}})^{-1}\boldsymbol{L}$ can be used as an instrument for the local average term $\boldsymbol{J\tilde{G}Y}$. 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, Patacchini & Zenou (2014) for more details.

2.3.5 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:

$$\boldsymbol{Y} = \boldsymbol{f_y}(\boldsymbol{w_y}(\boldsymbol{G},\, \boldsymbol{Y}), \boldsymbol{X}, \boldsymbol{w_x}(\boldsymbol{G},\, \boldsymbol{X}), \boldsymbol{Z}) + \boldsymbol{\varepsilon} \tag{2.20}$$

and network-level outcomes:

$$\bar{\boldsymbol{y}} = \boldsymbol{f_{\bar{y}}}(\bar{\boldsymbol{w}}_{\bar{\boldsymbol{y}}}(\boldsymbol{G}), \bar{\boldsymbol{X}}, \bar{\boldsymbol{w}}_{\bar{\boldsymbol{x}}}(\boldsymbol{G},\, \bar{\boldsymbol{X}})) + \boldsymbol{u} \tag{2.21}$$

$\boldsymbol{f_y}(.)$ and $\boldsymbol{f_{\bar{y}}}(.)$ are functions that specify the shape of the relationship between the network statistics and the node- and network-level outcomes. When $\boldsymbol{f_y}(.)$ is simply a linear index in its argument, Equation 2.22 remains nested in Equation 2.1. Though, in principle, the shape of $\boldsymbol{f_y}(.)$ should be guided by theory (where possible), through the rest of this Subsection, we take $\boldsymbol{f_y}(.)$ to be a linear index in its argument. $\boldsymbol{w_y}(\boldsymbol{G},\, \boldsymbol{Y})$ 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, $\boldsymbol{w_x}(\boldsymbol{G},\, \boldsymbol{X})$ 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

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

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. (2014) 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.

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 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 2.20 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:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \boldsymbol{\omega\beta} + \boldsymbol{X\gamma} + \mathbf{Z\eta} + \boldsymbol{\varepsilon} \tag{2.22}$$

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 2.4 and 2.5 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:

$$\begin{aligned} \boldsymbol{Y} = \alpha\boldsymbol{\iota} + \boldsymbol{w_y}(\boldsymbol{G},\, \boldsymbol{Y})\boldsymbol{\beta} + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\gamma}} + \boldsymbol{w_x}(\boldsymbol{G},\, \tilde{\boldsymbol{X}})\tilde{\boldsymbol{\delta}} \\ + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{L}\boldsymbol{\nu} + \boldsymbol{\varepsilon} \end{aligned} \tag{2.23}$$

$\boldsymbol{w_y}(\boldsymbol{G},\, \boldsymbol{Y})$ 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 2.3.2, 2.3.3 and 2.3.4 above, though it is not immediately obvious how such an IV could be constructed. Future research is needed to shed light on these issues.

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 2.21, where the shape of the function $\boldsymbol{f_{\bar{y}}}(.)$ 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 $\boldsymbol{f_{\bar{y}}}(.)$, this implies the following equation:

$$\bar{\boldsymbol{y}} = \phi_0 + \bar{\boldsymbol{\omega}}\boldsymbol{\phi}_1 + \bar{\boldsymbol{X}}\boldsymbol{\phi}_2 + \bar{\boldsymbol{w}}_{\bar{\boldsymbol{X}}}(\boldsymbol{G},\, \bar{\boldsymbol{X}})\boldsymbol{\phi}_3 + \boldsymbol{u} \tag{2.24}$$

where the variables are as defined after Equation 2.21. 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 2.5.3, the difficulties and costs involved in collecting network data often mean that in practice researchers have data for a small number of networks only.

2.3.6 Experimental Variation

Subsections 2.3.2 to 2.3.5 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 2.3.2 above. The issues related to using experimental variation to uncover the parameters of the local 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:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \beta\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\gamma} + \tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\delta} + \boldsymbol{L}\tilde{\boldsymbol{\nu}} + \boldsymbol{\varepsilon} \tag{2.25}$$

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

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

[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.

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 $\boldsymbol{Y}$:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \beta\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\gamma} + \tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\delta} + \rho\boldsymbol{t} + \boldsymbol{L}\tilde{\boldsymbol{\nu}} + \boldsymbol{\varepsilon} \tag{2.26}$$

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 2.3.2 above to account for the network-level fixed effect leads to the following specification:

$$\boldsymbol{J}\boldsymbol{Y} = \alpha\boldsymbol{J}\boldsymbol{\iota} + \beta\boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{J}\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\delta} + \rho\boldsymbol{J}\boldsymbol{t} + \boldsymbol{J}\boldsymbol{\varepsilon} \tag{2.27}$$

We can use instrumental variables to identify β as long as the deterministic part of the right hand side of Equation 2.27, $[\mathrm{E}(\boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{Y}), \boldsymbol{J}\boldsymbol{X}, \boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{X}]$ 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 2.26 under the assumption that $|\beta| < 1$, to construct instruments:

$$\begin{aligned}\mathbb{E}[\boldsymbol{J}\tilde{\boldsymbol{G}}\boldsymbol{Y}] = \boldsymbol{J}\tilde{\boldsymbol{G}}\sum_{s=0}^{\infty}\beta^s\tilde{\boldsymbol{G}}^s\alpha\boldsymbol{\iota} + \boldsymbol{J}(\tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\gamma} + \beta\tilde{\boldsymbol{G}}^2\boldsymbol{X}\boldsymbol{\gamma} + ...) + \boldsymbol{J}(\tilde{\boldsymbol{G}}^2\boldsymbol{X}\boldsymbol{\delta} + \beta\tilde{\boldsymbol{G}}^3\boldsymbol{X}\boldsymbol{\delta} + ...) \\ + \boldsymbol{J}(\rho\tilde{\boldsymbol{G}}\boldsymbol{t} + \beta\rho\tilde{\boldsymbol{G}}^2 t + ...)\end{aligned} \tag{2.28}$$

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 2.26. It can thus be used as an instrument for $\tilde{\boldsymbol{G}}\boldsymbol{Y}$, either 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

[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.

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:

$$\boldsymbol{Y} = \alpha\boldsymbol{\iota} + \beta\tilde{\boldsymbol{G}}\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\gamma} + \tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{\delta} + \rho\boldsymbol{t} + \tilde{\boldsymbol{G}}\boldsymbol{t}\mu + \boldsymbol{\varepsilon} \tag{2.29}$$

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 2.3.2.[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 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:

$$\boldsymbol{Y} = b_1\boldsymbol{\iota} + b_2\tilde{\boldsymbol{G}}\boldsymbol{t} + \boldsymbol{X}\boldsymbol{b}_3 + \tilde{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{b}_4 + b_5\boldsymbol{t} + \boldsymbol{u} \tag{2.30}$$

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

[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, Patacchini, Zenou & Lee (2014).

[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.

underlying this social effect, or the value of some 'deep' structural parameter.

2.3.7 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 $y_{i,g}$.

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 $\boldsymbol{G}_i$.

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

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 2.3.6. [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]

Instrumental Variables

In the first example above, the outcome y was determined by an equation of the form of Equation 2.1, where the network $\boldsymbol{G}$ was determined potentially by some of the observables already in Equation 2.1 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 2.1 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 in real world settings. It is also unlikely to be repre-

[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] Booij et al. (2015) 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.

sentative 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 & 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

[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.

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 cannot be used. We discuss further in Section 2.4 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.

Jointly model formation and social effects

In our second example at the beginning of Subsection 2.3.7 we considered the case where the outcome y was determined by an equation of the form of Equation 2.1, 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 & Lee (2014), and Goldsmith-Pinkham & 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

[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.

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 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 2.4, is the restriction that links are formed conditionally independently.

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.

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

2.4 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 Subsection 2.3.7, 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 2.3, we will obtain biased parameter estimates. One potential solution to this problem – proposed in different contexts by Goldberg & Roth (2003), Popescul & Ungar (2003), Hoff (2009), and Chandrasekhar & 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.

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

Finally, we saw in Section 2.3 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. 2015; Jackson et al. 2012), and agricultural practices (Conley & 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 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 2.4.1 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 2.4.2, 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 2.3.7. 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

[47] Through Section 2.4 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.

[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).

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 2.4.3, 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.

2.4.1 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

[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 2.4.3.

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

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.

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.

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

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

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

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 2.4.3.

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.

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 & 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 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]

[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).

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 & 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.

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 & 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 & 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 & 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 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

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

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 & 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

$$\Pr_{\boldsymbol{\theta}}(\boldsymbol{G}) = \frac{1}{\kappa(\boldsymbol{\theta})} \exp\left\{\sum_A \theta_A S_A(\boldsymbol{G})\right\} \tag{2.31}$$

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

[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.

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 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

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

[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.

assumption that we have made (see Equation 2.31), 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.

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

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

[69] Introduced to the social networks literature by Strauss & 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.

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 on the quadratic assignment procedure (QAP): a type of permutation test for correlation between variables. For more details see Appendix 2.7.2.

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 & Mourifié 2013; Goldsmith-Pinkham & 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 & Mourifié (2013) and Goldsmith-Pinkham & 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

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

use Markov Chain Monte Carlo (MCMC) maximum likelihood (Geyer & Thompson 1992, Snijders 2002, Handcock 2003). As noted earlier, the key difficulty with direct maximum likelihood estimation of Equation 2.31 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 & 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]

[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.

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 & 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 2.31 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.

Chandrasekhar & 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 $\tilde{\kappa}(\boldsymbol{\theta}) = \sum_{\boldsymbol{s}'} K_{\boldsymbol{S}}(\boldsymbol{s}')\exp\{\boldsymbol{\theta s}'\}$ 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 & 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 2.31, and the predicted

[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.

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.

2.4.2 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 2.3.7). 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 2.3.7 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 (meeting homophily). However, they all have the reduced form implication that more similar agents are more likely to be linked.[75]

Fafchamps & 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

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

[75] In Subsection 2.4.3 below, we consider homophily in more detail, and structural models that try to separate these causes of observed homophily.

are modelled as directed, then the model takes the general form

$$G_{ij} = f\left(\lambda_0 + (\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j})\boldsymbol{\lambda}_1 + \boldsymbol{x}_{2i}\boldsymbol{\lambda}_2 + \boldsymbol{x}_{3j}\boldsymbol{\lambda}_3 + \boldsymbol{w}_{ij}\boldsymbol{\lambda}_4 + u_{ij}\right) \tag{2.32}$$

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 & 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 & 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 & 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 2.4.1, which may also be used in this circumstance.

2.4.3 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.

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

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]

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 2.32 – 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 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.

[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.

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 2.32, 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 & 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

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

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.

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 2.3 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 2.5.3). 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 &

[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.

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. 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

[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.

[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.

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 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 & 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 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 2.4.1. 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.

[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.

[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.

2.5 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.

2.5.1 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 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

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

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).

2.5.2 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

[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.

[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.

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. (2014) 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

[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.

multiple subgroups (*e.g.* classes within the university). Network structure is then imposed by assuming that 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 2.3 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

[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.

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 2.3 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 data on voting and social networks used in Fafchamps & 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. Nonethe-

less, this method can collect, at reasonable costs, complete information on local neighbourhoods, which is needed to apply the methods outlined in Section 2.3 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.

2.5.3 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

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

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 2.5.3 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 2.3.2, 2.3.3 and 2.3.4) and network statistics (as in Subsection 2.3.5). 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 2.5.3 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 2.5.4 outlines the various methods that have been developed for this purpose.

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 2.3 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 & 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 & 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 2.6) 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 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 2.3.3, 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

[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.

include the model-based and likelihood-based corrections outlined in Subsection 2.5.4.

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 2.3.5, 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 2.4, 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 & Valente (2003), Lee et al. (2006), Kim & Jeong (2007) and Chandrasekhar & Lewis (2011) have estimated the magnitude of sampling induced bias in statistics such as 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

[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 2.5.4.

[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.

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 & Lewis (2011) reports a median sampling rate of 25% of nodes in a network. Table 2.1 below summarises findings from these papers for various commonly used network statistics.

Figure 2.4: Sampled networks with different sampling rates

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 2.7.1 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 2.5 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.

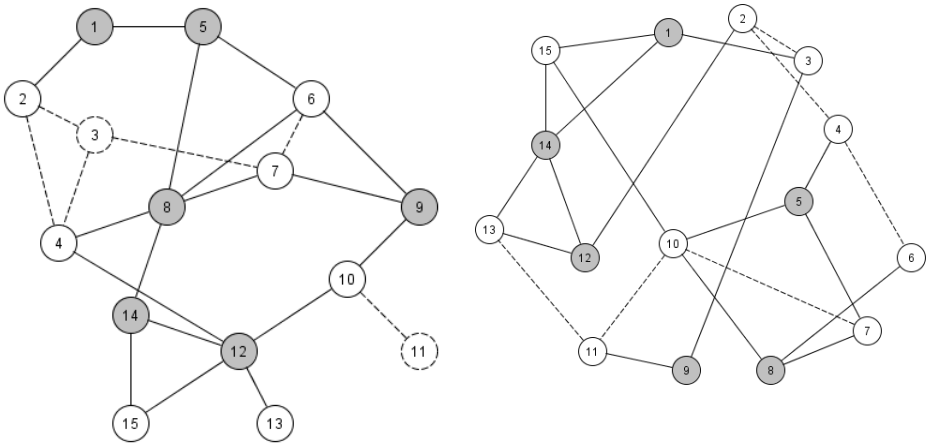
3. *The magnitude of error in network statistics due to sampling varies with the sampling method.* Different sampling methods result in varying magnitudes of errors in network statistics. Lee et al. (2006) compare data sampled via induced subgraph sampling, random sampling of nodes, random sampling of 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 re-

[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).

[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.

Figure 2.5: 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

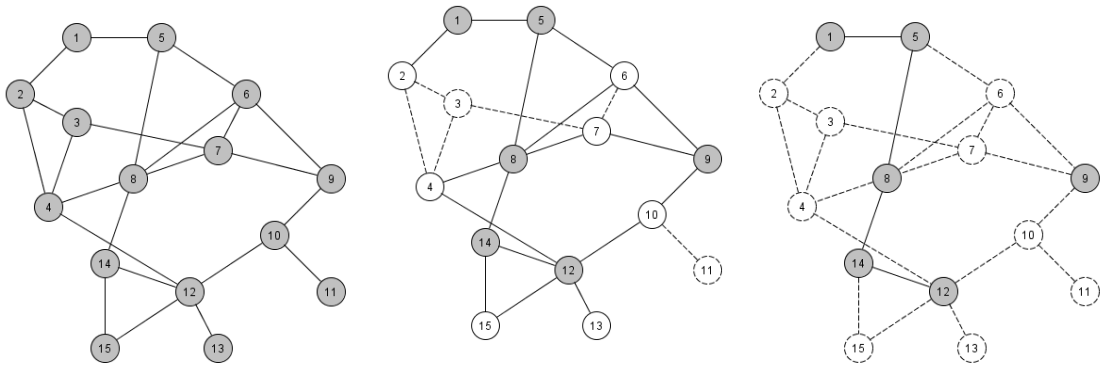
Notes to Figure: This figure displays the star subgraphs obtained from sampling 40% of nodes in a network with a uniform random topology (panel (a)) and a small world topology (panel(b)). The unshaded nodes represent nodes that were not sampled, and the dotted lines represent nodes and edges on which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

gardless of whether or not the nodes involved in the edges were sampled. Consider again the network graph considered in panel (a) of Figure 2.4 above, and displayed again in panel (a) of Figure 2.6 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 2.6 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 mea-

[103] Average degree is calculated as above, including all nodes and edges in the sample, *i.e.* those with firm lines in Figure 2.6.

Figure 2.6: Sampling with star and induced subgraphs



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.

sured statistic; *i.e.*, the measurement error is not independent of the true network statistic. Chandrasekhar & 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 2.2 summarises the findings from the literature on the effects of random sampling of nodes on parameter estimates.

5. *Top-coding of edges or incorrectly specifying the boundary of the network biases network statistics.* Network data collected through surveys often place an upper limit on the number of edges that can be reported. Moreover, limiting the network

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

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 2.1 and 2.2 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 2.6 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/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 & 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 & Valente (2003);[b]Source: Lee et al. (2006); [c]Source: Kim & 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. Bias magnitude large even when 50% nodes sampled.	Attenuated (–) (more than star subgraphs). Bias magnitude very large and falls with sampling rate.
Clustering Coefficient	Scaling (+) and attenuation (–); |scaling| > |attenuation|. Very large biases, which fall with sampling rate.	Attenuation (–), falls with sampling rate. 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 & 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 & Lewis (2011)

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 & 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 & Fafchamps (2014), who propose a method for identifying and correcting misreported edges.

2.5.4 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.

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 *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 & 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

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

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.

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 2.4.1 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 2.3 above. Chandrasekhar & 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 2.3.

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 2.4.1 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 & 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 2.1), though parameters of network-level models, such as Equation 2.5 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.

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 & 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 & 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 & 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.

2.6 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, identifi-

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

cation 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.

2.7 Appendix

2.7.1 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

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

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.

- **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.

- **Core-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.

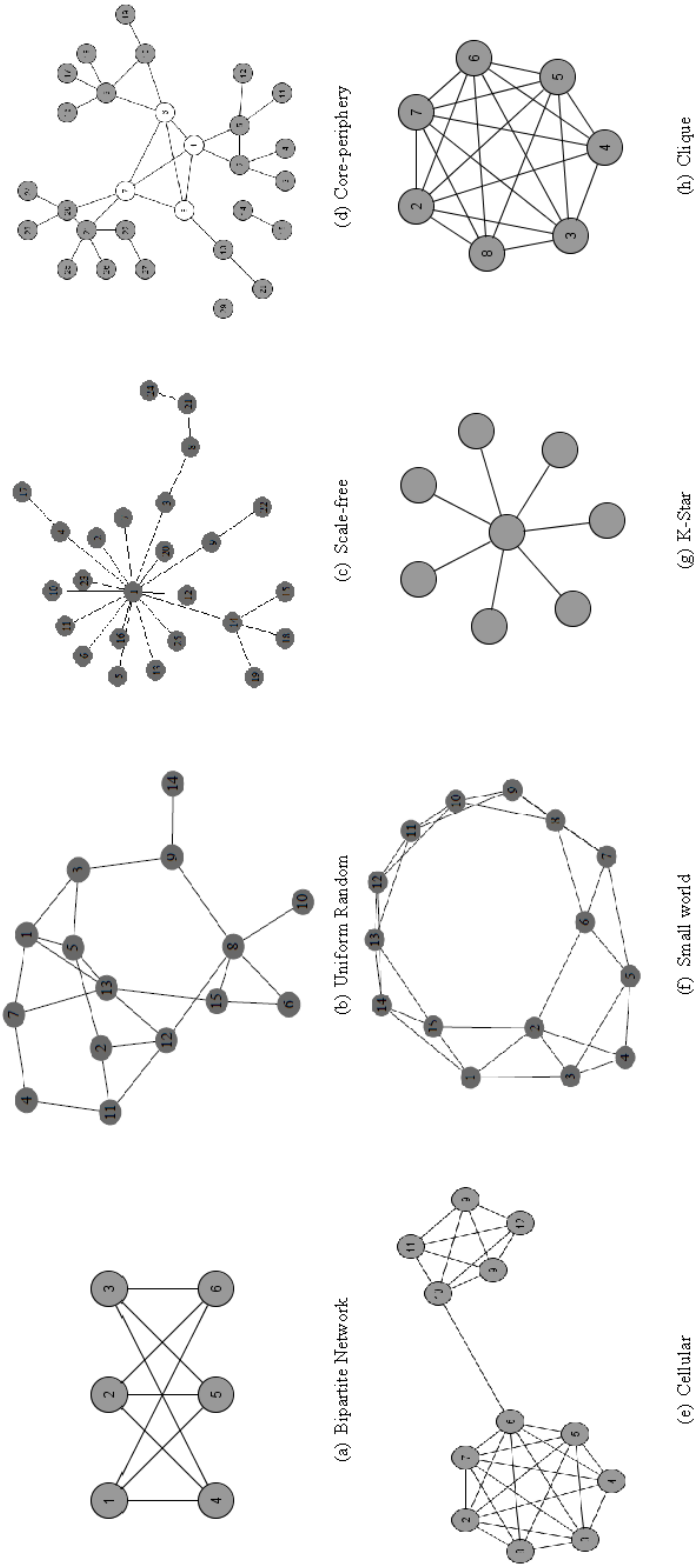


Figure 2.7: Network Topologies

- **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.

2.7.2 Quadratic Assignment Procedure

The Quadratic Assignment Procedure (QAP) was developed originally by Mantel (1967) and Hubert & 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 $\mathbf{y}(\boldsymbol{G}) = \{y_i(\boldsymbol{G}_g)\}_{i \in \mathcal{N}_g}$ and $\mathbf{x}(\boldsymbol{G}) = \{x_i(\boldsymbol{G}_g)\}_{i \in \mathcal{N}_g}$ 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.

Chapter 3

Socially Close and Distant Connections in Risk Sharing

3.1 Introduction

Risk is a salient fact of life in rural areas of developing countries. To cope with the consequences of this risk, households have to rely on informal arrangements with social connections (family and friends) in the absence of well-functioning formal credit and insurance markets and poor government capacity. Indeed, social connections have been shown to provide a high level of, though not complete, insurance.[1] Effective provision of insurance requires social connections to be able to effectively monitor and enforce informal arrangements, while also having sufficiently uncorrelated income processes so as to be able to provide help when needed. However, they vary on dimensions related to effective insurance provision (e.g. economic similarity, connection strength), thereby leading to heterogeneity in informal insurance outcomes across households and social networks.

In this chapter, I investigate theoretically and empirically how one feature of social connections – social distance – affects risk sharing when informal arrangements cannot

[0]I am grateful to Orazio Attanasio and Imran Rasul for their comments and guidance on this project. I also thank Marcos Vera-Hernandez, Monica Costa-Dias, Kim Scharf, Sarah Smith, Mushfiq Mobarak, Michele Tertilt, Robert Townsend, Antonio Cabrales, Arun Advani, Laura Abramovsky, Sonya Krutikova, Sonia Bhalotra, Pablo Branas-Garza and participants at the IFS work-in-progress seminar, Middlesex University, RES Conference, DIAL Conference, ESWC (Montreal) and the EEA Congress in Mannheim for useful comments and suggestions. Richard Audoly and Simon Robertson provided excellent research assistance. Funding from the ESRC Grant ES/K00123X/1 is gratefully acknowledged.

[1]For example, Rosenzweig & Stark (1989); Townsend (1994); Fafchamps & Lund (2003); Attanasio & Szekely (2004); Angelucci et al. (2015) among others. Social connections are defined to be either other households in the same village, or members of the same sub-caste or ethnic group; or extended family members.

be perfectly enforced. In such a setting, socially close connections – direct connections – should be better able to enforce informal arrangements, making them more valuable for risk sharing than socially distant (indirect) connections. However, they may offer fewer opportunities for risk sharing – they may be more economically similar and thus have more positively correlated income processes relative to distant connections, and also be fewer in number given costs of forming connections – thereby undermining their effectiveness in providing risk sharing. Thus, a trade-off may emerge between risk sharing opportunities and enforcement, which influences the relationship between risk sharing and socially close and distant connections.

To study the effects of this trade-off on the relationship between risk sharing and socially close and distant connections, I specify a simple theoretical model of risk sharing in networks based on Ambrus, Mobius & Sziedl (2014).[2] The model incorporates both imperfect enforcement of informal arrangements, and varying opportunities for risk sharing from socially close and distant connections. The latter arise from allowing incomes of socially close connections to be more positively correlated than those of distant connections; and from variation in the number of households at different social distances. I use this set-up to obtain comparative statics of how risk sharing and welfare vary with the number of socially close and distant connections in a network, as opportunities for risk sharing change. It is not possible to obtain the comparative statics analytically, so I numerically simulate the model to obtain qualitative predictions that are then verified empirically.

The theoretical analysis indicates that when enforcement concerns dominate, risk sharing (and welfare) increases with the number of socially close connections. Conversely, when opportunities for risk sharing are particularly important, risk sharing and welfare fall (increase) with the number of socially close (distant) connections. For parameter values where both concerns are relevant, the trade-off between enforcement and risk sharing opportunities generates an inverse-U shaped relationship between the extent of risk sharing (and welfare) and the number of socially close connections in a network. Networks with few socially close and many socially distant connections have low enforcement, which leads to low risk sharing; while networks with very high numbers of socially close connections and few or no distant connections have strong enforcement, but limited opportunities for risk sharing, which dampens risk sharing thereby leading to the inverse-U shaped relationship.

The empirical analysis draws on data on within-village extended family networks in rural Mexico. The extended family network forms a crucial source of informal insurance

[2] A network is a collection of all households connected either directly or indirectly through social connections, and the connections between them.

in this setting (Angelucci et al. 2015), making it a particular relevant network to study. The data is exceptionally detailed, with information on within-village, cross-household extended family (specifically parent, child and sibling) connections of the household head and his spouse, and a panel of socio-economic variables, for *all* households in over 500 poor, marginalised villages. The former allows me to overcome a key empirical challenge: identifying socially close and distant connections. I define these according to a network-theoretic notion of social distance: two households are considered to be socially close if there is a direct family connection (sibling, parent, child) between them; and socially distant if there is an indirect (e.g. sibling's spouse's sibling; or uncles, cousins) connection between them. The census of all households in the village allows me to calculate accurate measures of these within the village. This is particularly important since network measures constructed from a sample of the network are subject to substantial non-classical measurement error, which in turn generates large biases in regression estimates (Chandrasekhar & Lewis 2011).

In a first step, I investigate how risk sharing opportunities vary with social distance, making use of information on the occupation of the household head, as well as of household income. I document that the heads of socially close households are more likely to be engaged in the same occupation than those of socially distant households. This similarity in occupation choice also translates into similarities in income processes: incomes of socially close households are more positively correlated than those of socially distant households. Moreover, socially distant connections are more numerous on average than socially close connections. Both these findings indicate that socially distant connections provide more risk sharing opportunities in this context.

The next step of the analysis considers the implications of this variation in risk sharing opportunities on the relationship between risk sharing and the average number of socially close and distant connections in a household's network. Risk sharing is measured as the response of changes in household log consumption to fluctuations in household log income net of network-level aggregate resources. Consumption is a particularly apt measure of risk sharing, since it provides a summary measure of all risk sharing instruments used by households. Moreover, though this measure of risk sharing has been commonly used in the literature, (Townsend 1994 among others), it can also be motivated from the theoretical model.

The availability of panel data at the household level, as well as data on a large number of within-village extended family networks allows me to at least partially account for unobserved variables that might be correlated with both my measures of the number of socially close and distant connections and the risk sharing measure. This is important, since though the extended family network can be considered to be at least

partially exogenous (households do not choose their siblings or parents), choices related to marriage, household formation and migration make the within-village extended family network endogenous. The longitudinal dimension at the household level allows me to difference out fixed household level unobserved variables that might be correlated with both the within-village extended family network (e.g. ethnicity) and risk sharing. I also include network-time fixed effects to account for common network-level unobservables (e.g. unobserved local market conditions), including those that vary over time.

In addition, I conduct additional robustness checks to show that the findings are unlikely to be driven by systematic variation in network structure by wealth; or by measurement error in the network. Finally, the availability of such a large number of within-village extended family networks (unusual in the networks literature given the costs of collecting accurate information on social connections), allows me to conduct valid inference and obtain efficient estimates.

The findings indicate that households in networks with more socially distant connections achieve better risk sharing than those with few distant connections: increasing the number of socially distant connections by one standard deviation (23 households) from the sample average ($\tilde{}$ 20 households) reduces the response of household log consumption to fluctuations in log income by 20%. By contrast, the number of socially close connections has no effect on risk sharing. These results are not driven by wealthy households having few socially close connections and many distant connections within the village: the data indicate no significant correlation between wealth and the number of a household's socially close and distant connections within the village. Finally, they are also robust to measurement error in the network: changing the assumptions on who is identified to be a family connection doesn't alter the conclusions.

Thus, networks with more socially distant connections, which offer more opportunities for risk sharing provide higher informal insurance than the more close-knit networks with fewer socially distant connections and many socially close ones, highlighting the importance of sufficient risk sharing opportunities for the successful functioning of social network based informal insurance. These result indicate 'the strength of weak ties', to borrow the term proposed by Granovetter (1973), for the effective functioning of risk sharing arrangements in extended family networks. Granovetter (1973), who coined this term when studying information flows, argued that weak 'acquaintance' ties are valuable since they facilitate the flow of (new) information between closely knit groups of individuals. In the context of risk sharing, socially distant connections are valuable since they provide a large number of less positively correlated income streams, thereby improving opportunities for risk sharing.

My findings are also important for the design of effective policies. Decomposing the

effects of socially close and socially distant connections provides suggestive evidence on the key constraints facing informal insurance, and thus where policy intervention might be beneficial. This is important to know since well intentioned programs such as those that for example aim to improve risk sharing against aggregate shocks, could crowd out informal risk sharing in settings with imperfect enforcement of contracts, and reduce overall welfare (Attanasio & Rios-Rull 2000). My results indicate that insufficient opportunities for risk sharing limit the extent to which social connections can help households cope with the consequences of risk. Thus, policies that allow socially close connections to diversify their income sources might positively impact risk sharing.

Related Literature This paper contributes to a number of literatures. First, it adds to a growing literature investigating how variation in the network architecture affects informal risk sharing patterns.[3] A number of theoretical studies have shown that various measures of network structure, such as the length of cycles, the presence of common connections (support), how close-knit a network is (viscosity), and the extent to which a network spreads out (expansiveness), relate to whether a network can sustain informal risk sharing in the presence of frictions such as imperfect enforcement, and imperfect information (Bloch et al. 2008; Jackson et al. 2012; Ali & Miller 2013; Ambrus et al. 2014). Empirically, studies by Krishnan & Sciubba (2009), Ligon & Schechter (2012), Kinnan & Townsend (2012) and Chandrasekhar et al. (2014) have considered the implications of network architecture and household position in the social network on risk sharing arrangements and patterns in Ethiopia, Thailand and India. This study builds on this literature by incorporating an important driver for insurance – risk sharing opportunities – and considering how this relates theoretically and empirically to the relationship between the extent of risk sharing and the number of connections at different social distances. A closely related paper is Angelucci et al. (2015), which uses the same data to investigate how extended family connections affect consumption and investment decisions of households in the context of a conditional cash transfer programme. They document that the presence of extended family networks influences households' consumption and investment decisions, and also uncover heterogeneity in these effects by the architecture of the underlying network, though they do not shed light on the drivers of this heterogeneity.

Second, it contributes to our understanding of how social distance affects economic outcomes in poor, rural economies. The bulk of this literature has focused on the ef-

[3]More generally, it contributes to our understanding of informal risk sharing arrangements in developing countries. Key contributions to this literature include Townsend (1994), Ligon (1998), Ligon et al. (2003) and Kinnan (2014) among others.

fects of social distance on alleviating constraints to formal and informal contracts and arrangements. For example, Fisman et al. (2012) document that Indian bank officers make more loans to clients from the same caste, and these perform well, suggesting that socially close individuals are able to more effectively share information. Breza & Chandrasekhar (2015) show, using a field experiment, that socially close peer monitors encourage households to save more, while Chandrasekhar et al. (2014), show that socially close ties can cooperate without external enforcement in a lab-in-the-field experiment, while distant ties are unable to do so.

Finally, the paper also contributes to our understanding of how extended family networks, an extremely influential institution in developing countries, affect household outcomes. In particular, they have been shown to play critical roles in shaping risk sharing outcomes (Foster & Rosenzweig 2001; Fitzsimons et al. 2015), facilitating investments (Angelucci et al. 2010; Baland et al. 2015) and help with job search (Luke & Munshi 2006; Magruder 2010 and Wang 2013). This paper enhances our understanding of the features of these networks that enhance and limit the effective provision of informal insurance.

The rest of the paper is structured as follows: Section 3.2 outlines the theory that guides the empirical analysis. Section 3.3 describes the data used, including details on how extended family connections are identified. Section 3.4 then details the empirical model while Section 3.5 displays the results and conducts some robustness checks. Finally, Section 3.6 concludes.

3.2 Conceptual Framework

To guide the empirical analysis, I lay out a simple, stylised model of risk sharing in networks that builds on Ambrus et al. (2014) and embeds the following features (i) imperfect enforceability of informal arrangements; and (ii) differing opportunities for risk sharing from socially close and socially distant connections. I use the model to generate comparative statics on the relationship between risk sharing and welfare and the number of socially close and distant connections.

3.2.1 Setting

K households are embedded in a pre-existing network, represented as a graph $G = (N, L)$, which consists of a set of households, $N = \{1, ..., K\}$ and a set of links or connections between households, $L = \{(i, j)\}_{i \in N; j \in N}$. If i and j are directly linked, then $(i, j) \in L$. Links are taken to be undirected so that $(i, j) \in L$ implies that $(j, i) \in L$. Each household has a value associated with each connection, denoted by x_{ij}

for the value to a household i of its connection with a household j, which is determined outside the model. I interpret x_{ij} to be the expected utility value of future transfers that i expects to receive from j.

Socially close and distant connections are defined according to a graph-theoretic measure of social distance. For any two households i and j in the same network, the social distance d_{ij} is defined as the number of links that i has to go through to get to j in the network. I take socially close households to be those for whom $d_{ij} = 1$; and socially distant to be those for whom $1 < d_{ij} < \infty$.[4]

Households face a risky endowment, y_i, which for simplicity, I assume can take two values: h, with probability p, and l with probability $(1-p)$; and with $h > l$; $0 < p < 1$. They can share this risk through bilateral transfers, denoted by t_{ij} which represents the net transfer from i to j, with their direct (or socially close) connections. There are no transaction costs in this model, so it is natural to impose that $t_{ij} = -t_{ji}$, which means that the net transfer i makes to j is equivalent to the net transfer j receives from i. There is no storage in the model. Household consumption is thus calculated as $c_i = y_i - \sum_{ij \in L} t_{ij}$. Households gain utility from their own consumption, c_i and from the value of their connections, $x_i = \sum_{j:ij \in L} x_{ij}$. I assume that the utility of consumption and from connections is additively separable, which yields the following objective function:

$$u(c_i) + v(x_i)$$

where the functions $u(.)$ and $v(.)$ are assumed to be increasing and concave in their arguments.[5]

Transfer arrangements cannot be perfectly enforced in this setting, and so need to be self-sustaining. This is achieved by the following punishment mechanism: if a household i doesn't make a transfer to j, it loses the associated link value x_{ij}. This implies a connection-specific incentive compatibility constraint of this form:

$$u(c_i) + v(x_i) \geq u(c_i + t_{ij}) + v(x_i - x_{ij}) \quad \forall (i,j) \in L$$

Since the incentive compatibility constraint is connection-specific, households with more than 1 socially close connection will face multiple incentive compatibility con-

[4] $d_{ij} = \infty$ if i and j are not in the same network.

[5] The model is static so as to keep it tractable. In a dynamic model, one would need to keep track of changes to the network structure in all possible continuation values. This is an extremely complex object, which expands greatly the space of possible continuation values (e.g. there are $2^{\frac{K(K-1)}{2}}$ possible undirected network structures for a network with K households), making it extremely computationally challenging to solve the model.

straints. Moreover, the net transfers from i to all its socially close connections appear in each of its incentive compatibility constraints. This feature complicates analytic derivation of optimal transfers, other than in very specific cases (e.g. where consumption and connection value are perfect substitutes as in Ambrus et al. (2014)).

Households can observe all the endowments received by all other households, and all transfers made and received. Thus, there are no issues of imperfect information. Overall, this environment is consistent with village-based extended family networks, households are able to closely monitor each other, and share information, but may not be able to perfectly enforce informal arrangements.

Endowment Processes Across Households and Risk Sharing Opportunities Opportunities for risk sharing depend on the correlations in endowments of households embedded in the same network. Denote by $R = [r_{ij}]_{i \in N; j \in N}$ the matrix of pairwise endowment correlations for all pairs of households in a network, with the diagonal set to 1. When endowments are identically and independently distributed across households, a widely made assumption which I will consider to as a benchmark assumption, all off-diagonal terms in R are set to 0 and each additional connection, whether socially close or distant, would offer the same opportunity for risk sharing.

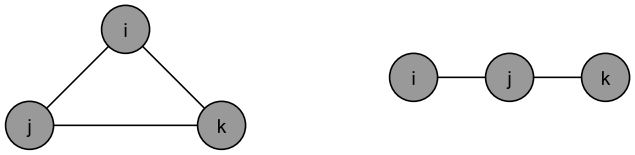
I introduce variation in the opportunities for risk sharing from socially close and distant connections by allowing the pairwise correlation in endowments, r_{ij}, to depend on social distance. Specifically, I assume that the pairwise correlation in endowments of socially close households i and j is positive, and more so than that for two socially distant connections, i and k: $r_{ij} > r_{ik}$. Though optimal risk sharing would imply that households select as risk sharing partners those with uncorrelated or negatively correlated income streams, closely connected households in the empirical setting studied in this paper have, on average, positively correlated income streams (as will be shown in Section 3.3), making this a suitable assumption. Studies from other settings provide further support for this assumption: Fafchamps & Gubert (2007) document that risk sharing connections tend to be geographically close (and hence be likely to have positively correlated incomes) in rural Philippines, while in India, households sort into occupations by sub-caste (Munshi & Rosenzweig 2006), a crucial institution for informal risk sharing (Mobarak & Rosenzweig 2014; Munshi & Rosenzweig 2016).

The underlying network structure influences which other households one's endowment is correlated with. Specifically, households in close-knit networks will be more likely to experience similar endowments (when the pairwise correlation in endowments is positive), than in more loosely connected networks.[6] To illustrate this, consider the

[6] The underlying network architecture could also generate feedback effects when pairwise endowment

two networks displayed in Figure 3.1. The network in the left panel is more close-knit than that on the right panel, and households on average have more socially close connections in that network. Next, introduce a positive correlation, $r_{ij} = \delta \ \forall (i,j) \in L$, with $0 \leq \delta \leq 1$, in the endowments of socially close connections in the two networks, allowing the correlation to decline geometrically with social distance. In the network in the left panel, $r_{ij} = r_{jk} = r_{ik} = \delta$, while in that in the right panel, $r_{ij} = r_{jk} = \delta$ and $r_{ik} = \delta^2$. Thus, the endowments of i and k will be more positively correlated in the network on the left panel, than that on the right panel. This difference increases the chances of states where all households experience the same endowment (and so where no risk sharing is possible) in the network in the left panel, relative to that in the right panel. Thus, with positive correlations in the endowments of socially close households, the underlying network structure will also influence the extent to which endowment realisations are correlated with one another, and so affect opportunities for risk sharing.

Figure 3.1: Example Networks



Model Solution Ambrus et al. (2014) show that this problem can be re-cast as that of a utilitarian social planner choosing bilateral transfers so as to maximise a weighted sum of households' expected utility (equation 3.1) subject to an aggregate budget constraint (3.5) and a set of link-specific incentive compatibility constraints (equation 3.2).

$$\underset{\{t_{ij}\}_{(i,j)\in L}}{max} \sum_{i \in L} \lambda_i \{u(c_i) + v(x_i)\} \tag{3.1}$$

subject to

$$u(c_i) + v(x_i) \geq u(c_i + t_{ij}) + v(x_i - x_{ij}) \ \ \forall \, (i,j) \in L \tag{3.2}$$

$$t_{ij} = -t_{ji} \tag{3.3}$$

correlations are non-zero. If a fraction of a household's socially close connections are also directly connected with one another, the positive correlations among their endowments generate a feedback effect on the household's own endowment, and so on.

$$c_i = y_i - \sum_{j:j\in N_i(g)} t_{ij} \tag{3.4}$$

$$\sum_{i\in N} c_i \leq \sum_{i\in N} y_i \tag{3.5}$$

where λ_i is a positive planner weight such that $\sum_{i\in N}\lambda_i = 1$.

Obtaining an analytical solution to this problem through the usual Karush-Kuhn-Tucker (KKT) conditions is not possible since they a complicated system of nonlinear simultaneous equations when any incentive compatibility constraint binds.[7] Ambrus et al. (2014) instead characterise the optimal solution in terms of the marginal social welfare gain of providing additional transfers to households.[8] The optimum solution for a given state of the world is such that the network partitions into 'risk sharing islands', where within an island, households equate their marginal social gain. On the border of the islands, there will be households for whom at least one incentive compatibility constraint binds in either direction. Different states of the world can partition the same network into different risk sharing islands, depending on the distribution of the endowment realisation across households in different network positions, and the value of x_{ij}.

That there is no closed form solution to the optimal transfers vector, or optimal consumption, poses a challenge to obtaining the types of predictions needed to guide the empirical analysis. To make progress, I solve the model numerically for a wide range of parameters, and use the simulations to generate qualitative predictions to verify in the data.

3.2.2 Comparative Statics

I use numerical simulations to shed light on how enforcement constraints, and opportunities for risk sharing affect how risk sharing varies with the number of socially close and distant connections in a household's network. In the model, enforcement constraints are embedded in the incentive compatibility constraints (3.2). Opportunities for risk sharing are allowed to vary in two ways: (i) allowing for more positive correlations in the endowment processes of socially close connections; and (ii) by varying the number

[7] When no incentive compatibility constraint binds, the KKT conditions are much simpler and yield an analytical solution.

[8] The marginal social welfare gain and optimal solution are fully defined and described in Appendix 3.7.1.

of socially close and distant connections.[9]

The specific questions to answer through the simulations are:

1. Benchmark case: For a given network size, what is the relationship between risk sharing (and welfare) and the number of socially close connections in a household's network when endowments are identically and independently distributed across households?

2. How does the extent of risk sharing (and welfare) change when opportunities for risk sharing for socially close and distant connections are changed by:

   (a) allowing for more positively correlated endowment streams for socially close households?

   (b) increasing the number of households in the network?

Measuring the Extent of Risk Sharing and Welfare

To answer these questions, I need a metric by which to measure the extent of risk sharing. The optimality conditions derived by Ambrus et al. (2014) imply one possible measure, based on household consumption and endowment realisations. This measure has the advantage of being easy to compute empirically, as long as panel data on consumption and income are available. The optimality conditions of Ambrus et al. (2014) indicate that at the optimum, the network will partition into a set of state-specific endogenous risk sharing islands. Within the islands, households will equate their marginal social gain, Δ_i, which, if the household is unconstrained in all of its incentive compatibility constraints, is simply a function of its (weighted) marginal utility of consumption, $\lambda_i u(c_i)$. If, however, the household is constrained in any of its incentive compatibility constraints, Δ_i will be a weighted sum of the household's own (weighted) marginal utility of consumption, and that of the connection with whom his incentive compatibility constraint binds the most. Thus the marginal social gain is related to households' marginal utility of consumption.

When no incentive compatibility constraint binds in all states – which corresponds with the benchmark of perfect risk sharing – there will be one risk sharing island only in the network and all households will equate their (weighted) marginal utility of consumption across all states (denoted by the subscript s below). This means that the ratio of a household's marginal utility of consumption across any two states will be a function of the ratio of aggregate network resources in the two states. That is,

[9] Note that the endowments of socially distant connections are also likely to be positively correlated, but the extent of the correlation will be lower than that for socially close connections.

$$\frac{u^{'}(c_{is})}{u^{'}(c_{is'})} = \frac{u^{'}(c_{js})}{u^{'}(c_{js'})} = \frac{\mu_s}{\mu_{s'}}$$

where μ_s is a multiplier associated with the aggregate resource constraint, and $u^{'}(.)$ represents the marginal utility of consumption.

However, when some incentive compatibility constraint binds so that there is more than one risk sharing island, this equality will no longer hold. In particular, incentive compatibility constraints are likely to bind when a household gets a good endowment draw (h) and its connections (close, or potentially even distant ones) receive a poor endowment draw (l). Thus, the cross-sectional distribution of the ratio of marginal utilities will be correlated with the cross-sectional distribution of the ratio of endowment realisations, even after accounting for aggregate resources. This observation forms the basis for the measure of risk sharing.

Assuming that $u(.)$ is of the constant relative risk aversion (CRRA) form, $u(c_{is}) = \frac{c_{is}^{1-\rho} - 1}{1-\rho}$, where $\rho \neq 0$ is the relative risk aversion parameter, and taking logs of marginal utilities implies that $\Delta log(c_{is}) = \Delta log(c_{js}) = \Delta log(\mu_s)$ when there is one risk sharing island only. Thus $\Delta log(c_{is})$ should move with aggregate network resources only. However, when there is more than one risk sharing island, $\Delta log(c_{is})$ will be correlated with $\Delta log(y_{is})$ even after accounting for changes in aggregate resources. The extent of this correlation will relate to the extent to which the incentive compatibility constraints bind. Thus, this correlation can be used as a measure of risk sharing: the closer the correlation is to 0, which is the perfect risk sharing benchmark, the higher the extent of risk sharing.

The conceptual framework also implies a second measure, which also provides and indication of welfare: the household's and planner's expected utility of consumption, $Eu(c_i)$ and $\sum_{i \in N} \lambda_i Eu(c_i)$ respectively. Given that the function $u(.)$ is concave, households would gain higher utility from having a smoother consumption stream across states of the world, and so better risk sharing should increase both of these. This measure is empirically challenging to compute since it requires knowledge of the underlying endowment process. The theoretical analysis will include both measures, while the empirical analysis will draw on the first measure only.

Details of the simulations

I numerically solve the model for the optimal consumption vector in all possible states of the world for given network structures. I focus on simulating the model for all non-

isomorphic[10] connected networks of sizes 3 and 5, for which only a relatively small number of connected networks need to be considered, and for which there is variation in the numbers of socially close and distant connections across networks.[11'12]

I make the following assumptions for the simulations:

A1. Utility is of a Constant Relative Risk Aversion (CRRA) form: $u(c_i) = \frac{(c_i)^{1-\rho} - 1}{1-\rho}$, where ρ is the coefficient of relative risk aversion.

A2. Planner weights are equal, i.e. $\lambda_i = \frac{1}{n}\ \forall i$.

