Social Interactions IV: Identification

Steven N. Durlauf

University of Chicago

Bonn 2018

Linear Social Interactions Models

Lawrence E. Blume

William A. Brock

Steven N. Durlauf

Rajshri Jayaraman

Journal of Political Economy 2015

Basic Objective

This paper provides a systematic analysis of identification in linear social networks models.

This is both a theoretical and an econometric exercise in that it links identification analysis to a rigorously delineated model of interdependent decisions.

Overview of Contributions

1. Microeconomic foundations of linear social interactions models are fully delineated in a Bayes-Nash framework.

2. Disparate identification findings are placed in a common framework.

3. Identification is shown to be generic when analyst has complete priori knowledge of social structure, the standard case in the literature.

4. Identification is possible with partial a priori knowledge.

5. Effects on endogeneity of networks are discussed.

Microfoundations

Linear social interactions models may be understood as Bayes-Nash equilibrium strategy profiles.

For this game, the set of players is V of size $N < \infty$.

Each individual is described by a vector of characteristics (x_i, z_i). $x_i \in \mathbf{R}$ is observable while $z_i \in \mathbf{R}$ is not. An individual's type vector t_i is a vector $(x, z_i) \in \mathbf{R}^{N+1}$

ρ is an exogenous prior distribution on types and is common knowledge.

Each individual chooses an $\omega_i \in \mathbf{R}$. Individual i's utility is a function of his type, his action, and the actions of others in the population. His payoff function is

$$U_i\left(\omega_i, \omega_{-i}\right) =$$
$$\left(\gamma x_i + z_i + \delta \sum_j c_{ij} x_j\right)\omega_i - \frac{1}{2}\omega_i^2 - \frac{\phi}{2}\left(\omega_i - \sum_j a_{ij}\omega_j\right)^2$$

This specification embodies two types of social interactions

The term $\delta \sum_j c_{i,j} x_j$ is known as contextual effect. It is a conventional externality, network average characteristics are a kind of group capital or public good.

The term $-\frac{\phi}{2}\left(\omega_i - \sum_j a_{ij}\omega_j\right)^2$ is known as an endogenous or peer effect. It represents a general version of conformity.

Key difference is that contextual effects involve characteristics, endogenous effects involve choices.

Sociomatrices

The strength of the direct impact of the choices and characteristics of others on each member of the population are summarized by the sociomatrices A and C, whose elements are taken from the individual utility functions.

These sociomatrices characterize what is meant by social structure.

Assumptions for Theoretical Analysis

T.1. $\phi \geq 0, A$ and C are non-negative, each row sums to either 0 or 1, and for all i, $a_{ii} = 0$.

T.2. Second moments of ρ exist.

T.1. Restricts the payoff function.

Non-negativity delimits the types of social interactions under study.

Row-summability allows for a meaningful distinction between intensity of social preferences and social structure.

$a_{ii} = 0$ formalizes the conformity interpretation for endogenous effects.

T.2. ensures that expected utility is well-defined for a large set of potential strategies.

A strategy for individual i is a function that assigns an action to each of his possible types, a function $f_i : \mathbf{R}^{N+1} \to \mathbf{R}$.

A Bayes-Nash equilibrium BNE of the game is a strategy profile $(f_i)_{i \in V}$ such that each f_i maximizes $E\left(U_i\left(\omega_i, \omega_{-i}\right) \middle| x, z_i\right)$ where the expectation is taken with respect to the strategies f_{-i} and the common prior ρ.

Theorem 1. Assume the sociomatrices satisfy T.1. For any prior distribution ρ satisfying T.2, there exists a unique BNE. The equilibrium strategy profile can be written in the form

$$f(x,z) =$$

$$\frac{1}{1+\phi}\left(I-\frac{\phi}{1+\phi}A\right)^{-1}(\gamma I+\delta C)x+\mu(x,z)+\frac{1}{1+\phi}z$$

$\mu(x,z)$ captures the role of higher order beliefs. $\mu(x,z)=\mu(z)$ if z independent of x. We can treat as part of constant term and ignore.

