Collecting Network Data in Surveys

Arun Advani and Bansi Malde[1]

September 11 2013

[1]We gratefully acknowledge funding from the ESRC-NCRM Node "Programme Evaluation for Policy Analysis" Grant reference RES-576-25-0042.

Objectives

- Lots of interest in collecting data on social networks, or measures related to social environments (e.g. social norms, social pressure, etc)
- Interest in answering questions of the following type:
  - What features of social networks influence the adoption and usage of new technologies?
  - Does the effectiveness of policy interventions depend on the underlying social structure?
  - Does information on new health practices given to a small subset the village diffuse to the rest of the village? How does this diffusion take place?
- Want to make sure we collect the best possible data for this purpose, but within budget constraints

Meeting Structure

- What do we know about collecting networks data?
  - Findings from Advani and Malde (2013), which surveys the literature on empirical methods for identifying and estimating social effects using network data.
  - By network data, we mean information on exact mapping of connections (which allows you to construct a network graph).
  - Special focus on:
    - Options for collecting network data
    - Measurement error in network measures

Advani and Malde (2013): Network Methods Review

- Brings together and reviews literature on empirical methods for identifying and estimating social effects using network data
  1. First considers how networks data can be collected, and common sampling methods
  2. Methods for estimating social effects, taking the network to be pre-determined or exogenously formed
  3. Dealing with endogenous network formation
  4. Measurement error in network data (sampling induced and other)
- Will only present findings related to (1) and (4).

Why we may want to collect detailed network data

- Eases identification
  - Can get around the reflection problem quite easily
- Can get more detailed measures of features of the underlying social structure that influence outcomes
  - For example, network centrality measures offer a way of identifying key households in a network
  - Expansiveness of a network measures its fragility (i.e. how likely it is to break into two or more factions)
- Impose weaker assumptions on the underlying interaction structure than with more aggregated data
  - E.g. all members of a peer group are uniformly linked with one another

Collecting Networks Data

- Involves collecting information on two inter-related objects - nodes and edges (or links) - within a pre-defined network
  - Nodes = individuals, households, firms (i.e. the economic agent of interest)
  - Edges/Links = connections between nodes (e.g. financial transactions, family links, friendships, neighbours, etc)

Graph

- First need to decide on the network to measure
  - Depends on the social dimension of interest
    - Friends, family, households one transacts/chats with, etc
  - Careful thinking needed on whether the measured network is that most relevant to the question of interest.
  - In practice, researchers need impose geographic boundaries on the scope of the network
    - No clear guidance on how to choose this

Collecting Networks Data

- Common ways of collecting networks data include:
  - Direct elicitation from nodes
  - Collection from existing sources
    - Imputing links from existing information on, e.g. group memberships, naming conventions, etc
  - Spatial networks constructed from GIS data

Direct Elicitation from Nodes

- Ask nodes to list all nodes they interact with on a given dimension (e.g. borrow/lend rice)
  - A variant asks them to list all interactions with nodes on a specified list.
  - Related method asks nodes for information on their links, and the connections of their links
- A few other things to decide on:
  - What network to elicit information on
    - Actual or potential interactions, e.g. who a household borrowed rice from, or who could they borrow rice from if they needed it?
  - Existence of links only or strength of links
    - Existence $\Longrightarrow$ e.g. "Who do you chat with [X]?"
    - Strength $\Longrightarrow$ e.g. "How often do you chat with [X]?"
  - Recall duration: Trade-off between recall error and frequency of interaction

Collecting Networks Data

- Types of data that one could have, in terms of network coverage include:
  - Census of all nodes and links
  - Census of all nodes, and sample of links from each node
  - Random sampling
    - Random sample of nodes and all links between sampled nodes (induced subgraph)
    - Random sample of nodes and all links of sampled nodes (star subgraph)
    - Random sample of links and nodes of sampled links
  - Link tracing and snowball sampling

Census of all nodes and all links

- Collect information on the full network
- Pros:
  - Obtain accurate picture of whole network, including local neighbourhoods
  - Allows accurate computation of any desired network statistics
- Cons:
  - Expensive
  - Infeasible for large networks

Census of nodes and sample of links

- Information on all nodes, but only a (possibly non-random) sample of links
  - Censoring of # of links that can be reported in a survey
  - Common in practice
- Pros:
  - Relatively accurate measure of network, and typically possible to do some correction for this depending on extent of censoring
- Cons:
  - Relatively expensive
  - Infeasible for large networks
  - Some measurement error introduced by censoring

Random Sample of Nodes or Links

- Construct network based on links collected from a random sample of nodes
  - Similarly for random sampling of links (ignored here since not commonly used in Economics)
  - In practice, random sample of nodes is drawn from some list of all nodes in the network, e.g. census of households in a village
- Pros:
  - Cost-effective
- Cons:
  - Whole network structure not observed $\Longrightarrow$ measurement error in network structure
  - Generates a non-random sample of links $\Longrightarrow$ measurement error is non-classical
  - Standard statistical results for inference may not hold

Snowball Sampling

- Used to obtain data on "hard-to-reach" populations
- Data collected via the following process:
  - Start with an initial (possibly random) sample of nodes from the sample of interest, N_0.
  - Elicit all or a sample of the links of these nodes, denoted L_1. New nodes are called N_1
  - Then interview the nodes the initially sampled nodes (N_0) report being linked to (N_1) and elicit their links L_2
  - And so on until a specific target is reached
- Pros:
  - Cheap way of getting accurate information on a node's local network neighbourhood
  - Useful for "hard-to-reach" populations
- Cons:
  - Even if you start with a random initial sample of nodes, end up with a non-random sample of nodes and links
  - Whole network usually not sampled $\Longrightarrow$ measurement error in network structure