A3. Link values are constant across the network. That is $x_{ij} = \bar{x}\ \forall (i,j) \in L$

A4. $v(x_i) = x_i$

Assumption A3 implies that each socially close connection has the same value associated with it. Given these assumptions, I can solve the optimisation problem described by the Equations (3.1) - (3.5) for a given network to obtain optimal transfers, and through this optimal consumption.[13] Given the optimal consumption vector, I can calculate the measure of risk sharing – the correlation between $\Delta log(c_{is})$ and $\Delta log(y_{is})$ net of changes in aggregate network resources –, as well as the welfare measures – expected household consumption utility, $Eu(c_i)$, and the social planner's weighted expected consumption utility $\sum_{i \in N} \lambda_i Eu(c_i)$. To shed light on the specific questions outlined at the start of this subsection, I compare these measures across networks with varying numbers of average socially close and distant connections for different values of correlations in the pairwise endowment.

Simulation Results

I start by fixing network size, and consider how risk sharing, and planner and household expected utility of consumption vary with the average number of socially close connections in a network,[14] in the benchmark case where endowments are identically and independently distributed (i.i.d.) across households.

To illustrate the implications of the model, I focus on networks with 5 households,

[10] Two networks are isomorphic if relabelling of nodes in one network generates the other network.

[11] Simulating the model for a representative sample of networks of larger sizes is complicated, since the number of non-isomorphic connected networks of size n is not known. Methods exist (e.g. McKay 1983) to calculate these for small n, but they indicate an exponentially fast increase in the number of possible connected network structures as n increases. For example, for $n = 5$, there are 21 connected networks possible, but this increases to 11,117 for n=8 and over 11 million for $n = 10$.

[12] In an ongoing extension, I simulate the model for network structures similar to those in the data.

[13] Note that there can be multiple possible transfers vectors that maximise the planner's expected utility. However, the optimal consumption vector will be unique since the optimisation problem involves maximising a concave function on a convex constraint set.

[14] Since network size has been fixed, the effects for socially distant connections will be the inverse of those for socially close connections.

and fix parameter values for h, l, ρ and $\bar{x}$ at 4.0, $l = 1$, $\rho = 2$, $p = 0.4$ and $\bar{x} = 0.1$.[15] Figure 3.2 displays how risk sharing, and planner and household expected utility of consumption vary with average number of socially close connections when endowments are i.i.d. across households. The Figure indicates that as the average number of socially close connections in a network increases, risk sharing improves, as indicated by declining correlations between $\Delta log(c_{is})$ and $\Delta log(y_{is})$ net of changes in network aggregate resources. Planner expected utility of consumption increases as well, as does household expected utility of consumption.

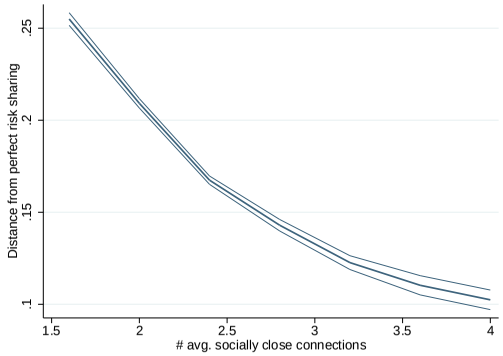
The underlying intuition for these patterns is as follows: in networks with many socially close connections on average, a household experiencing a low endowment can expect to receive more transfers before any incentive compatibility constraint towards it binds. By contrast, in networks with fewer socially close connections on average, for the same state of the world, a household experiencing a low endowment can expect to receive direct transfers from fewer close connections. Distant connections could provide indirect transfers, but these will be more limited: indirect transfers need to be made through an intermediary household, which will only pass on transfers until its incentive compatibility constraint with the household in need binds. This transfer amount will be $\leq$ the amount that could be transferred to the household had the connection been socially close rather than distant. As a result, households in networks with more socially close connections on average experience better risk sharing than those with fewer socially close connections.

I next consider the consequences of varying opportunities for risk sharing by introducing positive correlations, r_{ij}, in the endowments of socially close households. I allow r_{ij} to take values between 0 and 0.3.[16] Figure 3.3 displays how risk sharing and planner expected utility of consumption vary with the average number of socially close connections in a network; and how household expected utility of consumption varies with the household's number of socially close connections. It does so for different levels of pairwise correlation in the incomes of socially close households. Overall, risk sharing and planner and household expected utility worsen as the pairwise endowment correlation increases. This is because as the pairwise correlation increases, the probability of states where no or little risk sharing is possible also increases, leading to a reduction in

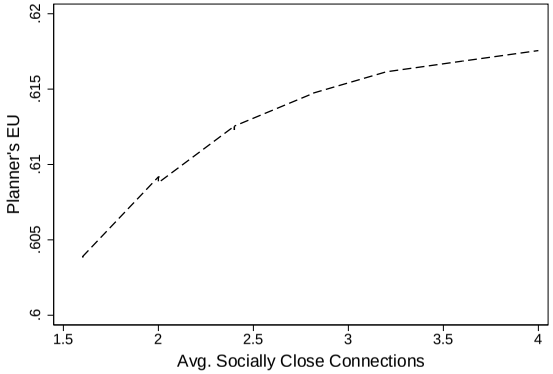
[15] These values have been chosen to ensure that some incentive compatibility constraint binds. When no incentive compatibility constraint binds, all networks achieve perfect risk sharing, and similar levels of expected consumption utility.

[16] The algorithm used to simulated correlated binary draws first converts the correlation between binary endowments into a correlation for a joint normal process. The resulting covariance matrix needs to be positive. However, this is not the case for all values of the binary correlation. In particular, for values above 0.3 the covariance matrix is not positive definite, and hence no correlated draws can be simulated for correlations > 0.3.

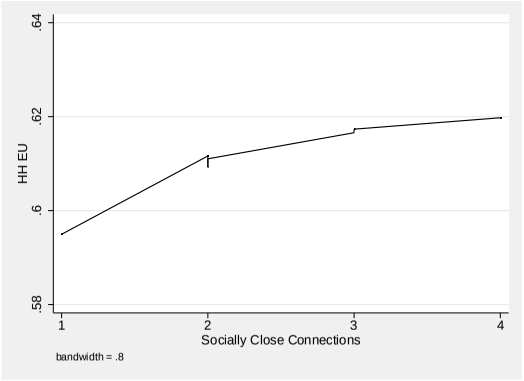
Figure 3.2: How Risk Sharing and Welfare vary with average number of socially close connections, Networks of Size 5



(a) Risk Sharing



(b) Planner's Expected Utility of Consumption



(c) Household's Expected Utility of Consumption

Notes to Figure: This Figure plots how household risk sharing (panel a), the planner's expected utility of consumption (panel b) and households' expected utility of consumption (panel c) vary with the network average number of socially close connections (household socially close connections in panel c) for networks with 5 households, when endowments are i.i.d. across households in the network. In panel (a), lower values of the risk sharing correlation imply better risk sharing; while higher values of expected consumption utility imply higher welfare in panels (b) and (c).

risk sharing and expected utility of consumption.

Moreover, as the pairwise endowment correlation increases, such that opportunities for risk sharing from socially close connections fall faster relative to those from socially distant connections, a trade-off emerges between enforcement concerns and risk sharing opportunities, which generates an inverse U-shaped relationship between the extent of risk sharing and the number of average socially close connections: networks with a moderate number of socially close connections on average obtain better risk sharing than networks with few or very many socially close connections.[17] A similar, starker relationship is obtained for the planner's, and household's expected utility of consumption.

Finally, I consider another margin for changing network-based risk sharing opportunities: changing the size of a network by adding households as either socially close or distant connections. Classical models of risk sharing imply that larger groups should achieve higher risk sharing, since additional households introduce less correlated endowment streams, and thus expand opportunities for risk sharing. However, models of risk sharing in groups with limited commitment and coalitional deviations (Genicot & Ray 2003) indicate that adding households to a network might lead to worse risk sharing, or even be unable to sustain risk sharing, since the additional households might destabilise existing risk sharing groups. It is thus important to consider how additional households affect risk sharing in this model, and whether these effects vary by where the additional household is added, i.e., whether it is socially close to all other households, or socially distant to some households.

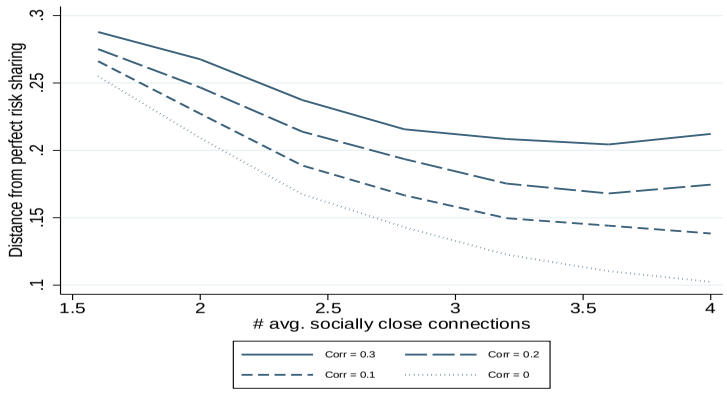
To assess the implications of this, I investigate how the planner's expected utility changes when the network size is increased from 3 to 5 households. Figure 3.4 displays the simulation results for the case where endowments are i.i.d. across households.[18] It shows that larger networks achieve higher expected utility, regardless of whether the new household is added as a socially close household to all other households, or as a socially distant connection to some households. However, the increase in expected utility is mildly higher if the new household is socially close rather than socially distant. Thus, increasing opportunities for risk sharing by increasing the number of households in the network improves risk sharing.[19]

[17] Note that lower values of the risk sharing measure imply better risk sharing.

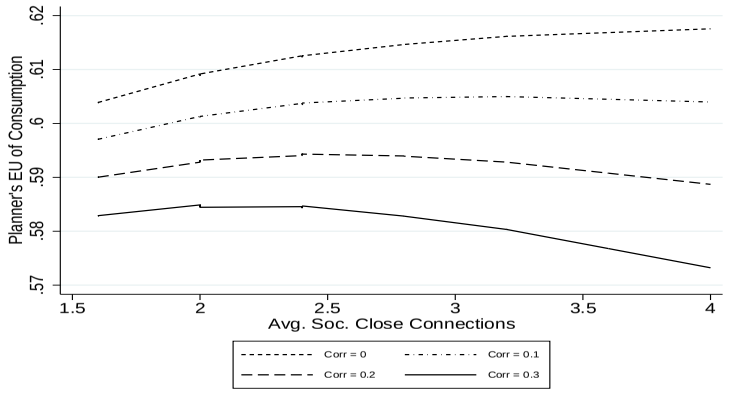
[18] Introducing positive pairwise correlations in endowments of socially close connections yields a similar picture.

[19] Though not shown here, the correlation between $\Delta log(c_{is})$ and $\Delta log(y_{is})$ net of changes in network aggregate resources also falls as the size of the network increases.

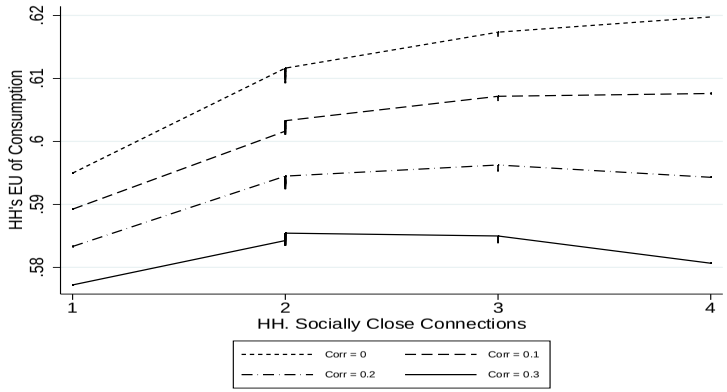
Figure 3.3: How Risk Sharing and Welfare vary with average number of socially close connections, Networks of Size 5



(a) Risk Sharing



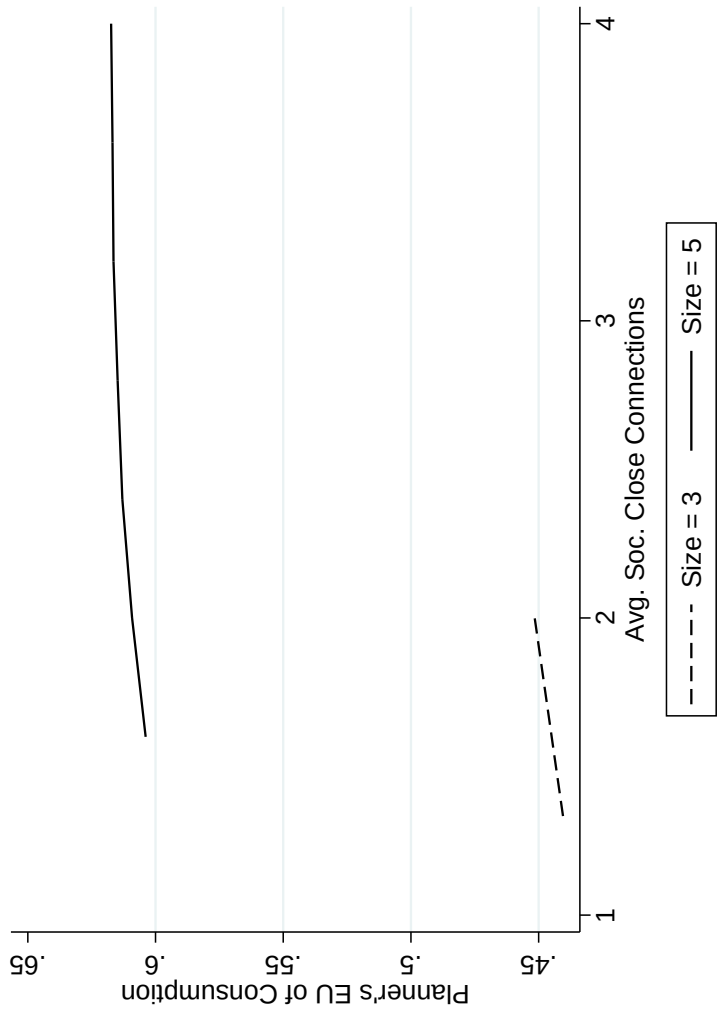
(b) Planner's Expected Utility of Consumption



(c) Household's Expected Utility of Consumption

Notes to Figure: This Figure plots how household risk sharing (panel a), the planner's expected utility of consumption (panel b) and households' expected utility of consumption (panel c) vary with the network average number of socially close connections (household socially close connections in panel c) for networks with 5 households, when endowments of socially close connections are allowed to be more positively correlated. In panel (a), lower values of the risk sharing correlation imply better risk sharing; while higher values of expected consumption utility imply higher welfare in panels (b) and (c). The Figures are plotted for $h = 4$, $l = 1$, $\rho = 2$, $p = 0.4$, $\bar{x} = 0.1$.

Figure 3.4: Planner's Expected Utility as Network Size Changes



Notes to Table: This Figure plots how the planner's expected utility of consumption varies with the average number of socially close connections in a network as the number of households in a network changes from 3 to 5. The Figure is plotted for $h = 4$, $l = 1$, $\rho = 2$, $p = 0.4$, $\bar{x} = 0.1$.

Insights from Alternative Parameter Values The analysis thus far has focused on one set of parameter values. Do the insights gleaned thus far extend to other parameter values, or are there any that are specific to the values in the example? To investigate this, I vary the values of h and l, the gap between which can be thought of as a proxy for the amount of uncertainty faced by a household in autarky. As this gap falls, imperfect enforcement concerns also fall, since the value of transfers needed to equate households' marginal utilities (and thus achieve perfect risk sharing) also falls. Instead, opportunities for risk sharing become more important, eventually leading to a monotonic, negative relationship between the planner's expected utility of consumption and average number of socially close connections. This is illustrated in Figure 3.5 illustrates when values of l are increased from 1 to 1.8 and then 2.6 with h fixed at 4.

Summary of implications from simulations The simulations thus imply the following qualitative predictions:

1. (Benchmark Case) For a given network size, when endowments are i.i.d. across households, networks and households with more socially close connections will achieve higher risk sharing and welfare.

2. When opportunities for risk sharing fall more for socially close connections compared to socially distant ones, the latter become more important for risk sharing. A trade-off emerges between risk sharing opportunities and enforcement, yielding an inverse U-shaped (U-shaped) relationship between the extent of risk sharing and the number of socially close (distant) connections in a household's network.

3. Improving opportunities for risk sharing by adding new households to a network improves risk sharing and welfare.

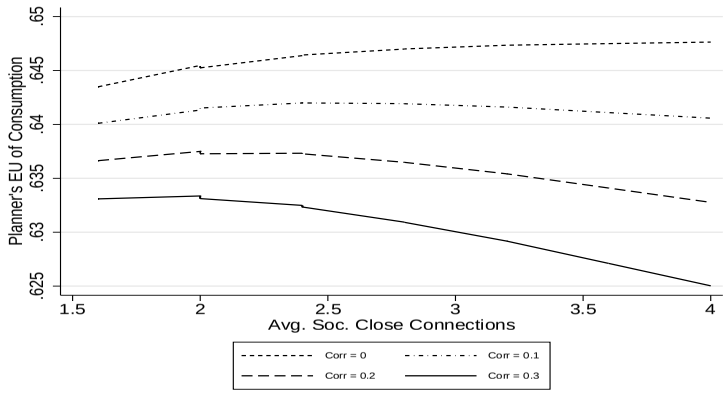
I now investigate whether there is support for these observations empirically in data on within-village extended family networks in rural Mexico.

3.3 Context and Data

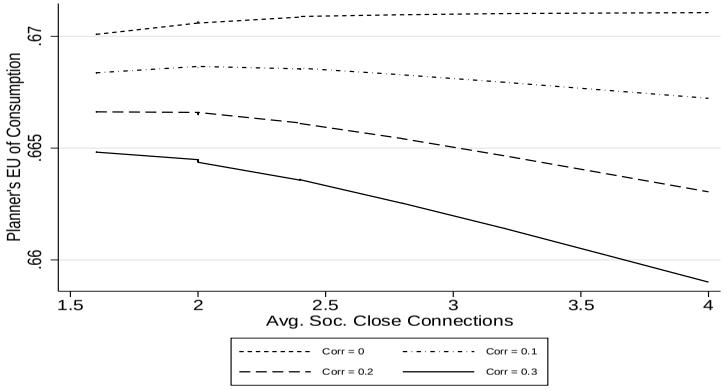
3.3.1 Context

The empirical setting is a set of poor, marginalised villages in rural Mexico, which were targeted by a conditional cash transfer anti-poverty programme, PROGRESA (later called *Oportunidades*, and now called *Prospera*). These villages are small (47 households on average), isolated – the closest city with at least 100,000 inhabitants is around 62 km away on average – and have limited access to formal markets: in the data (described

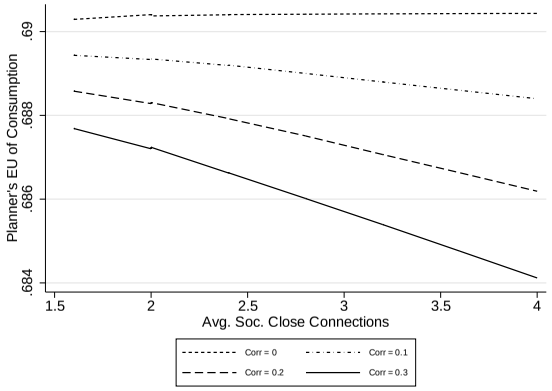
Figure 3.5: Planner's expected utility of consumption, varying values of h and l, networks of size 5



$h = 4;\ l = 1$



$h = 4;\ l = 1.8$



$h = 4;\ l = 2.6$

Notes to Table: This Figure plots how the planner's expected utility of consumption varies with the average number of socially close connections in a network for different values of l, for networks of size 5. The Figures are plotted for $h = 4$, $\rho = 2$, $p = 0.4$, $\bar{x} = 0.1$.

further below), only 3% of villages have a post office, 25% a public phone, while fewer than 20% have a government subsidised Diconsa shop, and only 36% have a grocery shop. Households in these villages are poor – only 40% have dwellings with good flooring materials, and 7% have access to piped water in their dwelling. A large proportion of them (70%) rely on rain-fed agriculture as their main source of income, and are subject to significant risk: around 35% (25%) of households in the data experienced a crop loss in 1998 (1999).

Despite facing significant income risk, the data indicates limited ex-ante smoothing of income: the vast majority of households (79%) engage in only one occupation. Among those engaged in agriculture, most grow one crop – corn – only. This is consistent with the presence of liquidity or insurance constraints, which prevent households from diversify into higher-return but riskier and unfamiliar crops (e.g. Karlan et al. 2014). Risk reducing technologies such as irrigation are uncommon: $< 10\%$ of households have irrigated plots. Moreover, another common income diversification strategy – migration is not very common in this context: data from October 1998 indicates that only around 7.5% of households report having a household member who had migrated for work in the 5 years preceding the survey, compared to 16% reported by Davis et al. (2002) for a broader set of rural villages in Mexico.

Extended Family Networks are Important for Risk Sharing Households thus face risky income streams, with limited recourse to formal financial instruments to help cope with the consequences of this risk. Instead, informal tools, which rely on pre-existing social connections play a crucial role for risk sharing. Existing evidence indicates that the extended family, in particular, plays an important role in providing insurance. Angelucci et al. (2015) show that households in this sample (which is the same as that used in their paper) rely on their within-village extended family connections to share risk, and cannot reject perfect risk sharing among these networks: households with within-village family connections have consumption streams which are uncorrelated with their incomes, net of aggregate network level shocks. By contrast, consumption and income co-move when the village is taken to be the relevant risk sharing group. Descriptive analysis of interhousehold transfers sent and received by sample households in the month prior to the survey also support the importance of the extended family for risk sharing in this context: the bulk of transfers (91% of monetary transfers; 89% of the volume of monetary and in-kind transfers) received by households are from relatives, while around 70% of monetary transfers sent to other households are to relatives. Thus extended family networks play an important role in informal risk sharing in this context, and will be the network within which I study the varying roles

of socially close and distant connections for risk sharing.

I focus specifically on a household's within-village extended family network. This will probably constitute only a subset of a household's extended family, since a fraction is likely to reside in other villages or towns.[20] These members might also be able to assist households with risk sharing. However, note that these villages are relatively isolated and have poor infrastructure, making it costly to send transfers or other help from outside the village. Moreover, it is easier for within-village extended family connections to offer in-kind help, such as labour-sharing. Given these issues, it is plausible that the within-village extended family connections are more effective in helping households deal with idiosyncratic risk, while outside village connections might be more valuable for coping with village level aggregate shocks, as well as very large idiosyncratic shocks.

3.3.2 Data

The empirical analysis draws on rich panel data collected to evaluate the Progresa cash transfer programme. Data was collected on *all* households in 506 rural villages in 7 states across Mexico over the period 1997-2003.[21] Baseline data was collected in Fall 1997, and follow up data was collected on 6-monthly intervals from May 1998 to November 2000, and then again in 2003, thus providing a relatively long panel.[22] The panel dimension of the data is crucial for the analysis, since it allows me to construct a measure of risk sharing, and also to account for fixed unobserved variables that might generate spurious correlations between the measures of risk sharing and social distance in the analysis. Moreover, the surveys also collected detailed socio-economic information, including data on income from numerous sources, consumption, household demographics, occupational and labor supply choices of all household members aged $\geq$ 8 years and migration.

The surveys did not, however, directly elicit information on inter-household extended family connections. However, I can identify such connections by exploiting the Mexican naming convention whereby individuals have 2 surnames – one from the father's parental lineage and the other from the mother's parental lineage – to identify extended family links across households within villages. I do so by applying an algorithm simi-

[20] Unfortunately, data limitations force me to restrict attention to within-village networks. To my knowledge, no dataset contains information on individuals' or households' entire social network. Indeed, collecting such data without implementing any geographic boundary is likely to be prohibitively costly, and infeasible, even in developing country settings. Despite this limitation, the data used in this paper provides a detailed picture of the within-village extended family network, along with a panel of socio-economic variables including income and consumption, which is particularly suited to the study of risk sharing.

[21] 320 villages were randomly chosen to receive the intervention, with a further 186 villages remaining as control villages. I pool together data from all the villages in the analysis.

[22] A further round of data, not used in this analysis, was collected in 2007.

lar to that used by Angelucci et al. (2009), which is described in more detail below.[23] Three features of this dataset make it particularly useful for studies related to networks. First, I have available information on the exact paternal and maternal surnames of the household head and spouse for 2 survey rounds (October 1998, and November 1999). Second, the surveys interviewed *all* households in a village. This means that I can apply the algorithm described in more detail below to identify all family links within a village, and obtain a complete picture of the structure of within-village extended family links as identified by the algorithm. Having a census of all households in the village is particularly important for the latter, since missing data on households or connections between households can generate severe non-classical measurement error in measures of the network, as well as in regression estimates (Chandrasekhar & Lewis 2011). Finally, the detailed socio-economic variables available help in improving the accuracy of the algorithm.[24]

Identifying Network Connections

I use a modified version of the algorithm applied by Angelucci et al. (2009) to identify within village extended family links using data from the October 1998 survey, which is the first round for which the names information is available.[25] The algorithm exploits the Mexican naming convention whereby individuals have 2 surnames – one from the father's parental lineage and the other from the mother's parental lineage – to identify extended family links across households within villages. For example, the wealthiest Mexican, Carlos Slim Helu, is known by his given name, Carlos, his paternal surname, Slim, and maternal surname, Helu. I will use the surnames of the head and spouse of a household to identify cross-household links. Since each individual has 2 surnames, couple-headed households will have 4 surnames that will be used for this purpose.

Figure 3.6 provides an illustration of the matching algorithm. The Figure displays 5 households, with the surnames of the head of household displayed in blue boxes and those of the spouse displayed in red boxes. H indicates the head of household and S the spouse of the head. The head of household 1 has paternal surname F1, and maternal surname M1, while his spouse has paternal surname F2 and maternal surname M2. Their children would have F1 as their paternal surname and F2 as the maternal surname, which is the surname combination of the head of household 2 and the spouse

[23] Algorithms based on surname combinations have also been used by Cruz et al. (2015). Information on surnames has also been used to study intergenerational mobility (Guell et al. 2015; Clark 2014)

[24] As will be described below, the algorithm makes use of information on age, and information from the household roster to reduce the likelihood of identifying spurious connections.

[25] Empirical results are very similar when I construct networks from the information in the 1999 survey or when I pool together both rounds and apply the algorithm to the pooled data.

of household 3. Hence, there is a parent/child link between households 1 and 2; and households 1 and 3. Moreover, siblings would have the same paternal and maternal surnames, as is the case for the head of household 2 and the spouse of household 1. By contrast, the head and spouse of household 5 have surname combinations that do not match with any of the other households, indicating that they do not have any sibling or parent/child connections with any of the other 4 households.

I combine the information from surname combinations with age restrictions to identify sibling and parent/child links within a village. Restricting links to be within the same village helps reduce the likelihood of identifying spurious links.[26] Sibling groups are identified as follows: two individuals are identified to be part of the same sibling group if they share the same paternal and maternal surnames, and if the age difference between the oldest and youngest 'sibling' is ≤ 30 years.[27] For parental ties, two households are identified to be related via parental/filial ties if the paternal surname of the (male) head and (female) spouse corresponds with the paternal and maternal surnames of the head or spouse of the other household. In addition, I impose the condition that the mother must be at least 15 years older than her eldest child, and at most 45 years older than her youngest child.[28]

Descriptive Statistics of the Identified Connections

The results of the algorithm are displayed in Tables 3.1 and 3.2. The algorithm identifies at least one household-level family link for almost 80% of couple-headed households (households where both the head and spouse are present) and for 44% of non-couple headed households. On average, couple-headed households have just over 3 family connections within the village, including 2.67 sibling links, 0.31 parental links and 0.33 child links. Non-couple headed households are not only less likely to have a family connection, but also have fewer connections - 1.21 on average.

[26] In addition, the use of the combination of two surnames also greatly reduces the likelihood of spurious links being identified.

[27] This differs from the algorithm used by Angelucci et al. (2015) who impose a weaker condition that any two individuals identified to have the same paternal and maternal surnames are siblings if the age difference between them is at most 30 years. Their algorithm thus allows for cases where two individuals identified to be siblings may have siblings who are not identified to be each other's siblings, thus leading to errors in the identified network structure.

[28] I experimented with a looser upper age cutoff for mothers, with little effect on the estimated parameters. In Section 3.5.2, I show that tightening the age cut-offs applied has little effect on the parameter estimates.

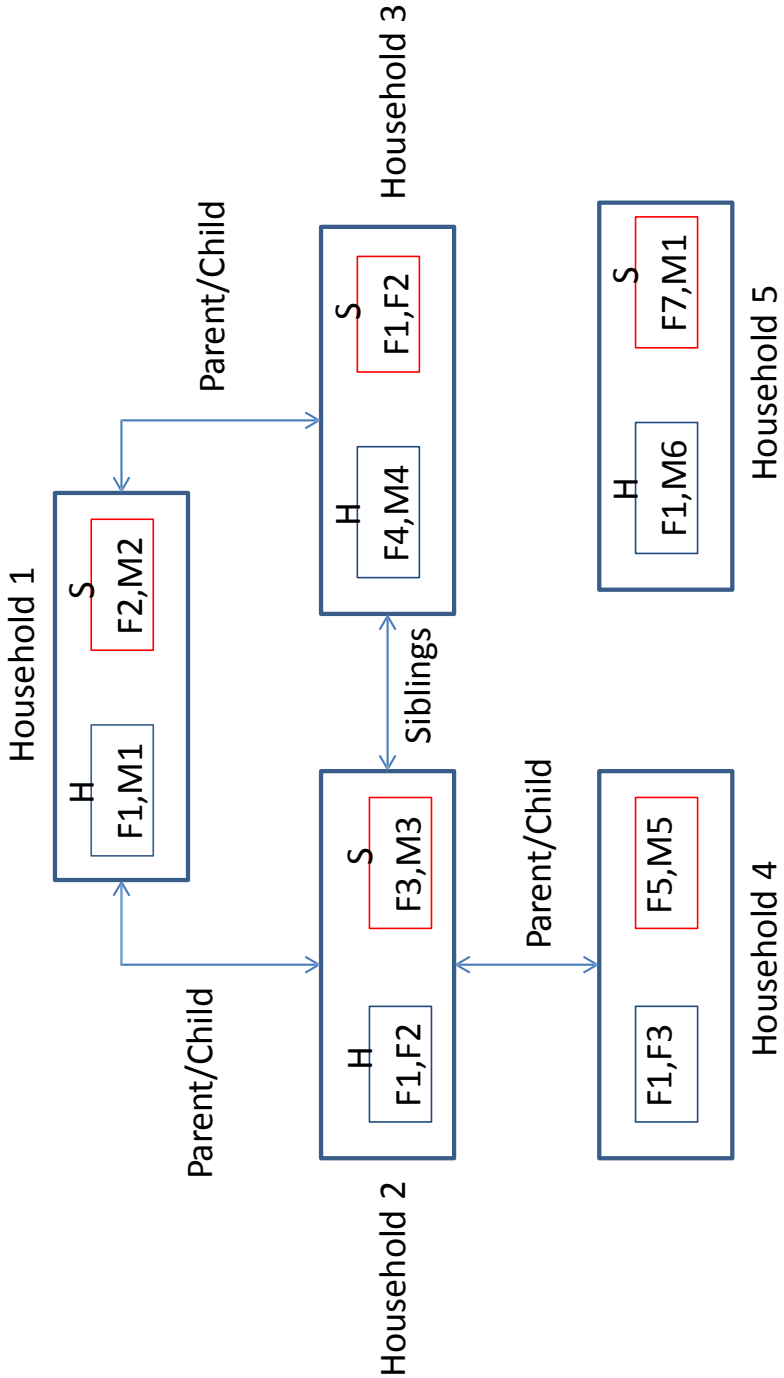


Figure 3.6: Matching Algorithm

Table 3.1: Any connections of couple-headed and non-couple-headed households

	Any Link	Any Parental Link	Any Child Link	Any Siblings
Couple-Headed Households	0.797	0.246	0.163	0.707
	[0.007]	[0.006]	[0.003]	[0.008]
N	19,143	19,143	19,143	19,143
Non-Couple-Headed Households	0.444	0.061	X	0.428
	[0.010]	[0.004]	X	[0.010]
N	4,428	4,428	4,428	4,428

Notes to Table: The table includes all households in the October 1998 round of data for whom surname information was available. Couple-headed households are those with a co-resident spouse, while non-couple-headed households are those without a co-resident spouse. All links are inter-household connections within the household identified by the algorithm described in Section 3.3.2. Standard errors clustered at the village level are in square brackets. The algorithm doesn't identify, by definition, any child links for non-couple-headed households.

Table 3.2: Number of connections of couple-headed and non-couple-headed households

	Number of Links	Number of Parental Links	Number of Child Links	Number of Siblings
Couple-Headed Households	3.317	0.313	0.330	2.672
	[0.155]	[0.009]	[0.010]	[0.145]
N	19,143	19,143	19,143	19,143
Non-Couple-Headed Households	1.210	0.076	X	1.134
	[0.065]	[0.006]	X	[0.063]
N	4,428	4,428	4,428	4,428

Notes to Table: The table includes all households in the October 1998 round of data for whom surname information was available. Couple-headed households are those with a co-resident spouse, while non-couple-headed households are those without a co-resident spouse. All links are inter-household connections within the household identified by the algorithm described in Section 3.3.2. Standard errors clustered at the village level are in square brackets. The algorithm doesn't identify, by definition, any child links for non-couple-headed households.

A detailed discussion of the algorithm performance and measurement error associated with a similar algorithm can be found in Angelucci et al. (2009). They show that the average number of identified links is within the range of those reported by similar households in the Mexican Family Life Survey, which directly elicited this information. Moreover, the proportion of individuals and households for whom implausible numbers of sibling links, and/or multiple possible parental links are identified is very small, which is reassuring.[29] Finally, analysis in Angelucci et al. (2009) indicates that the identified networks are correlated with observed characteristics in ways that are reasonable, and can be explained by economic models. In addition, I conduct some sensitivity analysis of our parameter estimates by varying the age restrictions in the algorithm, and find that the qualitative results continue to hold for all the alternatives considered. These results are shown in Section 3.5.2.

For the analysis, I retain households in networks with at most 100 households.[30]

[29] In one village, a large proportion of individuals had similar surname combinations, which reduced greatly the power of the algorithm in identifying family links. I thus drop this village from the subsequent analysis.

[30] I impose an upper limit on network size so as to alleviate potential biases arising from spurious

This implies a final sample of just over 16,000 households in close to 2500 networks in 501 villages.

Identifying Socially Close and Socially Distant Connections

I use a network theory based definition of social distance to define socially close and distant connections. To identify these, I use the connections generated by the algorithm to construct the map (or network graph) of cross-household extended family links within the same village. Two households are considered to be part of the same network if there exists a path through the network for one household to get to the other: essentially, they are part of the same network if they are connected either directly or indirectly through sibling, parent and child links. Based on the network graph, I can identify the socially close and socially distant connections for each household. Socially close connections are those with whom a household has a direct link: siblings, parents and adult children of the household head and spouse. Socially distant connections, by contrast, are those households that are part of the same network, but to whom the household is only indirectly connected. In this context, they are the siblings and parents of one's siblings' spouses; or intergenerational connections such as grandparents, uncles and aunts and cousins.

Table 3.3 displays descriptive statistics of measures of the network structure for households in the estimation sample. It focuses particularly on variables relating to social distance: the size of the network, numbers of socially close and socially distant connections (at the network- and household-levels), and the network average path length.

The table indicates that the average (median) household is in a network with 24.3 (14) households, of whom 3.6 (3) are socially close connections, and 19.7 (9) are socially distant. Thus, households are in networks with around 6 times more socially distant connections than close connections on average. Overall, the networks are closely knit, with an average shortest path length across networks of 2.52. Finally, the table also indicates that there is substantial variation in these measures of network structure across households.

connections identified by the algorithm, which would be more likely in particularly large networks. This condition leads to dropping around 1000 households in 6 networks.

Table 3.3: Descriptives of network structure

Measure	Mean	Std Dev	Median	Min	Max
Size	24.34	24.62	14	2	96
Avg. Socially Close Connections	3.64	2.45	3.36	1	25.09
Avg. Socially Distant Connections	19.70	23.18	9.29	0	84.56
Avg. Path Length	2.52	1.35	2.24	1	6.52
HH Socially Close Connections	3.45	2.76	3	1	16
HH Socially Distant Connections	19.89	23.51	9	0	94
N	16,053				

Notes to Table: The Table includes 16053 households in 2487 extended family networks with between 2-100 households constructed from family connections identified using the algorithm described in Subsection 3.3.2. Size captures the number of socially close and socially distant connections + 1 for each household in the network. Note that I trim households with outlying values of degree (the top 1% of the degree distribution).

3.3.3 Do socially close and distant connections offer different opportunities for risk sharing?

A central argument of this paper is that opportunities for risk sharing are important for the effective functioning of informal insurance arrangements; and this is a margin along which socially close and distant connections might vary for two reasons: (i) they may vary in their economic similarity; and (ii) they may vary in number. I now verify whether this is the case in the data. When risk sharing opportunities are important, households should choose to form risk sharing connections with those households who have uncorrelated or even negatively correlated income streams to their own. However, enforcement frictions suggest sharing risk with connections with whom one interacts frequently, e.g. family, who might also be similar on other dimensions, as I document below.

The descriptive statistics of the network architecture displayed in Table 3.3 in the previous section offer some initial evidence that supports the hypothesis that socially distant connections might offer more opportunities for risk sharing than socially close connections. They indicate that households have on average almost 6 times as many socially distant connections as socially close ones.

However, socially distant connections might also provide more opportunities for risk sharing if they are more economically different than socially close connections. This might happen for a few reasons: first, socially close connections (parents, adult non-resident children and siblings) might have similar endowments and abilities relative to socially distant connections, leading to similar occupation choices, and hence similar income processes. These could further be reinforced by assortative mating in the marriage

market. Moreover, production technologies and inputs that socially close households have access to might also be more similar in quality than those of socially distant connections. For example, if the main source of land for agriculture is through allocations from one's parents, two brothers will be more likely to farm neighbouring plots, which are likely to be of similar quality, than two cousins and therefore face similar localised shocks (e.g. pests).

Second, a large literature has documented that labour markets in village economies are far from perfect (see, for example, Bandiera et al. 2015), and trusted contacts – socially close connections – might be important for finding jobs. Thus, a household might have a higher probability of being in a similar occupation as its socially close connections. Moreover, credit and liquidity constraints, and a lack of occupation-specific skills might prevent individuals from choosing an occupation that is different from that of their parents and siblings, who are likely to be able to overcome these constraints for their specific occupation: for example, parents would be able to show their adult children how to grow specific crops, and also be able to provide them with land, seeds and other inputs.[31] Put together, these reasons imply that socially close connections might be more economically similar than distant connections.

To investigate whether this is the case, I study the household head's main occupation choice as reported in the October 1998 survey for socially close and distant connections.[32] Occupation is likely to be highly correlated with a household's income process – households in the same occupation are likely to be subject to similar risks and shocks – and is thus an important margin to consider. In particular, I ask whether the heads of households that are socially close are more likely to be in a similar occupation than heads of households who are more socially distant.

I do so by computing, for each household, the proportion of the heads of household of their socially close and socially distant connections that are engaged in the same occupation as the household head and use pairwise t-tests to evaluate whether these proportions are statistically different from one another. Table 3.4 displays these statistics.

The table indicates that around 57% of households' socially close connections' heads are engaged in the same occupation as the household head, compared to just over 52% of heads of socially distant connections. Much of this variation comes from households engaged in agriculture: for these households, a higher proportion of their socially close

[31] Bianchi & Bobba (2013) document that insurance constraints prevent households in this context from diversifying occupation within household.

[32] This is the first survey for which I observe occupations for all households for whom I can impute the network connections.

Table 3.4: Similarity in Occupation Choices of the Head of a Household and that of its Socially Close and Distant Connections

Variable	Socially Close	Std Dev	Socially Distant	Std Dev	Diff.
Both in same occupation	0.569	0.014	0.525	0.018	0.045***
Both in agriculture	0.652	0.014	0.623	0.016	0.028***
Both in non-agricultural occupations	0.288	0.018	0.276	0.021	0.012

Notes: *** p<0.01; ** p<0.05; * p<0.1. Standard errors clustered at the village level. Socially close connections are those who are directly connected to a household, while socially distant connections are those who are at a social distance of 2 or greater from the household. The table displays the proportion and std. deviations of a household's socially close and distant connections whose heads are engaged in the same occupation as the household head, including a breakdown by whether the head is engaged in an agricultural or non-agricultural occupation.

connections are also engaged in agriculture compared to their socially distant connections. By contrast, for households engaged in non-agricultural occupations, a marginally higher proportion of socially close connections are engaged in non-agricultural occupations relative to socially distant connections, though this difference is not statistically significant.

A natural question is whether these differences in occupation choices among socially close and socially distant households are significant enough to translate into differences in correlations in the income processes of these types of connections. Households could be engaged in the same occupation, and (theoretically) still face uncorrelated income processes, because, for example, they make production choices in a manner that makes incomes orthogonal to one another. The detailed data on income available for multiple survey rounds allows me to shed light on this question.

Specifically, for each pair of households in the same network, I calculate the pairwise correlation in their incomes. I then regress this pairwise correlation on the social distance between the two households using a specification of the following form:

$$Corr(y_{in}, y_{jn}) = \alpha_0 + \alpha_1 1(d_{ijn} > 1) + \nu_n + \xi_{ijn} \tag{3.6}$$

where $Corr(y_{in}, y_{jn})$ is the pairwise correlation in income of households i and j, $1(d_{ij} > 1)$ takes the value of 1 if households i and j are socially distant, and 0 if they are socially close, and ν_n is a network fixed effect which captures all network-level time invariant unobservables that may be correlated with both the pairwise correlations and social distance. I adjust standard errors for correlations arising from the fact that the same households are part of many household pairs (or dyads) by calculating Huber-

Table 3.5: Pairwise income correlations and social distance

	$Corr(y_{in}, y_{jn})$
Socially Distant	-0.014**
	[0.006]
Constant	0.077***
	[0.005]
Observations	354,182
R-squared	0.029

Notes to Table: *** p<0.01, ** p<0.05, * p<0.1. Standard errors clustered at the village level in brackets. Dependent variable is the pairwise correlation in per capita income for households i and j in the same network. Observations are for pairs of households in the same network (or dyads).

White standard errors clustered at the village level.

The results, displayed in Table 3.5, indicate that on average, the raw pairwise income correlation for socially close households is positive at just under 0.08. Moreover, there is a statistically significant negative correlation between the pairwise income correlation and social distance: relative to socially close connections, a household's income is less positively correlated with the income of more distant connections.

Thus, socially distant connections are economically more different than socially close connections. Putting together this evidence with that highlighted earlier – that socially distant connections are more numerous than close connections – indicates that socially distant connections will provide more opportunities for risk sharing in this context. Moreover, this channel is likely to be relevant in risk sharing networks in a range of contexts. Fafchamps & Gubert (2007), for example, document that geographic proximity, which facilitates enforcement, is strongly correlated with the presence of a risk sharing tie in the Philippines, and actual gifts; though they find no role for income correlation or social distance.

3.3.4 Social Distance and Household Income Fluctuations

I now consider how household income fluctuations vary with the number of socially close and distant connections in a household's network. I focus specifically on changes over time in household log income, $\Delta log(y_{int})$, and the time-series variance of log income, $Var_i(log(y_{int}))$. Tables 3.6 and 3.7 display the correlations for $\Delta log(y_{int})$ and $Var_i(log(y_{int}))$ respectively. The tables indicate very small and statistically insignificant correlations between household income fluctuations and the number of household and average network socially close and distant connections, thereby suggesting that

Table 3.6: Household income fluctuations and network characteristics

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: $\Delta\, log(c_{int})$				
Avg. Soc. Close Connections	-0.0012				
	[0.0020]				
Avg. Soc. Distant Connections		0.0002			
		[0.0002]			
Size			0.0002		
			[0.0002]		
HH. Soc. Close Connections				-0.0005	
				[0.0012]	
HH. Soc. Distant Connections					0.0002
					[0.0002]
Observations	43,308	43,308	43,308	43,308	43,308
R-squared	0.0160	0.0160	0.0160	0.0159	0.0160

Notes to Table: Standard errors clustered at the village level in brackets. All regressions include survey round dummies. *** p<0.01, ** p<0.05, * p<0.1.

while network characteristics (specifically social distance) affect the correlations in incomes of connected households, they are not associated with higher or lower variability in a single household's incomes.

3.4 Empirical Framework

I now introduce the empirical framework applied to investigate how risk sharing varies with the number of socially close and distant connections in a household's network. To answer this question, I use the first measure of risk sharing implemented in the numerical simulations in Section 3.2: the correlation between $\Delta log(c_{is})$ and $\Delta log(y_{is})$, net of aggregate network resources. This measure has been widely used in the literature on consumption smoothing (e.g. Townsend (1994)), and can also be motivated from the theoretical framework. As outlined above, this measure will be 0 when the network provides perfect risk sharing, and > 0 when risk sharing is partial. Empirically, I observe households experiencing different states of the world at different time periods. I thus assume that each time period offers a snapshot of a different state of the world. This

Table 3.7: Household Income Variance and Network Characteristics

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: $Var_i\, log(y_{int})$				
Avg. Soc. Close Connections	-0.0057				
	[0.0080]				
Avg. Soc. Distant Connections		0.0005			
		[0.0008]			
Size			0.0004		
			[0.0007]		
HH Soc. Close Connections				0.0071	
				[0.0056]	
HH Soc. Distant Connections					0.0003
					[0.0008]
Observations	14,569	14,569	14,569	14,569	14,569
R-squared	0.0001	0.0001	0.0001	0.0002	0.0000

Notes to Table: Standard errors clustered at the village level in brackets. *** p<0.01, ** p<0.05, * p<0.1.

assumption will be reasonable as long as income is not persistent over time in the data. This is likely to hold in this setting, since the gap between surveys is at least 6 months.

