C.I.R.P.É.E

Centre Interuniversitaire sur le Risque, les Politiques Économiques et l'Emploi

Cahier de recherche/Working Paper **09-13**

The Econometrics of Social Networks

Yann Bramoullé
Bernard Fortin

Avril/April 2009

Bramoullé : Department of Economics and CIRPÉE, Université Laval
ybramoulle@ecn.ulaval.ca
Fortin: Department of Economics and CIRPÉE, Université Laval
bernard.fortin@ecn.ulaval.ca

We thank Habiba Djebbari and Marcel Fafchamps for helpful comments.

Abstract:
In a social network, agents have their own reference group that may influence their behavior. In turn, the agents' attributes and their behavior affect the formation and the structure of the social network. We survey the econometric literature on both aspects of social networks and discuss the identification and estimation issues they raise.

Keywords: Social network, peer effects, identification, network formation, pair-wise regressions, separability, mutual consent

JEL Classification: D85, L14, Z13, C3

Introduction

Economists are becoming increasingly aware of the importance and ubiquity of social networks. At a general level, a social network represents any pattern of relationships between agents. Salient examples include friendship networks among adolescents, coauthorship networks among scientists, and trade networks between countries. In economics, a new body of theoretical work explores 1) how social networks influence outcomes and 2) how in turn this affects network formation (Jackson 2008). At the empirical level, the literature is still scarce but is expanding at a rapid pace. We focus here on the econometrics of social networks. Following the theoretical work, we divide our presentation in two parts. First, we discuss the issues raised by the analysis of the effects of social networks on outcomes. Second, we look at network formation. Throughout, we assume that the network is binary[1] and observed at one point in time.[2] We also leave aside the critical and understudied issue of sampling.

The Effects of Social Networks on Outcomes

Researchers suspect that in many contexts, the behavior of individuals is affected by others. Economists have long tried to obtain reliable estimates of such *peer effects*, but the task is not easy. Recent papers have introduced social networks to the analysis of peer effects. They have looked at many outcomes. Among others, we can mention welfare participation (Bertrand *et al.* 2000), employment of war veterans (Laschever 2005), informal insurance against illness (Dercon and De Weerdt 2006), educational choices (De Giorgi *et al.* 2007), obesity (Trogdon *et al.* 2008), academic achievement (Lin 2008, Calvó-Armengol *et al.* forthcoming), and recreational activities (Bramoullé *et al.*, forthcoming). Overall, these studies show that social networks offer a fresh perspective on the issue.

In general, the identification of peer effects raises three main challenges (Manski 1993). First, the researcher must determine the appropriate reference groups. Who is affected by whom? The collection of comprehensive information on interactions provides a direct answer to this question. By definition in a social network, each agent has his own specific reference group. Indeed, the papers mentioned above rely on original data sets possessing detailed information on social structures, such as the *Add Health* data.

Second, unobserved attributes that are correlated between peers may generate a problem of confounding variables (spurious correlation). For instance, individuals in the same reference group may face similar environments (*e.g.*, a student and her friends may have the same professor). Self-selection may also induce the presence of such *correlated effects*. Similar individuals tend to interact together, which makes the formation of the network endogenous. This endogeneity should be corrected for in the estimation of peer effects.

[1] The network is binary if two agents are either connected or unconnected. Alternatively, the links could differ in strength, *e.g.*, Marmaros and Sacerdote (2006).

[2] Repeated observations of the network through time opens up interesting econometric possibilities, *e.g.*, Fafchamps *et al.* (2008).

Third, simultaneity in peer behavior may hinder identification of exogenous effects, *i.e.,* the influence of peer attributes, from endogenous effects, *i.e.*, the influence of peer outcomes. This is the *reflection problem* studied by Manski (1993). Even in the absence of correlated effects, distinguishing exogenous and endogeneous effects is impossible in the context of a *linear-in-means* model when agents interact in groups, *i.e.,* the social network is partitioned in groups and individuals are affected by all others in their group but by none outside of it. Only a composite social effect can be identified. Moreover, this latter effect cannot generally be identified in the presence of correlated effects. In contrast all effects can be identified, under certain conditions, when the social network has a richer structure.[3]

