Obtaining network data when it is not easily accessible

Report

Patrick Altmeyer

28 May, 2021

Motivation

Both papers address the issue that network data is often very expensive, difficult or simply infeasible to collect. The primary complication in collecting network data stems from the fact that it is typically necessary to collect data on each individual in the network and for each of them gathering information about their relationships to other individuals.

Contribution

Previous work on network data accessibility focused on central nodes or specific aspects of the network data. The methodology proposed in Breza et al. (2020) can instead be used to recover the entire distribution of the network. De Paula, Rasul, and Souza (2019) propose a parsimonious frequentist approach to extract information about social networks from observational panel data.

Brief summary

Breza et al. (2020)

To solve the issue of inaccessibility of network data, Breza et al. (2020) propose that researchers collect aggregated relational data (ARD). They show that network data derived from ARD comes with a significant cost reduction.

The ARD methodology can be summarized as follows: collect ARD on a subset of all nodes as well as a set of covariates describing individual characteristics of all nodes in the network. Essentially the approach then boils down to forming the network on a latent space based both on the ARD and the covariates. Links between nodes in the latent space are formed based on their pair-wise distances in the latent space: the closer any two nodes are together, the more likely they are to form a link.

To demonstrate that their approach works, the authors show that they can use networks derived from ARD to replicate findings of previous studies that used expensive network data. For example, the first study they look at investigates how peer effects affect saving behaviour of individuals. It finds that individuals who are aware that their saving behaviour may be monitored are more likely to save. Using the network derived from ARD Breza et al. (2020) manage to produce estimates very close to those obtained in the original study (Figure Breza et al. (2020)). The authors further show that the total data collection costs went down from $189,164 for the original study to $34,512 using ARD.

One limitation of their proposed approach is that conclusions about the true existence of individual links cannot be made. Therefore the authors confess that if knowledge about individual links is required, researchers

TABLE 1—LOG TOTAL SAVINGS ACROSS ALL HOUSEHOLD ACCOUNTS REGRESSED ON MONITOR SIGNALING VALUE

	log total ending savings	
	(1)	(2)
Signaling value of monitor with full network data (q_{ij}), standardized	0.254	
	(0.0869)	
Predicted signaling value of monitor with ARD (q_{ij}), standardized		0.185
		(0.0925)
Observations	422	422
Number of villages	30	30

Note: Standard deviation of village-level block bootstrap in parentheses.

Figure 1: Replication of results from previous study.

will still have to specifically gather that information. Another concern the authors point to is the fact that the network formation is parametric and therefore subject to specific assumptions.

De Paula, Rasul, and Souza (2019)

The approach proposed by De Paula, Rasul, and Souza (2019) does not rely on any explicit data on network links at all:

> "[. . .] global identification of the en- tire structure of social networks is obtained, using only observational panel data that itself contains no information on network ties" — De Paula, Rasul, and Souza (2019)

Panel data here does not necessarily include a time dimension, but could instead involve a cross-section of individuals $n = 1, ..., N$ observed across some other discrete cross-section of instances $t = 1, ..., T$. With such a panel at hand De Paula, Rasul, and Souza (2019) propose estimating a canonical structural model of social interactions:

$$y_{i,t} = \rho_0 \sum_{j=1}^{N} W_{0,ij} y_{jt} + \beta_0 x_{it} + \gamma_0 \sum_{j=1}^{N} W_{0,ij} x_{jt} + \alpha_i + \alpha_t + \epsilon_{it} \tag{1}$$

where α_i, α_t are entity fixed effects and the main parameter of interest $W_{0,ij}$ measures the causal impact of individual j on the outcome of individual i. The authors show how W_0 along with the parameters measuring exogenous and endogenous peer effects (ρ_0 and γ_0, respectively) can be identified.

In order to test their proposed methodology the authors simulated random network structures through synthetic data and check how well their approach approximates the true parameters of interest (W_0, ρ_0, γ_0). Finally, they also apply their methodology to a real data set on tax competition between US states. Through the simulation exercises they demonstrate that their approach can accurately recover network structures. The empirical exercise on real data provides some evidence that their proposed model can also help uncover previously undisclosed findings from data.

References

Breza, Emily, Arun G Chandrasekhar, Tyler H McCormick, and Mengjie Pan. 2020. "Using Aggregated Relational Data to Feasibly Identify Network Structure Without Network Data." *American Economic Review* 110 (8): 2454–84.

De Paula, Áureo, Imran Rasul, and Pedro Souza. 2019. "Identifying Network Ties from Panel Data: Theory and an Application to Tax Competition." *arXiv Preprint arXiv:1910.07452*.