Rather than using a direct measure of risk sharing such as inter-household transfers, my measure relies on household consumption, which is advantageous from a measurement perspective. In particular, it does not rely on knowledge of the exact tools employed by households to share risk, which can be tricky to capture accurately in standard household surveys.[33] Instead, consumption should capture the net benefits of all the different tools utilised by households, thereby providing a more accurate summary measure of a household's risk sharing position.

I use this measure to shed light on how risk sharing varies with the average number of socially close and socially distant connections in a household's network. My main empirical specification, given in Equation 3.7, regresses changes in per-capita log consumption for a household i in a network n, $\Delta log(c_{int})$, on a vector of network-time dummies (which capture changes in network-level aggregate resources), μ_{nt}, changes in per-capita log household income $\Delta log(y_{int})$, and the changes in per-capita log household income interacted with the number of socially close and socially distant connections

[33]For example, Comola & Fafchamps (2015) show that households may respond to such questions in a strategic manner; while Mtika & Doctor (2002) uncover qualitative evidence of substantial underreporting of transfers (monetary and in-kind) among extended family connections where these transfers are frequent.

denoted by $w_i(G_n)$.

$$\Delta log(c_{int}) = \mu_{nt} + \beta_1 \Delta log(y_{int}) + f(w_i(G_n)) * \Delta log(y_{int})\beta_2 + \gamma \Delta X_{int} + \epsilon_{int} \tag{3.7}$$

The theory indicated that, depending on underlying parameter values, the relationship between the number of socially close and distant connections and risk sharing might be non-linear. The function $f(.)$ allows $w_i(G_n)$ to affect risk sharing in a non-linear fashion. The model in Section **??** indicates that $f(.)$ may be U-shaped (inverse U-shaped). Thus I allow $f(.)$ to be quadratic.The specification also controls for time-varying household characteristics (particularly household demographics) that are related to both changes in per-capita log consumption and log income, which are included in the vector, X_{int}. If risk sharing were perfect across all networks, I would expect the sum of the coefficients $\beta_1 + \beta_2 f^{'}(w_i(G_n)) = 0$ for all households, where $f^{'}(w_i(G_n))$ is the first derivative of $f(w_i(G_n))$. If risk sharing is partial, the sum of these coefficients will be > 0. Moreover, improvements in risk sharing from socially close or distant connections would imply that the marginal effect, $\beta_2 f^{'}(w_i(G_n)) < 0$.

Note that $w_i(G_n)$ does not enter the regression on its own, since the specification is in terms of first differences, and the measures of the number of socially close and socially distant connections are constant over time.[34]

To assess how the number of socially close and distant connections affect risk sharing, I first define $w_i(G_n)$ as a scalar in the average number of socially close or the average number of socially distant connections in the network. To ease comparisons across coefficients, I standardise the variables for the number of socially close and distant connections by subtracting the mean of each variable's distribution and dividing by the standard deviation of the variable's distribution.

The theoretical framework indicated that better enforcement of informal arrangements – offered by networks with more socially close connections here – and more opportunities for risk sharing – offered by networks with more socially distant connections in this setting – should both yield better risk sharing. If this is the case empirically, the marginal effect $\beta_2 f^{'}(w_i(G_n))$ will be < 0.[35]

The framework also allows me to study the relative importance, empirically, of limited commitment (or imperfect enforcement) frictions and risk sharing opportunities for social connections to be effective in providing informal risk sharing. I can do this be defining $w_i(G_n)$ to be a vector of the average number of socially close and dis-

[34] This is a reasonable assumption as I study households over a relatively short period of time, over which there are few changes in the status of the household head and spouse.

[35] It is not possible to include the size of the network as an additional control variable in regressions with the number of socially distant connections, since these variables are highly correlated in the data.

tant connections in the household's network, and studying the relative magnitudes of the marginal effects of the two coefficients. Denoting the marginal effect of socially close connections to be $\beta_{2,1}f^{'}(\#(d_{ijn}=1))$ and that of socially distant connections as $\beta_{2,2}f^{'}(\#(d_{ijn}>1))$, I expect $\left|\beta_{2,1}f^{'}(\#(d_{ijn}=1))\right| > \left|\beta_{2,2}f^{'}(\#(d_{ijn}>1))\right|$ if limited commitment frictions are more important than risk sharing opportunities. However, if risk sharing opportunities are more important than limited commitment frictions, $\left|\beta_{2,1}f^{'}(\#(d_{ijn}=1))\right| < \left|\beta_{2,2}f^{'}(\#(d_{ijn}>1))\right|$. Finally, the theory also indicated that networks with more households should achieve better risk sharing. I study this by defining $w_i(G_n) = K_n$, where K is the size of network n. If larger networks provide more risk sharing, the coefficient $\beta_2 f^{'}(K_n) < 0$.

In terms of inference, I cluster standard errors at the village level, which allows for correlations in the unobserved errors for households in the same, as well as different, extended family network(s) within the same village. Valid inference using this method requires a large number of independent clusters, a feature that is satisfied in my sample.[36]

A remark is at hand on identification. A key concern hampering causal interpretation of the coefficient (vector) β_2 is that the average number of within-village socially close and socially distant connections of a network might be correlated with unobserved variables that are also correlated with my measure of risk sharing. Focusing on extended family connections alleviates, at least partially, endogeneity concerns since households do not choose their sibling and parent/child connections. However, the number of these connections residing in the village might be endogenous as a result of fertility, marriage, migration (for work) and household formation choices made depending on unobserved variables that are correlated with risk sharing. The availability of household panel data allows me to further partially (though not completely) alleviate this issue. In particular, my key estimation equation is in first differences, which purges out any household-level unobservables that are fixed over time (for example, unobserved preferences), that may be correlated with the number of socially close and distant connections and risk sharing choices. Moreover the network-time dummies, not only absorb aggregate network shocks, but also account for fixed unobserved variables at the network level, such as village size and amenities, that might also be correlated with both the number of socially close and socially distant connections and the dependent variable through channels other than risk sharing. Finally, in Section 3.5.2, I present some robustness checks

[36] A concern may be that extended family networks in neighbouring villages might not be completely uncorrelated. Ignoring these correlations may yield standard errors which are too small. To assess the importance of this concern, I conduct some robustness analysis where I conduct inference using standard errors clustered at the municipality (which is a higher administrative level than a village) level. There are 191 municipalities in my sample. Inference remains unchanged and the main results still hold.

which suggest that biases arising from endogeneity of the network are unlikely to be driving the empirical findings.

3.5 Results

In this section, I present the results of the empirical analysis. I first show how the amount of risk shared varies with the average number of socially close and socially distant connections in a household's network, before outlining analyses undertaken to probe the robustness of the findings.

3.5.1 Risk Sharing and Socially Close and Socially Distant Connections

I estimate Equation 3.7 with, in turn, the average number of socially close connections and average number of socially distant connections in a household's network, before including both variables entering together to shed light on the relative importance of socially close and distant connections on risk sharing in this setting. The theoretical analysis in Section 3.2 indicated that the relationship between the extent of risk sharing, as measured by the correlation of $\Delta log(c_{int})$ and $\Delta log(y_{int})$ net of aggregate network shocks, is potentially non-linear with respect to the average number of socially close and distant connections. I incorporate this in the specification by allowing for $f(w_i(G_n)) = \{w_i(G_n),\ w_i(G_n)^2\}$ in addition to $f(w_i(G_n)) = w_i(G_n)$. To ease comparison of magnitudes of coefficients across the different measures, I standardise each of the network measures to have a mean of 0 and standard deviation of 1. Table 3.8 reports the results for this specification. A negative marginal effect on the interaction term(s), $f(w_i(G_n)) * \Delta log(y_{int})$, indicates improvements in risk sharing, while a positive coefficient indicates the converse.

The first column of Table 3.8 indicates that households embedded in networks with more socially close connections do not achieve more risk sharing than those in networks with fewer socially close connections. The coefficient is relatively small in magnitude, and statistically insignificant from 0. Adding the quadratic term does not reveal any nonlinearity as can be seen from the second column of the table. However, more socially distant connections are associated with an improvement in risk sharing, as is evident from Column 3. This provides some initial evidence that opportunities for risk sharing are important. As with socially close connections, the estimates suggest no nonlinearity in this relationship: the quadratic term is far from statistically significantly different from 0. Further evidence on the importance of risk sharing opportunities comes from

Table 3.8: Risk Sharing and Network Average Socially Close and Socially Distant Connections

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent Variable: $\Delta\, log(c_{int})$							
$\Delta\, log(y_{int})$	0.0390***	0.0392***	0.0401***	0.0403***	0.0401***	0.0400***	0.0401***	0.0404***
	[0.0037]	[0.0038]	[0.0037]	[0.0037]	[0.0037]	[0.0037]	[0.0037]	[0.0037]
$\Delta\, logy_{int}$ **interacted with:**								
Av. Socially Close (St.)	-0.0053	-0.0080			-0.0004	0.0061		
	[0.0045]	[0.0097]			[0.0054]	[0.0125]		
Av. Socially Close-sq (St.)		0.0035				-0.0066		
		[0.0103]				[0.0103]		
Av. Socially Dist. (St.)			-0.0084**	-0.0122	-0.0082**	-0.0150		
			[0.0033]	[0.0117]	[0.0036]	[0.0149]		
Av. Socially Dist.-sq (St.)				0.0039		0.0056		
				[0.0109]		[0.0125]		
Size (St.)							-0.0085**	-0.0139
							[0.0034]	[0.0131]
Size-sq (St.)								0.0053
								[0.0124]
Observations	43,308	43,308	43,308	43,308	43,308	43,308	43,308	43,308
R-squared	0.4106	0.4106	0.4107	0.4107	0.4107	0.4107	0.4107	0.4107
F-stat	1.366	0.659	6.529**	3.301**	3.309**	1.783	6.269**	3.234**

Notes: ***p<0.01, ** p<0.05, p<0.1. Standard errors clustered at the village level in brackets. Dependent variable is changes over time in per-capita household log consumption. $log(y_{int})$ is per capita household log income. All regressions include network-time dummies and controls for changes in household demographics. St. indicates that the variable has been standardised to have mean 0 and std. dev. 1. 'sq' indicates square of variable. Observations at top 1

the regression in Column 5 which includes both socially close and distant connections in the same specification. When both variables are pooled together linearly (Col. 5), more socially distant connections are still associated with improved risk sharing, while the coefficient on average socially close connections becomes even smaller and remains statistically insignificant.[37] This indicates that opportunities for risk sharing are more important than the additional enforcement that socially close connections can provide for the effective functioning of extended family network based risk sharing in this context. Throughout, I find no evidence of a nonlinear relationship between risk sharing and the number of socially close or distant connections in the household's network, thus indicating that, empirically, there is no trade-off between risk sharing opportunities and enforcement concerns in this context.[38]

In terms of magnitude, the coefficient in Column 5 indicates that the changes in log consumption of households in networks with an average number of socially distant connections that is 1 standard deviation (23 households) greater than the sample mean for that variable (just under 20 households) fluctuates 20% less in response to changes in household log income relative to that of households with just under 20 households (sample mean of the average number of socially distant connections).

Finally, Columns 7 and 8 shed light on how risk sharing varies with the total number of households in the network, whether they are socially close or distant connections. At with socially close and distant connections, no non-linearity is apparent from the coefficients reported in Column 8. The coefficient on the interaction term in Column 7 is negative and statistically significant from 0, indicating that larger networks indeed provide more risk sharing. The coefficient is small in magnitude, but meaningful relative to the baseline level of consumption smoothing: for households in the largest network in the sample the sum of coefficients $\beta_1 + \beta_2 * w_i(G_n)$ is 0.0164, indicating that if household income increases by 10%, household consumption will increase by approximately 0.164%. Thus, household consumption is almost perfectly smoothed in this network. The magnitude though still raises important questions on the capacity of social connections in helping households bear risk.[39] It should be noted though that

[37] Table 3.8 displays the results for each of the variables standardised by the sample mean and standard deviation. These are useful for comparing the total contribution of each type of connection to risk sharing. However, there are many more socially distant connections on average than socially close ones (at household- and network-level), and so to accurately assess the marginal contribution of each type of connection to risk sharing, one would want to compare the coefficients on the unstandardised values. These indicate that the coefficient on average socially distant connections is still larger than that associated with average socially close connections.

[38] Estimating the shape of this relationship non-parametrically using locally weighted regression further confirms the linearity of this relationship, as shown in the Appendix.

[39] There are reasons to believe that the effect might be larger in magnitude than that identified here. Classical measurement error in income is likely to attenuate the coefficient estimate towards 0. Endogeneity of the network might also bias upward the estimated coefficient: for example, if

this test might miss welfare losses that are not reflected in consumption which the network might be particularly helpful in alleviating. In particular, households might choose to smooth consumption in response to shocks they experience, at the expense of productive investments such as livestock or longer-term human capital investments such as education. Indeed, Angelucci et al. (2015) find evidence that the extended family facilitates household investments in schooling in response to a conditional cash transfer programme, which provided substantial transfers to a subset of households in this setting, while Shim (2015) shows that households in this setting make sub-optimal schooling choices in the absence of informal risk sharing instruments.

I obtain similar results when I use the household's number of socially close and socially distant connections (shown in Table 3.9): no non-linear effect is detected; and socially close connections have no effect on a household's risk sharing, while more socially distant connections improve risk sharing.

To summarise the findings, networks and households with more socially distant connections achieve better risk sharing, while socially close connections have no effect on risk sharing in this setting. These findings suggest that sufficient opportunities for risk sharing are necessary for social connections to be effective in providing risk sharing; and these are more important than the additional enforcement provided by socially close connections relative to socially distant connections within extended family networks.[40]

poorer households are more likely to have larger families, migrate less and marry within the village, they will have more socially close and socially distant connections within the village. By contrast, richer households might have smaller networks, but may be better able to self insure and thus have consumption streams that are less correlated with income. Households with small networks might thus appear to be receiving more risk sharing from their network than they actually are, thereby biasing the coefficient estimate to be smaller in magnitude than it actually is. Unfortunately no suitably strong instrument for $\Delta log(y_{int})$ or the number of socially close or socially distant connections is available in my data to resolve these problems.

[40] Ideally we would also want to disentangle the effect of socially distant connections on risk sharing to assess how much of it is driven by households having, on average, more socially distant connections ('size effect') and socially distant connections having less positively correlated income streams ('correlation effect'). One way of doing this is to investigate the extent to which effects of socially distant connections on risk sharing are concentrated in networks where incomes of socially close (and all) households are more positively correlated. I implemented this strategy by allowing for a triple interaction term with the median network-level pairwise income correlation for socially close households, and all households in Equation 3.7 along with an interaction term for the median network-level pairwise income correlation. Unfortunately, the pairwise correlations are too noisy to yield any statistically significant results (p-values on interaction terms with the correlations are > 0.6). However, the estimated coefficients have the expected signs: more positive pairwise correlations among socially close households worsens risk sharing, and more so in networks with higher average socially close connections. More socially distant connections in such networks improve risk sharing. The 'size effect' is statistically significant and dominates the 'correlation effect'.

Table 3.9: Risk Sharing and Household-Level Socially Close and Socially Distant Connections

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent Variable: $\Delta\, log(c_{int})$							
$\Delta\, log(y_{int})$	0.0387***	0.0356***	0.0401***	0.0386***	0.0401***	0.0363***	0.0401***	0.0404***
	[0.0037]	[0.0045]	[0.0038]	[0.0052]	[0.0037]	[0.0056]	[0.0037]	[0.0037]
$\Delta\, logy_{int}$ **interacted with:**								
HH Socially Close (St.)	-0.0044	-0.0118*			-0.0016	-0.0075		
	[0.0048]	[0.0067]			[0.0049]	[0.0072]		
HH Socially Close sq. (St.)		0.0064				0.0053		
		[0.0047]				[0.0047]		
HH Socially Distant (St.)			-0.0083**	-0.0105*	-0.0080**	-0.0093		
			[0.0033]	[0.0060]	[0.0034]	[0.0064]		
HH Socially Distant sq. (St.)				0.0018		0.0014		
				[0.0036]		[0.0037]		
Size (St.)							-0.0085**	-0.0139
							[0.0034]	[0.0131]
Size sq. (St.)								0.0053
								[0.0124]
Observations	43,308	43,308	43,308	43,308	43,308	43,308	43,308	43,308
R-squared	0.4106	0.4106	0.4107	0.4107	0.4107	0.4108	0.4107	0.4107
F-stat	0.843	1.564	6.228	3.195	3.133	2.042	6.269	3.234
p-value	0.359	0.210	0.0129	0.0418	0.0444	0.0874	0.0126	0.0402

Notes to Table: *** p<0.01, ** p<0.05, * p<0.1. Standard errors clustered at the village level in brackets. Dependent variable is the changes over time (t) in the per capita log consumption of a household i in a network n. $log(y_{int})$ is the per capita log income of a household i in network n at time t. Income includes labour earnings, asset returns and institutional transfers other than the Progresa cash transfer. All regressions include network-time dummies, and control for changes in household composition over time. St. indicates that the variable has been standardised to have a mean of 0 and standard deviation of 1. Observations in the top 1% of the socially close connections distribution are dropped..

3.5.2 Robustness

Alternative Explanations

Throughout the paper thus far, I take socially close connections to provide better enforcement, and socially distant connections as being valuable since they provide more risk sharing opportunities. However, there could be unobserved variables potentially related to the endogenous formation of the within-village extended family network, which are correlated with both the number of within-village socially close and distant connections as well as the measure of risk sharing, biasing the estimated coefficients reported in Tables 3.8 and 3.9. In particular, though individuals and households cannot choose all their family connections, they might make decisions relating to fertility, migration, marriage and household formation in a manner that affects the number of socially close and socially distant connections within the village. Moreover, there might be unobserved variables that correlate both with these decisions and thereby with the number of connections and the risk sharing measure, yielding an omitted variables bias.

Though I am unable to definitively rule out that the findings are not biased by the endogeneity of the within-village extended family network, analysis in this section rules out one important confounding factor – wealth. Wealthier households might be better able to self-insure (and thus have consumption that is less responsive to income fluctuations), and could also have fewer socially close, but many socially distant connections. This would bias upwards the coefficient related to socially close connections in Table 3.8, and bias downwards that on the number of socially distant connections. If wealth is indeed biasing the results, we should expect wealthy households to have few socially close, but many socially distant connections within the village. I verify whether this is the case in the data, by regressing separately household's number of socially close and socially distant connections on a household asset index, calculated based on ownership of various durables, and a vector of household- and village-level controls. Table 3.10 displays the findings. It indicates no significant correlation between the asset index and numbers of socially close and distant connections; suggesting that the findings are not driven by this channel.

Nonetheless, other unobserved variables could be correlated with the number of socially close connections and the risk sharing measure, invalidating its use as an indicator for better enforcement. To assess the importance of these biases, I use another strategy to study the importance of enforcement constraints in this context. Specifically, we expect the household's within-village extended family network to be a particularly important source of insurance in villages where fewer alternative options, e.g. no/fewer isolated households, or other extended family networks within the village, are available.

Table 3.10: Correlations between network structure and household and network variables

	(1)	(2)	(3)	(4)	(5)	(6)
	HH Socially Close		HH Socially Distant		Network Size	
HH Asset Index	0.0489	0.0214	0.5407	-0.3393	0.5896	-0.3179
	[0.0473]	[0.0447]	[0.8352]	[0.5633]	[0.8675]	[0.5919]
HH size	0.0484***	0.0518***	0.2615**	0.3816***	0.3099**	0.4335***
	[0.0105]	[0.0104]	[0.1148]	[0.1001]	[0.1206]	[0.1060]
Head works in agriculture	0.2618***	0.2370***	1.2521	0.5786	1.5140	0.8156
	[0.0827]	[0.0804]	[1.0573]	[0.8260]	[1.1029]	[0.8721]
Head age	0.0505***	0.0518***	-0.0524	-0.0310	-0.0019	0.0207
	[0.0110]	[0.0109]	[0.0947]	[0.0817]	[0.1005]	[0.0876]
Head age-sq	-0.0006***	-0.0006***	-0.0000	-0.0002	-0.0006	-0.0008
	[0.0001]	[0.0001]	[0.0009]	[0.0008]	[0.0009]	[0.0008]
Head speaks indigenous dialect	0.4530***	0.4112**	0.6685	-1.6852	1.1215	-1.2740
	[0.1639]	[0.1696]	[2.5242]	[1.9899]	[2.6170]	[2.0906]
Village Size		0.0314***		0.9007***		0.9321***
		[0.0056]		[0.0808]		[0.0831]
Village Size sq.		-0.0002***		-0.0035***		-0.0037***
		[0.0000]		[0.0004]		[0.0004]
Village has Diconsa shop		0.0875		2.7644		2.8519
		[0.1790]		[2.6047]		[2.7280]
Village has grocery shop		0.0768		1.6194		1.6962
		[0.1575]		[2.1794]		[2.2882]
Observations	15,096	15,082	15,096	15,082	15,096	15,082
R-squared	0.0192	0.0371	0.0037	0.3055	0.0051	0.2954

Notes to Table: *** p<0.01, ** p<0.05, * p<0.1. Standard errors clustered at the village level in brackets.

Table 3.11: Risk Sharing and Outside Options

	(1)	(2)
	Dependent Variable: $\Delta\, log(c_{int})$	
$\Delta\, log(y_{int})$	0.0387***	0.0399***
	[0.0063]	[0.0056]
$\Delta\, logy_{int}$ **interacted with:**		
Number Other Family Networks	0.0012	
	[0.0009]	
Number Isolated HHs		0.0004
		[0.0003]
Observations	43,308	43,308
R-squared	0.1337	0.1337
F-stat	1.985	1.953
p-value	0.160	0.163

Notes to Table: *** p<0.01, ** p<0.05, * p<0.1. Standard errors clustered at the village level in brackets. The variables 'Number Other Family Networks' and 'Number Isolated HHs' are calculated at the village level.

In these villages, we would expect x_{ij} in the model to be relatively high, and households embedded in these networks would achieve better risk sharing. I construct two measures of households' outside options – the number of isolated households within the village, and the number of other extended family networks – and use these to study how a household's risk sharing varies with the quality of its outside options. I do so by estimating the regressions of the following form:

$$\Delta log(c_{int}) = \mu_{nt} + \beta_1 \Delta log(y_{int}) + Opt_n * \Delta log(y_{int})\beta_2 + \gamma \Delta X_{int} + \epsilon_{int} \tag{3.8}$$

where Opt_n is a proxy for the outside option for a household in network n. Table 3.11 reports the findings for this regression. The coefficients on the interaction terms of both measures of the outside option are positive but small and not statistically significantly different from 0, thereby providing further evidence that enforcement concerns are less important in within-village extended family networks in this context.

Table 3.12: Scenarios considered for the sensitivity analyses

	Age Diff Siblings	Max age mother	Min. age mother
Benchmark	30	45	15
Scenario 1	25	45	15
Scenario 2	20	45	15
Scenario 3	30	40	15
Scenario 4	25	40	15
Scenario 5	20	40	15

Measurement Error in the Network

Another concern that could invalidate the findings is that of measurement error in the network (Chandrasekhar & Lewis 2011).[41] This concern is particularly salient here since I am inferring the network. The descriptive analysis in Section 3.3.2 and the study by Angelucci et al. (2009) have shown that the obtained family connections fall within reasonable ranges, and are correlated in expected ways with other socio-economic variables. Nonetheless, we might be concerned that the algorithm identifies spurious connections, subsequently biasing the estimated coefficients. To assess the importance of such a bias, I consider the sensitivity of the estimated parameters to alternative, more stringent age cut-offs in the algorithm described in Table 3.12. The results from this sensitivity analysis for the results displayed in Col. 3 of Table 3.9 are shown in Table 3.13.

The table indicates that the coefficients exhibit remarkable robustness to different assumptions on the age cutoffs. The biggest change in coefficient values appear when the age cutoff for siblings is reduced to 25 years (scenarios 1 and 4): the coefficient for socially close connections becomes more negative, while that for socially distant connection falls in magnitude but remains statistically significant at the 5% level of significance. Nonetheless, the qualitative conclusion that more socially distant connections yield better risk sharing remains valid under all the different assumptions.

A final concern is that the algorithm might miss identifying some connections, making small networks appear to be smaller than they actually are. To assess whether this affects the estimates, I drop the very small networks (of size < 5) from the sample and re-estimate the specifications. I find that the estimates are qualitatively similar to those

[41] As explained above, classical measurement error in income could also bias the estimates, particularly those reported in Tables 3.8 and3.9. This could be easily corrected if a suitably strong instrument was available, which is unfortunately not the case in my data.

Table 3.13: Sensitivity analysis of parameter estimates to alternative age assumptions in the algorithm

	Dependent Variable: $\Delta\, log(c_{int})$					
	Benchmark	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
$\Delta\, log(y_{int})$	0.0401***	0.0391***	0.0391***	0.0395***	0.0388***	0.0390***
	[0.0038]	[0.0038]	[0.0039]	[0.0038]	[0.0038]	[0.0039]
$\Delta\, logy_{int}$ **interacted with:**						
Soc. Close Connections (St.)	-0.0016	-0.0036	0.0007	-0.0028	-0.0032	-0.0006
	[0.0049]	[0.0045]	[0.0045]	[0.0048]	[0.0046]	[0.0044]
Soc. Distant Connections (St.)	-0.0080**	-0.0068**	-0.0076**	-0.0079**	-0.0066**	-0.0067**
	[0.0034]	[0.0034]	[0.0035]	[0.0034]	[0.0034]	[0.0034]
Observations	43,308	43,092	42,690	43,223	42,285	42,554
R-squared	0.4107	0.4132	0.4198	0.4140	0.4172	0.4237

Notes to Table: *** p<0.01, ** p<0.05, * p<0.1. Standard errors clustered at the village level in brackets. Dependent variable is the changes over time (t) in the per capita log consumption of a household i in a network n. $log(y_{int})$ is the per capita log income of a household i in network n at time t. Income includes labour earnings, asset returns and institutional transfers other than the Progresa cash transfer. All regressions include network-time dummies, and control for changes in household composition over time. St. indicates that the variable has been standardised to have a mean of 0 and standard deviation of 1. Observations in the top 1% of the socially close connections distribution are dropped. The scenarios are as described in Table 3.12.

reported in Table 3.9.

3.6 Conclusion

This chapter studies the role of socially close and distant connections in providing informal risk sharing in the context of village-based extended family networks in rural Mexico. It uses a simple theoretical model with limited enforcement of arrangements and differing opportunities for risk sharing by social distance, to show that the relationship between risk sharing and the number of socially close and distant connections in a household's network is influenced by a potential trade-off in enforcement and risk sharing opportunities. Socially close connections are better able to enforce arrangements, while distant connections may provide more opportunities for risk sharing. Numerical simulations of the theoretical framework indicate that when enforcement concerns dominate, risk sharing (and welfare) increases with the number of socially close connections. Conversely, when opportunities for risk sharing are particularly important, risk sharing and welfare fall (increase) with the number of socially close (distant) connections. When both concerns are relevant, the trade-off between enforcement and risk sharing opportunities generates an inverse-U shaped relationship between the extent of risk sharing (and welfare) and the number of socially close connections in a network.

The chapter then empirically verifies these qualitative predictions using panel data on over 16,000 households embedded in a large number of village-based extended family networks in rural Mexico. The data contains information on cross-household connections through sibling, parent and child relationships of the head and spouse of the head of the household for every pair of households within a village. This allows me to overcome the key empirical challenge of identifying socially close and distant connections of a household by applying a network-theoretic definition of socially close and distant connections. This measure defines as socially close connections siblings, parents and children of the head/spouse; and as socially distant connections, the families of one's siblings' spouses, or aunts, uncles and cousins. In a first step, it documents that socially close connections offer more risk sharing opportunities: they are more likely to be engaged in the same occupation and have more positively correlated income processes; and are fewer in number.

In a second step, it considers how this variation in risk sharing opportunities, along with imperfect enforcement, shape the relationship between risk sharing varies and the average number of socially close and distant connections in a household's network. Measuring risk sharing as the extent of the correlation between changes in household log consumption in response to fluctuations in log income, net of aggregate network

resources, it finds that households with more socially distant connections in their networks achieve better risk sharing. More socially close connections have a small and statistically insignificant effect on risk sharing. The findings highlight the importance of sufficient risk sharing partners with less correlated income streams, which has surprisingly received less attention in recent literature, for the effective functioning of social network based insurance. In addition, they highlight the 'strength of weak ties' in a risk sharing context.

The findings are important for the effective design of policies. Understanding how informal arrangements work, and factors affecting how well they function can shed light on where government intervention would be most beneficial. My findings suggest that sufficient opportunities for risk sharing are crucial for social connections to be able to provide risk sharing. Thus, policies that expand such opportunities, by for example, encouraging income diversification opportunities within a village might indirectly also improve household risk sharing.

The findings from this chapter raise some further questions: first, though the chapter documents variation in risk sharing opportunities, it did not study the drivers of this variation, which are important to understand for effective policy design. Second, the chapter considered only a sub-set, albeit an important one, of the whole extended family network. Contributions from the outside village extended family network will also influence risk sharing arrangements (Rosenzweig & Stark 1989), as well as decisions related to marriage and migration, thereby shaping the structure of the within-village extended family network. Understanding the interactions of these choices is left to future work.

3.7 Appendix

3.7.1 Additional Details on Model

Optimality Conditions

Here I provide more details of the optimality conditions in the theoretical framework as derived by Ambrus et al. (2014). I provide a short summary of the conditions and their implications for risk sharing patterns. The interested reader is directed to the paper by Ambrus et al. (2014) for details on the full derivation.

Define Δ_i to be the marginal benefit to the planner of transferring an additional dollar to i. When a household i is unconstrained, this will be equivalent to the household's marginal utility, $\lambda_i u'(c_i)$. However, when i is constrained in any of his incentive compatibility constraints, this is not the case, since increasing c_i will also relax any binding incentive compatibility constraints for i, making it optimal for the planner to transfer part of the additional dollar to connections of i for whom i's incentive compatibility constraints were previously binding. Thus when a household i is constrained, the marginal social welfare gain is defined in a recursive manner as follows.

For every j such that the incentive compatibility constraint from i to j binds, denote

$$\delta_{ij} = \lambda_i u'(c_i)\frac{u'(c_i + t_{ij})}{u'(c_i)} + \Delta_j \left[1 - \frac{u'(c_i + t_{ij})}{u'(c_i)}\right] \tag{3.9}$$

δ_{ij} measures the marginal social gain of an additional dollar to i under the assumption that i optimally transfers a fraction of the dollar to j. If many incentive compatibility constraints for i bind, the marginal social welfare gain is maximised if part of the dollar is transferred to the household j where it would be most productive, either because j has the highest marginal utility of consumption among all of i's connections, or because one of j's (direct or indirect) connections has a very high marginal utility of consumption (i.e. has a very low consumption). Defining $\delta_{ii} = \lambda_i u'(c_i)$, the marginal social welfare gain of transferring an additional dollar to an agent i can be defined formally as:

$$\Delta_i = max\{\delta_{ij} | \, j : \text{ the IC constraint from } i \text{ to } j \text{ binds}\} \tag{3.10}$$

The following proposition (from Ambrus et al. 2014) specifies the optimal allocation in terms of the planner's marginal social gain.

Proposition 1 (Proposition 13, Ambrus et al. 2014): Assume that the marginal rate of substitution between consumption and connection value, MRS_i is concave in c_i for

every i. A transfer arrangement t is constrained efficient iff there exist positive $(\lambda_i)_{i \in N}$ such that for every i, $j \in N$ one of the following conditions holds:

1. $\Delta_j = \Delta_i$
2. $\Delta_j > \Delta_i$ *and the IC constraint binds for* t_{ij}.
3. $\Delta_j < \Delta_i$ *and the IC constraint binds for* t_{ji}.

Proof: See the online appendix to Ambrus et al. (2014).

This proposition implies that for each state of the world, the network partitions into endogenous 'risk-sharing islands'. Within the islands, no incentive compatibility constraint binds, and condition (1) holds so that households equate their marginal social gains. On the borders of islands, incentive compatibility constraints bind and the marginal social gains are not equated.

3.7.2 Identifying Network Links - Algorithm Details

In this section, I outline the detailed algorithms used to identify parental and sibling relationships across households living in the same village. These relationships are only identified for the head and spouse of each household.

Identifying Sibling Links

I combine information from surname combinations with age restrictions to identify sibling groups within a village. Siblings should share the same paternal surname and maternal surname. In addition, I assume that the age difference between the oldest and youngest identified sibling cannot be more than 30 years. The algorithm proceeds as follows:

1. Form the super set of all 'potential siblings'. This is done by applying the following rule: two individuals are potential siblings if they have the same surname combination. Note that this super set will include all the siblings of an individual i and those of i's siblings.[42]
2. Order, by age, all potential siblings starting from the youngest to the eldest. Do this as shown in the Table below.
3. Calculate the age difference between the oldest sibling and the youngest. If this is ≤ 30 years, then the group of potential siblings are siblings.

[42] In a small proportion of households (<0.5%), the head and spouse both had the same surname combination. In this case, I dropped the spouse from the sample on which the algorithm was run.

Table 3.14: Example of an individual's potential sibling group

Order	Age	Age Gap
1	21	.
2	24	3
3	28	4
4	43	15
5	60	17
6	62	2

Note: In this example, the algorithm would partition this group of potential siblings into two groups: {1,2,3,4} and {5,6}

4. If the age difference is > 30 years, then follow the following steps:

   (a) Compute the age gaps between consecutive siblings, by subtracting the age of the lower birth order sibling from that of the higher birth order sibling.

   (b) The partition the potential siblings into sibling groups in the following manner:

      i. Find the largest age gap and partition the super set of potential siblings into 2 at this point.

      ii. Calculate the age difference between the eldest and youngest siblings in these 2 groups.

      iii. If the age difference is $\leq$ in either of the sub-groups, then that group is a sibling group.

      iv. For sub-groups where the age difference > 30 years, repeat steps (i) and (ii) until (iii) is satisfied for all sub-groups.

Identifying Parent-Child Links

Using surname combinations, similarly, allows us to identify parent-child relationships. Since children take the paternal surname of the father and the paternal surname of the mother, households where the paternal surname of the (male) head and (female) spouse corresponds with the paternal and maternal surnames of an individual in another household are potentially related via parental/filial ties. I use the following set of rules, that also impose restrictions on the age difference between parents and their children to identify links of this type:

1. Find the super set of potential parent-child links based on the paternal surname of an individual i in a household h matching exactly the paternal surname of the head of a household k, and i's maternal surname matching exactly the paternal surname of the spouse of household k.[43]

2. I then impose the following age restrictions:

   (a) The age difference between a mother and her oldest child cannot be < 15 years.

   (b) The age difference between a mother and her youngest child cannot be > 45 years.

I also use rich information in the data to remove some spuriously identified parental links. In particular, I use information from the household roster to purge spurious parental links when the parents of the head or spouse are reported to be resident within the household of any one of the identified siblings.

3.7.3 Data Appendix

Consumption and Income Measures

Detailed consumption data was collected in the October 1998, May 1999, November 1999, November 2000 and 2003 surveys. Information was collected on the quantity consumed and purchased of approximately 36 food items, and expenditure on these in the week preceding the survey, along with expenditure on non-durable items such as clothing, shoes, toiletries, transport costs, utilities, fuel, etc in the month or 6 months preceding the survey. A locality survey further collected prices for foods from local shops. Total food consumption is computed by summing food expenditures and imputed values of non-purchased food. To value non-purchased food, I use median unitvalues at the locality level (computed by dividing expenditure on a certain food by the quantity purchased).[44] Total food consumption and the non-durable expenditure items are all converted to monthly values and added up to obtain a measure of monthly total non-durable consumption.

The surveys also collected information on labour earnings of all employed household members aged > 8 years, rental, pension and interest income, institutional transfers, business revenues and costs, inter-household transfers and in some rounds, remittances.

[43] Clearly, there will be some selection here as households that are not couple-headed cannot be identified by this algorithm as parents of individuals in other households.

[44] For foods that were not very commonly purchased, median unitvalues computed at higher levels of aggregation, such as municipality or state, were used.

To ensure that I have an income measure that is comparable across the different survey rounds, I use only the income components that were collected in the above 5 survey rounds. Thus, income is computed as the sum of labour earnings of all household members, rental, pension and interest income, business profits and institutional transfers (excluding the Progresa grant).[45]

Finally, I convert consumption and income values to October 1998 levels, and calculate per-capita values by dividing by the household size.

3.7.4 Other Empirical Results

Non-parametric Analysis

To shed light on the shape of the relationship between a household's risk sharing and the number of its socially close and distant connections, I estimate this relationship non-parametrically using locally weighted regression. In a first step, I obtain the residuals from regressions of $\Delta log(c_{int})$ and $w_i(G_n) * \Delta log(y_{int})$ on the other right-hand-side variables of Equation 3.7: $\Delta log(y_{int})$, ΔX_{int} and μ_{nt}. Thereafter, I estimate a non-parametric locally weighted regression of the residuals for $\Delta log(c_{int})$ on those for $w_i(G_n) * \Delta log(y_{int})$. Figure 3.7 displays the results of this analysis for (i) average number of socially close connections; (ii) average number of socially distant connections; and (iii) the total number of a household's socially close or socially distant connections.

The plots do not uncover any strong non-linearities in the relationship between risk sharing and the number of socially close and distant connections; suggesting that a linear relationship is a good approximation.

[45] Note that all of these components are converted into monthly terms to give a measure of monthly income.

Figure 3.7: Risk Sharing and Socially Close and Distant Connections, Locally Weighted Regression Coefficients

Notes to Figure: This Figure plots the coefficients from a locally weighted regression of the residuals from a regression of changes in household log consumption on network-time dummies, and changes in household demographics, on residuals from regressions of average number of socially close connections (top left), average number of socially distant connections (top right) and network size (bottom left) on network-time dummies, and changes in household demographics.

Chapter 4

Group Size and the Efficiency of Informal Risk Sharing

4.1 Introduction

Risk is a salient fact of life in rural areas of developing countries. Moreover, these contexts are characterised by market imperfections such as weak enforcement (also known as limited commitment), costly monitoring, poor infrastructure, and weak government capacity; which lead to missing or incomplete insurance and credit markets, and an absence of government social safety nets.[1] Instead, households rely on a variety of informal mechanisms, such as (informal) transfers and loans from relatives and friends, to deal with the consequences of risk (Besley 1995). Such mechanisms are usually based on social ties and groups, such as family or friendship, which are typically more effective in overcoming the aforementioned market imperfections (Rosenzweig (1988*b*), Rosenzweig & Stark (1989), Fafchamps & Lund 2003, Fafchamps & Gubert 2007, Angelucci

[0]This chapter is co-authored with Emla Fitzsimons and Marcos Vera-Hernandez. We thank the Mai Mwana team, especially Tambozi Phiri, Andrew Mganga, Nicholas Mbwana, Christopher Kamphinga, Sonia Lewycka, and Mikey Rosato for their advice, useful discussions, and assistance with data collection. We are grateful also to Julia Behrman, Senthuran Bhuvanendra, Lena Lepuschuetz, Carys Roberts and Simon Robertson for excellent research assistance. We thank Orazio Attanasio, Richard Blundell, Antonio Cabrales, Ethan Ligon, Imran Rasul and participants at the IFS work-in-progress seminar, IFS-UCL Phd conference and EDePo Conference for helpful comments and suggestions. We thank Garance Genicot for kindly sharing code for the model of risk sharing with coalition-proof arrangements. Financial support from the ESRC-NCRM Node 'Programme Evaluation for Policy Analysis' Grant ES/I03685X/1 is gratefully acknowledged. Malde also gratefully acknowledges funding from ESRC Future Research Leaders Grant ES/K00123X/1.

[1]A sizeable literature considers the implications of these imperfections on risk sharing: Kocherlakota (1996), Foster & Rosenzweig (2001), Ligon et al. (2003) and Dubois et al. (2008) consider those for the imperfect enforceability of contracts, while Ligon (1998) and Attanasio & Pavoni (2011) study issues related to moral hazard, and Kinnan (2014) highlights the importance of hidden income.

et al. 2015).[2] A sizeable literature finds that these informal mechanisms are remarkably effective in helping households share risk, though they are unable to perfectly protect household wellbeing. Recent, mainly theoretical work, however, suggests that certain features of these groups are likely to influence how effective they are in providing risk sharing (Bloch et al. 2008, Jackson et al. 2012).

This chapter aims to study how one important characteristic of informal risk sharing groups – their size (or number of households in the group) – affects the amount of risk sharing they achieve. We establish theoretical predictions and then test these predictions empirically in a setting characterised by almost no formal enforcement mechanisms. Theoretically, in an environment where informal arrangements need to be self-sustaining, two forces are at play in influencing the relationship between group size and risk sharing: on the one hand, when households are sufficiently patient and interactions are repeated, larger groups allow for more diversification of shocks, leading to higher gains from sharing risk. On the other hand, as shown in the seminal paper by Genicot & Ray (2003), when arrangements need to be robust to deviations by sub-groups, larger groups can be destabilised by smaller subgroups that are large enough to provide significant levels of risk sharing, meaning that stable groups that can sustain risk sharing are bounded from the top. This suggests that the relationship between group size and risk sharing is unclear. We extend the set-up of Genicot & Ray (2003) and use simulations to show that the relationship between group size and risk sharing is theoretically ambiguous. Thus, the exact nature of the relationship between group size and risk sharing is an empirical question.

Conceptually, it is important to distinguish between the actual and potential risk sharing group. Empirically, the former poses several challenges: first, it is difficult to measure accurately,[3] and second, it will be endogenous since individuals sort into groups on the basis on unobserved characteristics and shocks that are also correlated with risk sharing. To partially overcome this, much prior literature has taken the risk sharing group to be a village (e.g. Townsend 1994;1995). Though readily observable in a large number of socio-economic datasets, this definition is likely to be too broad, especially since villages can have 500 or more households. We instead focus on the sibship of the household head and spouse, a group that is predetermined.[4] To reflect the fact that not

[2]For example, relatives have numerous opportunities to interact with one another, thus reducing the costs of monitoring each others' actions. Moreover, they could use strategies such as shame or even ostracism (both of which are typically not feasible for formal insurance and credit providers to use) to punish renegers in informal arrangements.

[3]For example, self reports are subject to strategic behaviour as shown by Comola & Fafchamps (2015).

[4]A large literature has documented the importance of the extended family for risk sharing in developing countries. See for example, Rosenzweig (1988*b*),Rosenzweig (1988*a*); Stark & Lucas (1988); Rosenzweig & Stark (1989); Foster & Rosenzweig (2001); Fafchamps & Lund (2003); Fafchamps &

all members of this group will actually share risk amongst each other, in what follows, we refer to it as 'potential group size'.

In the context we study – Mchinji, Malawi – the crucial role of the family for risk sharing has been documented in the anthropology and sociology literatures (Phiri 1983; Munthali 2002; Mtika & Doctor 2002; Peters et al. 2008). This is also reflected in the data we use: 80% of transfers received by a household are from family. Thus, the number of siblings of the head and spouse are a relevant proxy for 'potential group size' in this setting. Moreover, historical well-documented social norms in Mchinji give an important role to the wife's brothers (relative to her sisters) in ensuring her household's wellbeing. Though an individual's sibship size is predetermined, it might still be correlated with unobserved factors that are related with risk sharing. The norms allow us to not only to obtain a more fine grained measure of potential group size, but also provide us with an important dimension of heterogeneity that helps us to allay concerns of such omitted variable bias. In particular, we can build placebo tests using the wife's sisters to ascertain that our findings are not explained by omitted variables associated with larger families.

To investigate the empirical relationship between group size and informal risk sharing, we draw on a rich longitudinal dataset which includes information on household consumption, crop loss incidence (and intensity) and the number of living siblings of the head and spouse (who we refer to interchangeably as husband and wife) to conduct the analysis. We consider how well protected a household's consumption is to idiosyncratic crop losses – an important source of risk in our predominantly agricultural setting – given the size of its extended family. Given the social norms previously mentioned, we define groups separately by relationship to the husband or wife (that is, we consider groups such as brothers of husband, brothers of wife, and so on). The correlation between changes in log household consumption and the incidence (and intensity) of household crop loss provides a measure for risk sharing (see Townsend 1994; Mace 1991; and Attanasio & Szekely 2004, among others).We find that households where the wife has many brothers achieve worse risk sharing in response to crop losses relative to households where the wife has few brothers. A similar, though slightly weaker, pattern is also found for households where the husband has many sisters.

A concern is that these findings could be a result of the fact that households where the wife has many brothers (or husbands have many sisters) are poorer, and therefore more vulnerable to shocks. However, the fact that we fail to find a similar relationship among households where the wife has many sisters, or households where the husband has many brothers alleviates this concern. Of course, such a comparison would form a valid

Gubert (2007); Witoelar (2013); Angelucci et al. (2015).

placebo test only if households where the wife (husband) has many sisters (brothers) are similar to those where the wife (husband) has many brothers (sisters). We confirm this is the case, by testing directly for differences in the age, education and ethnicity of the wife (husband) between households where the wife (husband) has many brothers (sisters) and few sisters (brothers). Additional robustness checks indicate that the findings are unlikely to be explained by households with larger numbers of siblings being more vulnerable to crop losses; or by increased competition for production resources (specifically land) among families with many male siblings.

Lastly, we confirm that our empirical findings are compatible with Genicot & Ray (2003). To do so, we calibrate the theoretical model using values (where available) from the data. The calibrated model yields similar patterns between risk sharing and group size as those found in the data, indicating that the threat of coalitional deviations can explain our findings.

The chapter contributes to a number of strands of literature: It relates to a small literature investigating the relationship between risk sharing and group size. A number of studies show that the optimal risk sharing groups are likely to be small in the presence of coalitional deviations (Genicot & Ray 2003, Dubois 2006 and Chaudhuri et al. 2010) and transaction costs (Murgai et al. 2002). However, when households can choose the risks they face, and have heterogenous risk preferences, larger groups may become stable, as shown theoretically by Wang (2015).