An Observationally Equivalent Structure

Suppose that

$$U_i\left(\omega_i,\omega_{-i}\right)=$$
$$\left(\gamma x_i+z_i+\delta\sum_j c_{i,j}x_j\right)\omega_i-\frac{1}{2}\omega_i^2+\phi\sum_j a_{ij}\omega_i\omega_j$$

This preference structure involves spillovers in, say, educational costs rather than conformity.

The first order conditions for this model are, in form, identical to the conformity model. Hence the two preference structures are observationally equivalent.

From a Theoretical to an Econometric Model

In order to render the theoretical model econometrically useful, additional assumptions are needed.

E.1. The support of x has dimension N.

E.2. For all i and j, $a_{ij} > 0$ iff $a_{ji} > 0$. For some i and j $a_{ij} > 0$

E.3. For all i and j, $c_{ij} > 0$ iff $c_{ji} > 0$. For some pair $i \neq j$ $c_{ij} > 0$

E.4. For all i and j, x_i and z_j are uncorrelated.

E.5. At least one of γ and δ is non-zero.

E.1 ensures uniqueness of the reduced form projection of outcomes on observable characteristics.

E.2 and E.3 shrink the size of the parameter space. This simplifies derivations and comes at little cost since magnitudes are not restricted.

E.4. is the standard exogeneity assumption. It means the higher order beliefs term is restricted to $\mu(z)$; we relax this later.

E.5. eliminates the special case in which x has no effect on outcomes.

Data Assumption

K.1. For all i, the analyst observes $\left(\omega_i, x_i\right)$

This simply means individual-level data are available.

Identification

Definition 1. A *structure* s for the linear social network model is a list $\langle \gamma, \delta, \phi, A, C, \rho \rangle$, where γ, δ and ϕ are utility parameters. A and C are peer- and contextual-effects sociomatrices, and ρ is the *a priori* probability distribution on $\mathbb{R} \times \mathbb{R}$. A *model* is a set of structures.

Definition 2. Utility parameters γ, δ and ϕ are *identified in a model M* by B if for all $s, s' \in M$, if $B(s) = B(s')$ then $(\gamma, \delta, \phi) = (\gamma', \delta', \phi')$.

Theorem 2. Under $T.1 - T.2$, $E.1 - E.5$, and $K.1$

i. B, $\mu(z)$ and $\gamma + \delta$ are identified.

ii. γ, δ and ϕ are not identified.

Hence additional information is needed to recover model primitives.

Relation to Extant Literature

Standard models in the literature are special cases of our framework. Note: the literature focuses on first order conditions rather than equilibrium strategy profiles.

Example 1: Linear-in-Means Model

The most common linear interaction model is the linear-in-means model, in which the population is partitioned into nonoverlapping groups g, no intergroup social interactions are present and unweighted averages summarize intragroup interactions.

Letting n^g denote the population size of group g, these restrictions may be expressed as

$$c_{ij} = \frac{1}{n^g} \text{ if } i, j \in g,$$

$$a_{ij} = \frac{1}{n^g - 1} \text{ if } i, j \in g,$$

$$c_{ij} = a_{ij} = 0 \text{ if } i \in g, j \notin g$$

which produces the first order condition

$$\omega_i = \frac{\gamma}{1+\phi} x_i + \frac{\delta}{(1+\phi)\left(n^g - 1\right)} \sum_j x_j + \frac{\phi}{(1+\phi)\left(n^g - 1\right)} \sum_{j \neq i} E\left(\omega_j \mid x\right) + \frac{1}{1+\phi} \varepsilon_i$$

This model has been shown to be identified when n^g varies.

In contrast, the Manski (1993) reflection problem for nonidentification holds for the large sample limit of this model.

$$\omega_i =$$
$$\frac{\gamma}{1+\phi} x_i + \frac{\delta}{(1+\phi)} \bar{x}^g + \frac{\phi}{(1+\phi)} E\left(\bar{\omega}^g \middle| x\right) + \frac{1}{1+\phi} \varepsilon_i$$

Comments

1. Identification does not carry over to large sample approximations, hence such approximations can be misleading when used in studying identification.

