PEER EFFECTS AND ENDOGENOUS SOCIAL INTERACTIONS

KOEN JOCHMANS*
UNIVERSITY OF CAMBRIDGE

September 11, 2020

Abstract

We introduce an approach to deal with self-selection of peers in the linear-in-means model. Contrary to the existing proposals we do not require to specify a model for how the selection of peers comes about. Rather, we exploit two restrictions that are inherent to many such specifications to construct intuitive instrumental variables. These restrictions are that link decisions that involve a given individual are not all independent of one another, but that they are independent of the link decisions between other pairs of individuals. We construct instruments from the subnetwork obtained on leaving-out all one's own link decisions in a manner that is reminiscent of the approach of Bramoullé, Djebbari and Fortin (2009) when the assignment of peers is assumed exogenous. A two-stage least-squares estimator of the linear-in-means model is then readily obtained.

Keywords: clustered sample, instrumental variable, linear-in-means model, network, self-selection

JEL classification: C31, C36

Introduction

The importance of acknowledging the existence of social interactions between agents in the estimation of causal relationships is now widely recognized. In a program-evaluation problem, for example, non-treated individuals can nonetheless benefit from the program

*Address: University of Cambridge, Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom. E-mail: kj345@cam.ac.uk.
Financial support from the European Research Council through grant n° 715787 (MiMo) is gratefully acknowledged.
The first public version of this paper dates from August 18, 2020 and is available as arXiv:2008.07886v1 [econ.EM].

through spillovers from treated units with whom they interact. Examples of this are detailed in Miguel and Kremer (2004), Sobel (2006), and Angelucci and De Giorgi (2009). A key concern when estimating models that feature peer effects is that agents may self-select their peers, and do so based on (unobserved) factors that equally feature in the equation of interest, thus creating an endogeneity problem. Randomized assignment to peer groups has proven useful in circumventing this threat to identification (Sacerdote 2001 contains an early application of this strategy) but this is, of course, not possible in many situations. The literature has worked on approaches to deal with the self-selection problem in the linear-in-means model of social interactions as introduced in Manski (1993) and analyzed by Bramoullé, Djebbari and Fortin (2009) and Blume, Brock, Durlauf and Jayaraman (2015) for the case where individuals interact through a general, but fixed, network. The current paper is an addition to this growing body of work. A recent review is provided by Bramoullé, Djebbari and Fortin (2020).

To deal with endogeneity of the network Goldsmith-Pinkham and Imbens (2013) and Hsieh and Lee (2016) complete the linear-in-means model with a parametric specification of the link-formation process. Distributional assumptions on the unobservables allow to write down the likelihood of the full model. Arduini, Patacchini and Rainone (2015) and Johnsson and Moon (2019) (see also Auerbach 2019 for closely related work in a somewhat different context) weaken some of these requirements and propose two-step control-function approaches in the spirit of Heckman (1979) and Heckman and Robb (1985) for dealing with sample selection in cross-sectional data. These methods require data on large networks that are sufficiently dense. As such they are not well suited for the conventional sampling paradigm where we observe many networks, possibly of a small size; a typical example would be schools or classrooms.[1] Furthermore, they continue to be subject to the usual limitations of the control-function approach, which are its lack of robustness to misspecification of the link-formation process and its incapability of handling multi-dimensional sources of endogeneity.

[1] Manski (1993, p. 537) provides a discussion on the incompatibility of the linear-in-means model with different types of sampling schemes.

The limitations of the control-function approach can be sidestepped by taking an instrumental-variable route. Kelejian and Piras (2014), and later also Lee, Liu, Patacchini and Zenou (2020), followed such path, regressing link outcomes on exogenous variables that are presumed to drive the link decisions to cleanse them from endogenous factors. This is a general and simple technique although, of course, it presumes that such exogenous variables are available to the econometrician. However, the link predictions so constructed will tend to be poor predictors of actual link decisions unless the latter are mainly driven by the exogenous variables in question. A discussion on this is in Lee, Liu, Patacchini and Zenou (2020), and we equally observed this in our own numerical experiments (not reported on here further).