It also relates to the literature investigating risk sharing in the presence of coalitional deviations. Recent contributions have extended theoretically Genicot & Ray (2003) to characterise the optimal risk sharing contract when current transfers can depend on past transfers and shocks (Bold 2009); and to allow for savings, and the availability of formal and informal risk sharing institutions (Bold & Dercon 2014). Bold & Dercon (2014) also implement an empirical test of the model using data from funeral insurance groups in Ethiopia. However, they do not consider the relationship between risk sharing and group size.

Finally, the chapter contributes to the literature investigating the role of extended families in risk sharing in developing countries. Recent work has documented that market imperfections influence transactions and informal risk sharing arrangements within the family. For example, Foster & Rosenzweig (2001) document that limited commitment, tempered by altruism, is at play in rural India, while DeWeerdt et al. (2014) show that asymmetry of information among spatially dispersed extended family networks affects interhousehold transfer decisions in rural Tanzania. Baland et al. (2015) document that transfers among siblings in Cameroon follow a system of reciprocal credit, where older siblings support the education of younger siblings, with the expectation that the

younger siblings will reciprocate later. [5]. Our analysis complements this literature by considering how the size of extended family networks affects informal risk sharing.

The rest of the chapter is structured as follows. Section 4.2 lays out the conceptual framework, and shows that the relationship between the amount of risk shared and group size is theoretically ambiguous when coalitions can deviate. Section 4.3 provides details on the data, and the context, focusing particularly on norms governing extended family relationships in rural Malawi. Section 4.4 discusses the empirical specification; while Section 4.5 displays our main results and robustness checks. Section 4.6 outlines findings of the model calibration. Section 4.7 concludes.

4.2 Conceptual Framework

We consider optimal risk sharing in environments subject to imperfect enforceability of contracts. This assumption matches well our empirical setting – rural Malawi – where formal enforcement mechanisms are rarely available. We draw on the set-up in Genicot & Ray (2003), GR hereon, and add to their analysis by considering explicitly (using numerical simulations) the relationship between the extent of risk sharing and group size.

Households are part of a potential risk-sharing group (in our case, the family) of size n. They face a risky endowment, that takes on two values: h or l; $h \geq l$. The probability of drawing an endowment h in any period is π; $0 \leq \pi \leq 1$. Households are ex-ante identical, risk averse and gain utility from consumption. Household utility is increasing, concave and twice-continuously differentiable. There is no storage technology, and neither formal credit nor insurance is available.

To cope with the consequences of risk, households can make and receive transfers following a transfer rule that depends on the number of households in the group that receive the high endowment shock: When a household receives h, and $k-1$ other households also receive h, each household receiving h sends a transfer t_k to a common pool, which is then shared equally among those receiving l. Consumption for households receiving h is thus $h - t_k$, while that for those receiving l is $l + \frac{kt_k}{n-k}$.[6]

Households observe the endowments, consumption and transfers made and received by all other households in the group. However, this setting is subject to the imperfect enforceability of contracts. Thus, the transfer arrangement needs to be self-sustaining. In particular, it needs to be such that no individual or sub-group wants to deviate

[5] This literature also finds that social pressure to make transfers among kin leads to less optimal investment decisions, especially for women (Jakiela & Ozier forthcoming)

[6] Note that the transfer rule makes use of the fact that the group-level aggregate budget constraint for each period must be satisfied.

from the arrangement, i.e. it should be coalition-proof. The specific definition of coalition-proofness is as in Bernheim et al. (1987), which places a further restriction that sub-groups that deviate should themselves be robust to further deviations. Thus, arrangements need to be self-sustaining to deviations that are themselves credible.

Given the transfer rule, and the coalition-proofness condition, and focusing on stationary arrangements, the optimal risk sharing arrangement (i.e. transfer in each state) can be recovered from the solution to the following optimisation problem (expressed in per-period terms):

$$\max_{t_k} \; v(\boldsymbol{t}, n) = p^n u(h) + (1-p)^n u(l) + \sum_{k=1}^{n-1} p(k,n) \left[\frac{k}{n} u(h - t_k) + \frac{n-k}{n} u(l + \frac{kt_k}{n-k}) \right] \tag{4.1}$$

subject to

$$(1-\delta)u(h-t_k) + \delta v(\boldsymbol{t}, n) \geq (1-\delta)u(h) + \delta v^*(s) \quad \forall\, s \leq k \tag{4.2}$$

where δ is the discount factor, and $v^*(s)$ is the per-period expected utility a household could get by deviating to a stable sub-group of size s, and sharing risk in this sub-group in all subsequent periods. The incentive compatibility constraints in Equation (4.2) imply that the transfer arrangement should be such that the per-period discounted utility for households that achieve a good shock in the current period and make a transfer t_k to the common pool, and expect to achieve future expected utility of $v(\boldsymbol{t}, n)$ is greater than the utility it can achieve from deviating in a sub-group s where it consumes its endowment h this period and shares risk with the sub-group s in the future thus attaining an expected future utility of $v^*(s)$.[7]

When no incentive compatibility constraint binds, the first-best allocation, which equalises consumption for all households within the group for each state of the world, is achieved. By contrast, in autarky, when no risk sharing occurs, households consume their own endowment in each period, achieving a per-period expected utility of $pu(h) + (1-p)u(l)$.

Based on this set-up, GR show that a stable risk sharing arrangement may fail to exist for many group sizes, even for high values of the discount factor.[8] Moreover, they

[7] Note that this formulation assumes that in the period that an individual deviates, he consumes his endowment, regardless of the sub-group he deviates with; and shares risk with members of the subgroup in subsequent periods.

[8] In models where the risk sharing arrangement is sustained by ostracising individuals who deviate (i.e. deviating individuals revert to autarky in future periods), a stable arrangement may fail to exist when the discount factor is low. When arrangements need to be coalition-proof, however, a stable arrangement may fail to exist even if the discount factor is sufficiently high.

show that the size of stable risk sharing groups is bounded from above: essentially, large groups are not stable in the presence of coalitional deviations, since households receiving a good shock can deviate to form sub-groups within which they can still benefit from group-based insurance in the future. Thus, in larger groups, the outside option may potentially be better than in smaller groups (depending on the sizes of possible stable sub-coalitions). Thus, the transfer made by those receiving h will be lower than in arrangements sustained by ostracising a deviator to autarky in the future. This is because those receiving h need to be induced to remain in the group rather than deviate to a sub-group, which could provide higher utility than autarky. In some cases, no postive transfer may exist, leading to the non-existence of a stable risk sharing arrangement.[9]

Our contribution, relative to GR, is to show within the same set-up that the relationship between the amount of risk sharing and group size is ambiguous. The fact that a stable arrangement may not exist for many group sizes, complicates this exercise.[10] In particular, it is not possible to study this analytically. We instead use numerical simulations to shed light on the relationship.

We need to take a stand on how risk is shared in groups of size n where no stable risk sharing arrangement exists. One possibility is that households remain in autarky. However, this is not very satisfactory, especially since within this set-up, households can deviate from an autarky punishment by cooperating with subgroups of households. Thus, given that households are ex-ante identical in this setting, a natural assumption is that in cases where no stable arrangement exists for a group of size n, the group randomly partitions into stable subgroups in a manner so as to maximise the sum of expected utility,

$$\sum_{i=1}^{n} \sum_{s \in S} n_s * s * v_i(\boldsymbol{t}, s) \tag{4.3}$$

where S is the set of stable coalitions (or groups), and i indexes households in the group.[11] In other words, we assume that there exists a social planner who chooses a combination of stable sub-groups such that the sum of expected utility (as in Equation

[9] When arrangements can be non-stationary, a larger group could be stable. This is because only a sub-set, rather than all, of potential deviators need to be compensated to remain in the risk sharing arrangement. Nonetheless, GR show that the size of the largest stable group will still be bounded from the top (though it could be larger than the largest stable group under stationary arrangements).

[10] Moreover, as indicated by GR, the existence or not of a stable arrangement for groups of size greater than 2 is sensitive to parameter values.

[11] This need not be the only way by which the group partitions, particularly when households are allowed to be heterogenous. For example, partitions could emerge endogenously as in Ambrus et al. (2014), who allow for different transfers to be made between pairs of households embedded in a network.

(4.3)) is maximised, and then randomly sorts households into these sub-groups.[12,13] We can then calculate the expected utility of a household in the unstable potential group of size n as the weighted average of the expected utilities associated with the combination of stable subgroups (the actual risk sharing group) that maximises the potential group's expected utility, with weights calculated as the probability of being randomly assigned to a particular sub-group.

We evaluate the extent of risk sharing using two measures:[14]

- The household's weighted average expected utility,

$$\sum_{s \in S: s \text{ stable}} \pi_s v_i(\boldsymbol{t}, s) \tag{4.4}$$

  where π_s is the probability of being in the stable group s. This is the social planner's objective function. The value of this function increases as fluctuations in a household's consumption fall: a larger stable group will have a higher value of $v_i(\boldsymbol{t}, s)$ since (i) the probability of states where all households receive the same shock falls with group size, and so there is more scope for risk sharing; and (ii) households have concave utility.

- The weighted average expected difference in marginal utility between the two endowment realisations,

$$\sum_{s \in S: s \text{ stable}} \pi_s \sum_{k=1}^{s-1} p(k,s) \frac{k}{s} [u'(c^l_{k,s}) - u'(c^h_{k,s})] \tag{4.5}$$

  This measure captures the difference in marginal utility that a household expects between states where it receives h and those where it receives l. In a state where k households receive h, higher values of t_k (upto the value equating $c^l_{k,s}$ and $c^h_{k,s}$) will reduce the gap between $c^l_{k,s}$ and $c^h_{k,s}$, and so reduce the difference $u'(c^l_{k,s}) - u'(c^h_{k,s})$. If transfers are large enough such that $c^l_{k,s} = c^h_{k,s}$ for all states, perfect risk sharing is achieved and this measure will be 0. However, deviations from perfect risk sharing in any state of the world, in any of the stable sub-groups that the group

[12] In doing so, we assume that unstable groups are arranging themselves in a manner so as to generate the highest possible insurance for their members.

[13] Since households are ex-ante identical, we assume that the social planner places equal weight on each household when deciding how to allocate households in unstable groups to stable subgroups. However, this assumption can be relaxed easily to allow for arbitrary planner weights. However, note that the transfer rule, and thus expected utility, $v_i(\boldsymbol{t}, s)$, will be the same for all households.

[14] The measure used in the empirical analysis is slightly different and is based on ratios of the marginal utility of consumption.

can partition into, would lead to this measure being positive. Moreover, the greater the deviation from perfect risk sharing (i.e. the higher the gap between $c^l_{k,s}$ and $c^h_{k,s}$), the higher the value of this measure. Thus, lower values of this measure indicate better risk sharing.

We next use this set-up to assess the relationship between the extent of risk sharing, as measured by the expressions (4.4) and (4.5), and potential group size.

Simulations To simulate the model, we make some assumptions on the functional form of the utility function, and on parameter values. In the examples we show here, we use the same parameter values as in GR (Example 2).[15] Utility is assumed to be of the constant relative risk aversion form, i.e.

$$u(c) = \frac{c^{(1-\rho)} - 1}{(1-\rho)}$$

where ρ is the coefficient of relative risk aversion. n is assumed to be 8, which matches the largest group size in our data (see Section 4.3 below). ρ is assumed to be 1.6, $\delta = 0.83$, $h = 3$ and $l = 2$ as in GR. Finally, the probability of receiving the high endowment, $p = 0.4$. With this set of parameter values, only sub-coalitions of size 1, 2 and 3 are stable, as reported in GR and documented in the Table 4.1.

Given this set of stable sub-groups, we compute the two measures outlined in Equations (4.4) and (4.5) to evaluate the extent of risk sharing for each of the different potential group sizes. These are plotted in the left and right panels of Figure 4.1.[16] Weighted expected utility increases with group size for potential groups up to size 3 (which is expected as 3 is the largest stable group), before fluctuating in a zig-zag pattern. The fall with group size is a result of a breakdown in informal risk sharing: in a potential group of size 4, one household would be in autarky, while the other three households could cooperate together and benefit from risk sharing opportunities. The subsequent zig-zag style pattern arises from the combination of stable group sizes that is viable in larger unstable potential groups. A similar picture emerges for the second measure – the weighted average expected difference in marginal utility – (right panel, Figure 4.1), though the pattern is inverted since improvements in risk sharing are associated with decreases in this measure.

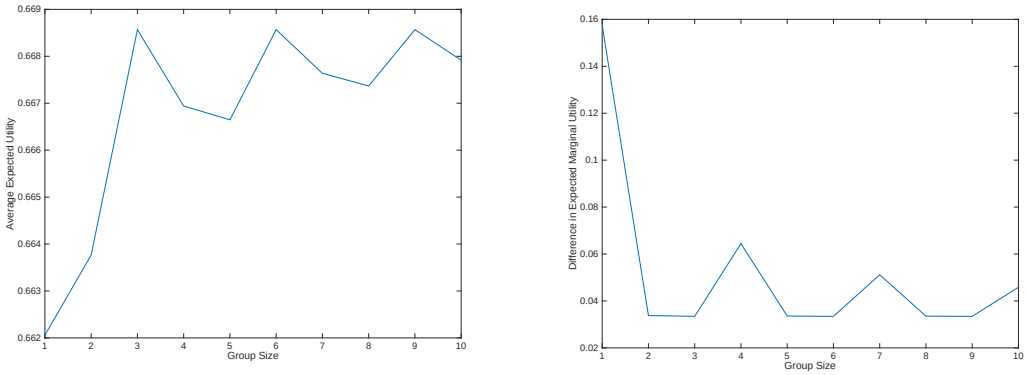
[15] We use these parameter values so as to illustrate what happens to the extent of risk sharing in a documented case where only a small number of potential group sizes is stable. In Section 4.6, we illustrate the patterns of risk sharing and potential group size that emerge when we set the parameter values to match our data.

[16] A detailed overview of the calculations that yield the Figure is in Appendix 4.8.1.

Table 4.1: Stable groups

Group Size	Parameter Set A
1	✓
2	✓
3	✓
4	×
5	×
6	×
7	×
8	×
9	×
10	×

Figure 4.1: Risk Sharing and Group Size - example from Genicot and Ray (2003)



(a) Weighted Avg Exp. Utility

(b) Weighted Avg. Exp. Diff. in Marginal Utility

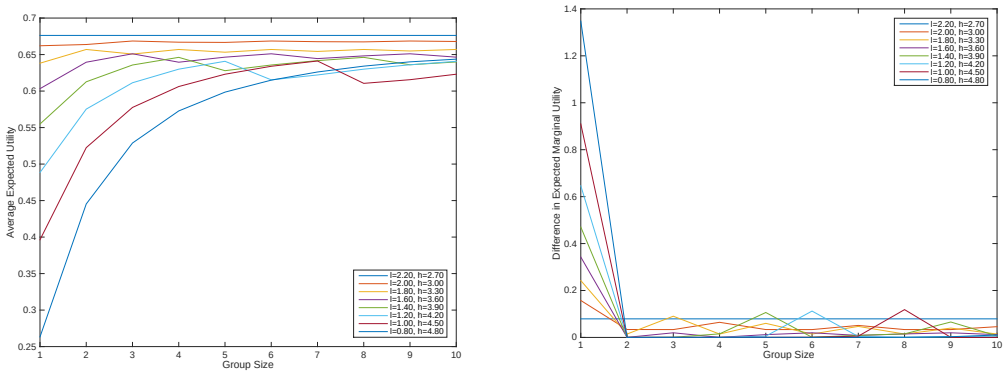
Notes: The Figure in panel (a) shows the relationship between weighted average expected utility and group size, while thatin panel (b)l shows the relationship between the weighted average expected difference in marginal utility and group size

To be noted, though, and documented by GR, is that the stability of groups is sensitive to parameter values. In particular, changing the parameters ρ, p, h and l a little can change which group sizes are stable.[17] This is displayed in the Figure 4.2, which plots the two measures of the degree of risk sharing for different levels of h and l. The values of these variables have been selected so as to have the same average endowment,

[17] From the repeated games literature, it is well known that groups of size 2 can be unstable for low levels of the discount factor, δ. The instability noted here for larger groups arises even when δ is high.

but different variances. A higher variance implies a greater need for insurance. The Figure indicates that as the need for insurance increases, larger potential groups become stable, and these groups achieve better risk sharing than smaller potential groups. This is best displayed by the line corresponding with the highest need for insurance ($l = 0.8$; $h = 4.8$), and is the lowest line in the left panel of Figure 4.2. This line is increasing monotonically, indicating that all group sizes are stable. By contrast, when the need for insurance is low ($l = 2.2$; $h = 2.7$), a case depicted by the top-most line in the left panel of Figure 4.2, no potential group of size > 1 is stable.[18]

Figure 4.2: Risk Sharing and Group Size - Example 2



(a) Weighted Avg Exp. Utility

(b) Weighted Avg. Exp. Diff. in Marginal Utility

Notes: The Figure in panel (a) shows the relationship between weighted average expected utility and group size, while that in panel (b) shows the relationship between the weighted average expected difference in marginal utility and group size

Thus, the simulations indicate even with a small set of parameter values, that there is a theoretically ambiguous relationship between group size and the extent of risk sharing in this model.[19] The nature of this relationship is thus an empirical question, which we now turn to.

[18] Average expected utility is nonetheless higher in this case (even in autarky) since the variance of the endowment is much lower in this case.

[19] We note that other models might also imply that the size of the optimal risk sharing group is smaller than the whole potential group. The presence of coordination costs that are increasing in group size could also yield a similar pattern, as shown by Murgai et al. (2002). However, to our knowledge, no work has characterised the relationship between the extent of risk sharing and group size.

4.3 Context and Data

Our empirical setting is Malawi, one of the poorest countries in Sub-Saharan Africa, with around three quarters of its population living on less than $1.25 a day. Over 80% of its population lives in rural areas, with subsistence agriculture providing the main source of income for a substantial proportion. Infrastructure in rural areas is very weak, with just one in sixteen households having access to electricity, and one in five households having access to piped water.[20] The main crops grown are maize, tobacco and ground nuts. Agriculture is mainly rain-fed, and agricultural production and income are thus highly dependent on unpredictable weather. Access to formal insurance and financial products and services is low, with only 3% of adults holding an insurance product and less than 20% a formal bank account.[21] Instead social connections, particularly family, are important for providing risk sharing, as we show below.

4.3.1 Data Description and Sample Selection

We use data from the Mai Mwana - IFS Economic Survey, a longitudinal survey collected in collaboration with the authors in Mchinji District to evaluate two randomised health interventions – a volunteer infant feeding counselling intervention and a women's group intervention.[22] The survey interviewed approximately 3000 women aged 17-43 and their households living in approximately 600 villages across the district. It collected detailed information on household consumption, adverse events, individual labour supply, health indicators, assets and demographics, and importantly for us, information on extended family networks within and outside the village. Two waves of data were collected, in 2008-09 and 2009-10. The panel dimension allows us to better control for household-level unobserved variables that are correlated with our measure of potential group size, crop losses, and risk sharing.

We restrict the analysis to the following sample: (i) Households living in control areas. (ii) Households where the main respondent was resident in the same village over both surveys. (iii) Households where the main respondent in our survey was either the head or the spouse. (iv) Villages with more than 1 household surveyed. Restriction (i) is imposed since the interventions could have altered risk sharing arrangements within the village, by for instance, altering social interactions or improving community cooperation (particularly in the case of the women's groups).[23] Restriction (ii) is imposed to

[20] Source: Malawi Population and Housing Census (2008).
[21] Source: Finscope Malawi (2009).
[22] See Lewycka et al. (2013) and Fitzsimons et al. (2014) for findings of the impact evaluation. The data is publicly available at http://discover.ukdataservice.ac.uk/catalogue?sn=6996
[23] Fitzsimons et al. (2013) find suggestive evidence of this.

allow us to correctly account for village-level aggregate shocks.[24] We impose restriction (iii) to ensure that we are studying the networks of individuals with relatively similar intrahousehold bargaining power in the sample. Finally, (iv) is imposed because we control for village fixed effects.

Table 4.2 displays some descriptive statistics of our analysis sample. It contains approximately 524 households living in 102 villages. Note that throughout what follows, we recode the male member of a couple (where available) to be the head, while the female member is designated to be the spouse. A note on terminology is in order: throughout the chapter, we will use head and spouse interchangeably with husband and wife. Both the head (husband) and spouse (wife) have low levels of education on average, with approximately 16% (7.4%) of husbands (wives) having some secondary schooling. Further, husbands are older than their wives by on average around 5 years. Households have on average just over 5 members, and most own their own dwelling and land. Despite this, households are in general poor, as indicated by their poor quality housing, and extremely limited access to water and sewerage infrastructure.

[24] Around 18% of the survey main respondents in the data migrated to another village between 2008-09 and 2009-10. The primary reason for migration was marriage. In additional analysis, we checked whether migration was systematically related with the crop loss, and found no evidence of this.

Table 4.2: Sample Descriptives

Variable	**N**	**Mean**	**Std. Dev.**
Husband has no education (yes=1)	477	0.140	0.348
Husband has some primary (yes=1)	477	0.222	0.416
Husband has completed primary (yes=1)	477	0.478	0.500
Husband has at least some secondary (yes=1)	477	0.159	0.366
Husband's years of education	477	5.157	3.514
Wife has no education (yes=1)	524	0.256	0.437
Wife has some primary (yes=1)	524	0.273	0.446
Wife has completed primary (yes=1)	524	0.397	0.490
Wife has at least some secondary (yes=1)	524	0.074	0.263
Wife's years of education	524	3.435	3.229
Age of Husband	478	37.464	10.110
Age of Wife	524	32.648	8.843
Household size	524	5.708	2.123
# of kids $<$ 6 years	524	1.403	0.958
# of kids aged 6-12 years	524	1.187	1.031
# individuals aged $>$ 12 years	524	3.115	1.347
Household owns dwelling (yes=1)	524	0.937	0.243
Household owns land (yes=1)	524	0.840	0.367
Household has good floor (yes=1)	524	0.099	0.299
Household has good roof (yes=1)	524	0.210	0.408
# of sleeping rooms	524	2.076	1.017
Household has access to piped water (yes=1)	524	0.078	0.269
Household has improved latrine (yes=1)	524	0.073	0.260

Notes to Table: The table includes households resident in the same village over both rounds of the IFS-Mai Mwana survey, and where the main respondent was married, and either the head or spouse of her household. Data for some husbands is missing if they are not living in the household at the time of the survey, but are still married to the wife.

4.3.2 Defining the Risk Sharing Group

Having described the data, we now discuss how we define the potential risk sharing group. As noted above, formal financial markets are almost absent in Mchinji, and there was no government safety net in place at the time of the surveys.[25] Instead, existing research in anthropology and sociology indicates that social connections, and in particular, extended family connections play a critical role in helping households deal with the consequences of risk and adverse events: for example, Trinitapoli et al. (2014) documents the role of older siblings in protecting educational investments of younger siblings, while Peters et al. (2008) and Munthali (2002) document the essential role played by the family in fostering and taking care of children orphaned by HIV/AIDS. We also find support for this in our data. In particular, looking at responses to a question on who households expect to receive informal monetary transfers, loans or gifts from, in the event of an income loss due to adverse idiosyncratic events (displayed in Table 4.3), we see that at the median, households expect to receive support from 2 family members and 1 friend. The average indicates the opposite pattern, though this is driven by a small number of households who can turn to a large number of friends.[26]

Table 4.3: Number of potential sources of support following adverse idiosyncratic event

Source of Support	Mean	Median	Std. Dev
Family	1.69	2	1.68
Friends	1.94	1	2.31
N	1048		

Notes to Table: This table shows the number of different individuals with a specific social relationship that a household expects to receive help from if it experiences an income loss as a result of an idiosyncratic adverse event.

Our data also allow us to look at the actual amounts of transfers, loans or gifts (monetary or in-kind) given to and received from family and friends (displayed in Table 4.4) in the year prior to the survey. The data indicates, on average, households give around 375 MK to family, and receive on average 321 MK. Their transactions with friends are of a much lower magnitude (two and a half times, in fact), with 113 MK given on average and 87 MK received from friends. These pieces of evidence thus confirm

[25] A cash transfer program, the Mchinji Cash Transfer, was being piloted in a small number of villages in Mchinji at the time of the survey. Less than 3% of households in our sample report receiving the transfer.

[26] 1% of households report being able to turn to 10 or more friends in case of an adverse event.

that the extended family is a critical source of risk sharing in this setting. Given the importance of family for risk sharing in this setting, we define 'potential group size' based on family.

Table 4.4: Transfers Given to and Received From Family and Friends

Source of Support	Support Given		Support Received		Support Given + Received	
	Mean	Std. Dev	Mean	Std. Dev	Mean	Std. Dev
Family	375.11	1485.83	321.22	1567.91	696.78	2378.13
Friends	113.59	677.72	87.65	599.74	201.24	919.48
N	1048		1048		1048	

Notes to Table: This table shows the amounts given to (left panel), received from (middle panel), and given to and received from (right panel) individuals with a specific social relationship by the household in the year prior to the survey for wave 1 and between surveys for wave 2. All amounts are in Malawi Kwacha. The exchange rate at the time of the survey was around US$1 = 140 MK.

Further anthropological evidence allows us to define the potential group more finely, and also suggests a placebo test to rule out any potential lingering endogeneity concerns related to this definition. Within the family, anthropological evidence suggests that a wife's brothers should play an important role in ensuring the well-being of her family. The predominant ethnic group in our sample, the Chewa, are a matrilineal and matrilocal ethnic group (Richards 1950, Phiri 1983, Mtika & Doctor 2002). Traditionally, under matriliny, society gives a special role to an individual's maternal family, resulting in a close bond between siblings, even after marriage. Moreover, a woman's brothers play a crucial role in supporting her family: The eldest brother is responsible for ensuring access for a woman's family to production resources, healthcare, and other things important for household welfare. As a result, children will consult with their maternal uncles as they are responsible for arranging marriages, ensuring the children have access to adequate land and other productive resources, as well as health care (Phiri 1983, Mtika & Doctor 2002).

The literature indicates that some practices may be less relevant today, while other aspects of matriliny have proved to be remarkably resilient over time. For instance, the practice of matrilocality – whereby the husband moves to the wife's home immediately after marriage – has waned somewhat in Mchinji, with about a half of couples in our sample living in the husband's village when interviewed, and the other half live in the wife's village of birth. At the same time, though, children are still considered to 'belong' to their mother's matriline, and the maternal relatives become their key

caretakers following her death Munthali (2002).

In terms of risk sharing arrangements, data on interhousehold transfers from the Family Transfers Project (collected within the Malawi Longitudinal Study of Families and Health) indicates that a wife's brothers remain an important source and recipient of transfers from a household: 33% (41%) of couples report having received (given) a material transfer from (to) the wife's brothers in the past growing season (which corresponds to a period of around 3-5 months). Moreover, they are less likely to receive material transfers from a wife's sisters (26% report receiving a material transfer), and received transfers are of lower magnitude (351 MK on average is received from brothers, compared to 119 MK from sisters).[27] The evidence thus suggests that the brothers of the wife are likely to still play an important role in risk sharing for the household. We thus define the potential risk sharing group to be the number of brothers (and separately, sisters) of the husband and wife.

4.3.3 Crop Losses

Measuring Crop Losses

Unexpected crop losses are used as our measure of shocks in the analysis.[28] Such crop losses could occur as a result of pests, variation in weather (whose effects could vary within a village by the type of soil, and other characteristics of the land), and other such factors. The first (second) survey collected information on whether the household experienced any crop loss in the year preceding the survey (or since the first survey); and if so, how much potential revenue was lost.[29] We use this information to construct two measures of crop loss: the first is a dummy variable defined to be 1 if the household experienced a crop loss event, thereby measuring the incidence of a crop loss; while the second is potential revenue lost normalised by a measure of 'permanent' consumption, thereby capturing the intensity of the crop loss.[30]

[27] These figures come from 220 observations, and are not adjusted for the number of siblings, or other variables.

[28] Crop losses have been used as a measure of adverse events by studies including Beegle et al. (2006).

[29] The exact questions were as follows: "*In the last year (since the last survey) did this household suffer from a bad harvest or crop loss?*" and *"How much potential revenue was lost as a result of the loss?"*

[30] We normalise the potential revenue lost by the household's permanent consumption to account for the fact that households that experience larger losses may be wealthier and better able to build up buffer stocks to deal with the consequences of risk. In this case, we would erroneously conclude that households are well insured. Household permanent consumption is measured as the part of household consumption predicted by the education of the female main respondent as measured in 2004. We also experimented with using household asset holdings in 2004 and quality of house in 2004, in addition to the education of the female main respondent, to predict household consumption. A concern with using past household assets, however, is that they may be correlated with a household's ability to currently smooth consumption, particularly if crop loss events are persistent. Results using this measure are

Crop losses are prevalent in this setting, as can be seen from Table 4.5: Around 24% of households in our sample experienced a crop loss over the 2-year period, losing on average, just over 3,700 MK. This amount corresponds to around one third of average monthly household food consumption. Among those who experienced a loss, the average loss is around 13,000 MK, which corresponds to 125% of average monthly household consumption. More crop losses were observed in the year prior to the 2008-09 survey relative to 2009-10, with the losses experienced in the former year being more severe in intensity.

Table 4.5: Crop Losses, By Year

	N	Mean	Std Dev
Overall Sample			
Crop loss incidence	1048	0.242	0.429
Income lost ('000s MK)	1044	3.756	19.337
2008-09			
Crop loss incidence	524	0.303	0.460
Income lost ('000s MK)	524	5.536	26.310
2009-10			
Crop loss incidence	524	0.181	0.386
Income lost ('000s MK)	520	1.962	6.891

Notes to Table: Sample includes households resident in the same village across the two surveys, and where the main respondent was married at the time of the survey and either the head or spouse of the head.

Finally, there is some persistence in crop losses among those who experienced a loss. From Table 4.6, we see that around 8% of households experience a crop loss in both survey rounds, which is higher than what we would expect if crop losses were independently distributed.[31]

available on request.

[31] Under the assumption that the crop loss distributions for the two years are independent, the probability of experiencing a crop loss in both survey rounds is the product of the probability of experiencing a crop loss in 2008-09 and the probability of experiencing a crop loss in 2009-10, which equates to around 5.4% of households.

Table 4.6: Persistence of Crop Losses

		Crop Loss in 2009-10		
		No	Yes	Total
Crop Loss in 2008-09	No	312	53	365
		[59.54]	[10.11]	[69.66]
	Yes	117	42	159
		[22.33]	[8.02]	[30.34]
		429	95	524
		[81.87]	[18.13]	[100]

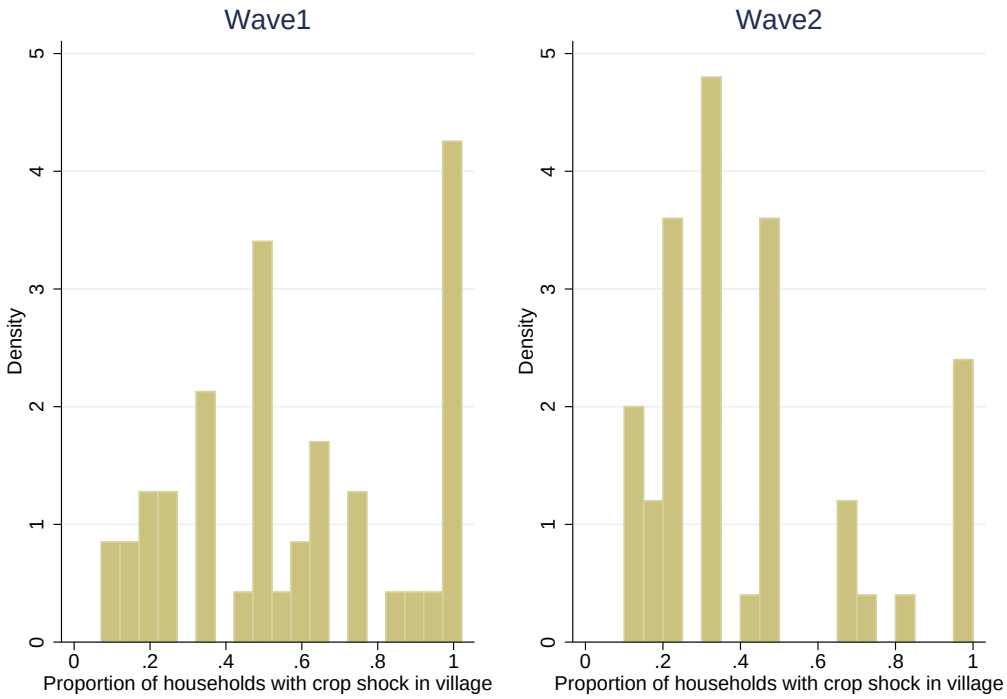
Notes to Table: Sample includes households resident in the same village across the two surveys, and where the main respondent was married at the time of the survey and is either the head or spouse of the head. Percentages in each category displayed in the parentheses.

Are crop losses idiosyncratic within the village?

Our objective is to investigate how the amount of idiosyncratic risk shared by a household varies with the size of its extended family. For our tests to have sufficient power, we require that there is sufficient variation within villages in the incidence of crop losses.[32] Such variation may arise as a result of differences in land quality, with some plots more resilient to poor weather relative to others; or due to variation in the crops grown (some crops and crop varieties may be more resilient to poor weather); or due to localised pests or crop diseases. Note that there was no drought or widespread flooding in Mchinji over the survey period. Nonetheless, we check here for the amount of idiosyncratic variation in our data. To do this, Figure 4.3 displays histograms of the within-village variation in the incidence of a crop loss, for each round of data. We see from the Figure that there are a number of villages with idiosyncratic variation in the incidence of crop losses.

[32] As we will show below, ideally we would like to be able to control for within-group shocks. However, we are unable to do this since we do not observe information on all members of the group. Controlling for aggregate village shocks allows us to partially account for common shocks experienced by group members in the village.

Figure 4.3: Variation in crop loss incidence within villages



Notes to Figure: The Figure plots a histogram for the proportion of households in each village that experienced a crop loss in wave 1 of the survey (left panel) and in wave 2 (right panel). For legibility of the graph, a peak at 0 with magnitude 10 has been omitted.

4.3.4 Measuring extended family networks

To investigate the relationship between the extent of risk sharing and the size of the extended family, we collected information in the survey on the numbers of siblings of the main respondent and her spouse. Data were collected on the numbers of siblings in the village and the number living[33], and on the location of residence of the respondent's mother and mother-in-law. We use the numbers of siblings as our measure of potential group size. The two surveys – conducted around a year apart – captured similar numbers of siblings for a large part of our sample. However, there were some discrepancies in

[33]The exact wording of the questions was as follows: Please tell me how many of the following categories of relatives are currently alive, regardless of where they live:
1. Sisters 2. Sisters-in-law 3. Brothers 4. Brothers-in-law
Please tell me how many of the following categories of relatives are currently living in this village:
1. Sisters 2. Sisters-in-law 3. Brothers 4. Brothers-in-law
Note that in our survey, sisters-in-law and brothers-in-law were translated in a manner so as to capture the siblings of one's spouse.

a sizeable minority (~30%) of observations, which could not be explained by naturally expected changes (e.g. deaths or divorce), and thus point towards reporting errors. To mitigate effects of such errors, we take the average of the reported information in both surveys as the preferred measure of potential group size. Moreover, we use information from the household roster, along with this data to construct variables for the number of siblings of the husband and wife living outside the household.

Tables 4.7 and 4.8 provide some descriptive statistics of sibling networks in this context. Virtually all households have a sibling link outside the household, and a lower, though sizeable proportion (~82%), has siblings within the same village. Households have on average 9.4 siblings outside the household, of whom close to 3 are within the same village. The high numbers of siblings (relative to Western contexts) reflects the high fertility rates in Malawi: the Total Fertility Rate[34] in rural areas was estimated to be around 7.6 in 1984, falling slightly to 6.7 by 2000. At the individual level, almost all husbands and wives have a living sibling, though roughly one-third of husbands and nearly half of wives do not have a sibling in the same village. On average, wives have more living siblings (~5) than husbands (~4.4), but both have similar numbers of siblings in the same village.

[34]This captures the average number of children that would be born to a woman over her lifetime if she were to experience the exact current age-specific fertility rate through her lifetime, and if she were to survive from birth to the end of her reproductive life.

Table 4.7: Any Family Links

	Any Sibling Link	Any Sibling Link of Husband	Any Sibling Link of Wife	Any Links			
				Husband		Wife	
				Brothers	Sisters	Brothers	Sisters
Alive	0.996	0.971	0.985	0.908	0.908	0.941	0.933
	[0.003]	[0.008]	[0.005]	[0.013]	[0.013]	[0.011]	[0.012]
In Same Village	0.819	0.666	0.531	0.534	0.517	0.418	0.437
	[0.021]	[0.024]	[0.028]	[0.025]	[0.022]	[0.021]	[0.030]

Notes to Table: The table includes households resident in the same village over both survey rounds, and where the main respondent is married, and is either the head or spouse of her household.

Table 4.8: Numbers of Family Links

	# of Sibs of Husband + Wife	# of Sibs of Husband	# of Sibs of Wife	Number of			
				Husband		Wife	
				Brothers	Sisters	Brothers	Sisters
Alive	9.418	4.422	5.162	2.281	2.267	2.519	2.740
	[0.172]	[0.098]	[0.113]	[0.064]	[0.068]	[0.069]	[0.079]
In Same Village	2.945	1.571	1.498	0.893	0.788	0.811	0.748
	[0.127]	[0.081]	[0.086]	[0.057]	[0.044]	[0.050]	[0.052]

Notes to Table: The table includes households resident in the same village over both survey rounds, and where the main respondent was married, and either the head or spouse of her household.

These patterns are in line with post-marital living patterns in this context. As mentioned already, though the Chewa were traditionally matrilocal, this seems to be waning in Mchinji, with roughly half of the wives in our sample moving to their husbands' village after marriage. Thus, roughly half the wives in our sample have a sibling in the same village, while two-thirds of husbands have a sibling in the same village. In terms of the type of sibling link, husbands and wives have similar numbers of brothers and sisters alive, though they have slightly more brothers than sisters living in the same village.

4.4 Empirical Model

Our objective is to understand how the amount of risk shared in the face of crop losses varies with the size of a household's family network. To do so, we require a measure of risk sharing, which can be computed in the available data. One measure implied by the model (assuming utility of the constant relative risk aversion form) is the deviation of changes in log consumption from the first-best allocation. Under the first-best allocation, where every group is stable, each household will consume an equal share of pooled resources. This means that changes in household-level log consumption should move one-to-one with aggregate group resources, and be uncorrelated with household-level idiosyncratic shocks. This is a well known result in the risk sharing literature (see, for example, Townsend 1994), which we use to construct our test for how risk sharing varies with the size of a household's family network.

Using consumption to construct our measure of risk sharing has the advantage of providing a useful summary measure of all the different risk sharing strategies employed by a household. Collecting reliable information on all the different methods used for risk sharing, and of the exact bilateral transactions between households in a group is very time-consuming and costly; and more vulnerable to measurement error: For example, Mtika & Doctor (2002) report that one reason why households in Malawi report few transfers to their parents is that respondents help out their parents all the time and do not remember all of the details of specific transactions; while Comola & Fafchamps (2015) show that there is a strategic behaviour in reporting bilateral inter-household transfers in rural Tanzania.

We next describe our estimation equation. The theoretical model did not suggest any clear prediction on the shape of this relationship. We thus begin by estimating a non-parametric relationship between group size and the extent of risk sharing. We do so using the following equation, which includes interaction terms with dummy variables for each potential group size value in the data:

$$\Delta log(c_{ivt}) = \alpha_0 + \alpha_1 \Delta(crop_{ivt}) + \sum_{n=1}^{N} \beta_n \Delta crop_{ivt} * 1(S_{iv} = n) + \Delta X_{ivt}\gamma + \sum_{n=1}^{N} \lambda_n \Delta crop_{ivt} * 1(F_{iv} = n) + \mu_{vt} + \Delta\epsilon_{int} \quad (4.6)$$

where $\Delta log(c_{ivt})$ is the change over time in log consumption for household i in village v at time t, $\Delta(crop_{ivt})$ indicates the change in crop loss incidence or intensity for household i between t and $t-1$, where the crop loss incidence and intensity are measured as explained in Section 4.3.3. The term $1(S_{ivt} = n)$ takes the value of 1 if the household has n brothers or sisters of the head or spouse and 0 otherwise. ΔX_{ivt} captures changes in household characteristics, such as household demographics, that could also affect changes in log consumption. The term $\sum_{n=1}^{N} \lambda_n \Delta crop_{ivt} * 1(F_{iv} = n)$ controls for direct effects of total sibship size of the husband or wife. μ_{vt} denote village-time dummies which capture village-level aggregate shocks. The coefficients of interest are β_n, while the sum of the coefficients $\alpha_1 + \sum_{n=1}^{N} \beta_n * 1(S_{iv} = n)$ indicates how well protected a household's consumption is against idiosyncratic crop losses. In line with the prevailing social norms in this context which indicate that a woman's brothers have an important role in helping out their sisters' households, we conduct the empirical analysis separately for the brothers and sisters of the head of a household and his spouse.

Ideally, we would like to control for group-level aggregate shocks, rather than just village-level aggregate shocks. However, we are unable to do so since we do not observe the crop losses or consumption of all members of the potential group. As a result, the group-level aggregate shock is an omitted variable, which will bias the estimates of interest if it is correlated with potential group size or crop loss incidence. To assess the consequences of this, we run some simulations where we generate data from a data generating process similar to that implied by the model in Section 4.2 (parameterised using values similar to those in the data), and use these to shed light on the direction and magnitude of the resulting omitted variable bias. The findings of this exercise are given in Subsection 4.5.2.

We include changes in crop loss, rather than crop loss in levels, as a measure of idiosyncratic shock for the following reason: assume we used the crop loss incidence between periods t and $t+1$ as the shock measure. The concern with this is that, in the absence of perfect risk sharing, a household may already have low consumption at

period t if it experienced a crop loss between periods $t-1$ and t. Moreover, assume it experiences another crop loss between t and $t+1$, and its consumption remains low at time $t+1$, resulting in little or no change in $\Delta log(c_{hvt})$. The household would then erroneously appear to be perfectly insured: so if crop losses are persistent (and there is some evidence of this for some households as seen in Section 4.3.3), we would erroneously conclude that households are perfectly insured since their consumption does not respond to a crop loss. For this reason, we define the shock measure as the difference in incidence (or intensity) of a crop loss between time periods $t-1$ and t and between t and $t+1$.[35]

This specification can shed light on the shape of the relationship between our measure of risk sharing and the size of a household's potential group. However, this approach, which is fully non-parametric in the number of siblings, might not have sufficient power to identify statistically significant effects. To improve power, we divide potential group size into three bands, the cutoffs of which are motivated by the findings from the nonparametric regression above, and use the following specification for the empirical analysis:

$$\Delta log(c_{hvt}) = \alpha_0 + \alpha_1 \Delta(crop_{hvt}) + \sum_{g=1}^{G} \beta_n \Delta crop_{hvt} * 1(NS_{g,hv} = 1) + \Delta X_{hvt}\gamma + \sum_{n=1}^{N} \lambda_n \Delta crop_{hvt} * 1(F_{hv} = n) + \mu_{vt} + \Delta\epsilon_{hnt} \quad (4.7)$$

where $1(NS_{g,hv} = 1)$ is a term that takes value 1 if the household's network size is within the cutoffs associated with band g, and 0 otherwise; and the rest of the variables are as defined above.[36]

4.5 Results

4.5.1 Main Specification

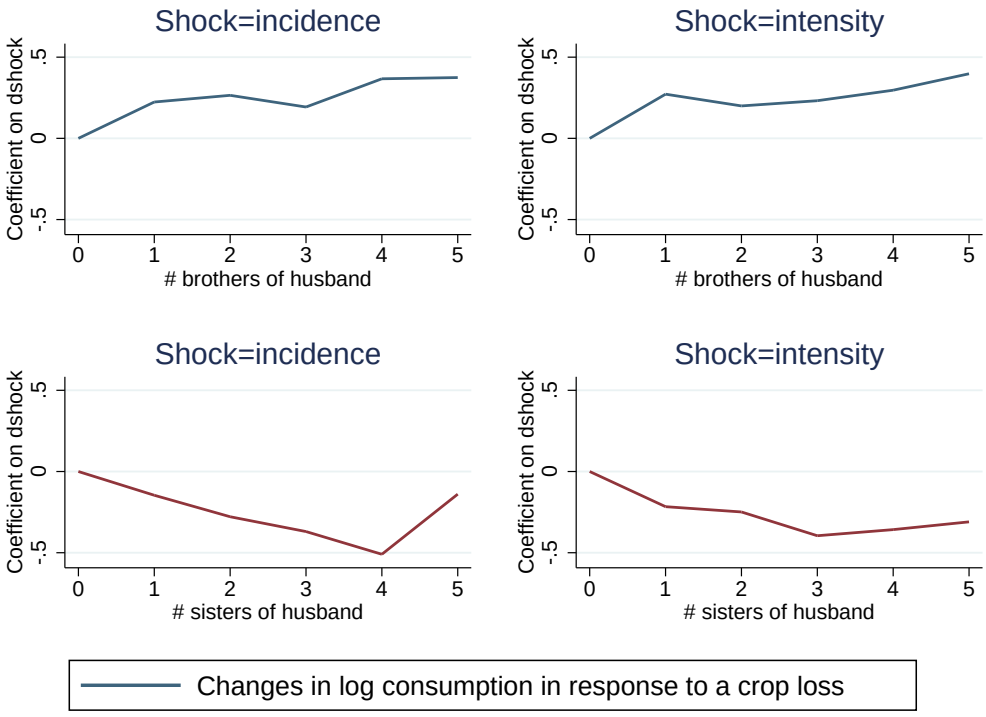
We first estimate Equation 4.6, separately for the brothers and sisters of the husband and wife. Figures 4.4 and 4.5 plot the coefficients from these regressions. We have extremely limited power in these specifications, and thus suppress the confidence intervals

[35] A further issue with focusing on incidence of rather than changes in crop losses is that we do not account for other risk faced by the household, which may affect both their consumption smoothing and the shocks they experience. To assess the importance of this issue in our context, we estimated specifications controlling for other idiosyncratic shocks experienced by the household (business shocks, theft, and marriage break-up) and found it made little difference to the key coefficients of interest.