2. Linear-in-means approach is a very strong assumption.

Neighborhood Generalizations of the Linear-in-Means Model

A second class of models, typically employed when data on network structure are available, associates with each agent i a group of others to whom he is directly connected. These others define the neighborhood. The effect of the neighborhood on an individual mimics the original linear-means model.

$$c_{ij} = \frac{1}{n^h} \text{ if } j \in h,$$

$$a_{ij} = \frac{1}{n^h} \text{ if } j \in h,$$

$$c_{ij} = a_{ij} = \frac{1}{n^h} = 0 \text{ if } j \notin h$$

Comments

1. This model also represents a substantive restriction on preferences.

2. Identification seems to typically hold, without any requirement on neighborhood size heterogeneity per se.

Questions

How do the positive identification findings link to Theorem 2?

Can disparate results be understood in a common framework?

Can identification be achieved under weaker preference assumptions?

These motivate the rest of the analysis.

Identification with Known Sociomatrices

The extant literature on identification is based on the assumption that the analyst knows the sociomatrices. This is why Theorem 2 does not contradict positive identification findings.

Formally

K.2. A and C are exogenous and known to the analyst *a priori.*

A Condition for Identification

Bramoullé, Djebbari, and Fortin (2009) provide a powerful identification requirement for the traditional linear-in-means model that provides a connection between identification and network structure. The next result extends this to our two-sociomatrix model.

Theorem 3. Suppose T.1-T.2, E.1-E.5, and K.1-K.2

i. If A,C and AC are distinct, then linear independence of I, A, C is necessary and sufficient for identification of γ, δ and ϕ are identified from the joint distribution of ω and x.

ii. If $A \neq C$ and $AC = C$, and $\gamma \neq 0$ then linear independence of I, A, and AC id necessary and sufficient for identification of γ, δ and ϕ are identified from the joint distribution of ω and x.

Characterizing the Failure of Identification

Identification will fail, according to Theorem 3, if

$$\lambda_1 + \lambda_2 c_{ij} + \lambda_3 \sum_j a_{ij} c_{ji} = 0 \text{ for all } i,$$

$$\lambda_2 c_{ij} + \lambda_4 a_{ij} + \lambda_3 \sum_j a_{ij} c_{ji} = 0 \text{ for all } i \neq j$$

One can thus consider the set of sociomatrices (A, C) such that this linear dependence condition holds.

Genericity of Identification

The identification literature on social interactions focuses on specific choices of A and C. Can something more general be said?

We focus on genericity.

Idea: Characterize set of A, C pairs that fulfill conditions for our econometric model. This set will have a certain dimension. Then characterize subset of original set in which identification of utility parameters does not hold. If the dimension of the latter set is lower than the former, identification is generic.

Blume, Brock, Durlauf, and Ioannides (2011) examine the case in which $A = C$ under much stronger error assumptions.

The relaxation of the error assumptions is relatively straightforward, albeit important for empirical relevance.

Relaxation of assumption $A = C$ leads to different results. In particular, one needs some additional structure beyond compatibility with T.1-T.2 and E.1-E.5 which do not arise under $A = C$.

Corollary 1.

For each sociomatrix A such that the network contains at least one component of size 3, there is a generic subset of contextual sociomatrices C such that γ, δ, and ϕ are identified.

Corollary 2.

Suppose that there are distinct individuals i and j who are connected by a series of links, some in the endogenous effects network and some in the contextual effects network, but who are not connected in either network in isolation. Then, γ, δ, and ϕ are identified.

Corollary 3

Suppose that there is a component of the contextual-effects matrix C such that the all c_{ij} are equal and assume that any component of of peer effect network is either a subset of or disjoint from the contextual effects component. Suppose there also exist two pairs of individuals $i \neq j$ and $k \neq l$ in the contextual effects component such that $a_{ij} \neq a_{kl}$. Then, γ, δ, and ϕ are identified.

Note: The linear-in-means model assumes all elements of A are equal and all elements of C are equal. A single deviation among the a_{ij}'s produces identification.

Why is Identification Generic under Such Conditions?

