Network endogeneity and ways to deal with it

Final report

Patrick Altmeyer

17 June, 2021

Introduction

In the existing literature network exogeneity is commonly assumed: network properties are assumed to be independent of individual characteristics. This assumption is restrictive and in fact often violated in real-world applications. The most commonly cited example is peer-effects in the context of academic performance, where the assumption of network exogeneity imposes that there are no unobserved factors that affect both the formation of friendships and scholarly outcomes Hsieh and Lee (2016). Clearly there are unobservable factors, for example concerning students' shared interests or family background, that may affect both the friendships they choose and how well they perform in school (Johnsson and Moon 2021). Other examples of endogenous networks come to mind: a central bank may be interested in studying the effect of interconnectivity in interbank money markets on the interest rates that individual market participants pay and receive. While there are observable variables such as size or geographical location of banks that may affect both their interconnectivity and the interest rates they face, it is easy to think of unobserved factors that lead to network endogeneity such as the banks' business models.

Unobserved individual heterogeneity is not unique to the study of peer effects. The literature on causal inference in social sciences has come up with a number of tools to deal with endogeneity, perhaps most notably the use instrumental variables (see for example Morgan and Winship 2015). Instrumental variables that affect the outcome only through the endogenous regressor are often hard to find and particularly so in the context of endogenous networks: linking this back to the example of students, it is difficult to think of an effective instrument without even knowing the mechanism of network formation (Hsieh and Lee 2016).

Another popular way to deal with endogenous regressors is commonly referred to as the control function approach originally introduced by Heckman and Robb Jr (1985). Instead of attempting to resolve the issue of endogeneity by instrumenting the endogenous regressors, the control function approach tackles endogeneity head-on by modelling the endogeneity in the disturbance term of the model. This report reviews a number of recent papers that use the control function approach to resolve the issue of network endogeneity. Johnsson and Moon (2021) is the most recent study that has emerged from that body of literature and will serve as the baseline for this review.

The remainder of this report is structured as follows: section briefly reviews the study of peer-effects in social networks. Section presents the control function approach and how it can be applied to endogenous networks. Section presents the approach proposed by Johnsson and Moon (2021) in some more details and extends the review to other existing approaches. Finally, section concludes.

Peer-effect models

The linear-in-means model was originally proposed by Manski (1993) under the premise that any individual's outcome tends to be affected both by her individual characteristics as well as the characteristics and outcomes of her friends. The dependence of individuals' outcomes on those of their peers can generate a social multiplier: a student who is surrounded by friends who perform well at school is likely to draw benefits from her peers' success (Carrell, Fullerton, and West 2009). Let $(\mathbf{x}, \mathbf{y}, \mathbf{G})$ denote the tuple of exogenous characteristics, endogenous outcomes and the network. In particular, let elements in $\mathbf{G}$ be defined as $g_{ij} = \frac{1}{d_i}$ if $j \in N_i$ where $d_i = |N_i|$ denotes the total degree of individual i, that is the size of her neighbourhood N_i. Then following Bramoullé, Djebbari, and Fortin (2020) the linear-in-means model can be formally defined as follows:

$$y_i = \alpha + \gamma x_i + \delta \frac{1}{d_i} \sum_{j \in N_i} x_j + \beta \frac{1}{d_i} \sum_{j \in N_i} y_j + \varepsilon_i \tag{1}$$

For an unbiased estimation of the parameters in (1) we need to have that $\mathbb{E}\left[\varepsilon_i | \mathbf{x}, \mathbf{G}\right] = 0$. In other words, linear-in-means models hinges on the assumption that both the individual characteristics $\mathbf{x}$ and the network $\mathbf{G}$ are strictly exogenous with respect to the outcome and the disturbance term ε_i is therefore free any endogeneity.

Studies that aim to provide credible estimates of causal effects from peers to outcomes must demonstrate that the assumption of strict network exogeneity is not violated: there should be no unobservable source of correlation between the network and the outcome of interest. Bramoullé, Djebbari, and Fortin (2020) review some of the most common tools that are available to researchers in order to do establish exogeneity. They can be summarized at a high level as follows:

1. Researchers may rely on **random peers** or **quasi-random peers**, that is randomly allocated peers in natural or controlled experiments Carrell, Sacerdote, and West (2013). A shortfall of this approach is that if peers are random, it is difficult to make sense of what peer effect exactly is being estimated, albeit causally.
2. Instead of assigning network connections at random, other researchers have instead relied on **random shocks**. This approach connects the linear-in-means model to the potential outcome framework (Morgan and Winship 2015) in which causality is established through interventions. It can be shown that as long as there is no selection bias with respect to treatment and the network of interactions is not affected by the treatment, then causal peer effects can be identified.
3. In the absence of exogenous variation, researchers can rely on **structural endogeneity**. Structural approaches attempt to model the source of endogeneity directly. Of course, the control function approach considered here falls into this category. In fact, the first paper to propose a structural approach to endogeneity uses the control function approach to introduce a latent variable that affects both the outcome and network formation (Goldsmith-Pinkham and Imbens 2013).
4. Finally, researchers have also made use of **panel data** to identify unobserved heterogeneity through individual fixed effects, thereby mitigating concerns around correlated effects.

In the following we will look at the third point in more detail.