[36] The exact cutoffs for the different bands are defined in Section 4.5.

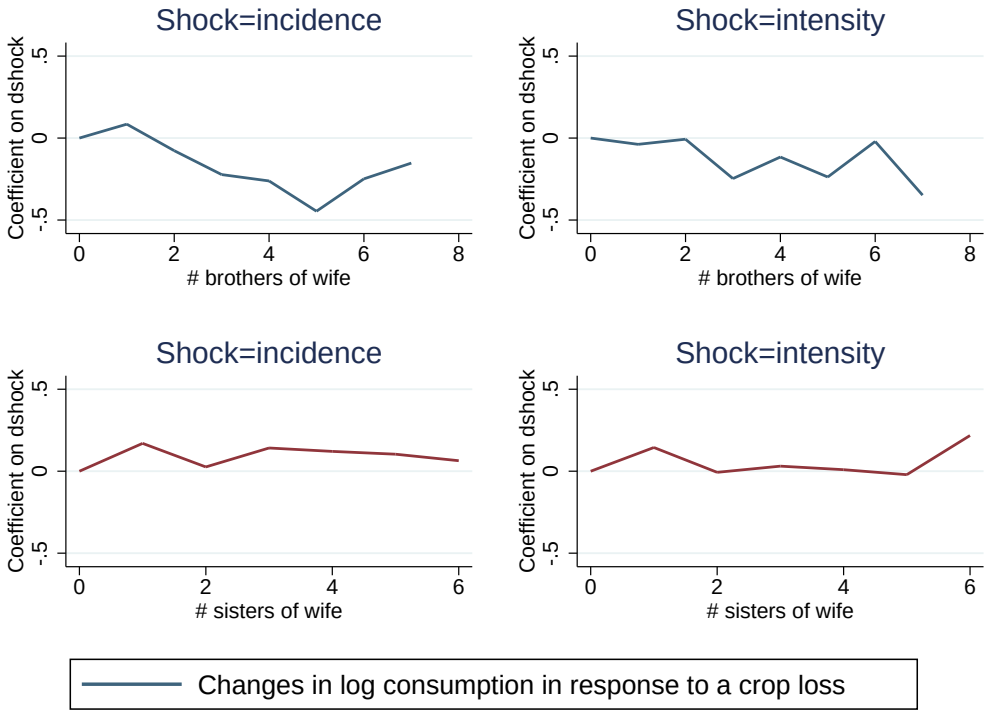
for these coefficients from the Figures. Despite the limitations in power, these Figures shed light on the possible shape of the relationship between informal risk sharing and potential group size in our data.

Figure 4.4: Risk Sharing by Number of Brothers and Sisters of Husband



Notes to Figure: The figures plot the correlation between changes in log consumption and household crop loss incidence (left panel) and intensity (right panel) for households with different numbers of brothers (top panel) and sisters (bottom panel) of the husband. The coefficient for zero brothers or sisters is normalised to 0, and lower values of the coefficient indicate worse risk sharing.

Figure 4.5: Risk Sharing by the Number of Brothers and Sisters of Wife



Notes to Figure: The figures plot the correlation between changes in log consumption and household crop loss incidence (left panel) and intensity (right panel) for households with different numbers of brothers (top panel) and sisters (bottom panel) of the wife. The coefficient for 0 brothers or sisters is normalised to 0, and lower values of the coefficient indicate worse risk sharing.

From the Figures, we can see that there are differences in the amount of consumption smoothing in the face of crop losses, by the size and types of family relations. In particular, Figure 4.4 indicates positive changes in log consumption (implying better protection of consumption) with larger numbers of brothers for the husband, and worse consumption smoothing with larger numbers of sisters of the husband. For the siblings of the wife, the Figures indicate that the consumption of households where the wife has a small number of brothers is almost perfectly smoothed, but worsens as the number of brothers increases. By contrast, no such relationship is seen for the number of sisters of the wife.

The analysis above suggests that there are nonlinearities in the relationship between the amount of risk shared in response to crop losses and the number of brothers and sisters of a household's head and spouse. Moreover, in line with the social norms suggested by the literature, the effects vary by type of sibling. In particular, for brothers

of the wife, and sisters of the husband, there is an initial improvement in consumption smoothing with network size, before worsening. However, we do not have sufficient power to obtain statistically significant estimates. To gain power, we thus pool together the number of siblings of a particular type into 3 groups: those with 0 siblings of a particular type, those with 1-2 siblings of that type, and those with 3 or more siblings of that type. These cutoff values are in line with the evidence presented in Figures 4.4 and 4.5 above, while also ensuring that each group has sufficient sample size to improve power. Table 4.9 presents the results for this specification, with our two measures for the crop loss shock: incidence and intensity. The top left panel displays the results pertaining to the brothers of the husband, while the top right panel displays these for the brothers of the wife. The bottom panel displays the results respectively for the sisters of the husband (left panel) and wife (right panel).

Table 4.9: Main results

	[1] $\Delta\, logc_{int}$	[2] $\Delta\, logc_{int}$	[3] $\Delta\, logc_{int}$	[4] $\Delta\, logc_{int}$
	Siblings of husband alive		**Siblings of wife alive**	
	shock = crop	**shock = Loss/Pred. Cons**	**shock = crop**	**shock = Loss/Pred. Cons**
Δshock	0.2114**	0.1915*	0.0126	0.0216
	[0.1003]	[0.0972]	[0.1294]	[0.2552]
No brothers*Δshock	-0.2306	-0.184	0.0264	0.0296
	[0.1690]	[0.1323]	[0.1482]	[0.0792]
$\geq$ 3 brothers*Δshock	0.0015	0.0614	-0.2578**	-0.1786***
	[0.1029]	[0.0492]	[0.1129]	[0.0615]
N	524	519	524	519
R-squared	0.3213	0.3348	0.3216	0.3366
Δshock	-0.1609	-0.2151	0.1095	0.1061*
	[0.1809]	[0.1403]	[0.1453]	[0.0618]
No sisters*Δshock	0.1594	0.2218**	-0.0783	0.0571
	[0.1231]	[0.0997]	[0.1246]	[0.3093]
$\geq$ 3 sisters*Δshock	-0.1522	-0.1274**	0.042	-0.0411
	[0.1039]	[0.0515]	[0.1068]	[0.0546]
N	524	519	524	519
R-squared	0.3126	0.3288	0.3253	0.3371

Notes: *** Significant at the 1% level; ** the 5% level; * the 10% level. Standard errors clustered at the village level in parentheses. Regressions pool together all households where a married head or spouse was surveyed, and who were resident in the same village for both survey rounds. All specifications control for village-time dummies and changes in household demographics. "Crop" indicates whether or not a household suffered a crop loss, while "Loss/Pred. Cons" measures the intensity of the crop loss as the income lost normalised by predicted household consumption.

The regression coefficients indicate that households where the wife has more than 3 brothers experience much worse risk sharing following crop losses than those where she has fewer than 3 (i.e. 0 or 1-2) brothers. We detect no such relationship for the

brothers of the husband. This finding is replicated across both our measures of crop losses - incidence and intensity. The coefficient estimates indicate that households where the wife has more than 3 brothers cut their consumption by approximately 26% when hit by a crop loss, while the intensity measure indicates that a crop loss of a magnitude equivalent to a month's consumption leads to a reduction in household consumption of approximately 18%.

The bottom panel of the table indicates worse risk sharing (significant only for the intensity measure) among households where the husband has any sisters, as can be evidenced by the positive coefficient on the interaction term for no sisters, and the negative coefficient on the interaction term for more than 3 sisters. No similar pattern is found for the sisters of the wife or the brothers of the husband. The absence of any significant differences in risk sharing by the number of sisters of the wife is consistent with the evidence showed in Subsection 4.3.2 which indicated that sisters of the wife are less important for risk sharing.

4.5.2 Robustness

The analysis in the previous subsection indicates that households where the wife (husband) has many brothers (sisters) achieve worse risk sharing following an idiosyncratic crop loss. In this Subsection, we outline various exercises undertaken to ascertain the robustness of this finding. In particular, we rule out that this finding is a result of being unable to account for unobserved common group shocks, or because larger networks are poorer, or because there is higher competition for resources among networks with many males, or because larger networks are more vulnerable to crop losses.

Aggregate Extended Family Shocks

As mentioned above, our data doesn't allow us to adequately account for common shocks at the extended family level. Such common shocks might be correlated with potential group size, hence biasing our estimates. For example, larger potential groups might be less vulnerable to a common group shock than smaller potential groups, leading to a negative bias in our coefficients of interest if the common group shock is not accounted for.

We use simulations to assess the magnitude and sign of this bias, under different assumptions on the magnitude of the common extended family level shock. We generate data under the assumption that risk is shared according to the model in Section 4.2 (augmented to allow for group-specific shocks), and parameters are set to match those in our data (where possible). In particular, we set the group size distribution to be

match the empirical distribution of brothers of the wife. Household income, y_i, consists of two components: y_h, an idiosyncratic household level component, and y_g - a common extended family component. Households' consumption rules are estimated numerically from the model in Section 4.2 augmented to allow for an independent group-level shock. For different levels of the common family shock, and randomly drawn idiosyncratic and group shocks, we assess how the coefficients on a specification similar to Equation 4.7 change when we add controls for common extended family shocks, rather than for common village shocks.[37] The findings are displayed in Table 4.10. The table indicates that all the coefficients are indeed biased, as expected. Moreover, the biases are sizeable, ranging from 10% of the true value to over 200%. In terms of the sign of the bias, α_1, the coefficient on the $\Delta crop$ variable is biased downwards, while β_2 (coefficient on $\Delta crop_{ivt} * 1(NS_{g,iv} = 1)$) is biased upwards in all but one case. By contrast, the coefficient on $\Delta crop_{ivt} * 1(NS_{g,iv} \geq 3)$,$\beta_3$, is biased upward. So, if anything, we are likely to be *underestimating* the negative effect of larger groups on risk sharing.

[37] Full details on the simulations and estimation equation are given in Appendix 4.8.2.

Table 4.10: Simulation Results to Assess the Sign and Magnitude of the Bias from Omitting Controls for Aggregate Extended Family Shocks

	Size of Group-Level Shock (% Avg. Annual HH Income)					
	0%	5%	10%	15%	20%	25%
Avg. $\hat{\alpha_1}$ (Group dummies)	-0.046	-0.183	-0.105	-0.105	-0.105	-0.117
Simulation std. error	[0.001]	[0.003]	[0.001]	[0.001]	[0.001]	[0.001]
Avg. $\tilde{\alpha_1}$ (Village dummies)	-0.119	-0.213	-0.160	-0.161	-0.161	-0.169
Simulation std. error	[0.003]	[0.003]	[0.003]	[0.001]	[0.003]	[0.004]
Avg. % Absolute Bias	160.7%	16.0%	51.6%	52.3%	52.4%	44.4%
Avg. $\hat{\beta_1}$ (Group dummies)	-0.106	0.090	-0.031	-0.086	-0.085	-0.099
Simulation std. error	[0.002]	[0.004]	[0.001]	[0.001]	[0.001]	[0.002]
Avg. $\tilde{\beta_1}$ (Village dummies)	-0.092	0.032	-0.044	-0.072	-0.072	-0.078
Simulation std. error	[0.004]	[0.004]	[0.004]	[0.004]	[0.004]	[0.006]
Avg % Absolute bias	13.4%	64.9%	42.9%	15.8%	15.8%	21.1%
Avg. $\hat{\beta_2}$ (Group dummies)	-0.074	0.037	-0.051	0.009	0.010	0.009
Simulation std. error	[0.002]	[0.004]	[0.002]	[0.003]	[0.002]	[0.003]
Avg. $\tilde{\beta_2}$ (Village dummies)	-0.033	0.040	-0.021	0.029	0.029	0.028
Simulation std. error	[0.003]	[0.003]	[0.003]	[0.004]	[0.004]	[0.005]
Avg % Absolute bias	56.2%	9.5%	58.6%	225%	217.5%	220%

Notes to Table: Data simulated with parameters to match those in data. Exact parameter values, and simulation details are explained in Appendix 4.8.2. The average annual household income is around 56000 MK. $h_h = 61223$MK and $l_h = 46475.64$MK. α_1 is the coefficient associated with $\Delta crop$, β_1 is that associated with No sibling of that type * $\Delta crop$; and β_2 is that associated with ($\geq$3 siblings)*$\Delta crop$.

Are larger networks poorer?

An important concern is that our findings may be driven by unobserved factors that drive both network size and changes in log consumption. One such set of factors relates to the fact that households with larger family networks may be poorer. Larger families

have long been observed to be poorer in a variety of contexts. This could make them less able to provide support to other family members when they need it, thus leading to worse risk sharing. We provide evidence that our results are not driven by the fact that larger families are poorer.

First, we fail to find similar results for the sisters of the wife, and for the brothers of the husband. If the findings were being driven by a family size effect, rather than being the effect of having many brothers, we would expect to find that households with many sisters are also less well protected from crop loss events. Of course, this argument is only valid as long as households with many sisters and those with many brothers are not different in other dimensions. To assess whether this is the case, we test whether households where the wife has ≥ 3 brothers and < 3 sisters are different to households with ≥ 3 sisters, but < 3 brothers, focusing on dimensions that are less likely to have changed as a result of recent shocks experienced by households. The findings from this analysis are displayed in Table 4.11 (4.12) for the husband (wife).

Table 4.11: Comparing characteristics of households where husband has ≥ 3 brothers with those where he has ≥ 3 sisters

	$\geq$3 sis of husband	sd	$\geq$3 bros of husband	sd	p-val of diff
Husband's Characteristics					
Years of education	4.815	0.380	5.257	0.329	0.391
Age	37.865	0.923	37.269	0.814	0.632
Chewa	0.931	0.027	0.945	0.028	0.527
Wife's Characteristics					
Years of education	3.404	0.282	3.609	0.274	0.582
Age	33.685	0.839	33.027	0.663	0.525
Chewa	0.978	0.014	0.973	0.014	0.768

Notes: ** Significant at 5% level; * at the 10% level. Sample includes households where the wife has 3 or more brothers and less than 3 sisters or 3 or more sisters and less than 3 brothers.

As can be seen from the tables, we find few differences in the small set of observable characteristics of the husband and wife in these two types of households. In particular,

there are no significant differences in the amount of education of the husband or wife in households where the husband (wife) has ≥ 3 brothers and those where he (she) has ≥ 3 sisters. Though males typically have a higher level of education than females, there is no difference in education levels by the sex composition of the individual's sibship.[38]

Table 4.12: Characteristics of households where wife has ≥ 3 brothers with those where she has ≥ 3 sisters

	$\geq$3 sis of wife	sd	$\geq$3 bros of wife	sd	p-value of diff
Husband's Characteristics					
Years of education	5.202	0.322	5.519	0.377	0.517
Age	37.487	0.807	37.404	1.051	0.954
Chewa	0.915	0.035	0.886	0.052	0.392
Wife's Characteristics					
Years of education	3.760	0.277	3.872	0.326	0.794
Age	33.241	0.592	32.456	0.925	0.489
Chewa	0.962	0.017	0.972	0.016	0.352

Notes: ** Significant at 5% level; * at the 10% level. Sample includes households where the husband has 3 or more brothers and less than 3 sisters or 3 or more sisters and less than 3 brothers.

Number of Brothers and Competition for Resources

Another concern is that there might be more competition for production resources among families with many males: essentially, if land is passed down to males only, and there are many males in a particular family, each male would receive a smaller land plot, and thus would be less able to help their sisters' households when they face idiosyncratic shocks. The land descent system in Mchinji is considered to be a mixed one: some households practice a patrilineal system and pass on land to males, whereas others practice a matrilineal system and pass on land to females. We provide some

[38]The differences in education levels by gender are likely to be driven by gender differences in the economic returns to education rather than due to explicit gender discrimination by parents. To our knowledge, there is no evidence of sex discrimination in investments in children at either the pre-natal or post-natal stage. Indeed, when we analyse the effects of a randomised infant feeding counselling intervention in this context by gender, we find no differences in nutritional investments in children by gender. These results are available on request.

suggestive evidence to rule out this channel. In particular, though we do not have information on the landholdings of siblings of the husband or wife, we can look at whether there are any differences in the size of land between households where the husband has many brothers and few sisters compared to those where the husband has many sisters but few brothers. If the patrilineal form of land descent is more dominant in our sample (which we do not believe it to be), we would expect households where husbands have many brothers to have smaller plots of land than households where the husband has many sisters. Examining the data, we see that households where the husband has 3 or more brothers and fewer than 3 sisters have on average 2.9 hectares of land, whereas those where the husband has 3 or more sisters and fewer than 3 brothers have on average 2.7 hectares of land. This difference is not statistically significant, thus providing suggestive evidence that the empirical findings are unlikely to be driven by this channel.

Potential Group Size and Incidence of Shocks

A second concern is that larger extended families could be more vulnerable to crop loss events, particularly if they are poorer. In that case, the deficiencies in risk sharing detected above may be a consequence of poverty, rather than a breakdown of risk sharing due to unstable coalitions.

To see if this is the case, we consider how the incidence and intensity of crop losses vary with potential group size. To do so, we regress the crop loss and intensity variables on our network size variables, pooling data from both survey rounds. Table 4.13 displays these results. The table does not indicate that households where the wife has many brothers are more vulnerable to crop loss events compared to households where the wife has fewer brothers. Thus, we can rule out that our finding of poor risk sharing among these households is driven by this channel. Interestingly, we find a negative coefficient for households where the wife has 3 or more sisters: such households are less likely to be affected by a crop loss incident, though there is no difference detected in the intensity of the crop loss.

Table 4.13: Network size and crop loss incidence

	[1] **crop loss incidence**	[2] **crop loss intensity**	[3] **crop loss incidence**	[4] **crop loss intensity**
	Siblings of husband alive		**Siblings of wife alive**	
No brothers	-0.0571	-0.0752	-0.0058	0.0012
	[0.0452]	[0.0721]	[0.0471]	[0.0902]
≥3 brothers	-0.014	-0.0527	0.0033	-0.0599
	[0.0262]	[0.0424]	[0.0275]	[0.0428]
N	1131	1131	1131	1131
R-squared	0.0244	0.0200	0.0289	0.0216
No sisters	0.0036	-0.0083	-0.0290	-0.0633
	[0.0548]	[0.0731]	[0.0522]	[0.0818]
≥3 sisters	-0.0075	-0.0203	-0.0628**	-0.0216
	[0.0285]	[0.0384]	[0.0314]	[0.0391]
N	1131	1131	1131	1131
R-squared	0.0262	0.0198	0.0306	0.0191

Notes: *** Significant at the 1% level; ** the 5% level; * the 10% level. Standard errors clustered at the village level in parentheses. Regressions pool together all households where a married head or spouse was surveyed and who were resident in the same village for both survey rounds. "Crop loss incidence" is a dummy variable that indicates whether the household experienced a crop loss in the previous year (or since the last survey), while "Crop loss intensity" is the size of the crop loss normalised by predicted household consumption.

4.6 Calibration

The empirical results show that households where the wife (husband) has a large number of brothers (sisters) achieve worse risk sharing outcomes compared to households where the wife (husband) has fewer brothers (sisters). The theory indicates that the relationship between risk sharing and potential group size is ambiguous and sensitive to parameter values: for some combination of parameters, larger potential groups can offer better risk sharing, while for others, they offer worse risk sharing. To investigate whether the findings can be explained by the theory, we conduct a calibration exercise to see if the model can reproduce the empirical findings when parameter values are set

to be similar to those in our data.

We parameterise the value of the high and low endowment as follows: From the data, we obtain the average annual household income from agriculture for all households in the sample, $\bar{y}$. This is equivalent to a weighted average of the high and low endowment states, where the low endowment state is taken to be the high endowment state less the crop loss (in nominal terms, without normalising for predicted consumption):

$$\bar{y} = p * h + (1 - p)(h - crop) \tag{4.8}$$

We obtain the values for $\bar{y}$, p and *crop* from the data, and use the formula 4.8 to back out the values for h and l respectively. Table 4.14 displays the resulting parameters. In addition to these parameters, we also need to specify a value for the coefficient of relative risk aversion, ρ and the discount factor, δ. We set $\rho = 1.5$ and $\delta = 0.95$. The value for δ, which is lower than that typically estimated for developed countries, is within the range estimated for India by Ligon et al. (2003).

Table 4.14: Parameter values for calibration

Parameter	Value
h	61223.64MK
l	46475.64MK
p	0.63
δ	0.95
ρ	1.5

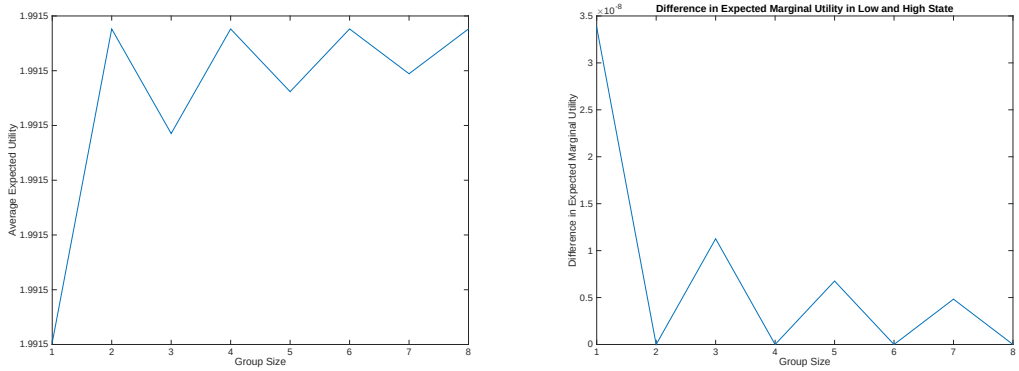
Note to Table: This table displays the parameter values used to calibrate the theoretical model. The values for the high and low endowments, h and l are in Malawi Kwacha (MK). The exchange rate at the time of the survey was roughly US$1 = 140MK.

Figure 4.6 plots the value for average expected utility and group size. What is striking is that weighted expected utility increases with potential group size initially, but then falls before increasing again in a zigzag pattern. This pattern can be explained by the fact that given the parameter values, only groups of size 1 and 2 are stable. Larger potential groups would then sort randomly into the smaller stable groups, for example, groups of size 3 would sort into groups of size 1 and 2. Since expected utility under autarky is lower than in a group of size 2, this results in a drop in average expected

utility for a potential group of size 3. In fact, such an argument holds for all odd-sized potential groups, while even-sized potential groups would sort into subgroups of 2 and attain the same average expected utility as a group of size 2.

Importantly, the drop in expected utility when moving from a potential group of size 2 to 3 matches the pattern found in the data, suggesting that threats of coalitional deviations may be a possible explanation for the worse risk sharing for households where the wife has many brothers.

Figure 4.6: Calibration Findings - Average Expected Utility and Network Size



(a) Weighted Avg Exp. Utility

(b) Weighted Avg. Exp. Diff. in Marginal Utility

Notes: The Figure on the left panel shows the relationship between weighted average expected utility and group size, while that on the right panel shows the relationship between the weighted average expected difference in marginal utility and group size

4.7 Conclusion

In this chapter, we study the relationship between group size and the extent of risk sharing in a setting with limited commitment and coalitional deviations. In such environments, two forces are at play in determining the relationship between group size and risk sharing: on the one hand, larger groups allow for more effective diversification, and hence better risk sharing. On the other hand, they are more vulnerable to deviations by sub-groups (coalitions) of households who can renege on the informal arrangement and continue sharing risk in the smaller subgroup. Thus, risk sharing groups will be bounded from the top (GR). We extend the model of GR and use simulations to show that the relationship between risk sharing and group size is theoretically ambiguous. The nature of this relationship is thus an empirical question.

We investigate this question empirically using data from rural Malawi, and over-

come the challenge posed by the fact that the size of the actual risk sharing group is endogenous, by considering potential group size, and focusing on a group likely to be exogenous – siblings of the household head and spouse. Evidence from the anthropological and sociological literatures indicate that the extended family is a crucial risk sharing institution in the setting we study. Moreover, historical, well-documented norms at play in this context indicate a much more important role for a wife's brothers in providing risk sharing, than for her sisters. These norms highlight an important source of heterogeneity in risk sharing patterns, and also allow us to construct placebo tests to alleviate concerns that unobserved factors correlated with our measures of group size and the efficiency of risk sharing are driving our findings.

We consider how well protected a household's consumption is to idiosyncratic crop losses – an important source of risk in this setting – allowing the effects to vary by the size of the family network of the husband and wife (defined separately by gender of sibling). In line with the literature on informal risk sharing, we measure the degree of risk sharing by the correlation between changes in household log consumption and idiosyncratic crop losses (net of aggregate shocks at the village level). A first non-parametric specification, which places no restrictions on the shape of the relationship between the degree of risk sharing and potential group size, indicates that this relationship is non-linear. However, these estimates are extremely imprecise.

To increase power, we divide group size into three bands, the boundaries of which are informed by the non-parametric analysis. Estimates from this specification indicate that households where the wife (husband) has many brothers (sisters) achieve worse risk sharing relative to households where they have fewer brothers (sisters). We fail to find a similar relationship for the wife's sisters (brothers), which indicates that the relationship is unlikely to be driven by the fact that households where the husband/wife have many siblings are poorer. Moreover, we show that these households are not more susceptible to crop losses, suggesting that the findings are not driven by this margin either. We also provide suggestive evidence to rule out other channels including higher competition for production resources among extended families with many male siblings. A calibration exercise, where we parameterise our theoretical framework using information from the data (where available), indicates that the empirical patterns could be produced by the theory.

Thus, larger potential risk sharing groups need not yield better risk sharing outcomes, indicating a role for governments and other actors to implement policies and mechanisms to better protect household wellbeing.

4.8 Appendix

4.8.1 Details of Model Simulation Calculations

In this section, we provide a step-by-step overview of the calculations that yield Figure 4.1 above. Given the specific parameter values associated with this particular example, groups of size, $N = 1, 2$ and 3 are found to be stable, while those of sizes, $N = 4 - 10$ are found to be unstable. The social planner randomly assigns households in a group of a specific size, N to stable subgroups of sizes $s_1, s_2, ...s_J$ in a manner so as to maximise total expected utility, while ensuring that all households are assigned to some stable sub-group. For groups of size, $N = 1, 2$ and 3, the social planner has no need to reassign households to stable sub-groups of a smaller size, s_j. Thus the average expected utility, and expected difference in marginal utility, for households in groups of these sizes can be recovered from Equations 4.4 and 4.5 by setting $\pi_s = 1$ for its group size and 0 for all other other stable group sizes and evaluating these equations at the optimal transfer. The calculated values are given in the Table 4.15 here.

Table 4.15: Expected Utility, and Expected Difference in Marginal Utility for Stable Groups, Example 1

Group Size	Expected Utility	Expected Difference in Marginal Utility
1	0.66206	0.15745
2	0.66377	0.03378
3	0.66857	0.03344

For groups of other sizes, we need to solve for the combination of stable sub-groups that maximises total expected utility when the social planner randomly assigns households to the stable subgroups. In this example, the optimal allocation of sub-groups for a group of size 4 is 1 sub-group of size 3 and 1 sub-group 1. Since households are randomly allocated into these sub-groups, each household has a $\frac{1}{4}$ chance of being in the sub-group of size 1 and $\frac{3}{4}$ of being in a group of size 3. The associated weighted average expected utility is thus

$$\frac{3}{4} * 0.66857 + \frac{1}{4} * 0.66206 = 0.66694$$

For a group of size 5, the optimal sub-groups are one of size 3 and one of size 2.

Each household now has a $\frac{3}{5}$ chance of being in the group of size 3 and a $\frac{2}{5}$ chance of being in a group of size 2. The corresponding weighted average expected utility is

$$\frac{3}{5} * 0.66857 + \frac{2}{5} * 0.66206 = 0.66665$$

Note that the weighted average expected utility for a group of size 5 is lower than that for a group of size 4 because the probability of being in the higher utility sub-group of size 3 is higher in the latter case than in the former. This probability difference offsets the increased expected utility from being in a sub-group of size 2 in the former case relative to being in one of size 1 in the latter case. Table 4.16 summarises these calculations for groups of sizes 4 - 10 in this example.

Table 4.16: Details of calculation for unstable groups, Example 1

Group Size	Prob. of being in stable subgroup of size:			Weighted Avg. EU	Weighted Avg. Expected Diff in MU
	1	2	3		
4	$\frac{1}{4}$	0	$\frac{3}{4}$	0.66694	0.06444
5	0	$\frac{2}{5}$	$\frac{3}{5}$	0.66665	0.03357
6	0	0	1	0.66857	0.03344
7	$\frac{1}{7}$	0	$\frac{6}{7}$	0.66764	0.05115
8	0	$\frac{1}{4}$	$\frac{3}{4}$	0.66737	0.03352
9	0	0	1	0.66857	0.03344
10	$\frac{1}{10}$	0	$\frac{9}{10}$	0.66792	0.04584

4.8.2 Details of Simulations to Assess the Sensitivity of Parameter Estimates to Aggregate Extended Family Shocks

A concern is that our estimates might be biased since we are unable to suitably control for group-level shocks. We use simulations to assess the sensitivity of our parameter estimates to biases arising from this issue. Here we provide some details on the set-up of the simulations.

1. First, we generate a set of households and assign them to groups and villages. Villages contain multiple groups, and groups can span across multiple villages. Groups have different sizes, with the distribution of group sizes (total, and in the village) selected to match those found in the data.

2. We set the income process as follows: household income is composed of two com-

ponents, a household-level component, $y_i = \{h_i, l_i\}$ and a group-level component, $y_g = \{h_g, l_g\}$. We select the values of h_i and l_i to be similar to those in our data. For the group-level shock, we set $h_g = 0$ and vary the values of l_g to be $\pi\bar{y}_i$, where $\bar{y}_i$ is the household's expected income. The probability of h_i is set to p, $0 < p < 1$; and that of h_g is set to π; $0 < \pi < 1$. Throughout, we set $p = 0.63$ and $\pi = 0.06$. The former probability is derived from our data, and the latter corresponds to the probability of a village-level aggregate shock in the data.

3. We extend the Genicot and Ray (2003) model to allow for common group level shocks (that are independent of the household-level shock), and given the values of h_i, l_i, h_g and l_g, and other parameters, compute the set of stable group sizes and derive the optimal transfer. We use the same consumption rule as in GR, and use the optimum transfers to calculate consumption under different states.

4. Given the set of stable group sizes, we allocate households in a potential group of size S to stable groups so as to maximise the total expected utility of the potential group. Since we assume the households are all homogenous, this amounts to a random allocation of households to stable groups.

5. We then randomly draw realisations of y_i for each household, and y_g for each group.

6. Given the stable group, and the realisations of y_i and y_g, we use the consumption rule computed in (3) above to assign consumption to each household.

7. We repeat (5) and (6) to attain a panel of shock and income realisations.

8. We then run specification 4.9, allowing first for the term μ_t^n to be a group-level dummy, and then for it to be a village-level dummy. We obtain the coefficients β_1, β_2 and β_3.

$$\Delta log(c_{ivt}) = \alpha_0 + \alpha_1 \Delta(crop_{ivt}) + \beta_1 \Delta crop_{ivt} * 1(NS_{g,iv} = 1) + \beta_2 \Delta crop_{ivt} * 1(NS_{g,iv} \geq 3) + \mu_t^n + \Delta\epsilon_{int} \quad (4.9)$$

9. Repeat steps 4-8 100 times. Table 4.10 displays the results for different levels of the common group shock.

Chapter 5

Nutrition, Information and Household Behavior: Experimental Evidence from Malawi

5.1 Introduction

Since Becker (1965)'s seminal contribution, economists have long recognized that many goods are not directly bought in the market, but are produced at home using a combination of market and non-market goods. The home production framework has been particularly fruitful in studying the production of health, in particular child health (Grossman 1972, Rosenzweig & Schultz 1983, Gronau 1987 and 1997). An important implication of such models is that households make choices given their knowledge of the (child) health production function. Consequently, deficiencies in knowledge lead to suboptimal household choices and thereby distorted levels of child health. Establishing empirically the consequences of deficiencies in knowledge on household behavior has, however, been challenging because knowledge is endogenous and is usually either unob-

[0]This chapter is co-authored with Emla Fitzsimons, Alice Mesnard and Marcos Vera-Hernandez. We thank the Mai Mwana team, especially Tambozi Phiri, Andrew Mganga, Nicholas Mbwana, Christopher Kamphinga, Sonia Lewycka, and Mikey Rosato for their advice, useful discussions, and assistance with data collection. We are grateful also to Julia Behrman, Senthuran Bhuvanendra, Lena Lepuschuetz and Carys Roberts for excellent research assistance. We also thank the editor and two referees, as well as Orazio Attanasio, Richard Blundell, Irma Clots, Colin Cameron, Esther Duflo, Markus Goldstein, Michael Kremer, Manoj Mohanan, Grant Miller, Amber Peterman, Ian Preston, Gil Shapira, Alessandro Tarrozi, Patrick Webb, and participants at numerous seminars and conferences for useful comments and discussions. The authors acknowledge financial support from the ESRC/Hewlett Joint Scheme under Grant reference RES-183-25-008; ESRC-NCRM Node 'Programme Evaluation for Policy Analysis' Grant reference RES-576-25-0042; and from Orazio Attanasio's ERC Advanced Researcher Grant, Agreement No. 249612 - IHKDC.

served or proxied by education which also affects child health through other channels including earnings.

In this chapter, we overcome this challenge by exploiting an intervention, implemented through a cluster randomized trial, aiming to improve mothers' knowledge of the child health production function in rural Malawi. The intervention solely provided information on child nutrition to mothers, thus yielding a clean source of identification. Our contribution is twofold. First we assess whether the intervention improved child nutrition and consequently health. Second, drawing on a simple theoretical model, we investigate how other household choices change to accommodate the improved knowledge of the production function. In so doing, we assess whether non-health choices, particularly parental labor supply, might be affected by parents' knowledge of the child health production function.

In the context we study, rural Malawi, mothers have many misconceptions about child nutrition. To take some examples, it is common practice to give porridge diluted with unsterilized water to infants as young as one week; the high nutritional value of groundnuts, widely available in the area, is not well-known; and widespread misplaced beliefs include that eggs are harmful for infants as old as 9 months, and that the broth of a soup contains more nutrients than the meat or vegetables therein. This evidence suggests that important changes can be expected if these misconceptions are corrected.

The intervention we study delivered information in an intense manner: trained local women visited mothers in their homes once before the birth of their child and four times afterwards, and provided information on early child nutrition on a one-to-one basis. Moreover, the fact that the intervention had been running for at least 3 years when outcome data were collected, allows for a sufficient time-frame for practices to change. This lapse also allows us to measure medium-term impacts, which is important since interventions often perform much better in the short- rather than medium-term (Banerjee et al. 2008 and Hanna et al. 2016).

Consistent with gains in knowledge, we find evidence of improvements in infants' diets and household food consumption, particularly an increase of protein-rich foods and of fruit and vegetables. We also find that household food consumption increases, and there is suggestive evidence that it might have been partially financed through increased labor supply. Overall, the findings are consistent with households learning that some relatively costly foods are more nutritious than they previously believed, and adjusting their labor supply so as to facilitate increases in their children's intake of them. Indeed, we show that households adjust their behavior on several margins including child diet inputs and labor supply, making their response more complex than simply changing the composition of consumption while keeping total consumption constant.

We find that the intervention improved children's physical growth, particularly height, a widely used indicator of long-term nutritional status. This finding is particularly important for policy: child malnutrition is a severe and prevalent problem in developing countries (de Onis et al. 2000), that leads to poor health and excess child mortality (Bhutta et al. 2008, Pelletier et al. 1994) and is also linked to poor human capital outcomes later on in life.[1]

The chapter deals carefully with the increasingly important issue of inference in cluster randomized trials when the number of clusters is small. It is well known that in this situation, standard statistical formulae for clustered standard errors based on asymptotic theory (cluster-correlated Huber-White estimator) provide downward biased standard error estimates (Donald & Lang 2007, Wooldridge 2004, Bertrand et al. 2004, Cameron et al. 2008). We use two leading methods for inference in this case - randomization inference (Fisher 1935, Rosenbaum 2002) and wild-cluster bootstrap-t (Cameron et al. 2008). Furthermore, we assess their performance in our data using Monte Carlo experiments, and find that both methods perform relatively well. Presenting the performance of these two methods side-by-side is of interest for many empirical applications, given the increasing trend in randomized trials with a small number of clusters.

Lewycka et al. (2013) studies the effect of this intervention on exclusive breastfeeding and infant mortality. Our paper addresses a different question, whether improving knowledge of the health production function affected consumption, labor supply, nutritional practices and child nutrition to the age of around 5 years. We also use a different dataset; they interview mothers until their child is six month old, while we rely on a representative sample of women of reproductive age, and their households. More details about the design of the intervention can be found in Lewycka et al. (2010).

Our work contributes to a number of strands of literature. First, it adds to the discussion on the effects of health information on behavior (Dupas 2011*b*).[2] The evidence is mixed: Madajewicz et al. (2007), Jalan & Somanathan (2008) and Dupas (2011*a*) find that providing information on, respectively, the arsenic or fecal concentration of water; and the risks of contracting HIV improves associated practices; while Kamali et al. (2003), Kremer & Miguel (2007) and Luo et al. (2012) find that health behaviors relating to, respectively, HIV, deworming and anemia do not respond to health education. This paper departs from these studies by not only considering a multifaceted

[1] See, among others, Behrman (1996), Strauss & Thomas (1998), Glewwe et al. (2001), Alderman et al. (2001), Behrman & Rosenzweig 2004, Schultz (2005), van den Berg et al. (2006), Hoddinott et al. (2008), Maluccio et al. (2009), Banerjee et al. (2010), Currie et al. (2010), van den Berg et al. (2009), Maccini & Yang (2008), Currie (2009), van den Berg et al. 2010, Lindeboom et al. (2010), Currie & Almond (2011), Barham (2012), Bhalotra et al. (2016).

[2] For the case of education, see for instance Jensen (2010).

information intervention, but also by studying household responses on a wider range of margins than those directly targeted by the intervention. In doing so, this is one of the first papers to investigate how behaviors not directly related to the topic of an information campaign adjust to it.

Second, it contributes to the literature evaluating the effects of nutrition information interventions on nutrition practices and child health. Morrow et al. (1999) and Haider et al. (2000) find improvements in excluding breastfeeding within small scale randomized controlled trials in Mexico and Bangladesh respectively; while Alderman (2007), Linnemayr & Alderman (2011); and Galasso & Umpathi (2009) find improvements in child weight-for-age, an indicator for medium-term health status, using non-experimental methods. Our paper builds on these by studying the effects on a range of measures of child health, health practices, and other margins of household behavior, all identified through a randomized controlled trial.

Finally, it relates to the literature investigating the causal effects of parental education on child health. In developed countries, Currie & Moretti (2003) and McCrary & Royer (2011) find respectively, decreased incidence of low birth weight and modest effects on child health of increased maternal schooling in the US, while Lindeboom et al. (2009) find little evidence that parental schooling improves child health in the UK. For developing countries, Brierova & Duflo (2004) and Chou et al. (2010) find that parental schooling decreases infant mortality in Indonesia and Taiwan respectively. However, it is difficult to disentangle whether the effect of education is working through changes in knowledge of the child production function, or through increased income and hence access to more and better quality care. Related to this, Thomas et al. (1991) and Glewwe (1999) find that almost all of the impact of maternal education on child's height in Brazil and Morocco can be explained by indicators of access to information and health knowledge.

The rest of the paper is structured as follows. Section 5.2 provides background information on rural Malawi and describes the experimental design and data, section 5.3 describes the theoretical framework, while section 5.4 sets out the empirical model. Our main results are presented in section 5.5. Section 5.6 rules out alternative potential explanations behind our findings, and section 5.7 concludes.

5.2 Background and Intervention

5.2.1 Background

Malnutrition in the early years (0-5) is one of the major public health and development challenges facing Malawi, one of the poorest countries in Sub-Saharan Africa. The

2004 Malawi Demographic and Health Survey (DHS) Report indicates an under-five mortality rate of 133 per 1000, and under-nutrition is an important factor driving this: Pelletier et al. (1994) estimate that 34% of all deaths before age 5 in Malawi are related to malnutrition (moderate or severe). Moreover, 48% of Malawian children aged < 5 years suffer from chronic malnutrition, a rate that is the second highest in sub-Saharan Africa.

Poor feeding practices are at least partly responsible for these extreme malnutrition indicators. Over half of all infants aged < 6 months are given food and/or unsterilized water (DHS, 2004), which can lead to gastrointestinal infections and growth faltering (Haider et al. 2000, Kalanda et al. 2006) and is contrary to World Health Organization (WHO) recommendation of exclusive breastfeeding for the first six months of an infant's life. Furthermore, porridge diluted with unsterilized water is often given in large quantities to infants as young as one week (Bezner-Kerr et al. 2007). In terms of nutrition for infants aged > 6 months, their diets - rich in staples such as maize flour - frequently lack the necessary diversity of foods to provide sufficient amounts of energy, proteins, iron, calcium, zinc, vitamins and folate: in our sample, 25% of children aged 6-60 months did not consume any proteins over the three days prior to the survey, with a further 30% consuming just one source of protein. Poor nutritional practices are likely to be related to a lack of knowledge: for instance, only 15% of mothers in our sample knew how to best cook fish combined with the local staple so as to maximize nutritional value.

It is against this background that, in 2002, a research and development project called MaiMwana (Chichewa for "Mother and Child") was set up in Mchinji District, in the Central region of Malawi.[3] Its aim was to design, implement and evaluate effective, sustainable and scalable interventions to improve the health of mothers and infants. Mchinji is a primarily rural district, with subsistence agriculture being the main economic activity. The most commonly cultivated crops are maize, groundnuts and tobacco. The dominant ethnic group in the district is the Chewa (over 90% in our data). According to the 2008 Malawi census, socio-economic conditions are comparable to or poorer than the average for Malawi (in parentheses in what follows), with literacy rates of just over 60% (64%), piped water access for 10% (20%) of households and electricity access for just 2% (7%) of households.

[3] MaiMwana is a Malawian trust established as a collaboration between the Department of Pediatrics, Kamuzu Central Hospital, the Mchinji District Hospital and the UCL Centre for International Health and Development. See http://www.maimwana.malawi.net/MaiMwana/Home.html

5.2.2 The Intervention

In 2005, MaiMwana established an infant feeding counseling intervention in Mchinji District (ongoing at time of follow-up), to impart information and advice on infant feeding to mothers of babies aged < 6 months.[4] The intervention thus targets the very first years of life, a critical period for growth and development during which nutritional interventions are likely to be most beneficial (Schroeder Jr et al. 1995, Shrimpton et al. 2001, Victora et al. 2010). The information is provided by trained female volunteers ("peer counselors" hereon) nominated by local leaders. In practice, peer counselors are literate local women aged 23-50 years with breastfeeding experience.[5]

Each peer counselor covers an average population of 1,000 individuals, identifying all pregnant women within this population and visiting them five times in their homes: once before giving birth (3rd trimester of pregnancy) and four times afterwards (baby's age 1 week, 1 month, 3 months, 5 months). Although all pregnant women are eligible for the intervention and participation is free, in practice around 60% of them are visited by the peer counselors. Our data show that women who were visited by the peer counselor tend to be poorer: in particular, they were 4.8 percentage points (7.5 percentage points) less likely to have a floor (roof) built with good materials.

Regarding the content of the visits, exclusive breastfeeding is strongly encouraged in all visits. Information on weaning is provided from when the baby is 1 month old (visits 3-5) and includes suggestions of suitable locally available nutritious foods, the importance of a varied diet (particularly, the inclusion of protein and micronutrient-rich foods, including eggs) and instructions on how to prepare foods so as to conserve nutrients and ease digestion (for instance to mash vegetables rather than liquidize them; to pound fish before cooking it). Peer counselors were provided with a manual to remind them of the content relevant for each visit, and simple picture books to aid in explaining concepts.

Experimental Design

The evaluation is based on a cluster randomized controlled trial designed as follows (see Lewycka et al. (2010), Lewycka (2011), Lewycka et al. (2013)). Mchinji District was divided into 48 clusters by combining enumeration areas of the 1998 Malawi Population

[4] Though the intervention is predominantly focused on nutrition, it also touches on other issues such as birth preparedness, HIV testing and counseling, vaccinations, and family planning. Section 5.6 discusses how these aspects relate to our results.

[5] Peer counselors receive an initial 5 day and annual refresher training, and attend monthly meetings. They are not paid, but receive a bicycle, meeting allowances, registers, calendars and supervision forms. They are supervised by 24 government health surveillance assistants and 3 MaiMwana officers.

and Housing Census.[6] This was done in a systematic way, based on the contiguity of enumeration areas and respecting boundaries of Village Development Committees (VDCs), such that each cluster contained approximately 8,000 individuals. Within each cluster, the 3,000 individuals (equating to 14 villages on average) living closest to the geographical centre of the cluster were chosen to be included in the study.[7] The study population therefore comprises of individuals living closest to the geographical centre of the clusters and was selected in this way in order to limit contamination between neighboring clusters by creating a natural buffer area. 12 clusters were randomly selected to receive the infant feeding counseling intervention, with an average of three peer counselors per cluster. A further 12 serve as controls.[8]

Evaluation Sample Description

A census of women of reproductive age was conducted by MaiMwana in all clusters in 2004, before the intervention started ("baseline census" from hereon) in July 2005 (see Figure 5.1).[9] Approximately 3.5 years into the intervention, which was still in place, we drew a random sample from the baseline census in order to conduct the first follow-up survey.[10] Specifically, in 2008 we drew a random sample of 104 women of reproductive age (17-43), regardless of their child bearing status[11] from each of the 24 clusters, leaving us with a target sample of 2,496 women.