To see why, let us focus on a simple model inspired by Bramoullé *et al.* (BDF, forthcoming). The reference group of agent i ($i = 1, ..., n$) in a network is given by the set P_i with size n_i. Assume a linear-in-means model where y_i is the outcome of agent i, x_i is an attribute of i (the model can easily be generalized to many attributes), β and δ are resp. the endogenous and the exogenous social effect, with $|\beta| < 1$, and ϵ_i is a random term. Suppose first that the x_i 's are strictly exogenous (no correlated effects). The structural model can be written as:

$$y_i = \alpha + \beta \frac{\sum_{j \in P_i} y_j}{n_i} + \gamma x_i + \delta \frac{\sum_{j \in P_i} x_j}{n_i} + \epsilon_i, \quad \mathbb{E}[\epsilon_i \mid \mathbf{x}] = 0,$$

or, in matrix notation,

$$\mathbf{y} = \alpha \boldsymbol{\iota} + \beta \mathbf{G}\mathbf{y} + \gamma \mathbf{x} + \delta \mathbf{G}\mathbf{x} + \boldsymbol{\epsilon}, \quad \mathbb{E}[\boldsymbol{\epsilon} \mid \mathbf{x}] = \mathbf{0}, \tag{1}$$

where $\mathbf{y}$ is a $n \times 1$ vector of outcomes for the network l, $\mathbf{G}$ is an $n \times n$ *interaction matrix* with $G_{ij} = 1/n_i$ if i is affected by j, and 0 otherwise, and $\boldsymbol{\iota}$ is a $n \times 1$ vector of ones. This model can be derived from a choice-theoretic approach where individuals choose their outcome to maximize a quadratic utility and social interactions have reached a Nash equilibrium. Note also that the systematic part of (1) is similar to that of a spatial autoregressive (SAR) model (*e.g.,* Cliff and Ord 1981) extended to allow for exogenous effects.

BDF show that $\boldsymbol{\theta} = (\alpha, \beta, \gamma, \delta)$ is identified given the moment restriction $\mathbb{E}[\boldsymbol{\epsilon} \mid \mathbf{x}] = \mathbf{0}$ *and* restrictions on $\mathbf{G}$. More specifically, they show that the model is identified if and only if the matrices $\mathbf{I}, \mathbf{G}$, and $\mathbf{G}^2$ are linearly independent. Two particular cases illustrate this basic result. First, suppose that agents interact in groups. In this case, P_i is composed of all individuals in i's group. Then, $\mathbf{G}$ is block-diagonal with $\mathbf{G}^2 = \mathbf{G}$. Therefore, social effects are not identified. This corresponds to the Manski's (1993) case. Second, assume the presence of intransitivity in the network, *e.g.,* in a friendship network, some

[3] Other approaches can help identification. For instance, Brock and Durlauf (2001) exploit the non-linearities emerging from discrete choice models, Gaviria and Raphael (2001) and Trogdon *et al.* (*2008*) impose exclusion restrictions by assuming no exogenous or no endogenous effects, and Sacerdote (2001) exploit data where individuals are randomly assigned to groups.

friends of i's friends are not her friends. Their attributes will affect her outcome only through their effect on her friends' outcomes. In this case, the model is identified and can be estimated using 2SLS (Lin 2008).[4] Indeed, from an expansion series of $\mathbf{Gy}$ in the reduced form, it is clear that ($\mathbf{G}^2\mathbf{x}, \mathbf{G}^3\mathbf{x}$,...) can be used as valid instruments. These instruments have a direct interpretation. For instance, $\mathbf{G}^2\mathbf{x}$ represents the vector of the friends' friends mean attributes of each agent in the network. Intransitivity guarantees that this vector is not perfectly correlated with the exogenous regressors.[5]

