Economics Letters 118 (2013) 243–246

Contents lists available at SciVerse ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Estimation of a local-aggregate network model with sampled networks[☆]

Xiaodong Liu [*]

Department of Economics, University of Colorado, Boulder, CO 80309, USA

ARTICLE INFO

Article history:
Received 29 August 2012
Received in revised form
29 October 2012
Accepted 30 October 2012
Available online 6 November 2012

JEL classification:
C13
C21

Keywords:
Social networks
Local-average models
Local-aggregate models
Sampling of networks
Weak instruments

ABSTRACT

This work considers the estimation of a network model with sampled networks. Chandrasekhar and Lewis (2011) show that the estimation with sampled networks could be biased due to measurement error induced by sampling and propose a bias correction by restricting the estimation to sampled nodes to avoid measurement error in the regressors. However, measurement error may still exist in the instruments and thus induce the weak instrument problem when the sampling rate is low. For a local-aggregate model, we show that the instrument based on the outdegrees of sampled nodes is free of measurement error and thus remains informative even if the sampling rate is low. Simulation studies suggest that the 2SLS estimator with the proposed instrument works well when the sampling rate is low and the other instruments are weak.

1. Introduction

In recent years, the literature on identification and estimation of social network models has progressed significantly (see Blume et al. (2011), for a recent survey). In his seminal work, Manski (1993) introduces a linear-in-means social interaction model with endogenous effects, contextual effects, and correlated effects. Manski shows that this model suffers from the *reflection problem* and the above-mentioned effects cannot be separately identified. Bramoullé et al. (2009) generalize Manski's linear-in-means model to a *local-average* network model, where the endogenous effect is represented by the *average* outcome of an agent's friends. They provide identification conditions for the local-average model and suggest using the characteristics of indirect friends as an instrument for the endogenous effect. Liu and Lee (2010) consider a *local-aggregate* network model where the endogenous effect is given by the *aggregate* outcome of the friends. They show that in the local-aggregate model, the Bonacich centrality (Bonacich, 1987) can be used as an additional instrument to achieve identification and improve estimation efficiency.

The above-mentioned papers assume that the outcomes, covariates and connections of the agents in a network can be fully observed, which may be unrealistic in some practical applications. Sojourner (forthcoming) considers a linear-in-means model with missing data on covariates. He shows that random assignment of agents to peer groups can help to overcome the missing data problem. On the other hand, Chandrasekhar and Lewis (2011) consider the missing data problem on network connections. They show that the estimation of sampled networks could be biased due to the measurement error induced by sampling. They propose a simple bias correction by restricting the estimation to the sampled agents, whose friends are observed, to avoid measurement error in the regressors. However, measurement errors may still exist in the instruments. For the local-average model, the instrument based on the characteristics of indirect friends is less informative when the sampling rate is low and thus may induce the weak instrument problem. In this work, we show that, for the local-aggregate model with sampled network data, the instrument based on the number of direct connections, which is the leading-order term of the Bonacich centrality, has no measurement error and thus remains informative even if the sampling rate is low.

The rest of the work is organized as follows. Section 2 introduces basic concepts and notation. Section 3 discusses identification and estimation of the network model with sampled networks. Section 4 provides simulation evidence for the finite sample performance of the estimator. Section 5 concludes and generalizes the proposed estimator to estimate a network model with network fixed effects.

☆ I thank Lung-fei Lee and an anonymous referee for helpful comments. The remaining errors are my own.

* Tel.: +1 303 4927414; fax: +1 303 4928960.
E-mail address: xiaodong.liu@colorado.edu.

0165-1765/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.econlet.2012.10.037

2. Networks and sampling

A *network* is represented by a *graph* $G = (V, E)$ where V is a set of n *nodes* and E is a set of ordered pairs of nodes called *arcs*. An arc $e_{ij} = (i, j)$ is considered to be directed from i to j where j is called the *head* and i is called the *tail* of the arc.[1] We assume there is no arc that points to itself so that $e_{ii} \notin E$ for all $i \in V$.