The baseline census contains some socio-economic and demographic characteristics of these women and their households, as shown in the left hand panel of Table 5.1. Women are on average 24.5 years old, just over 61% of them are married, over 70% have some primary schooling but just 6% have some secondary schooling. Households are

[6] The District Administrative Centre was excluded because it is relatively more urbanized and less comparable to the rest of the District.

[7] The geographic centre was chosen to be the most central village in the cluster as shown on a cartographic map from the National Statistical Office, Malawi. See Lewycka (2011), pp. 122 for more details.

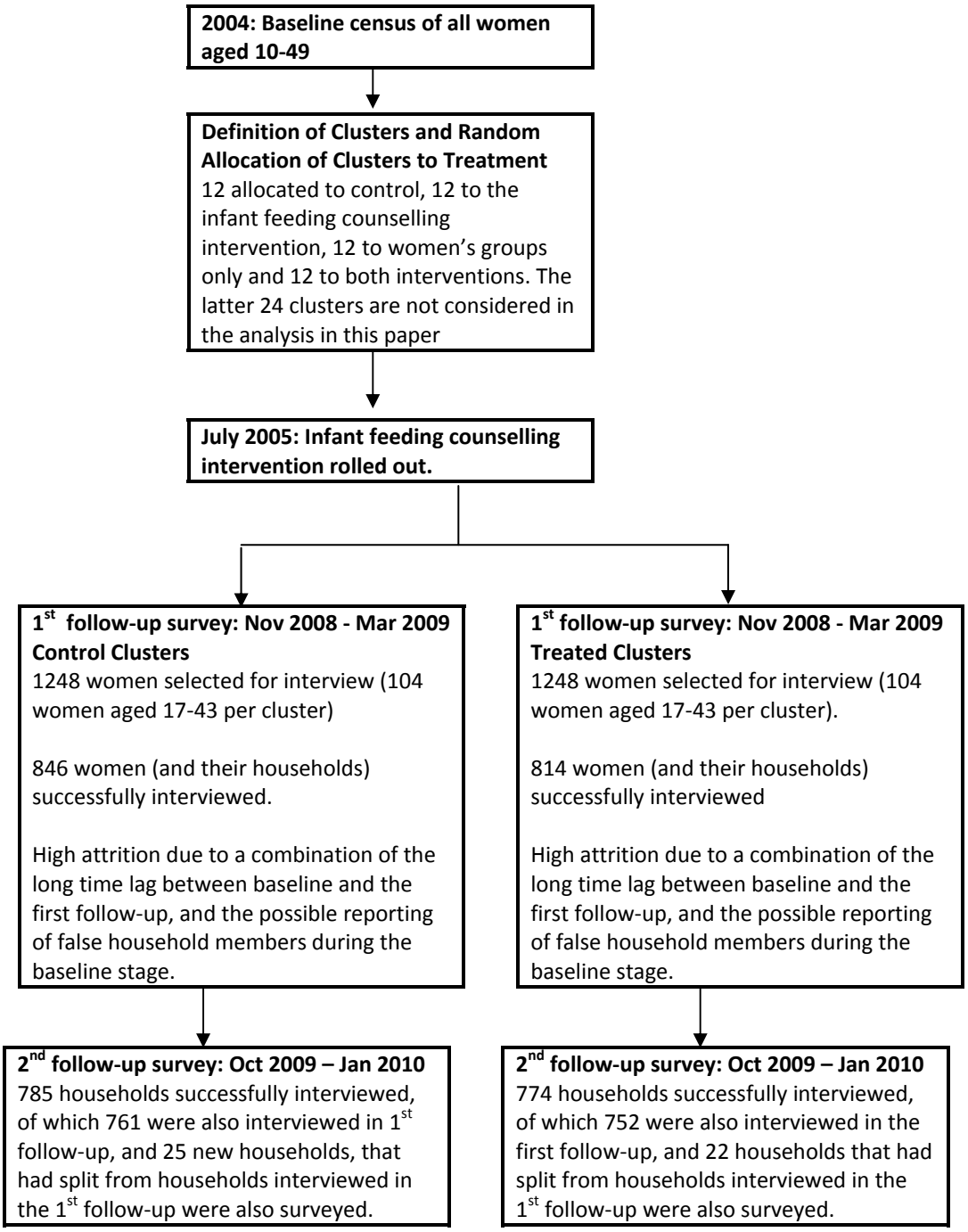
[8] Another 24 clusters were randomly assigned to receive a participatory women's group intervention, whereby women of reproductive age were encouraged to form groups to meet regularly to resolve issues relating to pregnancy, child birth and neo-natal health. Child nutrition was not a primary focus and so we exclude these clusters from this analysis (see instead Rosato et al. (2006), Rosato et al. (2009) and Lewycka et al. (2013)). MaiMwana Project also improved health facilities across the District, which equally benefitted intervention and control clusters.

[9] Further details on this baseline census can be found in Lewycka et al. (2010). We take the intervention start date to be July 2005, the date by which the first 6-month cycle had been fully completed, in line with Lewycka et al. (2013).

[10] Data collection was carried out by MaiMwana in collaboration with the authors. Data were collected in Nov 2008-March 2009 (Oct 2009-Jan 2010) at first (second) follow-up using PDAs. To ensure that results were not driven by seasonality, field teams collected data in intervention and control clusters at the same time. The data are available for download at http://www.esds.ac.uk/(study 6996).

[11] This was done to avoid any potential bias arising from endogenous fertility decisions in response to the intervention. This turns out not to be an important concern, as we show in section 5.6.

Figure 5.1: Surveys and Timing of Data Collection



predominantly agricultural and poverty is high, as indicated by the housing materials and assets. The table also shows that the randomization worked well with the sample well-balanced across intervention and control clusters at baseline given that only 1 out of 25 variables turns out to be unbalanced.[12]

[12]Other welfare programs were operating in the District at the same time as this intervention. The potentially most important is the Mchinji Social Cash Transfer, providing cash transfers to the poorest 10% of households in the district. At follow-up, the intervention was in the pilot stage and only 2.5% of households in our sample (distributed evenly between intervention and control clusters) report having received it.

Table 5.1: Sample Balance

	Full Sample			Analysis Sample - Wave 1			Analysis Sample - Wave 2		
	Control Group	Difference: Treatment - Control	p-value	Control Group	Difference: Treatment - Control	p-value	Control Group	Difference: Treatment - Control	p-value
Woman's Characteristics									
Married (dv = 1)	0.615	-0.021	0.386	0.661	-0.034	0.184	0.654	-0.024	0.340
Some Primary Schooling or Higher	0.707	0.033	0.402	0.682	0.040	0.340	0.68	0.037	0.438
Some Secondary Schooling or Higher	0.066	0.010	0.535	0.060	-0.007	0.545	0.059	-0.006	0.607
Age (years)	24.571	-0.180	0.637	25.492	-0.429	0.376	25.397	-0.217	0.621
Chewa	0.948	-0.044	0.330	0.957	-0.050	0.246	0.959	-0.054	0.268
Christian	0.977	0.006	0.476	0.979	0.008	0.336	0.981	0.005	0.454
Farmer	0.661	-0.075	0.108	0.688	-0.060	0.128	0.678	-0.055	0.220
Student	0.236	0.015	0.438	0.204	0.022	0.274	0.208	0.017	0.410
Small Business/Rural Artisan	0.036	0.030	0.129	0.037	0.024	0.220	0.039	0.025	0.264
Household Characteristics									
Agricultural household	0.995	-0.005	0.471	0.995	0.002	0.591	0.995	0.003	0.500
Main Flooring Material: Dirt, sand or dung	0.913	-0.041	0.232	0.916	-0.027	0.474	0.916	-0.028	0.422
Main roofing Material: Natural Material	0.853	-0.018	0.697	0.857	-0.004	0.891	0.86	-0.008	0.861
HH Members Work on Own Agricultural Land	0.942	-0.057	0.124	0.950	-0.056	0.120	0.95	-0.06	0.140
Piped water	0.011	0.040	0.314	0.009	0.032	0.340	0.01	0.034	0.440
Traditional pit toilet (dv = 1)	0.772	0.054	0.218	0.791	0.054	0.182	0.796	0.044	0.324
# of hh members	5.771	0.066	0.817	5.848	0.132	0.863	5.903	0.096	0.833
# of sleeping rooms	2.116	0.199	0.038*	2.152	0.166	0.128	2.174	0.155	0.136
HH has electricity	0.002	0.007	0.166	0.002	0.004	0.338	0.003	0.004	0.394
HH has radio	0.630	0.030	0.408	0.641	0.015	0.709	0.645	0.014	0.655
HH has bicycle	0.509	0.015	0.643	0.512	0.008	0.843	0.512	0.01	0.769
HH has motorcycle	0.008	0.001	0.925	0.007	0.002	0.779	0.008	0.003	0.685
HH has car	0.006	-0.002	0.612	0.007	-0.003	0.298	0.008	-0.004	0.302
HH has paraffin lamp	0.925	0.032	0.262	0.926	0.036	0.178	0.935	0.026	0.360
HH has oxcart	0.058	-0.015	0.204	0.059	-0.022	0.090+	0.06	-0.022	0.072+
N	1248	1248		846	814		785	774	

Notes to Table: p-values are computed using the wild cluster bootstrap-t procedure as in Cameron *et al.* 2008, explained in section 4.1. 'Full Sample' includes all women (and their households) originally drawn to be part of the 2008-09 survey. 'Analysis Sample - Wave 1' includes women (and their households) who were interviewed in 2008-09 (wave 1), while 'Analysis Sample - Wave 2' includes women (and their households) who were interviewed in 2009-10 (wave 2). ** p<0.01, * p<0.05, + p<0.1.

We assess the impact of the intervention over three and a half years after it began. While this has the benefit of allowing us to assess the effect of the intervention in the medium rather than short term, it also increases the risk of attrition. We succeeded in interviewing around two thirds of the sample drawn for the first follow-up survey: 65% in intervention clusters and 67% in control clusters. Apart from the time lapse between baseline and the first follow-up, two additional factors contributed to the attrition. First, the district of Mchinji is particularly challenging for the collection of panel data because respondents are known to report "ghost members" - fictitious household members - with the intention of increasing future official aid/transfers which may depend positively on household size (see Miller & Tsoka (2012) for "ghost members" and Gine et al. (2012) 2012 for problems relating to personal identification in Malawi). Hence, it is possible that some women listed in the baseline census were in fact "ghost members" and so could not be found by the field team in 2008. Second, an unexpected sharp drop of the British Pound against the Malawi Kwacha resulted in fewer resources to track women who had moved.

The middle panel of Table 5.1 shows that the balance on baseline characteristics is maintained in the sample of women who were found ("interviewed sample"). A small imbalance is detected on just 1 variable at the 10% level, suggesting that attrition between baseline and the first follow-up was not significantly different between intervention and control clusters. While this is reassuring, it could nonetheless be the case that there is differential attrition in terms of unobserved variables. We dispel these concerns in Appendix A. We conducted a second follow-up survey of these women one year later, in 2009-10, successfully interviewing around 92% of the women interviewed at first follow-up: 92.5% and 90% in intervention and control areas respectively. The baseline balance for this sample, displayed in the right hand panel of Table 5.1, is very similar to that for the first follow-up.

The surveys contain detailed information on household consumption; consumption of liquids and solids for each child in the household ($\leq$6 years); breastfeeding practices ($\leq$2 years); health for all individuals in the household, reported by main respondent; weights and heights of children ($\leq$6 years); labor supply ($\geq$6 years); and the main respondent's knowledge about child nutrition.

5.3 Conceptual Framework

In order to understand how information of the type provided by the intervention might affect household decisions, we present a simple theoretical model in which households care about adult consumption and leisure, and about the health of their child, which is a

function of the child's consumption of a combination of nutrition inputs. For simplicity we assume that this is a bundle of two inputs, C_1 and C_2. We also assume that households have 1 adult and 1 child. The adult chooses simultaneously the amounts to spend on each child consumption inputs, C_1 and C_2, adult consumption, A and leisure, L (or labor supply, $T - L$, where T is the total time endowment of the adult). The household optimisation problem is therefore:

$$\underset{\{C_1,C_2,A,L\}}{max} U(A,\ L,\ H) \tag{5.1}$$

subject to

$$A + p_1C_1 + p_2C_2 \leq w(T - L) \tag{5.2}$$

$$H = F(C_1, C_2) \tag{5.3}$$

where $U(.,.,.)$ captures the utility from adult consumption, leisure, and child health, p_1 and p_2 are the prices of child nutrition inputs relative to adult consumption, and w is the wage per unit of time.[13] The function$F(.,.)$ represents the health production function, which is increasing in both C_1 and C_2, and concave. Following Cunha et al. (2013) and Del Boca et al. (2014), we assume that both the utility function and the production function are Cobb-Douglas, that is, $U(A, L, H) = A^\alpha L^\beta H^\gamma$ and $H = C_1^\delta C_2^\theta$, with $\alpha,\ \beta,\ \gamma,\ \delta,\ \theta > 0$, and $\delta + \theta < 1$. We can therefore rewrite the optimization problem as:

$$\underset{\{C_1,C_2,A,L\}}{max} A^\alpha L^\beta C_1^{\gamma_1} C_2^{\gamma_2}$$

subject to:

$$A + p_1C_1 + p_2C_2 \leq w(T - L)$$

where $\gamma_1 = \gamma\delta$ and $\gamma_2 = \gamma\theta$.[14]

Households make their consumption and labor decisions under their own perception of the child health production function, $C_1^\delta C_2^\vartheta$, which might differ from the true one (see Cunha et al. 2013). This perceived production function depends on δ and ϑ, two

[13] We use a static, unitary model to draw out the key behavioral responses to the intervention in the simplest way. See Chiappori (1997) and Blundell et al. (2005), among others, for work that incorporates labor supply, household production and/or children within a collective framework. See Grossman (1972) for dynamic considerations of a health production function.

[14] We assume that the household cannot borrow, which is consistent with well-known credit constraints in developing countries, as discussed for instance in Dupas (2011*b*).

parameters that measure the household's perception of the returns to child nutrition inputs. Changes in these parameters will change γ_1 and γ_2.

To study the effect of the intervention, we differentiate the first order conditions with respect to γ_1 (see Appendix B), and find that: $\frac{dC_1}{d\gamma_1} > 0$, but that $\frac{dC_2}{d\gamma_1} < 0$, $\frac{dA}{d\gamma_1} < 0$, and $\frac{dL}{d\gamma_1} < 0$. This allows us to establish the following proposition:

Proposition 1: *If γ_1 increases, then C_1 and total household consumption increases, but C_2, A, and L decrease. Similarly, if γ_2 increases, then C_2 and total household consumption increases, but C_1, A, and L decrease.*

The intuition is as follows. If the perceived productivity of C_1, γ_1, increases, then more will be consumed of this input. Given the concavity of the utility function, this increase is better accommodated by a small decrease in all other arguments of the utility function (C_2, A, and L) rather than a large decrease in only one of them. Note that the increase in C_1 is not fully offset by the decrease in C_2 and A, because L also decreases, which implies that labor supply increases. As there is no borrowing or savings, the increase in labor supply implies an increase in overall household consumption.[15]

The intervention promotes the consumption of protein-rich foods, fruits and vegetables relative to others such as staples. If C_1 summarizes the goods that the intervention promotes, and C_2 summarizes the consumption of staples, then the effect of the intervention can be summarized in terms of increasing γ_1 but decreasing γ_2. Following Proposition 1, we expect an important composition effect (increase in C_1 and a decrease in C_2) but the predictions on labor supply, adult and total consumption are in principle ambiguous because these will depend on whether the γ_1 or the γ_2 effect dominates. This is ultimately an empirical issue that we study below.

5.4 Empirical Framework

5.4.1 Estimation and Inference

The randomized experiment provides a clean and credible source of identification to test the propositions emerging from the theoretical framework above. To do so, we estimate OLS regressions of the form

$$Y_{ict} = \alpha + \beta_1 T_c + X_{ict}\beta_2 + Z_{c0}\beta_3 + \mu_t + u_{ict}, \qquad t = 1, 2 \tag{5.4}$$

[15] Our simple model abstracts from differential labor supply responses of the mother and the father. In a two parent model, one could imagine that additional time devoted to the acquisition and preparation of more nutritious foods might be to the detriment of mother's labor supply and/or leisure.

where Y_{ict} includes outcomes for unit i (household or individual, depending on the outcome of interest) living in cluster c at time t ($= 1, 2$ for first and second follow-ups, 2008-09 and 2009-10, respectively).[16] In line with the model, the dimensions of household behavior likely to be affected include household and child consumption, labor supply, and child health; T_c is a dummy variable which equals 1 if the main respondent of our survey was, at the time of the baseline in 2004, living in a cluster that later received the intervention; X_{ict} is a vector of household/individual-level variables measured at time t including a quadratic polynomial in age and gender; Z_{c0} is a vector of cluster-level variables measured at baseline such as proportions of women with Chewa ethnicity, and proportions with primary or secondary schooling. μ_t is a vector of month-survey year dummies indicating the month of the interview, and u_{ict} is an error term which is uncorrelated with the error term of others living in other clusters ($E[u_{ict}u_{jwq}] = 0 \; \forall \, i \neq j, c \neq w$), but which may be correlated in an unrestricted way with that of others living in the same cluster, independently of the time period ($E[u_{ict}u_{jcq}] \neq 0$). Note that this correlation structure allows for the error term for individuals/households in the same cluster to be correlated over time, and also for the presence of spillovers within but not across clusters, which is reasonable for our case given the presence of large buffer areas in place between study areas in adjacent clusters, as discussed in section 5.2.2.

The treatment indicator, T_c, takes the value 1 if the respondent was living in a treatment cluster at the time of the 2004 baseline census, and 0 if living in a control cluster at that time. Therefore, we identify an intention-to-treat parameter. Moreover defining T_c on the basis of baseline rather than current residence circumvents any bias that might arise from selective migration from control to treatment clusters.

In terms of inference, standard statistical formulae for clustered standard errors based on asymptotic theory (cluster-correlated Huber-White estimator) provide downward biased standard error estimates if the number of clusters is small, thus over-rejecting the null of no effect (Wooldridge 2004, Bertrand et al. 2004, Donald & Lang 2007 and Cameron et al. 2008). This is a potential issue, as there are just 24 clusters. We use two approaches proposed to obtain valid inference: wild cluster bootstrap-t (Cameron et al. 2008) and randomization inference (Fisher 1935, Rosenbaum 2002).

To implement randomization inference, we follow Small et al. (2008) to account for covariates by regressing the outcome variable on all covariates, except for T_c, and applying the randomization inference procedure to the residuals from this regression. The test statistic is as follows:

[16] For binary outcomes, results using Probit models are very similar and are not reported.

$$\sum_{c:T_c=1} \frac{\hat{\nu}_{ict}}{N_1} = \sum_{c:T_c=0} \frac{\hat{\nu}_{ict}}{N_0}$$

where $\hat{\nu}_{ict}$ is the residual of the first-stage regression for household i in cluster c at time t, N_1 is the number of observations in treated clusters, while N_0 is that in control clusters. Randomization inference constructs the distribution for the test statistic for every possible permutation of the randomization across clusters.[17] In practice, given the large number of possible permutations ($2,704,156$ in our case), it is not possible to compute the test statistic for every possible permutation of the random allocation. We instead use $100,000$ randomly selected permutations to construct the distribution. The p-value is then constructed based on the proportion of test statistic values that are greater than the actual test statistic value.

In each of the estimation tables, we report clustered standard errors computed using the cluster correlated Huber-White estimator, as well as the p-values of tests of the null that the coefficient is zero computed using both wild-bootstrap cluster-t procedure and randomization inference. Moreover, in Appendix 5.8.3, we perform a Monte Carlo exercise where we compare the test size for these two approaches with the nominal test size, within data generating processes that incorporate the main features of our data (number of clusters, number of observations and intra-cluster correlation). The simulations indicate that both inference methods perform relatively well.

5.4.2 Outcomes

In line with the theoretical model, our outcomes of interest span six domains: health knowledge, child and household consumption, labor supply, and child health and morbidity. For child health and morbidity, which were the main focus of the intervention, we focus on children aged over 6 months, for whom the intervention would have completed. We pool data from the 2008-09 and 2009-10 follow-up surveys for the analysis. Details on the various measures within each domain are provided in Appendix 5.8.4. However, two points are worth highlighting here: first, child consumption is measured from maternal reports of the foods consumed by each child. Second, special care was taken to measure household consumption, rather than household expenditures. This is important in this context, since a large proportion of consumption is self-produced, rather than purchased from a market.

Within each domain, we have several outcome measures, meaning that we end up

[17] Randomization inference is non-parametric and exploits the randomization, rather than asymptotic results, for inference. A disadvantage, however, is that inference is conducted on a sharp null hypothesis of no effect for any unit in the data, rather than the more interesting hypothesis of null average effect.

with over 30 outcome variables. To limit the problem caused by multiple inference (the probability of rejecting a test is increasing in the number of tests carried out), we aggregate the multiple outcome measures within a domain into a summary index, following Anderson (2008).[18] The index is a weighted mean of the standardized values of the outcome variables (with outcome variables re-defined so that higher values imply a better/more desirable outcome), with the weights calculated to maximize the amount of information captured in the index by giving less weight to outcomes that are highly correlated with each other. Another benefit of averaging across outcomes is that power is increased by reducing measurement error. In Table 5.13 of Appendix 5.8.5, we report the outcomes used to compute the index associated with each domain.

By using a summary index, our results provide a statistical test for whether the intervention has a "general effect" on each of the six main domains being tested which is robust to concerns about multiple inference (Kling et al. 2007; Liebman et al. 2004). However, because it is not possible to assess the magnitude of the effect from the results using the index, we also report the results on individual outcome variables.

Descriptive statistics pertaining to the outcomes and the indices for households and individuals in the control clusters are provided in Table 5.14 in Appendix 5.8.5. The table indicates that maternal knowledge on infant nutrition is mixed: questions related to weaning and nutritious value of foods were mostly correctly answered, while those related to food preparation and feeding when the child/its mother were unwell were often incorrectly answered. The food intake information indicates poor feeding practices: almost half of infants aged < 6 months were given water, while each of the protein-rich foods was consumed by fewer than half of children aged > 6 months. Low consumption of protein-rich foods is also apparent from the data on household consumption. Labor supply rates are similar for males and females: over 80% have at least one paid job, while around 9% had an additional job, and work on average around 25 hours weekly. Finally, child health in this setting is very poor: the average child has a height-for-age z-score that is below -2 std deviations of the WHO benchmark (and thus is considered to be stunted); and the incidence of illness is relatively high.

5.5 Results

We first show the impacts on all six composite indices: pooled across waves in Table 5.2, and separated by wave in Table 5.3. The subsequent tables (Tables 5.4-5.9) display the

[18] While this helps to limit the problem of multiple inference, it does not address it fully because we still use 8 indices. Indeed, if the data on the 8 indices were independent, the Family Wise Error Rate would be at 40%. Adjusting for multiple inference within domains but not across domains is the most commonly used option (see for instance, Finkelstein et al. 2012)

impacts on the sub-components of each index for those indices which show an overall statistically significant effect.[19] Note that for ease of reading, each of Tables 5.4-5.9 reproduces, in its first column, the summary index from Table 5.2.

5.5.1 Overall Findings

Table 5.2 displays intervention impacts on all six composite indices, as described in section 5.4.2. For child level outcomes, we estimate the impacts on children born after the intervention began in July 2005, as these are the ones whose mothers were eligible to be visited by the peer counselor. This means that we consider impacts for children aged up to 4.5 years at the time of the second follow-up survey. Furthermore, since the intervention was ongoing at follow-up, we estimate impacts separately for children aged < 6 months (whose mothers were potentially being visited by the counselors at the time) and those aged > 6 months, and report impacts on health outcomes for the latter group only. For household and adult outcomes, we consider impacts on our entire sample, regardless of whether the household was directly exposed to the intervention; and of the household's fertility choices.

The key rationale underlying the intervention is that households are inefficient producers of child health because they do not have the correct knowledge. In other words, the child health production function that households optimize over is "distorted". In line with this, Column 1 of Table 5.2 reports that the intervention improved mothers' knowledge of child nutrition.[20] The effect is only significant at the 10% level, possibly due to the high intra-cluster correlation in this variable. These improvements in knowledge translated into improved child consumption for both children aged < 6 months and those aged > 6 months (columns 2 and 3 in Table 5.2).[21,22] The positive impacts on the latter group imply that benefits of the intervention were retained even once the peer counselor stopped visiting the household.

Though the intervention provides no monetary or in-kind resources, household food consumption could increase (see section 3). In line with this, column 4 of Table 5.2 shows that the intervention increases total household food consumption, measured using

[19] Tables E3 and E4 of Appendix E displays results for the sub-components of indices that do not show a statistically significant intervention impact.

[20] The knowledge index was constructed from questions designed in consultation with programme staff, and tailored to the content of the intervention. Though the questions were piloted, no formal validation exercise was conducted.

[21] Note child-specific consumption for children > 6 months is measured at second follow-up only.

[22] That the intervention improved both knowledge and child nutrition suggests that improving knowledge of the child health production function improves nutrition choices. One might want to test this mechanism directly using the intervention as an instrument for knowledge. Unfortunately, the intervention impact on knowledge is not sufficiently strong to allow us to do this without encountering a weak instrument problem.

Table 5.2: Main Results

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
	Main Respondent's Knowledge on Nutrition	**Child Food Consumption**		**Household Food Consumption**	**Labor Supply**		**Child Physical Growth**	**Child Morbidity (reversed)**
		< 6 months	**> 6 months**		**Adult Males**	**Adult Females**	**> 6 months**	**> 6 months**
T_c	0.167+	0.250*	0.143+	0.218*	0.262+	0.018	0.102*	-0.013
Standard Error	[0.085]	[0.098]	[0.074]	[0.082]	[0.131]	[0.165]	[0.036]	[0.102]
Wild Cluster Bootstrap p-value	{0.070}	{0.016}	{0.076}	{0.018}	{0.086}	{0.955}	{0.022}	{0.861}
Randomization Inference p-value	{0.065}	{0.028}	{0.099}	{0.037}	{0.062}	{0.903}	{0.035}	{0.920}
Observations	1512	151	1280	3200	3642	4138	2175	2356
R-squared	0.098	0.214	0.099	0.063	0.183	0.136	0.026	0.053
IntraCluster Correlation	0.169	0.041	0.085	0.087	0.146	0.140	0.021	0.150
Mean Control Areas	-0.040	-0.109	-0.054	-0.099	-0.135	-0.050	0.266	0.022

Notes to Table: Standard errors computed using the cluster-correlated Huber-White estimator are reported in square brackets, with clustering at the level of the cluster (at which treatment was assigned); wild cluster bootstrap-t p-values and randomization inference p-values in curly brackets. All regressions include controls for cluster-level average education and Chewa ethnicity, both measured in 2004, and dummies for month of interview. All regressions other than in column 4 include controls for age and age-squared. Outcome variables are summary indices of variables relating to that domain of outcomes. They are constructed as described in section 4.3. Higher values of the index in columns 9 and 10 indicate lower morbidity. The component variables for each index are outlined in Table E1 in the appendix. Sample of children includes all those born after the intervention began in July 2005, and were therefore aged 0-53 months at time of interview. Specific samples are as follows. Column 1: all households present in waves 1 and 2 with a female main respondent aged 15 years or more; column 2: all children at wave 2 aged <6 months (some components of food consumption for this group not measured at wave 1); column 3: all children at wave 2 aged 6-53 months (food consumption for this group not measured at wave 1); column 4: all households at waves 1 or 2; columns 5 (6): all adult males (females) aged 15-65 years at waves 1 or 2; columns 7, 8: all children at wave 1 or wave 2 aged 6-53 months. Note small discrepanices in samples between columns 8 and 10 due to missing values of outcome indicators. ** p<0.01, * p<0.05, + p<0.1.

the composite index, at 5% significance. The increase in household consumption might have been partially funded by improvements in adult labor supply, particularly of males (column 5); female labor supply is unchanged by the intervention (column 6). Although our model of section 3 already indicated that labor supply could increase, other factors may also explain increased consumption, including borrowing and/or drawing down savings. Increases in labor supply could also be due to a reduction in time devoted to caring for sick children.

A key policy question is whether the observed adjustments on various margins of household behavior (increased consumption and labor supply) improved child health. Column 7 shows that these changes in behavior translate into improved child physical growth for children aged $>$ 6 months. No significant effect is found on child morbidity.[23] Note though that given the substantial infant mortality reductions found by Lewycka et al. 2013, and under the assumption that weaker children are the ones more likely to survive as a result of the intervention (Deaton 2007, Bozzoli et al. 2009), the reported effects likely underestimate the true effect of the intervention on child health.

Table 5.3 shows the results by follow-up survey round ('wave'), which are of interest in order to see whether the effects are sustained over time. In general, the table shows that the point estimates share the same signs across both waves, and are not significantly different from each other. Notably, the point estimates of household food consumption, male labor supply, and child physical growth all show a tendency to be larger in wave 2 than in wave 1, and they are statistically significant in wave 2 only, although they are not significantly different from the wave 1 estimates.[24] The tendency for larger treatment effects on consumption and male labor supply in wave 2 may be due to some heterogeneity of treatment effect according to the time when the surveys were conducted. Wave 1 data were collected between mid November and the end of March, while wave 2 data were collected between October and the end of December. The level of the consumption and male labor supply index are the lowest in the October to mid November period, which is when the treatment effect is the highest.

While the composite indices allow us to assess the general impact of the intervention on each domain, their magnitudes cannot be interpreted, as the weighting used to build the index distorts the scale. To shed more light on the magnitude of the effects, we next

[23] We also considered the intervention impacts on child anthropometrics and morbidity for children aged $<$ 6 months who were undergoing the intervention at the time of the survey, and for whom these would be intermediary stage data. We find a positive, but statistically insignificant effect on both outcomes. Interestingly, we find that the prevalence of diarrhea decreases for children $<$ 6 months, consistent with the reduced intake of water and non-maternal milk for this group.

[24] Note that there are more children aged $>$ 6 months who would have been eligible for the intervention in wave 2 than wave 1 since the former includes children born between July 2005 and July 2009 while the latter includes children born between July 2005 and October 2008.

Table 5.3: Results by Wave

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
	Top Panel: Wave 1 Results							
	Main Respondent's Knowledge on Nutrition	**Child Food Consumption**		**Household Food Consumption**	**Labor Supply**		**Child Physical Growth**	**Child Morbidity (reversed)**
		< 6 months	**> 6 months**		**Adult Males**	**Adult Females**	**> 6 months**	**> 6 months**
T_c	0.195			0.156	0.183	-0.016	0.093	0.027
Standard Error	[0.136]			[0.113]	[0.135]	[0.163]	[0.045]	[0.103]
Wild Cluster Bootstrap p-value	{0.228}			{0.212}	{0.216}	{0.985}	{0.108}	{0.769}
Randomization Inference p-value	{0.143}			{0.206}	{0.244}	{0.920}	{0.107}	{0.853}
Observations	1512			1644	1790	2080	932	1061
R-squared	0.079			0.069	0.177	0.157	0.032	0.040
IntraCluster Correlation	0.156			0.141	0.140	0.183	0.026	0.175
Mean Control Areas	-0.054			-0.075	-0.119	-0.033	0.286	0.001
	Bottom Panel: Wave 2 Results							
	Main Respondent's Knowledge on Nutrition	**Child Food Consumption**		**Household Food Consumption**	**Labor Supply**		**Child Physical Growth**	**Child Morbidity (reversed)**
		< 6 months	**> 6 months**		**Adult Males**	**Adult Females**	**> 6 months**	**> 6 months**
T_c	0.152	0.250*	0.143+	0.305**	0.323*	0.050	0.112*	-0.051
Standard Error	[0.119]	[0.098]	[0.074]	[0.092]	[0.148]	[0.193]	[0.040]	[0.124]
Wild Cluster Bootstrap p-value	{0.273}	{0.016}	{0.076}	{0.002}	{0.036}	{0.877}	{0.022}	{0.743}
Randomization Inference p-value	{0.248}	{0.028}	{0.099}	{0.014}	{0.036}	{0.768}	{0.032}	{0.746}
Observations	1512	151	1280	1556	1852	2058	1243	1295
R-squared	0.043	0.214	0.10	0.050	0.184	0.125	0.028	0.043
IntraCluster Correlation	0.190	0.041	0.085	0.085	0.192	0.150	0.017	0.197
Mean Control Areas	-0.035	-0.109	-0.0541	-0.132	-0.148	-0.073	0.238	0.045

Notes to Table: Standard errors computed using the cluster-correlated Huber-White estimator are reported in square brackets, with clustering at the level of the cluster (at which treatment was assigned); wild cluster bootstrap-t p-values and randomization inference p-values in curly brackets. All regressions include controls for cluster-level average education and Chewa ethnicity, both measured in 2004, and dummies for month of interview. All regressions other than in column 4 include controls for age and age-squared. Those in Cols 5 and 6 also control for education. Outcome variables are summary indices of variables relating to that domain of outcomes. They are constructed as described in section 4.4. Higher values of the index in column 8 indicates lower morbidity. The component variables for each index are outlined in Table E1 in the appendix. Sample of children includes all those born after the intervention began in July 2005, and were therefore aged 0-53 months at time of interview. Specific samples are as follows. Column 1, both panels: all households present in both waves 1 and 2 with a female main respondent aged 15 years or more; column 2 bottom panel: all children at wave 2 aged <6 months; column 3, bottom panel: all children at wave 2 aged 6-53 months (food consumption for this group not measured at wave 1); column 4, top (bottom) panel: all households at wave 1 (2); column 5, top (bottom) panel: all adult males aged 15-65 years at wave 1 (2); column 6, top (bottom) panel: all adult females aged 15-65 years at wave 1 (2) columns 7, 8: top (bottom) panel: all children at wave 1 (2) aged 6-44 months (6-53 months). Note small discrepanices in samples between columns 7 and 8 due to missing values of outcome indicators. Knowledge index in wave 1 constructed with 3 questions asked in wave 1; and that in wave 2 with 4 questions asked in wave 2 only. ** p<0.01, * p<0.05, + p<0.1.

report and discuss findings for individual outcomes for the composite indices for which there is a statistically significant effect of the intervention. We note that the results on the index components must be considered exploratory and interpreted carefully since the Family Wise Error Rate is not being controlled for.

5.5.2 Nutritional Knowledge, Consumption and Labor Supply

The intervention resulted in improvements in the main respondent's knowledge of child nutrition. The index aggregates together the correct responses to 7 questions (reproduced in Appendix 5.8.6). Columns 2-8 of Table 5.4 report the impact of the intervention in terms of the proportion of respondents who correctly answered each of the 7 questions. The results show that the knowledge improvements are concentrated on breastfeeding practices when infants are ill, and on knowledge of food preparation practices. We note that the intra-cluster correlation coefficient is very high for most components of the index, which makes it particularly difficult to detect statistically significant differences.[25]

Improvements in child consumption were detected both for children below and above 6 months. For the former group, we see from Table 5.5 that the improvement comes from a reduction in non-maternal milk. There is also a reduction (though not statistically significant) in the consumption of water. Table 5.6 shows that improvements for the latter group are driven by substantially higher consumption of protein-rich beans in the three days prior to the interview. The intakes of meat and eggs (also protein rich) are also positive, although not statistically significant, most likely due to the reduced sample size (child food intake was collected at second follow-up only). Overall, these results indicate that the intervention significantly affected the composition of child nutritional intake.

We saw from Table 5.2 that the intervention resulted in improvements in overall household food consumption. Columns 2 – 5 of Table 5.7 show that the improvement is due to an increase in the consumption of proteins, and of fruit and vegetables. The effects are relatively large. Focusing on proteins, which are particularly important for child growth as shown by for example, Puentes et al. (2014), we decompose the effect on the extensive (i.e. moving from consuming no proteins to some proteins) and intensive margin (calculations available upon request). Around 26% of households in

[25]Note that the number of observations is lower than for other household level variables. This is because we combine wave 1 and wave 2 questions into a single index, to maximize its informational content, and drop households without a female main respondent aged 15 years or above. Note that the three questions in wave 1 are a subset of the seven questions asked in wave 2. We construct the index to include responses from wave 1 to the three common questions and the responses to the four questions unique to wave 2. This is because there was evidence of households having learnt or found out answers to the three questions carried over from wave 1 to wave 2.

Table 5.4: Knowledge

	Summary Index	Breastfeeding when infant has diarrhoea	Are biscuits or groundnuts/soy a more nutritious for kids aged 6 months-3 yrs?	From what age should solid foods be given infants?	How should an HIV positive woman feed her baby?	Is nsima or porridge more nutritious for an infant aged > 6 months?	What is the best way of cooking fish with porridge for an infant aged > 6 months?	Should eggs be given to an infant aged > 9 months?
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
T_c	0.167+	0.253+	-0.052	0.037	0.138	-0.101	0.067**	0.104
Standard Error	[0.085]	[0.115]	[0.041]	[0.026]	[0.150]	[0.078]	[0.019]	[0.069]
Wild Cluster Bootstrap p-value	{0.070}	{0.084}	{0.290}	{0.166}	{0.444}	{0.210}	{0.002}	{0.186}
Randomization Inference p-value	{0.065}	{0.028}	{0.222}	{0.292}	{0.399}	{0.179}	{0.008}	{0.192}
Observations	1512	1512	1512	1512	1512	1512	1512	1512
R-squared	0.10	0.10	0.05	0.04	0.04	0.07	0.04	0.02
IntraCluster Correlation	0.169	0.277	0.082	0.049	0.408	0.183	0.057	0.107
Mean, Control	-0.040	0.217	0.938	0.88	0.393	0.857	0.026	0.719

Notes to Table: All regressions include controls for age, age-squared, average cluster-level education and Chewa ethnicity, both measured in 2004, and dummies for the month of interview. Standard errors computed using the cluster-correlated Huber-White estimator are reported in square brackets, with clustering at the level of the the cluster (at which treatment was assigned); wild cluster bootstrap-t and randomization inference p-values in curly brackets. Sample includes all households with a female main respondent, at wave 2. "Summary Index" aggregates the measures in columns 2-8 using the method described in section 4.3. The variables in columns 2-8 are dummy variables equal to 1 if the respondent answered correctly. Questions in columns 2-6 and column 8 were multiple choice questions where respondents chose 1 correct answer from 3-5 options. Question in column 7 was an open-ended question, with interviewers marking correctly answered options. ** p<0.01, * p<0.05, + p<0.1.

Table 5.5: Child Food Intake, <6 months

	[1]	[2]	[3]
	Summary Index	**Water**	**Milk other than maternal**
T_c	0.250*	-0.107	-0.082*
Standard Error	[0.098]	[0.069]	[0.034]
Wild Cluster Bootstrap p-value	{0.016}	{0.122}	{0.012}
Randomization Inference p-value	{0.028}	{0.212}	{0.115}
Observations	151	151	151
R-squared	0.214	0.362	0.087
IntraCluster Correlation	0.0405	0.000	0.060
Mean, Control	-0.109	0.474	0.101

Notes to Table: All regressions include controls for age, age-squared, gender, average cluster-level education and Chewa ethnicity, both measured in 2004, and dummies for the month of interview. Standard errors computed using the cluster-correlated Huber-White estimator are reported in square brackets, with clustering at the level of the the cluster (at which treatment was assigned); wild cluster bootstrap-t and randomization inference p-values in curly brackets. Sample includes children at wave 2 aged less than 6 months. "Summary Index" aggregates the measures in columns 1-2 using the method described in section 4.3. "Water" is an indicator for whether the child had any water in the 3 days prior to the survey, "Milk other than maternal" is an indicator (measured in second follow up only) for whether the child had milk other than breastmilk in the 3 days prior to the survey. ** p<0.01, * p<0.05, + p<0.1.

control clusters report consuming no protein-rich foods in the 7 days prior to interview; hence there is clear potential for improvement in the extensive margin. Indeed, the extensive margin accounts for one third of the consumption increase.[26] The increase in the intensive margin corresponds to 210 grams of meat/poultry extra and 640 grams beans extra per child per month. To put these quantities in perspective, a toddler will usually consume 50 grams of beans in one portion, together with some vegetables and carbohydrates.

A number of factors are likely to explain this substantial increase in food consumption: first, the time span of the intervention is sufficiently long (it had already been up and running for over 3.5 years by the time consumption was first measured); second, the intervention was intensive, involving up to 5 one-to-one home visits; third, as seen from the labor supply results in Table 5.2, there was scope for labor supply to increase, and thereby fund at least some of the increased consumption.

Table 5.2 also showed that the male labor supply index increased as a result of the intervention. Looking at the sub-components of the index - probability of any

[26] The consumption increase coming from the extensive margin is calculated under the assumption that the households in the treated clusters induced to consume protein-rich foods as a result of the intervention all consume proteins equivalent to the average consumed by control cluster households with non-zero protein consumption. The increase on the intensive margin – corresponding to the rest of the consumption increase – is further decomposed into food quantities (beans and meat/poultry) under the assumption that the entire amount is consumed by children aged < 12 years only (who are, in control clusters, 2.4 per household on average), and households pay prices equivalent to the average cluster-level median unit values.

Table 5.6: Child Food Intake, >6 months

	Summary Index	Any beans	Any meat	Any fish	Any eggs	Any vegetables	Any fruit	Any nsima	Any porridge
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
T_c	0.143+	0.225**	0.089	0.006	0.025	-0.010	-0.009	0.025	0.096
Standard Error	[0.074]	[0.056]	[0.095]	[0.099]	[0.052]	[0.020]	[0.058]	[0.015]	[0.064]
Wild Cluster Bootstrap p-value	{0.076}	{0.002}	{0.474}	{0.925}	{0.655}	{0.723}	{0.941}	{0.144}	{0.208}
Randomization Inference p-value	{0.099}	{0.007}	{0.289}	{0.954}	{0.632}	{0.634}	{0.895}	{0.140}	{0.251}
Observations	1280	1280	1280	1280	1280	1280	1280	1280	1280
R-squared	0.099	0.067	0.021	0.012	0.010	0.142	0.153	0.144	0.035
IntraCluster Correlation	0.085	0.116	0.084	0.112	0.048	0.018	0.093	0.000	0.136
Mean, Control	-0.054	0.258	0.290	0.462	0.163	0.959	0.699	0.929	0.800

Notes to Table: All regressions include controls for age, age-squared, gender, average cluster-level Chewa ethnicity and education, both measured in 2004, and dummies for the month of interview. Standard errors computed using the cluster-correlated Huber-White estimator are reported in square brackets, with clustering at the level of the the cluster (at which treatment was assigned); wild cluster bootstrap-t and randomization inference p-values in curly brackets. Sample contains all children at wave 2 aged 6-53 months (data on child solid intake collected at wave 2 only). "Summary Index" aggregates the measures in columns 2-9 using the method described in section 4.3. The variables in columns 2-9 are dummy variables equal to 1 if the corresponding food was consumed by the child in the 3 days prior to the survey. ** p<0.01, * p<0.05, + p<0.1.

Table 5.7: Household Consumption

	[1]	[2]	[3]	[4]	[5]
		Per Capita Monthly Food Consumption for:			
	Summary Index	**Cereals**	**Proteins**	**Fruit and Vegetables**	**Other Foods**
T_c	0.218*	-9.768	129.15+	269.987*	60.701
Standard Error	[0.082]	[52.432]	[54.802]	[108.591]	[33.552]
Wild Cluster Bootstrap p-value	{0.018}	{0.863}	{0.066}	{0.044}	{0.126}
Randomization Inference p-value	{0.030}	{0.865}	{0.025}	{0.033}	{0.069}
Observations	3200	3200	3200	3200	3200
R-squared	0.063	0.117	0.02	0.195	0.025
IntraCluster Correlation	0.087	0.074	0.042	0.173	0.053
Mean Control Areas	-0.10	605.80	349.10	679.80	149.50

Notes to Table: Standard errors computed using the cluster-correlated Huber-White estimator are reported in square brackets, with clustering at the level of the cluster (at which treatment was assigned); wild cluster bootstrap-t and randomization inference p-values in curly brackets. Sample includes all households at waves 1 or 2. All regressions include controls for age, age-squared, average cluster-level education and Chewa ethnicity, both measured in 2004, and dummies for the month of interview. Coefficients in columns 2-6 are in terms of Malawi Kwacha. (The average exchange rate to the US Dollar was approx. 140MK = 1 US$ at the time of the surveys). "Food Index" is an index of the food items in cols. 2-5, constructed as described in section 4.3. "Cereals" includes consumption of rice, maize flour and bread, "Proteins" includes consumption of milk, eggs, meat, fish and pulses, "Fruit and Vegetables" includes consumption of green maize, cassava, green leaves, tomatoes, onions, pumpkins, potatoes, bananas, masuku, mango, ground nuts and other fruits and vegetables, "Other Foods" includes cooking oil, sugar, salt, alcohol and other foods. ** p<0.01, * p<0.05, + p<0.1.

Table 5.8: Male Labor Supply

	Males			
	[1]	[2]	[3]	[4]
	Summary Index	Works	Has at least 2 jobs	Weekly Hours Worked
T_c	0.262+	0.106	0.080**	4.310
Standard Error	[0.131]	[0.080]	[0.025]	[2.918]
Wild Cluster Bootstrap p-value	{0.074}	{0.272}	{0.010}	{0.240}
Randomization Inference p-value	{0.062}	{0.220}	{0.011}	{0.202}
Observations	3642	3642	3642	3642
R-squared	0.183	0.18	0.06	0.16
IntraCluster Correlation	0.146	0.213	0.033	0.100
Mean, Control	-0.135	0.819	0.094	25.740

Notes to Table: All regressions include controls for age, age-squared, average cluster-level education and Chewa ethnicity, both measured in 2004, and dummies for the month of interview. Standard errors computed using the cluster-correlated Huber-White estimator are reported in square brackets, with clustering at the level of the the cluster (at which treatment was assigned; wild cluster bootstrap-t and randomization inference p-values in curly brackets. Sample includes all males aged 15-65 years at waves 1 or 2. "Summary Index" contains the variables in columns 2-4 and is computed using the method described in section 4.3. "Works" in an indicator of whether individual had an income-generating activity at the time of the survey, "Has at least 2 jobs" is an indicator for whether individual has 2 income generating activities, "Weekly Hours worked" give the total hours worked in the week prior to the survey on both income generating activities. ** p<0.01, * p<0.05, + p<0.1.

work, probability of having at least two jobs, and the number of hours worked - Table 5.8 reports positive effects of the intervention on all three, though only statistically significant for the probability of having at least two jobs. However, the intra-cluster correlation for the number of hours worked is much higher than for the probability of having at least two jobs (0.10 vs. 0.036), which greatly reduces the power to detect a significant effect of the intervention on the former.