Measurement Error and why we should worry about it

- Two types of measurement error
  - Sampling induced
    - Measuring two interrelated objects - nodes and edges
    - Collecting links (nodes) from a random sample of nodes (links) generates a non-random sample of links (nodes)
    - So random sampling does not generate an accurate picture of the network structure or of related statistics
    - Missing data generates non-classical measurement error
  - Non-sampling measurement error
    - Censoring of the number of links, e.g. surveys may ask individuals to report their 3 best friends
    - Boundary specification, i.e. incorrectly restricting a network to be within a village, for instance
    - Miscoding, misreporting, non-response, etc

Sampling Induced Measurement Error

Broad facts emerging from literature on using sampled network data to compute network statistics and parameter estimates:

1. Network statistics computed from samples containing moderate (30-50%) and even relatively high (~70%) proportions of a network can be highly biased. Sampling a higher proportion of the network generates more accurate network statistics.
2. Measurement error due to sampling varies with the underlying topology.
3. The magnitude of error in network statistics due to sampling varies with the sampling method.
    - E.g. random sampling of nodes and edges under-estimates the proportion of nodes with high numbers of connections, while snowball sampling over-estimates them (since it misses nodes with low number of edges).
4. Parameters in economic models using mismeasured network statistics are subject to substantial bias and inconsistency.

Sampling Induced Measurement Error

Bias in network statistics and parameter estimates in star and induced subgraphs

Statistic	Network Statistic		Parameter Estimates	
	Star	Induced	Star	Induced
Avg. degree	-	-	+	+
Avg. path length	not known	+	-	-
Spectral gap	ambiguous	ambiguous	-	-
Clustering coeff	-	none/small	+	-
Avg graph span	not known	+	+/-	+/-
In-Degree	small	-	-	+
Out-Degree	-	-	-	+
Degree Centrality	not known	-	not known	not known

- Betweenness centrality and Bonacich centrality distributions are uncorrelated with truth under random sampling of nodes
- Magnitude of bias in parameter estimates can range from -100% to over 120% and may even lead to different signs.

Non-Sampling Measurement Error

Consequences include:

- Top-coding and boundary mis-specification lead to under-estimation of average degree and over-estimation of average path length; no bias in the clustering coefficient
- Missing and spurious nodes and miscoded edges don't bias degree or eigenvector centrality
- But, results above come from simulations which assume random errors, which may not match the actual missing data process.

Dealing with Measurement Error

Two broad ways of dealing with measurement error:

- Analytical Corrections
  - Correct network statistics using inverse probability weighting estimators
  - Applies to network statistics that can be expressed in terms of sums
  - And for sampling designs where the probability of being in the sample can be calculated:
    - Random sampling - can calculate exact sample inclusion probabilities for non-sampled quantity
    - Snowball sampling - exact probability can't be calculated, but there are methods to estimate these.

Dealing with Measurement Error

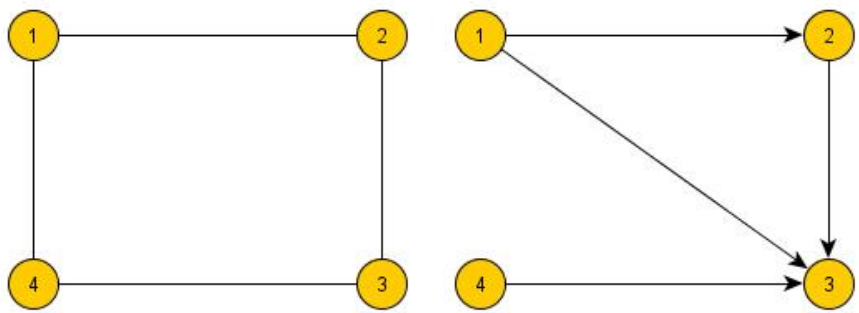
- Model-Based Corrections
  - Involves computing a statistical network formation model to predict missing links
    - Leading models are Exponential Random Graph Models (ERGMs)
    - Estimate a separate model per network
    - Need information on some characteristics that predict link existence for all nodes (e.g. head age, education, land ownership, wealth, caste, etc)

Summary

- Summarised methods for collecting networks data
  - Direct elicitation
    - What network to collect, boundary specification, etc
  - Impute from existing data
- Discussed ways of sampling
  - Necessary since it is costly and often infeasible to conduct a complete census
  - But, sampling leads to substantial measurement error and bias in network statistics measures and parameter estimates
- Ways of dealing with measurement error
  - Analytic corrections
  - Model based corrections - need some information on the non-sampled nodes to be able to predict links

Graph Theory Basics

- Mathematically, networks are represented as graphs. Graphs contain nodes and edges (links)
- Graphs can be directed or undirected; Links can be weighted (e.g. with frequency of interaction).
- Directed, unweighted graph G=(N,E)
- $j \in N$ is j is a node in the network; $ij \in E$ if i and j are linked.



Graph Theory Basics

- Network links can also be represented by an adjacency matrix

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix} \text{ or } \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

Network Representation and Measures

- Work in other disciplines has shown that features of the graph provide measures for concepts such as power, etc
- Well-known graph measures include:
  - Degree: Number of other nodes that a node is linked with
  - Clustering: If i is linked with j and k, what is the probability that j and k are also linked with one another?
  - Geodesic: Shortest path (number of links) that i has to travel through to get to j
  - Diameter: largest geodesic between any 2 links in the network
  - Centrality: captures a node's position in a network
    - Degree centrality – how connected a node is
    - Closeness centrality – how easily a node can reach others
    - Betweenness – the importance of a node in connecting others
    - Eigenvector centrality - how connected a node is to other important nodes

Back