The first order conditions for linear social interactions models have a very similar structure to linear simultaneous equations systems.

A priori knowledge of *A* and *C* reduces the number of unkwown parameters to 3.

Identification fails when there is too much symmetry in the system.

ii. aggregation

Classroom-level and village-level data often come aggregated. For example, an education data set may contain observation on mean outcome and mean characteristics of many classrooms.

Glaeser, Sacerdote, and Scheinkman (1996,2003) and Graham (2008) study this case.

For our context, we think of data as drawn from nonoverlapping groups. Utility parameters are constant across groups, but groups may differ in size and in values of sociomatrices.

Assumptions

K.1'. For all g, A^g and C^g are exogenous and known to the analyst a priori.

K.2'. For all g, the analyst observes $\left(\bar{x}^g, \bar{\omega}^g\right)$

K.3'. $\operatorname{var}\left(\bar{x}^g\right)$ and $\operatorname{var}\left(\omega^g\right)$ are observed.

E.6. For x_i^g and ε_i^g are iid.

This stronger assumption on the unobserved heterogeneity means that

$$Var\left(\bar{\omega}^g\right) = \frac{1}{\left(n^g\right)^2}\sum_{j\in g}\left(\sum_{i\in g} B_{ij}^g\right)^2 \sigma_x^2 + \left(\frac{1}{1+\phi}\right)^2 \frac{1}{n^g}\sigma_\varepsilon^2$$

which follows from

$$\bar{\omega}^g = \mu^\varepsilon + \frac{1}{n^g}\sum_{j\in g}\left(\sum_{i\in g} B_{ij}^g\right) x_j^g + \frac{1}{1+\phi}\frac{1}{n^g}\sum_{i\in g}\left(\varepsilon_i^g - \mu^\varepsilon\right)$$

Theorem 4. Under T.1-T.2, E.1-E.6, and K.1-K.3, Suppose that γ, δ, and ϕ nonzero. Suppose that $G = 5$ and that for each $g = 1...5$, $n^g \geq 3$ and no A^g is bistochastic. Then the set $J \subset \prod_g M_C^g$ of matrices $C^1...C^5$ such that $\nu_1...\nu_5$ does not identify γ, δ, and ϕ is of lower dimension than $\prod_g M_C^g$.

Theorem 4 says when data are in the form of group averages, second moments can be used to identify the utility parameters provided that the $\left(A^{g}, C^{g}\right)$ pairs fulfill a condition on C_{g} that, given A_{g}, holds generically.

In addition, one needs a certain degree of variation across the sociomatrices to allow for the different groups to provide distinct second moments from which the utility parameters can be backed out.

Identification with Mixed Individual-Level and Aggregate Data

We conclude this section by considering linear social interactions models that are based on a combination of individual-level and aggregate data.

A number of studies, including many in the important first generation of empirical social interactions research, combine individual-data from the Panel Study of Income Dynamics (PSID) with averages of individual outcomes based on the geographic identifiers in the PSID.

The sampling scheme for the PSID, when combined with aggregate information, produces regressions of the form

$$\omega_i = b_0 + b_1 x_i + b_2 \bar{x}^g + \eta_i$$

where g denotes the relevant level of aggregation.

This regression, to be interpretable as an equilibrium strategy profile, requires that the linear-in-means sociomatrices are the true ones, and that the aggregation level defines actual social groups.

Since the sampling scheme we describe provides no information on A_g and C_g, this equation represents an information reduction relative to the equilibrium best response function that describes ω_i. We have already showed is not identified when these matrices are unknown. This individual/aggregate regression is misspecified so its parameters will depend on $\gamma, \delta, \phi, A^g$ and C^g.

The one positive use of this equations is that if $b_2 = 0$, then neither peer nor contextual effects are present in the preferences of agents.

Identification with Partial Information on Social Structure

The assumption that the sociomatrices A and C are known is clearly very strong.

The common weights in the linear-in-means model are not theoretically motivated. Models employing empirically generated sociomatrices assume that these matrices are functions of observed adjacency matrices, which is also not theoretically motivated.