The finding that the intervention increases male labor supply is consistent with it being a margin with considerable scope for increase. Indeed, previous research in Malawi has shown that labor supply is upward sloping rather than fixed (Dimova et al. 2010; Goldberg 2016). In our data, only 12% of males in control clusters have a second

job, most of them in non-agricultural self-employment activities.[27] Moreover, there is considerable entry into and exit from secondary jobs: among those with (without) a secondary job at first follow-up, 33% (7%) have one by the time of the second follow-up, a year later. While an extensive literature has documented increases in labor supply in response to increases in uncertainty and income shocks in developing countries (Saha 1994, Kochar 1999, Rose 2001, Lamb 2003, Ito & Takashi 2009), this is the first paper to document labor supply responses to changes in the perceived child health production function.

Beyond the mechanism for the increase in labor supply indicated in section 5.3, important cultural features of Malawian society are likely to contribute to the increase in male, rather than female, labor supply. In particular, the main ethnic group in Mchinji - the Chewa - is a matrilocal and matrilineal group, where men usually move to their wives' villages on marriage, and wealth (predominantly land) is held by women and passed on through the matriline (Phiri 1983, Sear 2008). As a consequence, women have more power and authority than in patrilineal societies common across most of Africa and South Asia (Reniers 2003). Indicative of this empowerment, all three measures of labor supply - work participation, the likelihood of having two jobs and hours worked - are strikingly similar for males and females (last rows of Table 5.8 and Table 5.15).[28] Finally, mothers are generally the main caregivers of children. So the finding that male labor supply increases in response to the mother receiving information on child nutrition is in line with the cultural background, where females are relatively empowered.

5.5.3 Child Health

Table 5.2 documented improvements in child physical growth for children > 6 months. Looking at the sub-components of the physical growth index in Table 5.9, we see that the improvement in growth is due to an increase in the average height-for-age z-score by 0.27 of a standard deviation of the WHO norm.[29] This is an important increase, and corresponds in magnitude to 65% of the average effect size obtained with the direct provision of food in food-insecure populations (Bhutta et al. 2008). Interestingly, further analysis, documented in Table 5.17 of Appendix 5.8.5, indicates that the effects

[27] Over half of these second jobs involve employment in own/family business, a quarter involve work on the family farm, and the rest involve work as an employee in public/private sector (~20%) or on someone else's farm (<5%).

[28] This has been documented by others for the Malawian context including Goldberg 2016 and 2004 Malawi DHS Report (pages 34-36). In the matrilineal Khasi society (India), women and men also have similar labor supply profiles (Gneezy et al. 2009).

[29] As is common with anthropometric data from developing countries, the SD of the height-for-age z-score in our sample is larger than in the WHO Reference Population (in our case the SD is 1.5 instead of 1), and so this increase corresponds to a 18% of a SD increase using the SD for our sample.

on physical growth are much stronger for children aged 6-24 months.[30]

Clearly, we cannot disentangle whether the improvement in physical growth is due to the reduction in intake of liquids other than breast milk when the child was < 6 months, or to the improvement in child food intake after age 6 months, or a combination of both. Our key message is that households responded to the intervention by increasing consumption and working more, which is the first such finding in this literature.[31]

5.6 Alternative Explanations

We have argued, using the model of section 5.3, that consumption and labor supply will increase because the perceived productivity of child consumption (in terms of child health) increased as a result of the intervention. Here we consider 4 alternative explanations. First, we consider and rule out that the increases in adult labor supply are driven by improvements in adult health somehow generated by the intervention (Table 5.18 in Appendix 5.8.5). Second, parental investment in child nutrition could have increased as a result of decreased fertility caused by the intervention, potentially yielding an increase in child quality (Becker & Tomes 1976). The intervention could have reduced fertility by reducing infant mortality and consequently inducing households to demand fewer children; or through the family planning component of the intervention. Analysis of the intervention effects on family planning behavior and births to women in our sample (as reported in the Mai Mwana Health Surveillance System)[32] reveals very small and statistically insignificant effects, ruling out this channel (Table 5.19 in Appendix 5.8.5).[33]

Third, the reduction in infant mortality and improvement in child health could have affected parental labor supply, through changing the demand for childcare. It is plausible that if infant mortality declines and there are more surviving children, mothers in treated clusters may increase their time devoted to childcare, therefore working less,

[30] These patterns are consistent with two non-competing explanations: that the intervention did not work very well at the beginning and/or children in control clusters experienced catch-up growth at slightly older ages.

[31] We have also examined the heterogeneity of the effect of the intervention on the anthropometric and morbidity indices according to whether the mother has had more than one child since the intervention started. The interaction terms were far from statistically significant (p-value of 0.45 or larger).

[32] The MaiMwana Health Surveillance System interviews the mothers of all children born in the 24 clusters since 2005 at 1 month and 7 months of age, and thus provides a more complete picture of births in the study areas than cross-sectional surveys.

[33] Because the intervention decreased infant mortality, an alternative explanation for our findings is that the children who survive (a) tend to have worse health and (b) parents compensate for the worse health by providing them with more resources. Based on the results of Lewycka et al. (2013), we estimate that the marginal surviving children would be approximately 2.3% of the intervention sample, which is too small to explain the magnitude of the treatment effects if these were to be driven entirely by these marginal children.

Table 5.9: Child Anthropometrics

	[1]	[2]	[3]	[4]
	Summary Index	**Height for Age**	**Healthy weight for age**	**Healthy weight for height**
T_c	0.102*	0.274*	0.028	0.042+
Standard Error	[0.036]	[0.100]	[0.017]	[0.024]
Wild Cluster Bootstrap p-value	{0.022}	{0.022}	{0.120}	{0.132}
Randomization Inference p-value	{0.035}	{0.055}	{0.308}	{0.147}
Observations	2175	2175	2175	2175
R-squared	0.026	0.048	0.02	0.029
IntraCluster Correlation	0.021	0.023	0.018	0.014
Average, Control	0.266	-2.326	0.829	0.852

Notes to Table: Standard errors computed using the cluster-correlated Huber-White estimator are reported in square brackets, with clustering at the level of the the cluster (at which treatment was assigned); wild cluster bootstrap-t p-values in curly brackets. All regressions include controls for age, age-squared, gender, dummies for the month of interview and average cluster-level education and Chewa ethnicity, both measured in 2004. Sample includes children aged 6-53 months at waves 1 or 2. "Summary Index" contains the variables in columns 2-4 and is computed using the method described in section 4.3. "Height-for-Age"is a standardised z-score relative to the WHO reference population, "Healthy weight for age" is a dummy variable =1 if child's weight-for-age z-score is not more than 2 std deviations above or below the WHO reference population and "Healthy weight for height" is a dummy variable =1 if child's weight-for-height z-score is within 2 std deviations of the WHO reference population. ** p<0.01, * p<0.05, + p<0.1.

leading to fathers working more to compensate for this. However, as we showed in section 5.5.1, the intervention does not appear to have reduced female labor supply, suggesting that this mechanism is not at play in our context. Another potential channel through which labor supply may change as a result of improvements in children's health is through reducing the need for fathers to be at home to help take care of children, thus facilitating an increase in their labor supply

Finally, effects could also be driven by information provided by the intervention on issues other than infant feeding practices, e.g. vaccination of infants, promotion of HIV testing and hygiene practices. Though these could have improved child health, it is unlikely that they would improve household consumption and labor supply. Available evidence suggests that these other components would have had very modest or no effects. Lewycka et al. (2013) find mixed intervention effects on vaccination rates (BCG vaccination rates increased, while polio vaccination rates decreased). Moreover, vaccination rates in control clusters were high, leading to small intervention effects. Furthermore, they find that the intervention wasn't effective in improving antenatal HIV counseling and treatment. This is not surprising, since the intervention simply encouraged women to get tested for HIV, without any efforts to alleviate cost constraints or stigma effects related to being tested (Thornton 2008; Ngatia 2012; Derksen et al. 2014). Finally, our finding that the intervention did not reduce the prevalence of diarrhea for children aged between 6 and 53 months and adults (Tables 5.16 and 5.18) suggests that the component on hygiene information probably had limited success.

5.7 Conclusion

In this paper, we use exogenous variation in mothers' knowledge of the child health production function induced by a cluster randomized intervention in Malawi, to study empirically whether improving knowledge of the child health production function influences a broad range of household behaviors.

We first document that the intervention improved mothers' knowledge of nutrition. Using a simple theoretical model, we show that households should react to this improved knowledge by changing the composition of child food intake in favor of protein-rich foods, fruits and vegetables. The intervention could also increase household food consumption and adult labor supply, although the theoretical predictions are ultimately ambiguous. Our empirical results show that, indeed, both child's food intake and child nutritional status improved, and that ultimately both labor supply and food consumption increased.

We hypothesize that two issues might have contributed to the success of the in-

tervention. First, the provision of information was not merely a one-off event in the intervention areas, but a sustained activity, still in place, that serves to spread information and to remind households of the importance of child nutrition on an ongoing basis. This may also explain why households adjusted on non-health margins to adhere to advice provided by this nutrition intervention and may shed light on why some health information campaigns have been successful, while others have failed. Second, the main ethnic group in rural Malawi, the Chewa, is a matrilineal one, in which women are likely to have more bargaining power and authority within the household than women in patrilineal societies common in much of the rest of Africa and South Asia. This higher female empowerment might indicate that women are in a good position to implement the recommendations given by the counselors as well as to encourage fathers to work more. It is not clear whether such responses may emerge in other settings and we see this as an area worthy of further investigation.

5.8 Appendix

5.8.1 Attrition

We here address the potential concern that our results may be biased due to attrition between the baseline census (2004) and the two follow-up surveys (2008-09, 2009-10). Although attrition is related to observables (Table 5.10), the key is that it is the same in treatment and control (follow-up rates of 65% and 67% in intervention and control clusters respectively). Moreover we showed in Table 5.1 that both the sample drawn and the sample successfully interviewed are well-balanced along observed characteristics. However a concern might remain that attrition induced differences in unobserved variables, potentially biasing our findings.

In particular, our estimates on child physical growth (Table 5.9) could be biased upwards if households with worse health endowments were more likely to attrit from intervention than from control clusters. However, when we repeat the analysis in Table 5.9 for older children living in intervention clusters (born before July 2005, hence whose mothers were not eligible to receive the counselors' visits when they were young infants), we find that their health status is worse (though not significantly so) in intervention than in control clusters. This provides suggestive evidence that those who attrited from intervention clusters are, if anything, relatively healthier than those attriting from control clusters (results available upon request).

We also address the issue of attrition directly using a Heckman selection model (Heckman 1979). A first stage Probit model estimates the probability that a sampled woman (and therefore her household) was successfully interviewed in the follow-up

surveys as a function of the intervention and characteristics of the assigned interviewer at first follow-up (given that the majority of attrition occurred between baseline and first follow up). Estimates from the first stage yield an inverse-Mills Ratio, which enters as an additional regressor in the second stage - equation (5) augmented with the inverse Mills Ratio - thereby correcting for selection due to attrition.

The interviewer characteristics provide a source of exogenous variation in the first stage (see for instance Zabel 1998, Fitzgerald et al. 1998). Specifically, we use the number of children aged 0-3 in the interviewer's household and the size of the interviewer's plot of land, both of which proxy for the ease and intensity with which interviewers were able to track respondents. Individuals with young children may be more intrinsically motivated to take part in a study on child health, and/or they may know many other community members with young children; interviewers with a larger plot of land have a higher opportunity cost of time. Both of these factors turn out to be jointly strong predictors of whether or not a woman is interviewed (p-value of joint significance <0.01). A key identification assumption is that interviewer characteristics are uncorrelated with respondents' characteristics and outcomes. We believe this assumption to be reasonable in this context.[34]

Table 5.11 reports the estimates of the program effects for two outcomes, household consumption and main respondent's labor supply.[35] As can be seen, the selection corrected estimates (middle panel) are very close in magnitude to the OLS estimates reported earlier (repeated here in the top panel), thereby providing additional evidence that our results are not driven by attrition bias.

[34] A concern noted by Thomas et al. 2012 is that good interviewers may be assigned to the most difficult clusters. In our case this concern is not relevant due to the process through which interviewers were allocated to clusters. Clusters were paired so as to include an intervention and a control cluster in the pairing. Among potential interviewers residing in either of the two clusters, the best was selected as an interviewer to cover the pair of clusters (and hence the interviewer was not allocated to the area from a central pool). The fact that there was just 1 interviewer per pair of clusters makes it very unlikely that chosen interviewers were representative of the population of the cluster.

[35] The baseline census does not include information on men or individual children, so we do not know who attrited.

Table 5.10: Differences in characteristics between those who attrited and those who did not

	Non-attrited	Difference Attrited - Not Attrited	p-value
Woman's Characteristics in 2004			
Married (dv = 1)	0.646	-0.112	0.004**
Some Primary Schooling or Higher	0.704	0.053	0.068+
Some Secondary Schooling or Higher	0.055	0.042	0.001**
Age (years)	25.169	-1.904	0.002**
Chewa	0.934	-0.021	0.118
Christian	0.982	-0.008	0.184
Farmer	0.661	-0.104	0.002**
Student	0.213	0.087	0.002**
Small Business/Rural Artisan	0.050	0.005	0.555
Age less than 16 in 2004	0.142	0.068	0.000**
Household Characteristics in 2004			
Agricultural household	0.996	-0.010	0.088+
Main Flooring Material: Dirt, sand or dung	0.910	-0.046	0.001**
Main roofing Material: Natural Material	0.859	-0.044	0.062+
HH Members Work on Own Agricultural Land	0.925	-0.032	0.048+
Piped water	0.026	0.014	0.106
Traditional pit toilet (dv = 1)	0.818	-0.053	0.046*
# of hh members	5.837	-0.090	0.468
# of sleeping rooms	2.215	0.002	0.943
HH has electricity	0.004	0.002	0.651
HH has radio	0.646	-0.003	0.833
HH has bicycle	0.511	0.014	0.583
HH has motorcycle	0.006	0.006	0.210
HH has car	0.006	-0.002	0.330
HH has paraffin lamp	0.947	-0.016	0.044**
HH has oxcart	0.048	0.007	0.472
N	1594	902	

Notes to Table: + indicates significant at the 10% level, * indicates significant at the 5% level. p-values reported are computed using the wild cluster bootstrap-t procedure as in Cameron *et al.* 2008, explained in section 4.1. Non-attrited refers to women (and their households) actually interviewed in 2008-09 (and used in the analysis). Attrited refers to women (and their households) drawn to be part of the sample in 2008-09, but who were not interviewed.

Table 5.11: Heckman Selection Equation Results

	[1]	[2]
	Food Index	**Main Respondent Labor Supply**
Ordinary Least Squares		
T_z	0.218*	-0.077
Standard Error	[0.082]	[0.187]
Wild Cluster Bootstrap p-value	{0.018}	{0.769}
Randomisation Inference p-value	{0.037}	{0.659}
Observations	3200	2938
R-squared	0.063	0.088
IntraCluster Correlation	0.087	0.165
Mean Control Areas	-0.10	-0.03
Heckman Selection Model for Attrition		
T_z	0.216*	-0.096
Standard Error	[0.108]	[0.234]
Inverse Mills ratio	-0.683	-0.700
	[0.463]	[0.866]
Selection Equation (coefficients)		
T_z	-0.08	-0.061
	[0.141]	[0.141]
# children 0-3	0.221*	0.252**
	[0.092]	[0.090]
land size (acres)	-0.017	-0.015
	[0.014]	[0.015]
Observations	4986	4621

Notes to Table: Standard errors computed using the cluster-correlated Huber-White estimator are reported in brackets, with clustering at the level of the cluster (at which treatment was assigned); wild cluster bootstrap-t p-values in curly brackets. Standard errors for Heckman Selection model computed using a block bootstrap method. Regressions include controls for dummies for the month of interview and cluster-level education and Chewa ethnicity in 2004. Column 2 regression includes controls for age and age-squared. Sample in column 1, upper panel, includes all households at waves 1 or 2; sample in column 2, upper panel, includes all main respondents aged 15-65 in waves 1 or 2. Sample in column 1, lower panel, includes all households of women drawn to be surveyed in wave 1 or 2 regardless of whether surveyed; sample in column 2, lower panel, includes all women drawn to be surveyed in wave 1 or 2 regardless of whether surveyed. Households/women who attrited between the baseline and wave 1 are considered to have attrited in wave 2 as well. Excluded variables in the second stage of the Heckman Selection Model are "# children 0-3" (number of children of interviewer aged 0-3 at wave 1) and "land size(acres)" (interviewer's land size in acres at wave 1). ** p<0.01, * p<0.05, + p<0.1.

5.8.2 Proofs

Proof of Proposition 1

The optimization problem that the household solves is

$$\underset{\{A,L,C_1,C_2\}}{Max} A^{\alpha} L^{\beta} C_1^{\gamma_1} C_2^{\gamma_2}$$

$$s.t.: \quad A + p_1 C_1 + p_2 C_2 \leq w(T - L)$$

Given that the objective function is increasing in each argument, the budget constraint will be binding at the optimum. We use the budget constraint to solve for A and substitute in the objective function to obtain:

$$\underset{\{L,C_1,C_2\}}{Max} F(L, C_1, C_2)$$

where $F(L, C_1, C_2) \equiv (w(T - L) - p_1 C_1 - p_2 C_2)^{\alpha} L^{\beta} C_1^{\gamma\delta} C_2^{\gamma\theta}$. The first order conditions are:

$$F_{C_1}(L, C_1, C_2) \equiv -\frac{\alpha p_1}{w(T - L) - p_1 C_1 - p_2 C_2} + \frac{\gamma_1}{C_1} = 0 \tag{e}$$

$$F_{C_2}(L, C_1, C_2) \equiv -\frac{\alpha p_2}{w(T - L) - p_1 C_1 - p_2 C_2} + \frac{\gamma_2}{C_2} = 0 \tag{f}$$

$$F_L(L, C_1, C_2) \equiv -\frac{\alpha w}{w(T - L) - p_1 C_1 - p_2 C_2} + \frac{\beta}{L} = 0. \tag{g}$$

It will be useful to use how the different cross-derivatives relate to F_{LC_1}:

$$F_{c_1 c_2} = F_{LC_1} \frac{p_2}{w}, \tag{h}$$

$$F_{c_2 c_2} = F_{LC_1} \frac{p_2^2}{w p_1} - \frac{\gamma_2}{c_2^2}, \tag{i}$$

$$F_{c_2 L} = F_{LC_1} \frac{p_2}{p_1}, \tag{j}$$

$$F_{LL} = F_{LC_1} \frac{w}{p_1} - \frac{\beta}{L^2}. \tag{k}$$

Differentiating the first order conditions (e)-(g) with respect to γ_1, we get:

$$\begin{bmatrix} F_{c_1c_1} & F_{c_1c_2} & F_{c_1L} \\ F_{c_1c_2} & F_{c_2c_2} & F_{c_2L} \\ F_{c_1L} & F_{c_2L} & F_{LL} \end{bmatrix} \begin{bmatrix} dC_1 \\ dC_2 \\ dL \end{bmatrix} = - \begin{bmatrix} F_{c_1\gamma_1} \\ F_{c_2\gamma_1} \\ F_{L\gamma_1} \end{bmatrix} d\gamma_1,$$

where $F_{c_2\gamma_1} = 0$ and $F_{L\gamma_1} = 0$. Using Cramer's rule, we obtain that

$$\frac{dC_1}{d\gamma_1} = -\frac{F_{c_1\gamma_1}\left(F_{c_2c_2}F_{LL} - F_{c_2L}^2\right)}{|H|},$$

$$\frac{dC_2}{d\gamma_1} = -\frac{F_{c_1\gamma_1}\left(F_{c_1c_2}F_{LL} - F_{c_1L}F_{c_2L}\right)}{|H|},$$

$$\frac{dL}{d\gamma_1} = -\frac{F_{c_1\gamma_1}\left(F_{c_1c_2}F_{c_2L} - F_{c_1L}F_{c_2c_2}\right)}{|H|},$$

where

$$|H| = \begin{vmatrix} F_{c_1c_1} & F_{c_1c_2} & F_{c_1L} \\ F_{c_1c_2} & F_{c_2c_2} & F_{c_2L} \\ F_{c_1L} & F_{c_2L} & F_{LL} \end{vmatrix}$$

Note that $F_{c_1\gamma_1} > 0$, and that the second order condition ensure both $\left(F_{c_2c_2}F_{LL} - F_{c_2L}^2\right) > 0$, $|H| < 0$. Hence, we get that $\frac{dC_1}{d\gamma_1} > 0$. Using (h)-(k), the above comparative statics can be simplified as:

$$\frac{dC_1}{d\gamma_1} = \frac{F_{c_1\gamma_1}(F_{c_1L}\left(\frac{\beta p_2^2}{L^2 w p_1} + \frac{\gamma_2 w}{C_2^2 p_1}\right) - \frac{\beta\gamma_2}{C_2^2 L^2})}{|H|} > 0, \tag{l}$$

$$\frac{dC_2}{d\gamma_1} = -\frac{F_{c_1\gamma_1}F_{c_1L}\left(\frac{\beta p_2}{wL^2}\right)}{|H|} < 0, \tag{m}$$

$$\frac{dL}{d\gamma_1} = -\frac{F_{c_1\gamma_1}F_{c_1L}\left(\frac{\gamma_2}{c_2^2}\right)}{|H|} < 0, \tag{n}$$

where we have used that $F_{c_1L} < 0$.

Using the budget constraint, we have that

$$\frac{dA}{d\gamma_1} = -w\frac{dL}{d\gamma_1} - p_1\frac{dC_1}{d\gamma_1} - p_2\frac{dC_2}{d\gamma_1}, \tag{o}$$

which simplifies to

$$\frac{dA}{d\gamma_1} = \frac{F_{c_1\gamma_1}\left(\frac{\gamma_2\beta p_1}{LC_2^2}\right)}{|H|} < 0,$$

after substituting (l)-(n) into (o).

Denote total consumption by $TC = A + p_1C_1 + p_2C_2$. Using the budget constraint, and (n), we can conclude that $\frac{dTC}{d\gamma_1} = -w\frac{dL}{d\gamma_1} > 0$

5.8.3 Monte Carlo Simulation

Standard errors based on cluster-correlated Huber-White standard errors might be too small when the number of clusters is relatively small (Wooldridge 2004, Bertrand et al. 2004, Donald & Lang 2007, and Cameron et al. 2008). This might lead to over-rejection of the null hypothesis that the coefficient of interest is zero when it is correct. To deal with this issue, in the paper we report p-values for the null hypothesis of no effect using the two leading approaches for valid inference in this case: wild cluster bootstrap-t (Cameron et al. 2008) and randomization inference (Fisher 1935, Rosenbaum 2002). Since there is limited evidence on when these approaches are valid (knowledge on the performance of the wild bootstrap-t is based on simulations from a dataset with features which may not match those of the data we use), we here provide the results of a Monte Carlo simulation to estimate the test size (the probability that the null hypothesis is rejected when it is true) for a nominal significance level of 5%. We next provide the details of the Monte Carlo simulation.

We analyze 8 Data Generating Processes (DGPs), one for each of the columns in Table 5.2. In each DGP, the sample and covariates are the ones that we use to estimate the regressions in Table 5.2. The parameters of the DGP (coefficients multiplying the covariates, variance of the error term and intra-cluster correlation) are also the ones that we obtain when we estimate the regressions in Table 5.2. Hence, the results from the Monte Carlo simulation are indeed informative about our case. For each column of Table 5.2, we follow the steps below:

Step 1: Use OLS to estimate regression (5.4) in which the dependent variable, Y_{ict}, and the sample are the ones indicated in the heading of the corresponding column in Table 5.2. The estimates, $[\hat{\alpha}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\mu}_t]$, which are the same as those reported in Table 5.2, are saved and used in the steps below (except $\hat{\beta}_1$, which is discarded). Using the residuals from this OLS regression, we estimate the intra-cluster correlation and the variance of the error term $[\hat{\rho}_u, \hat{\sigma}_u^2]$.

Step 2: Obtain 24 draws (our number of clusters) from a standardized normal distribution $\left\{\tilde{\theta}_c\right\}_{c=1}^{24}$.

Step 3: Obtain N draws (number of observations) from a standardized normal distribution, $\{\tilde{\varepsilon}_i\}_{i=1}^{N}$.

Step 4: Using the parameter values of step 1, and the random draw from step 2 and 3, $[\hat{\alpha}_0, \hat{\alpha}_2, \hat{\sigma}_\varepsilon^2]$, we obtain simulated values for the dependent variable, $\tilde{Y}_{ict}$, under the assumption that the treatment effect is null, that is,

$$\tilde{Y}_{ict} = \hat{\alpha}_0 + 0 * T_c + X_{ict}\hat{\beta}_2 + Z_{c0}\hat{\beta}_3 + \hat{\mu}_t + \hat{\sigma}_\theta\tilde{\theta}_c + \hat{\sigma}_\varepsilon\tilde{\varepsilon}_{ict}$$

where $\hat{\sigma}_u^2 = \hat{\sigma}_\theta^2 + \hat{\sigma}_\varepsilon^2$ and $\hat{\rho}_u = \dfrac{\hat{\sigma}_\theta^2}{\hat{\sigma}_\theta^2 + \hat{\sigma}_\varepsilon^2}$.

Step 5: We use OLS to estimate regression (5.4),

$$Y_{ict} = \alpha + \beta_1 T_c + X_{ict}\beta_2 + Z_{c0}\beta_3 + \mu_t + u_{ict}$$

using the simulated dependent variable calculated in step 4. We use three different methods for inference (cluster-correlated Huber-White standard errors, wild cluster bootstrap-t, randomization inference) to obtain three different p-values for the null hypothesis that β_1 is zero. Under each method, we reject the null hypothesis at 5% significance if its respective p-value < 0.05.

Step 6: Repeat steps 2-5 1000 times, keeping T_c, X_{ict}, Z_{c0} and the parameters from step 1, $[\hat{\alpha}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\mu}_t, \hat{\rho}_u, \hat{\sigma}_u^2]$ fixed. Hence, the only differences across repetitions are the random draws from steps 2 and 3, and hence the simulated values of the dependent variable, which are used in step 5.

For each method (cluster-correlated Huber-White standard errors, wild cluster bootstrap-t, randomization inference), the estimated test size, π (reported in Table 5.12) is the number of repetitions where the null hypothesis is rejected over 1000, the number of simulations. A 95% confidence interval for the estimated test size can be computed using the formula $\pi \pm 1.96 * \sqrt{0.05 * 0.95/1000}$, where 1.96 is the 97.5% standard normal critical value. In Table 5.12, we report whether the estimated test size is significantly different from the nominal one (0.05).

The first row shows the test size when we use cluster-correlated Huber-White standard errors to form the t-statistic. As expected, the test sizes are considerably larger than 0.05 and hence the test clearly over-rejects the null. Randomization inference provides test sizes that are generally statistically close to the nominal test size, and if anything slightly below it. The results of the wild-t bootstrap procedure are also quite close to the nominal size, but slightly above it for some cases (although not by much). Because one inference procedure yield test sizes slightly above the nominal size and the other one slightly below, it is reassuring that we obtain very similar p-values for the different outcome variables across Tables 5.2-5.9. These results are informative for

other researchers not only because it extends the characteristics of the Data Generating Processes in which these procedures are shown to work, but also because it compares side by side the two leading approaches for carrying out inference with a small number of clusters, which, to our knowledge has not been done so far.

5.8.4 Outcome Measures

In this appendix, we detail the measures for each of our outcomes of interest.

Child Consumption We collected information on child-specific intake of liquids and solid foods, focusing on diet variety. These are reported by the main respondent, who is the mother in the majority (92%) of cases. For children under the age of 2, there are three measures of liquid intake - whether or not (s)he had maternal milk, other milk, or water in the 3 days prior to the survey. In the second follow-up survey, there are also data on whether or not certain foods were consumed in the 3 days prior to the survey by all children aged less than 6 years. We use whether the children had any porridge, nsima,[36] meat, fish, eggs or beans, and fruit or vegetables.

Food Consumption We collected information at the household level on the quantities consumed and purchased of over 25 different food items in the week preceding the survey, and the amounts spent on them. In 2009-10, information was also collected on conversion factors from the most-frequented markets and trading centres, which are used to convert non-standard measurement units (such as a heap of tomatoes) into standard measurement units (such as kilograms).

Food consumption aggregates are computed by summing up food expenditures and adding on the values of non-purchased food. To impute the latter, we first use conversion factors to convert quantities measured in non-standard units to standard units, and then use median unit values to impute their value.[37] Finally, we obtain per-capita consumption values by dividing the relevant value by household size.

[36] Nsima is a thick paste made from maize flour and is a staple food in Malawi. Apart from being difficult to digest for infants, nsima does not contain all of the nutrients required by infants. MaiMwana recommends giving porridge to infants, ideally mixed with vegetables or protein, rather than nsima.

[37] These conversion factors from the second follow-up were applied to data from both waves. Median unit values are computed by dividing expenditure on a certain good by the quantity purchased, and taking the median at the cluster level. In the small number of cases where there were insufficient observations within a cluster to reliably compute the median, it was taken at the district level instead. This method of imputation is similar to that used by Attanasio et al. (2013). As a robustness check, we also valued consumption using the market prices rather than the median unit values. This is not our preferred method, since most households rarely purchase the foods they commonly consume from the markets. Reassuringly, though, both methods yield a food consumption share of total non-durable consumption of 0.86.

Table 5.12: Monte Carlo Simulations

Test size ↓	[1] Main Respondent's Knowledge on Nutrition	[2] Child Food Intake	[3]	[4] Household Food Consumption	[5] Labor Supply	[6]	[7] Child Physical Growth	[8] Child Morbidity (reversed)
		< 6 months	> 6 months		Adult Males	Adult Females	> 6 months	> 6 months
Method								
Huber-White Clustered Standard Errors	0.093*	0.088*	0.072*	0.078*	0.081*	0.085*	0.086*	0.084*
Wild Cluster Bootstrap-t	0.048	0.061	0.061	0.065*	0.055	0.047	0.070*	0.051
Randomization Inference	0.039	0.046	0.052	0.034*	0.05	0.041	0.047	0.047
Intra-Cluster Correlation in Data	0.169	0.041	0.033	0.087	0.146	0.140	0.020	0.150

Notes to Table: Table reports test sizes from Monte Carlo simulations conducted using the 3 different inference methods above. Simulations conducted according to the procedure described in Appendix B. Nominal test size for each simulation is set at 0.05. * Indicates statistically different test size from 0.05 at the 5% level of significance. [a] For the outcome "Improvements in Child Physical Growth" for children aged < 6 months, the intra-cluster correlation for the outcome variable once the effects of covariates are removed was 0 and thus we did not conduct the Monte Carlo simulations for this outcome. Higher values of the index in the last two columns indicate less morbidity.

Adult Labor Supply Labor supply is measured in three ways: whether or not an individual is engaged in an income-generating activity; whether or not an individual has a secondary income-generating activity; and the total number of hours worked in the week preceding the survey (number of days worked in the week preceding the survey multiplied by the number of hours worked per day; set to zero for those not working).

Child Health Both physical growth and morbidity are used as indicators of child health. Physical growth is measured by height and weight. For height, we use the standardized height-for-age z-score. Unlike height, weight is non-monotonic because both having too high a weight and too low a weight is unhealthy and hence undesirable. Hence, we use whether the child has a healthy weight for his/her age, and whether he/she has a healthy weight for his/her height. Healthy weight for his/her age occurs when the weight-for-age z-score is within -2 standard deviations +2 standard deviations from the WHO norm. Healthy weight-for-height is defined in an analogous way. Child morbidity is maternal-reported and includes the prevalence of diarrhea, fast breathing, fever, chills, and vomiting in the 15 days prior to the survey.

5.8.5 Additional Tables

Table 5.13: Outcome Measures for Each Domain

Domain	Outcome Measures Constituting Index
Nutrition knowledge	See exact questions in Appendix 5.8.6
Child Liquid Intake	Water intake in 3 days preceding survey; Intake of milk other than maternal in 3 days preceding survey
Child solid intake	Intake of any proteins in 3 days preceding survey; intake of any staples (nsima or porridge) in 3 days preceding survey; intake of any fruit and vegetables in 3 days preceding survey
Household Food Consumption	Amounts (in kwacha) of cereals, proteins, fruit and vegetables and other foods
Adult Labor Supply	Whether or not the individual works; whether or not the individual has 2 jobs; hours worked
Child Physical Growth	Height for age z-score; whether the child has a healthy weight for age z-score; whether the child has a healthy weight for height z-score
Child Morbidity	Whether or not the child did not suffer from diarrhoea; vomiting; fast breathing; fever; and chills in the 15 days preceding the survey
Adult Health	Whether or not the adult can walk 5 kms easily; whether or not the individual can carry a 20 kg load easily; ability to carry out daily activities; whether or not the individual suffered from diarrhoea; fever; cough; chills; and vomiting in 30 days preceding survey

Table 5.14: Descriptive Statistics on Outcome Variables, Control Clusters

	Pooled		Wave 1		Wave 2	
Outcome Variable	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
Nutrition Knowledge (correct answer=1)						
Knowledge Index	-0.040	0.434	n/a	n/a	n/a	n/a
Breastfeeding when infant has diarrhoea	0.216	0.412	0.216	0.412	n/a	n/a
Biscuits or groundnuts/soya more nutritious for kids aged 6-36 months?	0.938	0.242	0.938	0.242	n/a	n/a
Age when solid foods should be given	0.880	0.325	0.880	0.325	n/a	n/a
Feeding baby when woman is HIV positive	0.394	0.489	n/a	n/a	0.394	0.489
Is nsima or porridge more nutritious for infant aged > 6 months	0.858	0.349	n/a	n/a	0.858	0.349
Best way of cooking fish with porridge	0.140	0.348	n/a	n/a	0.140	0.348
Should eggs be given to an infant aged > 9 months?	0.718	0.450	n/a	n/a	0.718	0.450
Child Food Intake, < 6 months						
Index	0.010	0.824	n/a	n/a	0.010	0.824
Water	0.474	0.503	n/a	n/a	0.474	0.503
Non-maternal milk	0.103	0.305	n/a	n/a	0.103	0.305
Child Food Intake, > 6 months						
Index	-0.001	0.489	n/a	n/a	-0.001	0.489
Any beans	0.256	0.437	n/a	n/a	0.256	0.437
Any meat	0.289	0.453	n/a	n/a	0.289	0.453
Any fish	0.461	0.499	n/a	n/a	0.461	0.499
Any eggs	0.160	0.367	n/a	n/a	0.160	0.367
Any vegetables	0.958	0.200	n/a	n/a	0.958	0.200
Any fruit	0.699	0.459	n/a	n/a	0.699	0.459
Any nsima	0.929	0.257	n/a	n/a	0.929	0.257
Any porridge	0.799	0.401	n/a	n/a	0.799	0.401
Household Consumption						
Food Index	-0.098	0.654	-0.076	0.664	-0.132	0.670
Per capita cereal consumption (MK)	605.911	379.674	731.243	403.121	471.466	299.458
Per capita fruit and vegetable consumption (MK)	679.831	585.218	572.906	537.757	794.530	612.081
Per capita protein-rich food consumption (MK)	349.086	483.191	370.902	525.027	325.684	432.968
Per capita other foods consumption (MK)	149.492	495.483	164.119	225.059	133.801	156.341
Male Labor Supply						
Index	-0.065	0.723	-0.085	0.721	-0.044	0.727
Works (yes=1)	0.818	0.386	0.825	0.380	0.812	0.391
Works in two jobs (yes=1)	0.094	0.292	0.096	0.294	0.092	0.289
Hours worked	25.728	20.341	24.550	17.978	26.858	22.327
Female Labor Supply						
Index	-0.051	0.719	-0.067	0.729	-0.032	0.712
Works (yes=1)	0.846	0.361	0.827	0.378	0.866	0.341
Works in two jobs (yes=1)	0.086	0.280	0.098	0.297	0.074	0.261
Hours worked	24.449	17.409	23.692	16.895	25.213	17.889
Child Anthropometrics, > 6 months						
Index	0.287	0.525	0.254	0.522	0.311	0.528
Height for age z-score	-2.326	1.499	-2.339	1.500	-2.315	1.499
Healthy height for weight (yes=1)	0.852	0.355	0.859	0.348	0.847	0.360
Healthy weight (yes=1)	0.829	0.377	0.785	0.411	0.863	0.344
Child Morbidity, > 6 months						
Index (reversed)	0.000	0.591	0.001	0.594	-0.001	0.577
Suffered diarrhoea (yes=1)	0.253	0.435	0.354	0.479	0.164	0.370
Suffered from vomiting (yes=1)	0.207	0.405	0.237	0.426	0.181	0.385
Suffered from fast breathing (yes=1)	0.100	0.301	0.112	0.315	0.090	0.287
Suffered fever (yes=1)	0.507	0.500	0.551	0.498	0.469	0.499
Suffered from chills (yes=1)	0.146	0.353	0.155	0.363	0.138	0.345

Notes to Table: The table includes data on control clusters only. Sample for knowledge index includes households present in both waves of the survey, with a female main respondent aged 15 years or more; Sample of children aged > 6 months includes those born after July 2005 (when the intervention began), and who would have been aged at most around 53 months at wave 2; Sample for Household Consumption includes all households; Sample for male (female) labor supply includes males (females) aged 15-65. Child food consumption data collected in wave 2 only. Knowledge index constructed from wave 1 responses to 3 questions, and wave 2 responses to 4 questions asked in this wave only.

Table 5.15: Index Components for Female Labor Supply

	[1]	[2]	[3]	[4]
	Summary Index	**Works**	**Has at least 2 jobs**	**Weekly hours worked**
		Adult Females		
T_z	0.018	-0.03	0.040	-1.740
Standard Error	[0.165]	[0.104]	[0.023]	[3.308]
Wild Cluster Bootstrap p-value	{0.915}	{0.799}	{0.120}	{0.633}
Randomization Inference p-value	{0.903}	{0.742}	{0.101}	{0.585}
Observations	4138	4138	4138	4138
R-squared	0.136	0.144	0.045	0.149
IntraCluster Correlation	0.14	0.222	0.0265	0.144
Mean, Control	-0.05	0.847	0.0867	24.54

Notes to Table: All regressions include controls for age, age-squared, cluster-level education and Chewa ethnicity in 2004, and dummies for the month of interview. Standard errors computed using the cluster-correlated Huber-White estimator are reported in square brackets, with clustering at the level of the the cluster (at which treatment was assigned); wild cluster bootstrap-t and randomisation inference p-values in curly brackets. ** p<0.01, * p<0.05, + p<0.1. Sample includes all females aged 15-65 years. "Summary Index" contains the variables in columns 2-4 and is computed as described in section 4.4. "Works" is an indicator of whether individual had an income-generating activity at the time of the survey, "Has at least 2 jobs" is an indicator for whether individual had at least 2 income generating activities, "Weekly hours worked" give the total hours worked in the week prior to the survey on both income generating activities.

Table 5.16: Index Components for Child Morbidity

	[1]	[2]	[3]	[4]	[5]	[6]
	Summary Index	**Suffered Diarrhoea**	**Suffered from Vomiting**	**Suffered from Fast Breathing**	**Suffered Fever**	**Suffered from Chills**
			> 6 months			
T_z	-0.013	-0.009	-0.026	0.028	0.027	0.003
Standard Error	[0.102]	[0.040]	[0.047]	[0.057]	[0.063]	[0.050]
Wild Cluster Bootstrap p-value	{0.861}	{0.861}	{0.631}	{0.683}	{0.691}	{0.913}
Randomisation inference p-value	{0.920}	{0.838}	{0.705}	{0.682}	{0.742}	{0.962}
Observations	2356	2356	2356	2356	2356	2356
R-squared	0.053	0.118	0.018	0.027	0.015	0.013
IntraCluster Correlation	0.150	0.034	0.082	0.140	0.081	0.111
Mean, Control	0.022	0.253	0.208	0.101	0.507	0.147

Notes to Table: Standard errors computed using the cluster-correlated Huber-White estimator are reported in square brackets, with clustering at the level of the cluster (at which treatment was assigned); wild cluster bootstrap-t p-values in curly brackets. ** p<0.01, * p<0.05, + p<0.1. All regressions include controls for age, age-squared, gender, dummies for the month of interview and cluster-level education and Chewa ethnicity in 2004. Sample includes children born after July 2005 and who were aged between 6 and 53 months at time of survey. Each column represents a different dependent variable which takes value 1 if the the child has suffered the condition specified in the column heading in the 15 days previous to the survey, as reported by the main respondent, and is 0 otherwise.

Table 5.17: Effects on Physical Growth, Children Aged 6-53 months, by age groups

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
	Summary Index	**Height for Age**	**Healthy weight for age**	**Healthy weight for height**	**Summary Index**	**Height for Age**	**Healthy weight for age**	**Healthy weight for height**
	6-24 Months				**24-53 months**			
T_c	0.176**	0.554**	0.025	0.083+	0.050	0.087	0.032	0.019
Standard Error	[0.049]	[0.118]	[0.025]	[0.038]	[0.038]	[0.126]	[0.031]	[0.020]
Wild Cluster Bootstrap p-value	{0.008}	{0.000}	{0.372}	{0.060}	{0.296}	{0.492}	{0.316}	{0.500}
Randomization Inference p-value	{0.008}	{0.000}	{0.448}	{0.087}	{0.338}	{0.624}	{0.429}	{0.461}
Observations	952	952	952	952	1,223	1,223	1,223	1,223
R-squared	0.049	0.124	0.041	0.029	0.026	0.017	0.014	0.023
IntraCluster Correlation	0.038	0.0217	0.0092	0.038	0.0118	0.0404	0.03	0.000138
Average, Control	0.248	-2.306	0.845	0.795	0.275	-2.339	0.817	0.893

Notes to Table: Standard errors computed using the cluster-correlated Huber-White estimator are reported in brackets, with clustering at the level of the the cluster (at which treatment was assigned); wild cluster bootstrap-t p-values in curly brackets. ** p<0.01, * p<0.05, + p<0.1. All regressions include controls for age, age-squared, gender, dummies for the month of interview and cluster-level education and Chewa ethnicity in 2004. Sample in Cols 1-4 (5-8) includes children born after July 2005 and who were aged between 6 and 24 (24 and 53) months at time of measurement. "Summary Index" in Col. 1(5) contains the variables in columns 2-4 (6-8) and is computed using the method described in section 4.3. "Height-for-Age"is a standardised z-score relative to the WHO reference population, "Healthy weight for age" is a dummy variable =1 if child's weight-for-age z-score is not more than 2 std deviations above or below the WHO reference population and "Healthy weight for height" is a dummy variable =1 if child's weight-for-height z-score is within 2 std deviations of the WHO reference population.

Table 5.18: Intervention Effects on Adult Health

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
	Summary Index	**Walk 5 kms Easily**	**Carry a 20 kg Load Easily**	**Unable to Carry Out Daily Activities**	**Suffered Diarrhoea**	**Suffered Fever**	**Suffered from Cough**	**Suffered from Chills**	**Suffered from Vomiting**
					Males				
T_z	-0.004	-0.042	0.025	0.066	0.005	0.071	0.019	0.018	0.012
Standard Error	[0.044]	[0.046]	[0.034]	[0.039]	[0.013]	[0.048]	[0.063]	[0.027]	[0.017]
Wild Cluster Bootstrap p-value	{0.873}	{0.442}	{0.486}	{0.162}	{0.711}	{0.302}	{0.791}	{0.565}	{0.521}
Randomisation Inference p-value	{0.899}	{0.445}	{0.503}	{0.180}	{0.738}	{0.204}	{0.778}	{0.632}	{0.594}
Observations	3760	3760	3760	3760	3760	3760	3760	3760	3760
R-squared	0.015	0.099	0.081	0.022	0.008	0.011	0.008	0.007	0.010
IntraCluster Correlation	0.074	0.120	0.0703	0.042	0.008	0.059	0.076	0.053	0.015
Mean, Control	0.004	0.904	0.918	0.287	0.052	0.236	0.269	0.086	0.099
					Females				
T_z	-0.019	-0.050	0.030	0.043	-0.004	0.090	0.025	0.014	0.005
Standard Error	[0.038]	[0.046]	[0.033]	[0.044]	[0.014]	[0.044]	[0.062]	[0.036]	[0.031]
Wild Cluster Bootstrap p-value	{0.693}	{0.352}	{0.386}	{0.382}	{0.831}	{0.118}	{0.739}	{0.731}	{0.869}
Randomisation Inference p-value	{0.685}	{0.353}	{0.480}	{0.419}	{0.839}	{0.103}	{0.722}	{0.782}	{0.913}
Observations	4252	4252	4252	4252	4252	4252	4252	4252	4252
R-squared	0.030	0.114	0.106	0.02	0.011	0.019	0.017	0.011	0.01
IntraCluster Correlation	0.060	0.110	0.067	0.042	0.010	0.048	0.085	0.077	0.048
Mean, Control	0.011	0.870	0.888	0.412	0.073	0.329	0.277	0.119	0.148

Notes to Table: All regressions include controls for age, age-squared, gender, dummies for the month of interview and cluster-level education and Chewa ethnicity in 2004. Standard errors computed using the cluster-correlated Huber-White estimator are reported in square brackets, with clustering at the level of the cluster (at which treatment was assigned); wild cluster bootstrap-t p-values in curly brackets. ** p<0.01, * p<0.05, + p<0.1. Each column represents a different dependent variable which takes the value 1 if the column heading is "Yes" according to the main respondent and 0 otherwise.