Here, in stead, we exploit two restrictions on network formation that are implicit in most network-formation models investigated in the literature to generate instrumental variables that are internal to the model, in the same vain as in a dynamic panel data model. These restrictions are that (i) link decisions of a given individual are dependent, but that (ii) link decisions involving any two distinct pairs of agents are (conditionally) independent. Condition (ii) limits the degree of the endogeneity problem. In turn, the implication of Condition (i) is that link decisions between any triple of individuals are informative about each other. Together, these conditions pave the way for the construction of instrumental variables.

Conditions (i) and (ii) are sufficiently general to cover both settings where networks are formed cooperatively or non-cooperatively, allow for the possibility of transfers between individuals, and accommodate homophily of unrestricted form, for example. They are satisfied in the models of Auerbach (2019) and Johnsson and Moon (2019), and of Arduini, Patacchini and Rainone (2015), for example, as well as in more general versions thereof. Importantly, our conditions do rule out interdependent link formation behavior such as that stemming from transitivity, where individuals are more likely to link if they have more connections in common. This is equally ruled-out in the control-function approach as, there, such a mechanism would result in an incomplete model, rendering the parameters of the network-formation model set- as opposed to point-identified, causing the approach to break down.

In the linear-in-means model the outcome of a given individual depends on the average outcome and the average characteristics of her peers, as well as on her own characteristics. When peer groups are exogenous only the first of these peer effects creates an endogeneity problem. The approach of Bramoullé, Djebbari and Fortin (2009), in essence, instruments the average peer outcome by the average characteristics of the peers of peers. We, in stead, are faced with a situation in which both types of peer effect are endogenous. We construct instrumental variables as follows. For each individual we set up the subnetwork obtained on removing all links in which this individual is involved. Under our conditions this leave-own-out network is exogenous and contains useful predictive information about the individuals own link behavior. Next, we instrument average peer characteristics by the average of these characteristics in the leave-own-out network. In the same way, we instrument average peer outcomes by the average of the characteristics of peers of peers in the leave-own-out network. Like in the exogenous case, the procedure can be iterated to involve characteristics of peers further away in the network. This is an intuitive extension of Bramoullé, Djebbari and Fortin (2009). The resulting two-stage least-squares procedure is standard to implement and generates the usual procedures to test for network endogeneity through a Durbin-Wu-Hausman test. Under a sampling scheme where we observe many independent networks, the estimator's asymptotic distribution follows from Hansen and Lee (2019).

Below we first set up the model, state our identifying assumptions, and motivate them by showing how they are implied in the settings of Auerbach (2019) and Johnsson and Moon (2019), and of Arduini, Patacchini and Rainone (2015) under much less assumptions than are maintained there. We then give our instruments as derived from the leave-own-out networks. After this we present the resulting two-stage least-squares estimator along with its large-sample distribution under a set of quite weak regularity conditions. Results from a simulation experiment show that this estimator performs well under in a setting that is not amenable to existing techniques.

1 Setup

Our asymptotics will involve data on many networks but, for now, it suffices to consider a single network.

Model Consider a network involving n agents. Let $\boldsymbol{A}$ denote its $n \times n$ adjacency matrix;

$$(\boldsymbol{A})_{i,j} = \begin{cases} 1 & \text{if } j \text{ is a peer of } i \\ 0 & \text{otherwise} \end{cases}.$$

The agents j for which $(\boldsymbol{A})_{i,j} = 1$ are called the neighbors of agent i. As usual, we do not consider agents to be linked with themselves, so matrix $\boldsymbol{A}$ has only zeros on its main diagonal. Note that we allow $\boldsymbol{A}$ to be asymmetric, thus covering both directed and undirected betworks. It will be useful to have a notational shorthand for the row-normalized adjacency matrix, $\boldsymbol{H}$, say;

$$(\boldsymbol{H})_{i,j} = \begin{cases} (\boldsymbol{A})_{i,j} \Big/ \sum_{j'=1}^{n} (\boldsymbol{A})_{i,j'} & \text{if } \sum_{j'=1}^{n} (\boldsymbol{A})_{i,j'} > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Recall that $\boldsymbol{H}$ corresponds to the transition matrix of a random walk through our network. Moreover, $(\boldsymbol{H})_{i,j}$ is the probability that, when taking a single step, starting at agent i, we arrive at agent j. In the same way, $(\boldsymbol{H}^2)_{i,j}$ is the probability of arriving in two steps, and so on.