This baseline model can be extended to the presence of correlated effects. Assume that the latter can be treated as component fixed effects. A component is a maximum set of indirectly related agents. Fixed effects can be interpreted as a two-step network formation: the agents of the same type join a club (the component) and then links between these agents are made at random. A number of papers treat correlated effects using fixed effects (*e.g.*, Clark and Loheac 2007, Lin 2008, Lee *et al.* 2008). In analogy with a panel model, one can get rid of the fixed effects through transformations in deviation from the agent' neighbors (local transformation) or from the agent' component (global transformation). Thus, under the local transformation, the model can be written as:

$$(\mathbf{I}-\mathbf{G})\mathbf{y} = \beta(\mathbf{I}-\mathbf{G})\mathbf{G}\mathbf{y} + \gamma(\mathbf{I}-\mathbf{G})\mathbf{x} + \delta(\mathbf{I}-\mathbf{G})\mathbf{G}\mathbf{x} + (\mathbf{I}-\mathbf{G})\boldsymbol{\epsilon},$$

and is identified if and only if the matrices $\mathbf{I}, \mathbf{G}$, $\mathbf{G}^2$, and $\mathbf{G}^3$ are linearly independent. As expected, this condition is more restrictive than in the absence of correlated effects. Identification now fails on some intransitive networks such as the star.

If correlated effects vary within components, however, a more elaborate study of network formation may be needed. Thus, we turn to the econometrics of network formation.

The Formation of Social Networks

Economists have only recently started to study network formation empirically. We discuss here the main features and limitations of the current econometric methods. Most existing studies rely on some form of *pairwise regressions*, see De Weerdt (2004), Fafchamps and Gubert (2007), Mayer and Puller (2008), Mihaly (2007), Santos and Barrett (2008), Udry and Conley (2004). The key idea is to consider the links themselves as the outcomes to be explained. Researchers then usually adapt standard empirical procedures for binary dependent variables.

Formally, a typical pairwise regression relies on the following econometric

[4] The model can also be estimated using conditional maximum likelihood if one is ready to impose more structure on the distribution of $\boldsymbol{\epsilon}$, see Lee *et al.* 2008.

[5] Surprisingly, intransitivity is not necessary to obtain identification. Social effects may be identified under group interactions if certain assumptions are satisfied, see Lee (2007) and BDF.

model:

$$\begin{aligned} Y_{ij} &= X'_{ij}\zeta + \varepsilon_{ij} \\ g_{ij} &= 1 \text{ if } Y_{ij} \geq 0 \text{ and } 0 \text{ if } Y_{ij} < 0 \end{aligned} \tag{2}$$

where Y_{ij} is the propensity to form link ij, X_{ij} is a vector of characteristics of link g_{ij} and ε_{ij} is a link-specific error term. This model can be applied to directed as well as undirected networks. When the network is undirected, $g_{ij} = g_{ji}$, and characteristics and propensities are defined for unordered pairs.

What are the choice-theoretic foundations of this model? The theory of network formation gives us some guidance here, see Jackson (2008). In general, individual i may derive some utility $u_i(\mathbf{g})$ from network $\mathbf{g}$ and this utility may depend on the network's structure in complex ways. Theorists have identified two natural assumptions to study network formation. Under mutual consent, a link is formed if both individuals agree to it. Alternatively, links may be formed unilaterally.

To tie back pairwise regressions to individual decisions, strong assumptions are needed. First and foremost, Model (2) relies on the *separability* of the utility function. The utility derived from the network is equal to the sum of the utilities brought by each link and these link-specific utilities are not affected by the structure of the network. Formally, $u_i(\mathbf{g}) = \sum_j v_i(g_{ij})$ and $Y_{ij} = v_i(g_{ij} = 1) - v_i(g_{ij} = 0)$. But separability is not enough. When the network is directed, model (2) is only consistent with separability and unilateral link formation. When the network is undirected, an additional assumption of symmetry must be imposed: $v_i(g_{ij} = 1) - v_i(g_{ij} = 0) = v_j(g_{ij} = 1) - v_j(g_{ij} = 0)$. Model (2) is then only consistent with separability and symmetry (under unilateral link formation or mutual consent).

Even under these stark assumptions, estimating model (2) requires first to answer three non-trivial questions.