It turns out that identification or partial identification holds under weaker information assumptions.

i. unknown peer-effects sociomatrices

We first consider the case wherein the contextual effects sociomatrix is *a priori* knowledge, but the peer-effects sociomatrix is unknown to the econometrician.

A priori knowledge of contextual effects is more natural than endogenous effects as the latter is psychological. In contrast, public goods games can provide a priori knowledge of C. Of course, plausibility will depend on "context."

Additional Assumptions

E.6. $-\gamma/\delta$ is not an eigenvalue of C.

E.7. $\phi > 0$.

E.6 ensures that B is nonsingular.

E.7. is for analytical convenience.

Theorem 5.

Assume that contextual-effects sociomatrix C is known *a priori*. Assume too that the peer-effects sociomatrix A is unknown but the peer-effects network is known *a priori*. Suppose that $N \geq 3$. Suppose that there are two distinct individuals j and i who are known to be unconnected in the peer-effects network. If $B_{ji}^{-1} \neq 0$, then the utility parameters γ, δ and ϕ are identified from the conditional mean of ω given x.

This theorem is of interest because under a weak type of partial knowledge on A, identification holds.

One can do even better using second moments.

Theorem 6.

Suppose only the contextual-effects sociomatrix C is known. If $N \geq 3$, then for each $C \in M_C$ there is a generic subset $S_A \subset M_A$ such that if $A \in S_A$, then γ, δ and ϕ are also identified in M'_C by the conditional mean of outcomes given characteristics.

Finally, one can even recover the A matrix.

Corollary 4.

Under the conditions of theorem 6, the peer-effects sociomatrix A is identified.

Why is this possible? The dimension of the set of peer effects matrices is $N(N-2)$. The dimension of the set $B((m : C = C^*))$ for a fixed C^* is no more than $N(N-1)+1$, but we can show it to be no less than $N(N-2)$. One needs to recover $N(N-2)+3$ parameters from B; the corollary shows this is possible.

ii. Identification with a priori “qualitative” network knowledge

Data sets such as AddHealth provide information on direct links between agents. They do not report intensities of the bilateral interactions.

These data provide adjacency matrix information. This provides a path to identification that is analogous to the use of exclusion restrictions in classical simultaneous equations analysis.

From the reduced form, self-consistency of beliefs means that for agents agents other than j, $E(\omega_j) = b_j x$. Therefore

$$\omega_i =$$

$$\frac{\gamma + \delta c_{ii}}{1+\phi} x_i + \frac{\delta}{1+\phi} \sum_{j \sim_C i} c_{ij} x_j + \frac{\phi}{1+\phi} \sum_{j \sim_A i} a_{ij} b_j x + \frac{1}{1+\phi} \varepsilon_i$$

$$= \pi_1 x_i + \sum_{j \sim_C i, j \neq i} \pi_{2j} x_j + \sum_{j \sim_A i} \pi_{3j} b_j x + \frac{1}{1+\phi} \varepsilon_i$$

where $j \sim A^i$ means $a_{ij} \neq 0$, etc.

Theorem 7.

Assume T.1–T.2, E.1, E.4, E.5, and K.1. Suppose that the only a priori information about A and C is, for some individual i, the sets $\{j : j \sim_A i\}$ and $\{j : j \sim_C i\}$. For an individual i, consider the following three conditions.

1. $\#\{j \nsim_C i\} + \#\{j \nsim_A i\} \geq N - 1$.
2. $N - 1 > \#\{j \nsim_C i\} \geq \#\{j \nsim_A i\}$.
3. $\#\{j \nsim_A i\} \geq \#\{j \nsim_C i\}$.

If conditions 1 and 2 are satisfied, then for each γ and δ there is a generic set of contextual-effects matrices C such that the utility parameters are identified. If conditions 1 and 3 hold, then there is a generic set of peer-effects matrices A such that the utility parameters are identified.