Control function approach

Without loss of generality consider a simplified version of the linear-in-means model in (1)

$$y_i = \delta \tilde{x}_i + \varepsilon_i \tag{2}$$

where $\tilde{x}_i = \frac{1}{d_i} \sum_{j \in N_i} x_j$. Assume that the assumption of strict exogeneity is violated: $\mathbb{E}[\varepsilon_i|\mathbf{x}, \mathbf{G}] = \mathbb{E}[\varepsilon_i|\tilde{x}_i] \neq 0$. Suppose there exists some exogenous variable z_i that could serve as an instrument, if only we were able to find it. We have that

$$\tilde{x}_i = \beta_z z_i + v_i \tag{3}$$

with $\mathbb{E}[v_i|z_i] = 0$ and $\mathbb{E}[\varepsilon_i|z_i, v_i] = \mathbb{E}[\varepsilon_i|v_i]$ where the latter implies that the instrument affects the outcome y_i only through its effect on the instrumented variables $\tilde{x}_i$ (exclusion restriction).

In practice, if we could find a valid instrument z_i then causal peer effects could be identified through two-stage least-squares. The control function approach instead makes use of the fact that the exclusion restriction implies the following:

$$\mathbb{E}[y_i|\tilde{x}_i, v_i] = \delta \tilde{x}_i + \mathbb{E}[\varepsilon_i|\tilde{x}_i, v_i] = \delta \tilde{x}_i + \mathbb{E}[\varepsilon_i|z_i, v_i] = \delta \tilde{x}_i + \mathbb{E}[\varepsilon_i|v_i] \tag{4}$$

Note here that $\mathbb{E}[\varepsilon_i|v_i]$ is just a function $h(v_i)$ which if modelled correctly can be used to control for the endogeneity in the disturbance term ε_i (Heckman and Robb Jr 1985).

In the context of the peer-effects model defined in (2) we can define both the instrument z_i and the control function more precisely. Following Johnsson and Moon (2021) let $\mathbf{D}_n$ denote the $N \times N$ adjacency matrix corresponding to the matrix of interactions $\mathbf{G}$, where links are formed as follows:

$$d_{i,j} = \mathbb{I}(g(v_i, v_j) \geq u_{ij})\mathbb{I}(i \neq j) \tag{5}$$

Here v_i, v_j are the unobserved individual characteristics corresponding to the residual term in (3), $g(\cdot, \cdot)$ is some function and $u_{i,j}$ is a link-specific component. We can think of the adjacency matrix $\mathbf{D}_n$ as a way to summarize the mechanism of network formation. In other words, $\mathbf{D}_n$ is precisely the instrument that we would like to observe, but in practice may have difficulty to find as already mentioned above. Provided though that we find a way to estimate the control function $h(v_i) = \mathbb{E}[\varepsilon_i|v_i] = \mathbb{E}[\varepsilon_i|\mathbf{D}_i, v_i]$ the problem of network endogeneity can be resolved. In particular, all that is left to do is to actually control for $h(v_i)$ in (2), namely

$$y_i = \delta \tilde{x}_i + h(v_i) + \epsilon_i \tag{6}$$

where now clearly $\mathbb{E}[\varepsilon_i|\tilde{x}_i, h(v_i)] = \mathbb{E}[\varepsilon_i|\mathbf{x}, \mathbf{G}, h(v_i)] = 0$.

Literature review

The problem with the model in (6) is that v_i is unobserved and its absence $h(v_i)$ cannot be estimated. The main contribution of Johnsson and Moon (2021) is the observation that by the Weak Law of Large Numbers (WLLN) the control function can be proxied by a function of the average node degree

$$h(v_i) = \mathbb{E}\left[\varepsilon_i | v_i\right] \approxeq \mathbb{E}\left[\varepsilon_i | \bar{k}_i\right] = h_*(\bar{k}_i) \tag{7}$$

where $\bar{k}_i = \frac{1}{N} \sum_{j \neq i}^{d_{ij}}$.

Conclusion

References

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin. 2020. “Peer Effects in Networks: A Survey.” *Annual Review of Economics* 12: 603–29.

Carrell, Scott E, Richard L Fullerton, and James E West. 2009. “Does Your Cohort Matter? Measuring Peer Effects in College Achievement.” *Journal of Labor Economics* 27 (3): 439–64.

Carrell, Scott E, Bruce I Sacerdote, and James E West. 2013. “From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation.” *Econometrica* 81 (3): 855–82.

Falk, Armin, and Andrea Ichino. 2006. “Clean Evidence on Peer Effects.” *Journal of Labor Economics* 24 (1): 39–57.

Goldsmith-Pinkham, Paul, and Guido W Imbens. 2013. “Social Networks and the Identification of Peer Effects.” *Journal of Business & Economic Statistics* 31 (3): 253–64.

Heckman, James J, and Richard Robb Jr. 1985. “Alternative Methods for Evaluating the Impact of Interventions: An Overview.” *Journal of Econometrics* 30 (1-2): 239–67.

Hsieh, Chih-Sheng, and Lung Fei Lee. 2016. “A Social Interactions Model with Endogenous Friendship Formation and Selectivity.” *Journal of Applied Econometrics* 31 (2): 301–19.

Johnsson, Ida, and Hyungsik Roger Moon. 2021. “Estimation of Peer Effects in Endogenous Social Networks: Control Function Approach.” *Review of Economics and Statistics* 103 (2): 328–45.

Manski, Charles F. 1993. “Identification of Endogenous Social Effects: The Reflection Problem.” *The Review of Economic Studies* 60 (3): 531–42.

Morgan, Stephen L, and Christopher Winship. 2015. *Counterfactuals and Causal Inference*. Cambridge University Press.

Sacerdote, Bruce. 2001. “Peer Effects with Random Assignment: Results for Dartmouth Roommates.” *The Quarterly Journal of Economics* 116 (2): 681–704.