1. *Who are the potential partners?* The default assumption is that every other individual in the population is a potential partner. In this case, any pair of individuals, connected or not, constitutes an observation. The number of observations is $n(n-1)$ for directed networks and $n(n-1)/2$ for undirected ones. In large populations, however, most pairs tend to be unconnected. The outcome to be explained has little variation and the interpretation of the method is problematic. It means that an individual considers whether to form a link or not with every other individual. Time and social constraints suggest, rather, that the set of potential partners is usually much smaller than the population at large. In some cases, individuals are naturally partitioned into particular sets (schools, villages) and assuming that individuals can only connect within these sets mitigates these concerns. But more research is likely needed to provide a good answer to this problem.[6]

[6] Mihaly (2007) assumes that the set of potential partners is unobserved and builds a simulated likelihood by picking at random many sets of potential partners. Another promising idea is to distinguish between meetings and links formed conditional on meetings, see Jackson & Rogers (2007) and Mayer & Puller (2008).

2. *How to include individual characteristics?* Link characteristics X_{ij} are often built from individual characteristics X_i and X_j. Researchers have used various ways to do that. We think that pairwise regressions for directed networks should generally include X_i, X_j, and $|X_i - X_j|$.[7] Including X_i and X_j separately is necessary to account for the potential effect of the characteristic on the propensity to initiate or receive a link. Including $|X_i - X_j|$ is natural given the prevalence of homophily in social networks.[8]

3. *What structure to impose on the error term?* Unobserved attributes of an individual likely affects all his linking decisions. Thus, we generally expect ε_{ij} to be correlated with ε_{ik} and this should be accounted for when computing standard errors. Introducing individual fixed effects provides one way to address this issue, see Udry and Conley (2004). Another way is to generalize the standard computations for robust covariance matrices. Fafchamps and Grubert (2007) build on Conley (1999) to derive appropriate formulas.

Once these three questions have been answered, the likelihood of network $\mathbf{g}$ can be computed and model (2) can be estimated through maximum likelihood. Logit procedures can be used. Pairwise regressions, however, have severe limitations. First as mentioned above, even under separability, model (2) is inconsistent with natural models of network formation such as mutual consent for directed networks. Comola and Fafchamps (2008) tackle this issue. They show that the appropriate likelihood can easily be computed and maximized as long as separability holds.

More importantly, separability is a strong assumption which is unlikely to hold in many settings. Especially, model (2) may fail to explain key *structural properties* of the network, such as low diameter, high clustering, and fat tails in degree distributions (Jackson 2008). At this stage, a main challenge for applied economists is to develop models which go beyond separability, explain these structural properties, and have sound microeconomic foundations. To do that, they should build on insights from game theorists, sociologists (Snijders *et al.* 2006), and physicists (Newman *et al.* 2006). Many interesting models of network formation have been proposed. The difficulty is now to implement these models empirically.[9]

Finally, a proper analysis of network formation shoud help to correct for the endogeneity of the network in the estimation of peer effects. Especially, economists should be able to develop two-step selection models à la Heckman in a network context.[10] Much research remains to be done on each step separately and on their combined estimation.

References

[7] For undirected networks, pairwise regressions should include $X_i + X_j$ and $|X_i - X_j|$.

[8] Homophily refers to the tendency of similar individuals to interact together.

[9] See Mayer and Puller (2008), Bramoullé and Rogers (2009) and Krishnan and Scubbia (2007) for first steps in this direction.

[10] Three recent papers attempt a joint estimation of network formation and network effects, see Weinberg (2006), Mihaly (2007) and Conti *et al.* (2009).