Key to Proof

Consider

$$\omega_i = \frac{\gamma}{1+\phi} x_i + \frac{\delta}{(1+\phi)} \sum_j c_{ij} x_j + \frac{\phi}{(1+\phi)} \sum_j a_{ij} E(\omega_j | x) + \frac{1}{1+\phi} \varepsilon_i$$

Replace $E(\omega_j | x)$ with $proj(\omega_j | x)$; conditions of theorem essentially require that regressors in this expression are linearly independent after this substitution. This is classical simultaneous equations!

Implications

Theorem 7 says that under certain conditions, holes in the network can enable identification: they serve as exclusion restrictions, which can be exploited to back out the utility parameters. Theorem 7's necessary condition is an order condition.

One way to understand this theorem is that the individual-level equation is similar to the second stage of a two-stage least-squares estimation procedure.

The differences between the classical results and ours lay in the fact that in the classical case identification of the second-stage parameter estimates is not an issue, and the rank and order conditions have to do with backing structural parameter estimates out of the second-stage estimates. Here, in contrast, the issue is identifying the parameters π_i from which the structural parameters need to be recovered..

Finally, note that typically $i \sim C^i$ and in this case the necessary condition is that the total number of links emanating from i in either network not exceed N.

This means that sparse networks are needed for identification based on adjacency matrices.

Our results also suggest a potentially serious limitation in current surveys, specifically Addhealth, which is arguably the most popular data set for the study of social network effects.

Its main draw is that high school students in its nationally representative sample are interviewed not only about the usual demographic and outcome variables of interest, but also about who their friends are.

Unfortunately, the data set's friendship questions are restricted in that each student is allowed to name up to 5 friends of each gender. This has important ramifications in view of the result in theorem 5, which indicates that it is more useful to know who is not someone's friend rather than who is.

Moreover, the restriction on the number of friends means that the failure to identify someone as a friend does not mean that there is a corresponding zero in the associated sociomatrices.

While the limitation on the number of friends that could be named in the interviews has long been understood as inducing measurement error in network structure, as far as we know, the effects of this limitation on identification per se have not been recognized.

iii. identification with aggregated social network data.

We conclude with an evaluation of the individual/group average pairing discussed above. Following our earlier discussion, we focus on the case in which the peer effects matrix is unobserved and.

K.1′′. C is exogenous and known to the analyst.

K.2′′. The analyst observes $\left(\bar{\omega}^g, \bar{x}^g, \omega_1, x_1\right)$.

E.8. x_i is i.i.d. across members of g

Under these information assumptions, available information are summarized by

$$E\left(\bar{\omega}\middle|x_i,\bar{x}^g\right)=\mu^g+b^gE\left(x\middle|x_i,\bar{x}^g\right)+b_{gi}x_i$$
$$E\left(\omega_i\middle|x_i,\bar{x}^g\right)=\mu_i+b_{-i}E\left(x\middle|x_i,\bar{x}^g\right)+b_{ii}x_i$$

Theorem 8.

Assume T.1–T.2 and E.1–E.4 and E.8, and suppose that $E\left(\bar{\omega} \middle| x_i\right)$. Assume that for some i, $B_{A\neg C}$ has full row rank, $\#\left\{j : j \sim_A i\right\}$. Then $\beta = \gamma + \delta$ is identified, and ϕ is not identified.

Theorem 8 shows that the assumption of a linear-in-means structure entails too great a loss of information to allow for identification of the utility parameters. As in other cases, if the projection of ω_i onto x_i and x_{-i} differs from the projection of ω_i onto x_i, then all one can say is that some sort of social interaction is present.

This is a cautionary message given the ubiquity of these models in empirical practice.

Endogeneity of Social Structure

A natural way to extend our model of social interactions to network formation is to make it the second stage of a two-stage game, in which networks are formed in the first stage and actions are determined in the second. For each possible network there exists a unique second stage equilibrium, and each individual's expected utility of this second-stage equilibrium is a value function for the network which gives payoffs for the first-stage game.

Self-selection and its attendant problems arise when both x and ε are available to individuals when choosing their first-stage action.

In these models each individual's utility of a given network is the expected utility of equilibrium outcomes with that network in the second stage, conditional on information available in the first stage.