Denote the set of m randomly sampled nodes by S. Following Chandrasekhar and Lewis (2011), there are two different sampling schemes. In the first sampling scheme, the researcher surveys a set of m nodes and asks each node to nominate his/her friends among the other $m - 1$ nodes sampled. The sampled network $G^{|S} = (S, E^{|S})$, where $E^{|S} = \{e_{ij}|e_{ij} \in E, i \in S, j \in S\}$, is called the *induced subgraph*. In the second sampling scheme, the researcher has information of all the nodes in V and survey a set of m nodes and asks each node to nominate his/her friends from the list of the n nodes in V. The sampled network $G^S = (V, E^S)$, where $E^S = \{e_{ij}|e_{ij} \in E, i \in S, j \in V\}$, is called the *star subgraph*. In this work, we focus on the estimation of a star subgraph. Fig. 1 gives an example of a star subgraph.

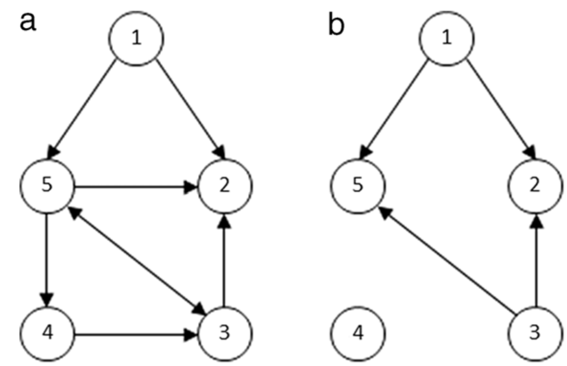

Fig. 1. (a) The full network; (b) a star subgraph with $S = \{1, 2, 3\}$.

3. Social network models with sampled networks

In a social network model, the connections in a network $G = (V, E)$ are represented by an $n \times n$ *adjacency matrix* $A = [a_{ij}]$ where $a_{ij} = 1$ if $e_{ij} \in E$ and $a_{ij} = 0$ otherwise. The model of the full network is given by

$$Y = \lambda_0 AY + \alpha_0 l_n + X\beta_0 + AX\gamma_0 + \epsilon. \tag{1}$$

Here, $Y = (y_1, \ldots, y_n)'$ where y_i is the observed outcome of the ith node. l_n is an $n \times 1$ vector of ones. $X = (x_1', \ldots, x_n')'$ where x_i is a $1 \times k$ vector of exogenous characteristics of the ith node. ϵ is an $n \times 1$ vector of i.i.d. innovations. According to Manski (1993), λ_0 captures the endogenous effect, where an agent's outcome may depend on the outcomes of his/her friends, and γ_0 captures the exogenous (contextual) effect, where an agent's outcome may depend on the exogenous characteristics of his/her friends. How to identify and estimate those two different effects has been a main interest for social interaction models.

If $(I_n - \lambda_0 A)$ is invertible, the reduced form equation of (1) is given by

$$Y = (I_n - \lambda_0 A)^{-1}(\alpha_0 l_n + X\beta_0 + AX\gamma_0 + \epsilon). \tag{2}$$

For identification and estimation of model (1), we need to find instruments for AY. As $(I_n - \lambda_0 A)^{-1} = I_n + \lambda_0 (I_n - \lambda_0 A)^{-1} A$, from (2), we have

$$\begin{aligned} E(AY|A, X) = {} & \alpha_0 A(I_n - \lambda_0 A)^{-1} l_n + AX\beta_0 \\ & + A^2 (I_n - \lambda_0 A)^{-1} X(\lambda_0 \beta_0 + \gamma_0). \end{aligned} \tag{3}$$

We will discuss the potential instruments implied by (3) in the following subsections.

Suppose we can observe (y_i, x_i) for all $i \in V$ and arcs e_{ij} if and only if $i \in S$ in the data. In other words, the sampled network can be represented by a star subgraph $G^S = (V, E^S)$. As argued by Chandrasekhar and Lewis (2011), this sampling scheme is quite common. For instance, consider the network data collected by Banerjee et al. (2011) from 43 villages in Karnataka, India, in order to study the diffusion of microfinance. The data collection process includes a full census that collected demographic data on all households in the villages and a follow-up survey of a subsample of villagers asking them to nominate their social connections with other villagers. The resulting sampled network can be considered as a star subgraph.