Bertrand, M., Luttmer E.F.P. and Mullainathan S. (2000): "Network Effects and Welfare Cultures", *Quarterly Journal of Economics*, 115, 1019-1056
Bramoullé, Y., Djebbari H. and Fortin B. (forthcoming): "Identification of peer effects through social networks", *Journal of Econometrics.*
Bramoullé, Y. and Rogers B. (2009): "Diversity and Popularity in Social Networks," mimeo.
Brock, W. and Durlauf, S. (2001): "Interaction-based Models", *Handbook of Econometrics, vol 5,* J. Heckman and Leamer E. (Eds), Amstersam: North-Holland.
Calvó-Armengol, A., Patacchini, E., and Zenou, Y. (forthcoming): "Peer Effects and Social Networks in Education", *Review of Economics Studies.*
Clark, A. and Loheac, Y. (2007): "It wasn't me, It was them! Social Influence in Risky Behaviour by Adolescents", *Journal of Health Economics*, Vol.26, no.4, pp.763-784.
Cliff, A. and Ord J. K. (1981): *Spatial Processes.* London: Pion.
Comola, M. and Fafchamps M.. (2008): "Testing Unilateral versus Bilateral Link Formation," mimeo.
Conley, T. (1999): "GMM estimation with cross-sectional dependence," *Journal of Econometrics*, 92, 1-45.
Conti, G., Galeotti, A., Mueller, G., and Pudney S.. (2009): "Popularity," mimeo.
Dercon, S. and De Weerdt, J. (2006): "Risk-Sharing Networks and Insurance against Illness", *Journal of Development Economics*, 81(2), 337-356.
De Giorgi, G., Pellizzari, M., and Redaelli S. (2007). " Be Careful of the Books you Read as of the Company You Keep: Evidence on Peer Effects in Educational Choices" (2833), IZA DP No. 2833.
De Weerdt, J. (2004): "Risk-Sharing and Endogenous Network Formation," ch.10 in *Insurance Against Poverty*, S. Dercon (ed.), Oxford University Press.
Fafchamps, M. and Gubert F. (2007): "Risk Sharing Networks in Rural Philippines," *Journal of Development Economics*, 71, 261-287.
Fafchamps, M., Goyal, S. and van der Leij M. (2008): "Matching and Network Effects," mimeo.
Gaviria, A. and Raphael, S. (2001): "School based Peer Effects and Juvenile Behavior", *Review of Economics and Statistics*, 83(2), 257-268.
Jackson, M. (2008): *Social and Economic Networks*, Princeton University Press.
Jackson, M. and Rogers B. (2007): "Meeting Strangers and Friends of Friends: How Random are Social Networks?" *American Economic Review*, 97(3), 890-915.
Krishnan, P. and Scubbia E. (2007): "Links and Architecture in Village Networks," *Economic Journal*, forthcoming.
Laschever, R. (2005): "The Doughboys Network: Social Interactions and Labor Market Outcomes of World War I Veterans", Mimeo, Northwestern University.
Lee, L.F., Liu, X., and Lin, X. (2008): "Specification and Estimation of Social Interaction Models with Network Structure, Contextual Factors, Correlation, and Fixed Effects", Mimeo, Department of Economics, Ohio State University.

Lee, L. F. (2007): "Identification and Estimation of Econometric Models with Group Interactions, Contextual Factors and Fixed Effects", *Journal of Econometrics,* 140(2), 333-374.
Lin, X. (2008): "Identifying Peer Effects in Student Academic Achievement by Spatial Autoregressive Models with Group Unobservables", Mimeo, Department of Economics, Tsinghyua University, Beijing.
Marmaros, D. and Sacerdote B. (2006): "How do friendships form?" *Quarterly Journal of Economics*, 121(1), 79-119.
Manski, C. (1993): "Identification of Endogenous Social Effects: The Reflection Problem", *Review of Economic Studies*, 60(3), 531-542.
Mayer, A. and Puller S. (2008): "The old boy (and girl) network: Social network formation on university campuses," *Journal of Public Economics*, 92, 329-347.
Mihaly, K. (2007): "Too Popular for School? Friendship Formation and Academic Achievement," mimeo.
Newman, M., Barabási, A.L., and Watts D. (2006): *The Structure and Dynamics of Networks*, Princeton University Press.
Sacerdote, B. (2001): "Peer Effects with Random Assignment: Results for Darmouth Roommates", *Quarterly Journal of Economics*, 116(2), 681-704.
Santos, P. and Barrett C.. (2008): "Identity, Interest and Information Search in a Dynamic Rural Economy," mimeo.
Snijders, T., Pattison, P., Robins, G., and Handcock M. (2006): "New specifications for exponential random graph models," *Sociological Methodology*, 99-153.
Trogdon J., Nonnemaker J. and Pais J. (2008): "Peer Effects in Adolescent Overweight", *Journal of Health Economics*, 27(5), 1388-1399
Udry, C. and Conley T. (2004): "Social Networks in Ghana," mimeo.
Weinberg, B. (2006): "Social Interactions and Endogenous Association," mimeo.