Suppose that x and ε are available at the outset of the first stage. The expected second state payoff will depend upon both of these variables, and so both will influence individuals' first-stage choices. Consequently, an individual i, observing that he is linked to j, and seeing x_j, can make an inference about the value of z_j that is informed by x_j.

Thus E.4 is violated as $\mu(x,z)$ is not independent of x and second-stage equilibrium will no longer be linear in x. (They are, however, still described by theorem 1.)

This is the selection problem.

It is not just a statistical issue — it affects the basic structure of equilibrium, because it affects inference not only of the econometrician but of individuals constructing the network.

We emphasize, however, that if either the public types or the private types relevant for the second-stage choice are not observed at the time the network is formed, then the missing variable cannot enter into the first-stage interim payoff functions, and so the linear structure of the second stage is maintained.

A formidable problem for the construction of the two-stage game is that there is simply no general theoretical model of network formation. Networks for business relations, job search and classroom friendships are formed according to very different rules, and vary greatly in the degree to which they are instrumental for the second-stage game. Particular network-formation games that appear from time to time in the literature are not compelling.

Perhaps the most interesting strategic approach to network formation is to imagine conditions that should be properties of equilibrium outcomes for many different games. This path, first travelled by Gale and Shapley (1962), leads to network stability concepts such as pairwise stability (Jackson and Wolinsky 1996) and pairwise-Nash stability (Calvó-Amrengol and Ilkiliç 2009).

A network is pairwise-Nash stable if and only if a) no individual wants to drop any links, and b) there is no missing link that if added would, ceteris paribus, be a Pareto improvement for the individuals it connects. It is neither a strictly cooperative nor a strictly non-cooperative concept. Stability expresses the idea that breaking relations is a non-cooperative activity while forming new relations involves mutual consent.

While pairwise-stability is a promising approach to modeling network formation, it has one fundamental problem.

Stable networks are not always guaranteed to exist. For some parameter values and realizations, the first-stage conditional expected values of playing the second stage in various network configurations might be such that the set of stable networks is empty. Nonexistence can be overcome by introducing random stable networks. But these may assume away the selection problem.

Second, there is a converse problem of multiplicity of pairwise stable networks.

Multiplicity means that there may not be sufficient empirical content in stability to address self-selection.

Third, this approach takes a dessicated view of the reason for network formation.

It is very sensible to think that networks are formed for factors other than the second stage of a game.

What Can be Done?

An alternative approach to modelling a two stage game is to study the second stage and account for the fact that in those cases where latent variables are believed to play a role in network formation, $\mu(x,z)$ in the equilibrium strategy profile (will no longer be independent of x; that is, the conditional expectation of regression residuals is no longer uncorrelated with the relevant variables.

Calculation of the full equilibrium of the two-stage game is not necessary for identification.

The source of the failure of identification is $E\left(\mu(x,z)\middle|x\right)$.

This nothing more than a control function.

To be clear, the robustness of identification to endogenous network formation exploits the quadratic game structure that leads to linear equilibrium strategies profiles. But this is true for general control function approaches; they break down when $E\left(\mu(x,z)\middle|x\right)$ is linear in x.

Where would instrumental variables approaches come into play, from the vantage point of our two stage game? Suppose that the researcher has available a vector of observable individual attributes v. From the vantage point of this two stage game, the critical question involves the timing by which this information is revealed.

If agents observe v by the outset of the second stage, then endogenous network formation means that one needs to analyze $E\left(z_i \middle| x, v\right)$

But this means that v no longer constitutes an instrument, since it is correlated with the errors in the regressions that emerge in the second stage of the game.

In this sense, the pro forma use of instruments on the grounds that they are associated with the payoffs of network formation and not behaviors conditional on the network, is invalid.

Once one introduces instruments to account for network endogeneity, one needs to account for their implications for the second stage regression errors, which will, outside of special cases, be present even if the payoff in the second stage is independent of the instrument.

Next Steps

In terms of future research, there are several obvious directions.

First, further mapping of assumptions/possibilities frontier is needed.

Second, estimation issues need to be addressed.

Third, control function approach needs to be elaborated.

Fourth, information in group composition and prices should be explored