Denote the corresponding adjacency matrix based on the sampled arcs by $A^* = [a_{ij}^*]$, where $a_{ij}^* = 1$ if $e_{ij} \in E^S$ and $a_{ij}^* = 0$ otherwise. The model with a sampled network is given by

$$Y = \lambda_0 A^* Y + \alpha_0 l_n + X\beta_0 + A^* X\gamma_0 + \epsilon. \tag{4}$$

For the estimation of model (4), we consider two different specifications of the network model, namely, the local-average model and the local-aggregate model.

3.1. The local-average model

For network models, it is quite common to row-normalize the adjacency matrix A such that the sum of each row of A is unity. Let $d_i = \sum_{j=1}^n a_{ij}$ denote the *outdegree* of node i (i.e. the number of tails adjacent to a node). The row-normalized A is given by $\bar{A} = [\bar{a}_{ij}]$ where $\bar{a}_{ij} = a_{ij}/d_i$.[2] With a row-normalized adjacency matrix, the network model is

$$Y = \lambda_0 \bar{A} Y + \alpha_0 l_n + X\beta_0 + \bar{A} X\gamma_0 + \epsilon,$$

where $\bar{A}Y$ and $\bar{A}X$ represent the average outcome and average characteristics of the connections respectively. Therefore, we call this model the local-average model.

We assume that $|\lambda_0| < 1$ so that $(I_n - \lambda_0 \bar{A})^{-1} = \sum_{j=0}^{\infty} (\lambda_0 \bar{A})^j$. As $\bar{A} l_n = l_n$, we have $\alpha_0 \bar{A}(I_n - \lambda_0 \bar{A})^{-1} l_n = \frac{\alpha_0}{1-\lambda_0} l_n$. Hence, it follows from (3) that

$$\begin{aligned} E(\bar{A}Y|\bar{A}, X) = {} & \frac{\alpha_0}{1 - \lambda_0} l_n + \bar{A}X\beta_0 \\ & + (\bar{A}^2 X + \lambda_0 \bar{A}^3 X + \cdots)(\lambda_0 \beta_0 + \gamma_0). \end{aligned}$$

If $\lambda_0 \beta_0 + \gamma_0 = 0$, $E(\bar{A}Y|\bar{A}, X)$ becomes a linear combination of l_n and $\bar{A}X$, and thus the local-average model cannot be identified. If $\lambda_0 \beta_0 + \gamma_0 \neq 0$, then $\bar{A}^2 X$ can be used as an instrument for $\bar{A}Y$ under the identification condition given by Bramoullé et al. (2009). Let $\bar{Z} = [\bar{A}Y, l_n, X, \bar{A}X]$, $Q_1 = [l_n, X, \bar{A}X, \bar{A}^2 X]$, and $P_1 = Q_1(Q_1'Q_1)^{-1}Q_1'$. The 2SLS estimator of $\delta_0 = (\lambda_0, \alpha_0, \beta_0', \gamma_0')'$ is given by $\hat{\delta}_n = (\bar{Z}'P_1\bar{Z})^{-1}\bar{Z}'P_1 Y$.

For a star subgraph, let $\bar{A}^*$ denote the row-normalized A^*. As $\bar{A}^*$ is misspecified, it introduces measurement errors to both regressors and instruments. Chandrasekhar and Lewis (2011) show that the 2SLS estimator for the local-average model with sampled networks is inconsistent because the measurement error in the instruments is correlated with that in the regressors. They propose a simple correction by estimating the model only with the sampled nodes. Let $\bar{a}_i$ ($\bar{a}_i^*$) denote the ith row of $\bar{A}$ ($\bar{A}^*$). As $\bar{a}_i^* = \bar{a}_i$ for $i \in S$, there is no measurement error in the regressors $[\bar{a}_i^* Y, 1, x_i, \bar{a}_i^* X]$

[1] In this work, we focus on the estimation of *directed* graphs. The estimators can be easily modified to estimate a *undirected* graph.

[2] For simplicity, we assume that $d_i > 0$ for all $i \in V$.