Table 5.19: Effects on Family Planning and Fertility

	[1]	[2]
	Use of any modern family planning method	**Number of children since intervention began**
T_c	0.023	-0.049
Standard Error	[0.052]	[0.040]
Wild Cluster Bootstrap p-value	{0.667}	{0.300}
Randomisation Inference p-value	{0.652}	{0.525}
Observations	2809	1655
R-squared	0.065	0.089
IntraCluster Correlation	0.036	0.014
Mean, Control	0.378	0.583

Notes to Table: Standard errors computed using the cluster-correlated Huber-White estimator are reported in square brackets, with clustering at the level of the cluster (at which treatment was assigned); wild cluster bootstrap-t p-values in curly brackets. All regressions includes controls for age, age-squared, and (family planning regression only) for cluster-level Chewa ethnicity and average cluster-level education, both measured in 2004, and dummies for the month of interview. "Number of children since July 2005" is the number of children born to the main respondent and surveyed at age 1 month since July 2005; Column 1 sample includes women 17-43 years old (when available, both waves responses are included). Sample in column 2 includes all main respondents in wave 1 linked to the Mai Mwana Health Surveillance System, which measures at age 1 month, all children born to these women since the start of the intervention. ** p<0.01, * p<0.05, + p<0.1.

5.8.6 Knowledge Questions

1. If an infant is being breastfed and suffers from diarrhoea, should the breastfeeding:
   (a) Continue as usual
   (b) Increase
   (c) Decrease
   (d) Stop and replace with another type of milk or liquid
   (e) Don't Know
2. Which of the following is most nutritious for infants between 6 months and 3 year?
   (a) Biscuits

(b) Groundnuts or soya

(c) They both have the same nutritional value

(d) Don't Know

3. When should you start to give some solid foods to the baby?

(a) From birth

(b) After 1 month old

(c) After 3 months old

(d) After 6 months old

(e) Don't Know

4. If a woman is HIV positive, how should she feed her baby?

(a) Exclusive breast feeding for 6 months, followed by early cessation

(b) Exclusive breast feeding for 6 months, followed by complementary feeding

(c) Complementary feeding from birth

(d) Don't Know

5. What is more nutritious for a child older than 6 months:

(a) Nsima

(b) Phala (porridge)

(c) Both are the same

6. Can you explain to me how best to cook fish with phala for a child older than 6 months (tick all those mentioned).

(a) Pound the fish

(b) Sieve the powder

(c) Add powder to flower/phala

(d) Use powder + flour to prepare phala

(e) None of the above

(f) Don't Know

7. Should eggs be given to an infant aged 9 months and above?

(a) Yes

(b) No

(c) Don't Know

Chapter 6

Conclusion and Future Work

This dissertation uses household micro-data, combined with economic theory, to study how informal insurance in extended family networks in developing countries varies with features of network structure; and the consequences of incorrect knowledge of the child health production function on health and non-health choices.

Chapter 2 provides an overview of methods to identify social effects – the effects of social networks on outcomes – in linear social effects models when networks data (detailed data on agents and exact interactions between them) is available. It first provides a common framework nesting the most widely used models in this class, and thereafter gives an overview of the theoretical models underlying each empirical specification. It then outlines methods to deal with one key source of endogeneity – network formation – including methods for specifying and estimating models of network formation. Networks are high-dimensional objects, which complicates this exercise. Thereafter, it tackles issues to do with measuring the network. It brings together literature from across many disciplines on the consequences of partial observation on the network on the accuracy of measured network statistics, and parameter estimates using these; and outlines methods proposed to deal with measurement error. This is a fast evolving literature, as methods are developed and adapted to analyse increasingly available detailed network data. The review also highlights areas for future work. These include developing network formation models that are feasible to compute, and in developing low cost ways of collecting accurate measures of network structure, including those that do not require a census of the network. Work by (Banerjee et al. 2016) makes some promising first steps.

Thereafter, Chapter 3 considers theoretically and empirically how risk sharing varies with the average number of socially close and distant connections in a household's network. Socially close connections are better able to enforce informal arrangements,

but may be more economically similar and hence offer fewer opportunities for risk sharing; thereby potentially generating a trade-off between these. Theoretically, when both enforcement and risk sharing opportunities are important for risk sharing, this trade-off generates a U-shaped (inverse U-shaped) relationship between risk sharing and the number of socially close and distant connections in a household's network. The chapter then studies this relationship empirically using data on within-village extended family networks. It documents that socially distant connections do indeed provide more opportunities for risk sharing in this context, and that these opportunities are particularly important for the effective functioning of within-village extended family network based insurance: networks with more socially distant connections provide more risk sharing. No relationship is found for socially close connections. This chapter provides some of the first evidence documenting that risk sharing opportunities vary with social distance, and highlights the importance of incorporating this variation in models of risk sharing in social networks. Moreover, the findings also raise the question of where this variation arises from. The chapter speculates on some reasons for this, including the presence of credit constraints and labour market imperfections preventing diversification across socially close households. A more complete analysis on exact drivers of this finding is left to future work. A second question raised by this chapter relates to the interaction between within- and outside- village extended family networks in risk sharing concerns. Risk sharing concerns are likely to influence location decisions for members of the same extended family network. Future work should consider this question, and consequently effects on overall risk sharing of the complete extended family network.

Chapter 4 studies the relationship between group size and informal risk sharing in rural Malawi, in a setting with imperfect enforcement and coalitional deviations. Building on (Genicot & Ray 2003), the chapter first shows that in such a setting, the relationship between risk sharing and group size is theoretically ambiguous. The question is empirically analysed using data from Malawi with information on sibship sizes. The chapter exploits a social norm among the largest ethnic group in the data – the Chewa – which indicates that a woman's brothers have responsibility for the wellbeing of her household to define the potential risk sharing group, and also construct a placebo test that alleviates concerns that estimates are biased by unobserved variables that might be correlated with risk sharing and group size. We find that households where the wife has many brothers are poorly insured against crop loss events. A calibration exercise indicates that the threat of coalitional deviations can explain the empirical findings. A natural question is whether such a relationship exists in other settings. This is left to future work.

Finally, Chapter 5 uses exogenous variation in mothers' knowledge of the child health production function, induced by a cluster randomised control trial in rural Malawi, to study whether improving knowledge influences health and non-health choices. The chapter uses a simple theoretical model to show that changing mothers' knowledge will influence households to change child consumption patterns towards foods it realises are more productive, such as proteins and fruits and vegetables. Household consumption and adult labour supply could increase, though the ultimate effect is ambiguous. Empirically, the chapter establishes that the intervention improved knowledge. In line with this, children's diets and nutritional status improved, as did household food consumption, and male labour supply. We hypothesise that two features of the context might have contributed to the success of the intervention: first, the provision of information was a continuous, rather than one-off, event within the community. Regular visits by counselors to different community members would have helped spread information, and also served as a reminder of the information, thereby making it more salient. Second, the main ethnic group in the study area – the Chewa – is a traditionally matrilineal group, in which women are more likely to have more bargaining power within the household, potentially making it easier for them to implement the information provided, and to encourage fathers to work more. Further work is needed on how intra-household dynamics influence households' responses to information interventions of the type studied in this chapter. This will undoubtedly help shed light on the likely success of such an intervention in other settings.

Bibliography

Alatas, V., Banerjee, A., Chandrasekhar, A. G., Hanna, R. & Olken, B. A. (2014), "'Network Structure and the Aggregation of Information: Theory and Evidence from Indonesia"', *mimeo, MIT* .

Alderman, H. (2007), '"Improving Nutrition through Community Growth Promotion: Longitudinal Study of the Nutrition and Early Childhood Development Program in Uganda"', *World Development* **35(8)**, 1376–1389.

Alderman, H., Behrman, J., Lavy, V. & Menon, R. (2001), '"Child Health and School Enrolment. A Longitudinal Analysis"', *The Journal of Human Resources* **36(1)**, 185–205.

Ali, N. & Miller, D. (2013), "'Enforcing Cooperation in Networked Societies"', *mimeo, University of Michigan* .

Ambrus, A., Mobius, M. & Sziedl, A. (2014), "'Consumption Risk Sharing in Social Networks"', *American Economic Review* **104(1)**, 149–82.

Anderson, M. (2008), '"Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects"', *Journal of the American Statistical Association* **103(484)**, 1484–1495.

Angelucci, M., De Giorgi, G., Rangel, M. & Rasul, I. (2009), "'Village Economies and the Structure of Extended Family Networks"', *The B.E. Journal of Economic Analysis and Policy* **9(1)**.

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

Angelucci, M., De Giorgi, G. & Rasul, I. (2015), "'Resource Pooling Within Family Networks: Insurance and Investment"', *mimeo, University College London* .

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

Attanasio, O., di Maro, V., Lechene, V. & Phillips, D. (2013), "'The effect of increases in food prices on consumption and welfare in rural Mexico"', *Journal of Development Economics* **104**, 136–151.

Attanasio, O. & Pavoni, N. (2011), "'Risk Sharing in Private Information Models with Asset Accumulation: Explaining the Excess Smoothness of Consumption"', *Econometrica* **79(4)**, 1027–1068.

Attanasio, O. & Rios-Rull, V. (2000), "'Consumption Smoothing in Island Economies: Can Public Insurance Reduce Welfare?"', *European Economic Review* **44(7)**, 1225–1258.

Attanasio, O. & Szekely, M. (2004), "'Wage Shocks and Consumption Variability in Mexico during the 1990s"', *Journal of Development Economics* **73(1)**, 1–25.

Azevedo, F. A., Carvalho, L. R., Grinberg, L. T., Farfel, J. M., Ferretti, R. E., Leite, R. E., Jacob Filho, W., Lent, R. & Herculano-Houzel, S. (2009), '"Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain"', *Journal of Computational Neurology* **513**, 532–541.

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

Baland, J.-M., Bonjean, I., Guirkinger, C. & Ziparo, R. (2015), "'The Economic Consequences of Mutual Help in Extended Families"', *mimeo, University of Namur* .

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

Bandiera, O., Burgess, R., Das, N., Gulesci, S., Rasul, I. & Sulaiman, M. (2015), "'The Misallocation of Labor in Village Economies"', *mimeo, University College London* .

Banerjee, A., Chandrasekhar, A., Duflo, E. & Jackson, M. (2013), '"The Diffusion of Microfinance"', *Science* **341**.

Banerjee, A., Chandrasekhar, A., Duflo, E. & Jackson, M. (2016), "'Gossip: Identifying Central Individuals in a Social Network"', *mimeo, MIT* .

Banerjee, A., Duflo, E. & Glennerster, R. (2008), "'Putting a band-aid on a corpse: incentives for nurses in the Indian public health care system"', *Journal of the European Economic Association* **5(2-3)**, 487–500.

Banerjee, A., Duflo, E., Postel-Vinay, G. & Watts, T. (2010), "'Long-Run Health Impacts of Income Shocks: Wine and Phylloxera in Nineteenth-Century France"', *Review of Economics and Statistics* **92(4)**, 714–728.

Barham, T. (2012), "'Enhancing Cognitive Functioning: Medium-Term Effects of a Health and Family Planning Program in Matlab"', *American Economic Journal: Applied Economics* **4(1)**, 245–273.

Becker, G. (1965), "'A Theory of the Allocation of Time"', *The Economic Journal* **75(299)**, 493–517.

Becker, G. & Tomes, N. (1976), "'Child Endowments and the Quantity and Quality of Children"', *Journal of Political Economy* **84**, S143–S162.

Beegle, K., Dehejia, R. & Gatti, R. (2006), '"Child Labor and Agricultural Shocks"', *Journal of Development Economics* **81**, 80–96.

Behrman, J. (1996), "'The Impact of Health and Nutrition on Education"', *The World Bank Research Observer* **11**(1), 23–37.

Behrman, J. R. & Rosenzweig, M. R. (2004), "'Returns to Birthweight"', *Review of Economics and Statistics* **86**(2), 586–601.

Bernheim, D., Peleg, B. & Whinston, M. (1987), '"Coalition-Proof Nash Equilibria I. Concepts"', *Journal of Economic Theory* **42**, 1–12.

Bertrand, M., Duflo, E. & Mullainathan, S. (2004), "'How Much Should We Trust Differences-in-Differences Estimates?"', *Quarterly Journal of Economics* **119(1)**, 249–275.

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

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

Besley, T. (1995), "'Non-Market Institutions for Credit and Risk-Sharing in Low-Income Countries"', *Journal of Economic Perspectives* **9(3)**, 115–127.

Bezner-Kerr, R., Berti, P. & Chirwa, M. (2007), "'Breastfeeding and mixed feeding practices in Malawi: timing, reasons, decision makers, and child health consequences"', *Food and Nutrition Bulletin* **28(1)**, 90–99.

Bhalotra, S., Karlsson, M. & Nilsson, T. (2016), "'Infant Health and Longetivity: Evidence from A Historical Intervention in Sweden"', *mimeo, University of Essex* .

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

Bhutta, Z. A., Ahmed, T., Black, R. E., Cousens, S., Dewey, K., Giugliani, E., Haider, B. & et al. (2008), "'What works? Interventions for maternal and child undernutrition and survival"', *The Lancet* **371(9610)**, 417–440.

Bianchi, M. & Bobba, M. (2013), "'Liquidity, Risk and Occupational Choices"', *Review of Economic Studies* **80(2)**, 491–511.

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

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

Blume, L. E., Brock, W. A., Durlauf, S. N. & Ioannides, Y. M. (2010), "Identification of Social Interactions", *in* J. Benhabib, A. Bisin & M. Jackson, eds, 'Handbook of Social Economics', Vol. 1B, North Holland.

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

Blundell, R., Chiappori, P.-A. & Meghir, C. (2005), "'Collective Labor Supply with Children"', *Journal of Political Economy* **113(6)**, 1277–1306.

Bold, T. (2009), '"Implications of Endogenous Group Formation for Efficient Risk Sharing"', *Economic Journal* **119(536)**, 562–591.

Bold, T. & Dercon, S. (2014), '"Insurance Companies of the Poor"', *CEPR Discussion Papers* **10278**.

Booij, A. S., Leuven, E. & Oosterbeek, H. (2015), Ability Peer Effects in University: Evidence from a Randomized Experiment, IZA Discussion Papers 8769, Institute for the Study of Labor (IZA).
URL: *https://ideas.repec.org/p/iza/izadps/dp8769.html*

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

Bozzoli, C., Deaton, A. & Quintana-Domeque, C. (2009), "'Adult Height and Childhood Disease"', *Demography* **46(4)**, 647–669.

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

Bramoullé, Y. & Kranton, R. (2007), '"Public Goods in Networks"', *Journal of Economic Theory* **135(1)**, 478–494.

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

Breza, E. & Chandrasekhar, A. (2015), "'Social Networks, Reputation and Commitment: Evidence from a Savings Monitors Field Experiment"', *mimeo, Columbia University* .

Brierova, L. & Duflo, E. (2004), "'The Impact of Education on Fertility and Child Mortality: Do Fathers Really Matter Less Than Mothers?"', *NBER Working Paper No. 10513* .

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

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

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

Cameron, C., Gelbach, J. & Miller, D. (2008), "'Bootstrap-Based Improvements for Inference with Clustered Errors"', *Review of Economics and Statistics* **90**, 414–427.

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

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

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

Chandrasekhar, A. G. & Lewis, R. (2011), '"Econometrics of Sampled Networks"', *mimeo, Massachusetts Institute of Technology* .

Chandrasekhar, A., Larreguy, H. & Kinnan, C. (2014), '"Social Networks as Contract Enforcement: Evidence from a lab experiment in the field"', *mimeo, Stanford University* .

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

Chaudhuri, A., Gangadharan, L. & Maitra, P. (2010), '"An Experimental Analysis of Group Size and Risk Sharing"', *Unpublished* .

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

Chiappori, P.-A. (1997), "'Introducing Household Production in Collective Models of Labor Supply"', *Journal of Political Economy* **105(1)**, 191–209.

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

Chou, S., Liu, J., Grossman, M. & Joyce, T. (2010), "'Parental Education and Child Health: Evidence from a Natural Experiment in Taiwan"', *American Economic Journal: Applied Economics* **2(1)**, 33–61.

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

Chuang, Y. & Schechter, L. (2014), "Social Networks In Developing Countries".

Clark, G. (2014), *"Surnames and the History of Social Mobility"*, Princeton University Press.

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

Comola, M. & Fafchamps, M. (2014), '"Estimating Mis-reporting in Dyadic Data: Are Transfers Mutually Beneficial?"', *mimeo, Paris School of Economics* **124**, 954–976.

Comola, M. & Fafchamps, M. (2015), '"The Missing Transfers: Estimating Misreporting in Dyadic Data"', *CEPR Discussion Paper No. DP10575* .

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

Conley, T. G. & Udry, C. R. (2010), '"Learning About a New Technology: Pineapple in Ghana"', *American Economic Review* **100**, 35–69.

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

Costenbader, E. & Valente, T. W. (2003), '"The stability of centrality measures when networks are sampled"', *Social Networks* **25**, 283–307.

Cruz, C., Labonne, J. & Querubin, P. (2015), "'Politician Family Networks and Political Outcomes: Evidence from the Phillipines"', *mimeo, Yale-NUS College* .

Cunha, F., Elo, I. & Culhane, J. (2013), "'Eliciting Maternal Expectations About the Technology of Cognitive Skill Formation"', *NBER Working Paper 19144* .

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

Currarini, S., Jackson, M. O. & Pin, P. (2010), '"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.

Currie, J. (2009), "'Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development"', *Journal of Economic Literature* **47(1)**, 87–122.

Currie, J. & Almond, D. (2011), Chapter 15 - human capital development before age five, Vol. 4, Part B of *Handbook of Labor Economics*, Elsevier, pp. 1315 – 1486.
URL: *http://www.sciencedirect.com/science/article/pii/S0169721811024130*

Currie, J. & Moretti, E. (2003), "'Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings"', *The Quarterly Journal of Economics* **118(4)**, 1495–1532.

Currie, J., Stabile, M., Manivong, P. & Roos, L. (2010), "'Child Health and Young Adult Outcomes"', *Journal of Human Resources* **45(3)**, 517–548.

Davis, B., Stecklov, G. & Winters, P. (2002), "'Domestic and international migration from rural Mexico: Disaggregating the effects of network structure and composition"', *Population Studies* **56**, 291–309.

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

de Onis, M., Frongillo, E. & Blossner, M. (2000), "'Is Malnutrition Declining? An Analysis of Changes in Levels of Child Malnutrition since 1980"', *Bulletin of the World Health Organization* **78(10)**, 1222–1233.

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

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

Deaton, A. (2007), "'Height, Health and Development"', *Proceedings of the National Academy of Sciences* **104(33)**, 13232–13237.

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

Del Boca, D., Flinn, C. & Wiswall, M. (2014), "'Household Choices and Child Development"', *The Review of Economic Studies* **81(1)**, 137–185.

Derksen, L., Muula, A. & Oosterhout, J. (2014), "'Love in the Time of HIV: Theory and Evidence on Social Stigma and Health Seeking Behavior"', *mimeo, University of Toronto* .

DeWeerdt, J., Genicot, G. & Mesnard, A. (2014), '"Asymmetry of Information within Family Networks"', *IZA Discussion Papers No. 8395* .

Dimova, R., Michaelowa, K. & Weber, A. (2010), "'Ganyu Labour in Malawi: Understanding Rural Households' Labour Supply Strategies"', *CIS Working Paper 52* .

Donald, S. & Lang, K. (2007), "'Inference with Differences-in-Differences and Other Panel Data"', *Review of Economics and Statistics* **89**, 221–233.

Dubois, P. (2006), '"Heterogeneity of Preferences, Limited Commitment and Coalitions: Empirical Evidence on the Limits to Risk Sharing in Rural Pakistan"', *Mimeo, Toulouse School of Economics* .

Dubois, P., Jullien, B. & Magnac, T. (2008), '"Formal and Informal Risk Sharing in LDCs: Theory and Empirical Evidence"', *Econometrica* **76(4)**, 679–725.

Ductor, L., Fafchamps, M., Goyal, S. & van der Leij, M. (2014), '"Social Networks and Research Output"', *Review of Economics and Statistics* **96(5)**, 936–948.

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

Dupas, P. (2011*a*), "'Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya"', *American Economic Journal: Applied Economics* **3(1)**, 1–34.

Dupas, P. (2011*b*), "'Health Behavior in Developing Countries"', *Annual Review of Economics* **3**, 425–449.

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

Erdős, P. & Rényi, A. (1959), '"On Random Graphs"', *Publicationes Mathematicae* **6**, 290–297.

Fafchamps, M. & Gubert, F. (2007), '"The Formation of Risk Sharing Networks"', *Journal of Development Economics* **83**, 326–350.

Fafchamps, M. & Lund, S. (2003), '"Risk-Sharing Networks in Rural Phillipines"', *Journal of Development Economics* **71(2)**, 261–287.

Fafchamps, M. & Vicente, P. (2013), '"Political Violence and Social Networks: Experimental Evidence"', *Journal of Development Economics* **101(C)**, 27–48.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K. & the Oregon Health Study Group (2012), "'The Oregon Health Insurance Experiment: Evidence from the First Year"', *The Quarterly Journal of Economics* **127(3)**, 1057–1106.

Finscope Malawi 2008 Report (2009), `http://www.finance.gov.mw/fspu/index.php/studies-reports-fspu/completed-studies/52-2008-finscope-malawi-survey/file`.

Fisher, R. (1935), *"The Design of Experiments"*, 9th Ed. MacMillan.

Fisman, R., Paravisini, D. & Vig, V. (2012), "'Social Proximity and Loan Outcomes"', *mimeo, Columbia University* .

Fitzgerald, J., Gottschalk, P. & Moffitt, R. (1998), "'An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics"', *Journal of Human Resources* **33(2)**, 251–299.

Fitzsimons, E., Malde, B., Mesnard, A. & Vera-Hernandez, M. (2014), '"Nutrition, information and household behaviour: experimental evidence from Malawi"', *IFS Working Paper Series* .

Fitzsimons, E., Malde, B. & Vera-Hernandez, M. (2013), '"Does increased interaction help consumption smoothing? Experimental Evidence from Malawi"', *mimeo, IFS.* .

Fitzsimons, E., Malde, B. & Vera-Hernandez, M. (2015), "'Group Size and the Efficiency of Informal Risk Sharing"', *IFS Working Paper* .

Foster, A. & Rosenzweig, M. (2001), '"Imperfect Commitment, Altruism and the Family: Evidence from Transfer Behavior in Low-Income Countries"', *The Review of Economics and Statistics* **83(3)**, 398–407.

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

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

Frank, O. (1980b), '"Sampling and Inference in a Population Graph"', *International Statistical Review/Revue Internationale de Statistique* **48(1)**, 33–41.

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

Frank, O. & Strauss, D. (1986), '"Markov Graphs"', *Journal of the American Statistical Association* **81**, 832–842.

Frantz, T. L., Cataldo, M. & Carley, K. (2009), '"Robustness of centrality measures under uncertainty: Examining the role of network topology"', *Computational and Mathematical Organization Theory* **15**, 303–328.

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

Galasso, E. & Umpathi, N. (2009), "'Improving nutritional status through behavioral change: Lessons from Madagascar"', *Journal of Development Effectiveness* **1(1)**, 60–85.

Genicot, G. & Ray, D. (2003), '"Group Formation in Risk-Sharing Arrangements"', *Review of Economic Studies* **70 (1)**, 87–113.

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

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

Gine, X., Goldberg, J. & Yang, D. (2012), "'Credit Market Consequences of Improved Personal Identification: Field Experimental Evidence from Malawi"', *The American Economic Review* **102(6)**, 2923–54.

Glaeser, E., Sacerdote, B. & Scheinkman, J. (1996), '"Crime and Social Interactions"', *Quarterly Journal of Economics* **115**, 811–846.

Glewwe, P. (1999), "'Why Does Mother's Schooling Raise Child Health in Developing Countries? Evidence from Morocco"', *The Journal of Human Resources* **34(1)**, 124–159.

Glewwe, P., Jacoby, H. & King, E. (2001), "'Early Childhood nutrition and academic achievement: a longitudinal analysis"', *Journal of Public Economics* **81**, 345–368.

Gneezy, U., Leonard, K. & List, J. (2009), "'Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society"', *Econometrica* **77(5)**, 1637–1664.

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

Goldberg, J. (2016), "'Kwacha Gonna Do? Experimental Evidence about Labor Supply in Rural Malawi"', *American Economic Journal: Applied Economics* **7(1)**, 129–149.

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

Graham, B. S. (2008), '"Identifying Social Interactions through Conditional Variance Restrictions"', *Econometrica* **76**, 643–660.

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

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

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

Gronau, R. (1987), *Home Production - A Survey*, Handbook of Labor Economics, Elsevier, chapter Home Production - A Survey, pp. 273–304.

Gronau, R. (1997), "'The Theory of Home Production: The Past Ten Years"', *Journal of Labor Economics* **15(2)**, 197–205.

Grossman, M. (1972), "'On the Concept of Health Capital and the Demand for Health"', *Journal of Political Economy* **80(2)**, 223–255.

Guell, M., Rodriguez Mora, J. & Telmer, C. (2015), "'The Informational Content of Surnames, the Evolution of Intergenerational Mobility and Assortative Mating"', *Review of Economic Studies* **82(2)**, 693–735.

Haider, R., Ashworth, A., Kabir, I. & Huttly, S. (2000), "'Effect of Community-Based Peer Counsellors on Exclusive Breastfeeding Practices in Dhaka, Bangladesh: a Randomised Control Trial"', *The Lancet* **356(9242)**, 1643–1647.

Hammersley, J. & Clifford, P. (1971), "Markov fields on finite graphs and lattices".

Handcock, M. S. (2003), "Assessing Degeneracy in Statistical Models of Social Networks", Technical report, CSSS Working Paper no. 39.

Hanna, R., Duflo, E. & Greenstone, M. (2016), "'Up in smoke: the influence of household behavior on the long-run impact of improved cooking stoves"', *American Economic Journal: Economic Policy* **8(1)**, 80–114.

Heckman, J. (1979), "'Sample Selection Bias as a Specification Error"', *Econometrica* **47(1)**, 153–161.

Hoddinott, J., Maluccio, J., Behrman, J., Flores, R. & Martorell, R. (2008), "'Effect of a nutrition intervention during early childhood on economic productivity in Guatemalan adults"', *The Lancet* **371**, 411–16.

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

Holland, P. W. & Leinhardt, S. (1977), "Notes on the Statistical Analysis of Network Data".

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

Hsieh, C.-S. & Lee, L.-F. (2014), '"A Social Interactions Model with Endogenous Friendship Formation and Selectivity"', *Journal of Applied Econometrics* .

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

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

Ito, T. & Takashi, K. (2009), "'Weather Risk, Wages in Kind, and the Off-Farm Labor Supply of Agricultural Households in a Developing Country"', *American Journal of Agricultural Economics* **91(3)**, 697–710.

Jackson, M. (2008), *"Social and Economic Networks"*, Princeton University Press.

Jackson, M. O., Rodriguez-Barraquer, T. & Tan, X. (2012), "'Social Capital and Social Quilts: Network Patterns of Favour Exchange"', *American Economic Review* **102(5)**, 1857–97.

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

Jakiela, P. & Ozier, O. (forthcoming), "'Does Africa Need a Rotten Kin Theorem? Experimental Evidence from Village Economies"', *Review of Economic Studies* .

Jalan, J. & Somanathan, R. (2008), "'The Importance of Being Informed: Experimental Evidence on Demand For Environmental Quality"', *Journal of Development Economics* **87(1)**, 14–28.

Jensen, R. (2010), "'The (Perceived) Returns to Education and the Demand for Schooling"', *Quarterly Journal of Economics* **125(2)**, 515–548.

Kalanda, B., Verhoeff, F. & Brabin, B. (2006), "'Breast and Complementary Feeding Practices in Relation to Morbidity and Growth in Malawian Infants"', *European Journal of Clinical Nutrition* **60**, 401–407.

Kamali, A., Quigley, M., Nakiyingi, J. & et al. (2003), "'Syndromic Management of Sexually-Transmitted Infections and Behaviour Change Interventions on Transmission of HIV-1 in Rural Uganda: A Community Randomised Trial"', *The Lancet* **361**, 645–652.

Karlan, D., Osei, R., Osei-Akoto, I. & Udry, C. (2014), "'Agricultural Decisions after Relaxing Credit and Risk Constraints"', *The Quarterly Journal of Economics* **129(2)**, 597–652.

Kelejian, H. H. & Piras, G. (2014), '"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.

Kim, P. & Jeong, H. (2007), '"Reliability of Rank Order in Sampled Networks"', *The European Physical Journal B* **55**, 109–114.

Kinnan, C. (2014), "'Distinguishing Barriers to Insurance in Thai Villages"', *mimeo, Northwestern University* .

Kinnan, C. & Townsend (2012), "'Kinship and Financial Networks: Formal Financial Access and Risk Reduction"', *American Economic Review, Papers and Proceedings* **102(2)**.

Kling, J., Liebman, J. & Katz, L. (2007), "'Experimental Analysis of Neighborhood Effects"', *Econometrica* **75(1)**, 83–119.

Kochar, A. (1999), "'Smoothing Consumption by Smoothing Income: Hours-of-Work Responses to Idiosyncratic Agricultural Shocks in Rural India"', *Review of Economics and Statistics* **81(1)**, 50–61.

Kocherlakota, N. (1996), '"Implications of Efficient Risk Sharing without Commitment"', *Review of Economic Studies* **63(4)**, 595–609.

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

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

Kossinets, G. (2006), '"Effects of Missing Data in Social Networks"', *Social Networks* **28**, 247–268.

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

Kremer, M. & Miguel, E. (2007), "'The Illusion of Sustainability"', *Quarterly Journal of Economics* **112(3)**, 1007–1065.

Krishnan, P. & Sciubba, E. (2009), "'Links and Architecture in Village Networks"', *Economic Journal* **119 (537)**, 917–949.

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

Lamb, R. (2003), "'Fertilizer Use, Risk, and Off-Farm Labor Markets in the Semi-Arid Tropics of India"', *American Journal of Agricultural Economics* **85(2)**, 359–371.

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

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

Lee, S. H., Kim, P. & Jeong, H. (2006), '"Statistical Properties of Sampled Networks"', *Physical Review E* **73(1)**.

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

Lewycka, S. (2011), "Reducing Maternal and Neonatal Deaths in Rural Malawi: Evaluating the Impact of a Community-based Women's Group Intervention", PhD thesis, University College London.

Lewycka, S., Mwansambo, C., Kazembe, P., Phiri, T., Mganga, A., Rosato, M., Chapota, H., Malamba, F., Vergnano, S., Newell, M.-L., Osrin, D. & Costello, A. (2010), "'A cluster randomised controlled trial of the community effectiveness of two interventions in rural Malawi to improve health care and to reduce maternal, newborn and infant mortality"', *Trials* **11:88**.

Lewycka, S., Mwansambo, C., Rosato, M., Kazembe, P., Phiri, T., Mganga, A., Chapota, H., Malamba, F., Kainja, E., Newell, M., Greco, G., Brannstrom, A., Skordis-Worrall, J., Vergnano, S., Osrin, D. & Costello, A. (2013), '"Effect of women's groups and volunteer peer counselling on rates of mortality, morbidity and health behaviours in mothers and children in rural Malawi (MaiMwana): a factorial, cluster-randomised controlled trial"', *The Lancet* **381**, 1721–35.

Liebman, J., Katz, L. & Kling, J. (2004), "'Beyond Treatment Effects: Estimating the Relationship Between Neighborhood Poverty and Individual Outcomes in the MTO Experiment"', *Working Paper 493, Industrial Relations Section, Princeton University.* .

Ligon, E. (1998), '"Risk Sharing and Information in Village Economies"', *Review of Economic Studies* **65(4)**, 847–864.

Ligon, E. & Schechter, L. (2012), "'Motives for Sharing in Social Networks"', *Journal of Development Economics* **99(1)**, 13–26.

Ligon, E., Thomas, J. & Worrall, T. (2003), "'Informal Insurance Arrangements with Limited Commitment: Theory and Evidence from Village Economies"', *The Review of Economic Studies* **69(1)**, 209–244.

Lindeboom, M., Nozal, A. & van der Klauw, B. (2009), "'Parental education and child health: Evidence from a schooling reform"', *Journal of Health Economics* **28**, 109–131.

Lindeboom, M., Potrait, F. & van den Berg, G. (2010), "'Long-run Effects on Longevity of a Nutritional Shock Early in Life: The Dutch Potato Famine of 1846-1847"', *Journal of Health Economics* **29(5)**, 617–629.

Linnemayr, S. & Alderman, H. (2011), "'Almost random: Evaluating a large-scale randomized nutrition program in the presence of crossover"', *Journal of Development Economics* **96(1)**, 106–114.

Liu, X. (2013), '"Estimation of a Local-aggregate Network Model with Sampled Networks"', *Economics Letters* **118**, 243–246.

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

Liu, X., Patacchini, E., Zenou, Y. & Lee, L.-F. (2014), '"Criminal Networks: Who is the Key Player?"', *Unpublished Manuscript* .

Luke, N. & Munshi, K. (2006), "'New Roles for Marriage in Urban Africa: Kinship Networks and the Labor Market in Kenya"', *Review of Economics and Statistics* **88(2)**, 264–282.

Luo, R., Shi, Y., Zhang, L., Zhang, H., Miller, G., Medina, A. & Rozelle, S. (2012), "'The Limits of Health and Nutrition Education: Evidence from Three Randomized-Controlled Trials in Rural China"', *CESifo Economic Studies* **58**(2), 385–404.

Maccini, S. & Yang, D. (2008), "'Under the Weather: Health, Schooling, and Economic Consequences of Early-Life Rainfall"', *American Economic Review* **99(3)**, 1006–1026.

Mace, B. (1991), '"Full Insurance in the Presence of Aggregate Uncertainty"', *Journal of Political Economy* **99(5)**, 928–956.

Madajewicz, M., Pfaff, A., van Geen, A., Graziano, J., Hussein, I., Momotaj, H., Sylvi, R. & Ahsan, H. (2007), "'Can information alone change behavior? Response to arsenic contamination of groundwater in Bangladesh"', *Journal of Development Economics* **84**(2), 731 – 754.

Magruder, J. (2010), "'Intergenerational Networks, Unemployment and Persistent Inequality in South Africa"', *American Economic Journal: Applied Economics* **2(1)**.

Maluccio, J., Hoddinott, J., Behrman, J., Martorell, R., Quisumbing, A. & Stein, A. (2009), "'The Impact of Improving Nutrition During Early Childhood on Education among Guatemalan Adults"', *The Economic Journal* **119(537)**, 734–761.

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

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

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

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

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

McCrary, J. & Royer, H. (2011), "'The Effect of Female Education on Fertility and Infant Health: Evidence from School Entry Policies Using Exact Date of Birth"', *American Economic Review* **101(1)**, 158–195.

McKay, B. (1983), "'Applications of a Technique for Labelled Enumeration"', *Congressus Numerantium* **40**, 207–221.

Méango, R. (2014), '"International Student Migration: A Partial Identification Analysis"', *SSRN Working Paper No. 2392732* .

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

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

Miller, C. & Tsoka, M. (2012), "'Cash Transfers and Children's Education and Labor among Malawi's Poor"', *Development Policy Review* **30(4)**, 499–522.

Mobarak, M. & Rosenzweig, M. (2014), "'Risk, Insurance and Wages in General Equilibrium"', *mimeo, Yale University* .

Moffitt, R. (2001), "Policy interventions, low-level equilibria, and social interactions", *in* S. Durlauf & H. P. Young, eds, 'Social Dynamics', MIT Press, Cambridge, pp. 45–82.

Morrow, A. L., Guerrero, M. L., Shults, J., Calva, J. J., Lutter, C., Bravo, J., Ruiz-Palacios, G., Morrow, R. C. & Butterfoss, F. D. (1999), 'Efficacy of home-based peer counselling to promote exclusive breastfeeding: a randomised controlled trial', *The Lancet* **353**(9160), 1226–1231.

Mtika, M. & Doctor, H. (2002), "'Matriliny, Patriliny and Wealth Flow Variations in Rural Malawi"', *African Sociological Review* **6(2)**, 71–97.

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

Munshi, K. & Rosenzweig, M. (2006), "'Traditional Institutions Meet the Modern World: Caste, Gender, and Schooling Choice in a Globalizing Economy"', *American Economic Review* **96**(4), 1225–1252.

Munshi, K. & Rosenzweig, M. (2016), "'Networks and Misallocation: Insurance, Migration, and the Rural-Urban Wage Gap"', *American Economic Review* **106**(1), 46–98.

Munthali, A. (2002), '"Adaptive Strategies and Coping Mechanisms of Families and Communities Affected by HIV/AIDS in Malawi"', *Draft paper prepared for the UNRISD project HIV/AIDS and Development* .

Murgai, R., Winters, P., Sadoulet, E. & de Janvry, A. (2002), '"Localized and Incomplete Mutual Insurance"', *Journal of Development Economics* **67**, 245–274.

National Statistics Office, M. & Macro, O. (2005), "Malawi Demographic and Health Survey 2004", Technical report, National Statistics Office, Malawi and OCR Macro. **URL:** *www.measuredhs.com/pubs/pdf/FR175/FR-175-MW04.pdf.*

Ngatia, M. (2012), '"Social Interactions and Individual Reproductive Decisions"', *Unpublished Manuscript* .

Office, N. S. (2008), *"Malawi Population and Housing Census"*, National Statistics Office.

Organisation, W. H. (2000), "Complementary Feeding: Family Foods for Breastfed Children", Technical report, World Health Organisation.

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

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

Pelletier, D., Frongillo, E., Schroeder Jr, D. & Habicht, J. (1994), "'A methodology for estimating the contribution of malnutrition to child mortality in developing countries."', *Journal of Nutrition* **124(10 Suppl.)**, 2106S – 2122S.

Peters, P., Walker, P. & Kambewa, D. (2008), '"Striving for Normality in a Time of AIDS in Malawi"', *The Journal of Modern African Studies* **46(4)**, 659–687.

Phiri, K. (1983), '"Some Changes in the Matrilineal Family System among the Chewa of Malawi since the Nineteenth Century"', *The Journal of African History* **24(2)**, 257–274.

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

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

Puentes, E., Wang, F., Behrman, J., Cunha, F., Hoddinott, J., Maluccio, J., Adair, L., Borja, J., Martorell, R. & Stein, A. (2014), "'Early Life Height and Weight Production Functions with Endogenous Energy and Protein Inputs"', *mimeo, Rice University* .

Reniers, G. (2003), "'Divorce and Remarriage in Rural Malawi"', *Demographic Research* **Special Collection 1, Article 6**, 175–206.

Richards, A. I. (1950), *"African Systems of Kinship and Marriage"*, Oxford University Press, chapter "Some Types of Family Structure Amongst the Central Bantu", pp. 207–251.

Rosato, M., Lewycka, S., Mwansambo, C., Kazembe, P. & Costello, A. (2009), "'Women's Groups' Perceptions of Neonatal and Infant Health Problems in Rural Malawi"', *Malawi Medical Journal* **21(4)**, 168–173.

Rosato, M., Mwansambo, C. W., Kazembe, P. N., Phiri, T., Soko, Q. S., Lewycka, S., Kunyenge, B. E., Vergnano, S., Osrin, D., Newell, M.-L. & de L Costello, A. M. (2006), "'Women's groups' perceptions of maternal health issues in rural Malawi"', *The Lancet* **368**(9542), 1180–1188.

Rose, E. (2001), "'Ex Ante and Ex Post Labor Supply Response to Risk in a Low-income Area"', *Journal of Development Economics* **64(2)**, 371–388.

Rosenbaum, P. (2002), *"Observational Studies"*, Springer Series in Statistics, Springer.

Rosenzweig, M. (1988*a*), "'Risk, Implicit Contracts and the Family in Rural Areas of Low-Income Countries"', *The Economic Journal* **98**, 1148–1170.

Rosenzweig, M. (1988*b*), "'Risk, Private Information and the Family"', *The American Economic Review* **78(2)**, 245–250.

Rosenzweig, M. & Schultz, T. (1983), "'Estimating a Household Production Function: Heterogeneity, the Demand for Health Inputs, and Their Effects on Birth Weight"', *Journal of Political Economy* **91(5)**, 723–746.

Rosenzweig, M. & Stark, O. (1989), '"Consumption Smoothing, Migration and Marriage: Evidence from Rural India"', *Journal of Political Economy* **97(4)**, 905–926.

Sacerdote, B. (2011), "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?", *in* E. Hanushek, S. Machin & L. Woessman, eds, 'Handbook of the Economics of Education', Vol. 3, Elsevier.

Saha, A. (1994), "'A Two-season Agricultural Household Model of Output and Price Uncertainty"', *Journal of Development Economics* **45(2)**, 245–269.

Schroeder Jr, D., Martorell, R., Rivera, J., Ruel, M. & Habicht, J. (1995), "'Age Differences in the Impact of Nutritional Supplementation on Growth"', *The Journal of Nutrition* **125(4 Suppl)**, 1051S – 1059S.

Schultz, T. (2005), *Population Policies, Fertility, Women's Human Capital and Child Quality*, Vol. 4 of *Handbook of Development Economics*, Elsevier, chapter Population Policies, Fertility, Women's Human Capital and Child Quality.

Schutz, B. (2003), *"Gravity from the Ground Up"*, Cambridge Univ Press.

Sear, R. (2008), "'Kin and Child Survival in Rural Malawi"', *Human Nature* **19(3)**, 277–293.

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

Sherman, S. (1973), '"Markov Random Fields and Gibbs Random Fields"', *Israel Journal of Mathematics* **14**, 92–103.

Shim, E. (2015), "'The Impact of Conditional Cash Transfer Programs under Risk Sharing Arrangements: Schooling and Consumption Smoothing in Rural Mexico"', *mimeo, UCSD* .

Shrimpton, R., Victora, C., de Onis, M., Costa Lima, R., Blossner, M. & Clugston, G. (2001), "'Worldwide Timing of Growth Faltering: Implications for Nutritional Interventions"', *Pediatrics* **107(5)**, e75.

Small, D., Ten Have, T. & Rosenbaum, P. (2008), "'Randomization Inference in a Group-Randomized Trial of Treatments for Depression: Covariate Adjustment, Noncompliance and Quantile Effects"', *Journal of the American Statistical Association* **103(481)**, 271–279.

Snijders, T. A. B. (2002), '"Markov Chain Monte Carlo Estimation of Exponential Random Graph Models"', *Journal of Social Structure* **3**, 1–40.

Stark, O. & Lucas, R. (1988), '"Migration, Remittances and the Family"', *Economic Development and Cultural Change* **36(3)**, 465–481.

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

Strauss, J. & Thomas, D. (1998), "'Health, Nutrition and Economic Development"', *Journal of Economic Literature* **36(2)**, 766–817.

Thomas, D., Strauss, J. & Henriques, M. (1991), "'How does mother's education affect child height?"', *Journal of Human Resources* **26(2)**, 183–211.

Thomas, D., Witoelar, F., Frankenberg, E., Sikoki, B., Strauss, J., Sumantri, C. & Suriastini, W. (2012), "'Cutting the costs of attrition: Results from the Indonesia Family Life Survey"', *Journal of Development Economics* **98(1)**, 108–123.

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

Thornton, R. (2008), "'The Demand for, and Impact of, Learning HIV Status"', *The American Economic Review* **98(5)**, 1829–63.

Topa, G. & Zenou, Y. (2015), "Neighborhood and Network Effects", *in* G. Duranton, V. Henderson & W. Strange, eds, 'Handbook of Regional and Urban Economics', Vol. 5A, Elsevier, chapter 9.

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

Townsend, R. (1995), '"Consumption Insurance: An Evaluation of Risk-Bearing Systems in Low-Income Economies"', *Journal of Economic Perspectives* **9(3)**, 83–102.

Trinitapoli, J., Yeatman, S. & Fledderjohann, J. (2014), '"Sibling Support and the Educational Prospects of Young Adults in Malawi"', *Demographic Research* **30**, 547–578.

van den Berg, G., Deeg, D., Lindeboom, M. & Potrait, F. (2010), "'The role of early-life conditions in the cognitive decline due to adverse events later in life"', *Economic Journal* **120**, F411–F428.

van den Berg, G. J., Lindeboom, M. & Lopez, M. (2009), "'Inequality in individual mortality and economic conditions earlier in life "', *Social Science & Medicine* **69**(9), 1360 – 1367.

van den Berg, G., Lindeboom, M. & Potrait, F. (2006), "'Economic Conditions Early in Life and Individual Mortality"', *The American Economic Review* **96(1)**, 290–302.

Victora, C., de Onis, M., Hallal, P., Blossner, M. & Shrimpton, R. (2010), "'Worldwide Timing of Growth Faltering: Revisiting Implications for Interventions"', *Pediatrics* **125**, e473.

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

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

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

Wang, S. (2013), "'Marriage Networks, Nepotism and Labor Market Outcomes in China"', *American Economic Journal: Applied Economics* **5(3)**, 91–112.

Wang, X. (2015), '"Risk Sorting, Portfolio Choice, and Endogenous Formal Insurance"', *mimeo, Duke University.* .

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

Witoelar, F. (2013), '"Risk Sharing Within the Extended Family: Evidence from the Indonesia Family Life Survey"', *Economic Development and Cultural Change* **62(1)**, 65–94.

Wooldridge, J. (2004), "'Cluster-Sample Methods in Applied Econometrics"', *The American Economic Review* **93(2)**, 133–138.

Zabel, J. (1998), "'An Analysis of Attrition in the Panel Study of Income Dynamics and the Survey of Income and Program Participation with an Application to a Model of Labor Market Behavior"', *Journal of Human Resources* **33(2)**, 479–506.