Let y_i and x_i denote scalar variables, observable for each agent. Our baseline model is

$$y_i = \alpha + \beta x_i + \gamma \Big(\sum_{j=1}^{n} (\boldsymbol{H})_{i,j} x_j \Big) + \varepsilon_i,$$

where ε_i is a mean-zero unobserved variable. Taking the regressor to be a scalar is done only for notational convenience. Here, β captures the direct effect of x_i on y_i while γ reflects an indirect, spillover, effect from the covariate values of the neighbors. In matrix form we can succinctly write

$$\boldsymbol{y} = \alpha \boldsymbol{\iota}_n + \beta \boldsymbol{x} + \gamma \boldsymbol{H} \boldsymbol{x} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\iota}_n = (1, \ldots, 1)'$ is the n-vector of ones, $\boldsymbol{x} = (x_1, \ldots, x_n)'$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$. An extension of the baseline specification that accommodates endogenous peer effects, where y_i also depends on $\sum_{j=1}^{n}(\boldsymbol{H})_{i,j}y_j$, gives rise to what we will call the full model,

$$\boldsymbol{y} = \alpha\boldsymbol{\iota}_n + \delta\boldsymbol{H}\boldsymbol{y} + \beta\boldsymbol{x} + \gamma\boldsymbol{H}\boldsymbol{x} + \boldsymbol{\varepsilon},$$

which is the workhorse linear-in-means model on general networks as studied in Bramoullé, Djebbari and Fortin (2009) and Blume, Brock, Durlauf and Jayaraman (2015). This is more complicated because $\boldsymbol{H}\boldsymbol{y}$ would be endogenous even if peer selection is exogenous. It will be useful to first deal with the baseline model. The extension to the full model will then be intuitive.

Restrictions Identification of the slope coefficients in our model is well-understood when the strict exogeneity condition $\mathbb{E}(\varepsilon_i|\boldsymbol{A},\boldsymbol{x}) = 0$ holds. Here we relax this restriction by allowing for dependence between the link decisions and the unobserved component in our model. We work with

$$\mathbb{E}(\varepsilon_i|\boldsymbol{A}_i,\boldsymbol{x}) = 0, \tag{1.1}$$

where $\boldsymbol{A}_i$ is the $(n-1)\times(n-1)$ adjacency matrix of the subnetwork obtained from $\boldsymbol{A}$ on deleting its ith row and its ith column. This condition implies unconditional moments that can be used in a two-stage least-squares procedure. For our instruments to be relevant we will presume that

$$\mathbb{E}((\boldsymbol{A})_{i,j}|\boldsymbol{A}_i,\boldsymbol{x}) \neq \mathbb{E}((\boldsymbol{A})_{i,j}|\,\boldsymbol{x}). \tag{1.2}$$

This condition states that the link decisions of a given agent are not independent of one another, conditional on the covariate, and is natural in our context. Before turning to our instrumental-variable approach we provide motivation and justification for the conditions in (1.1) and (1.2).

Network formation It is useful to start with the model for link decisions specified in Auerbach (2019) and Johnsson and Moon (2019). They stipulate that, for each pair of agents (i,j), with $i<j$,

$$(\boldsymbol{A})_{i,j} = (\boldsymbol{A})_{j,i} = \begin{cases} 1 & \text{if } h(\eta_i,\eta_j) > u_{i,j} \\ 0 & \text{otherwise} \end{cases}, \tag{1.3}$$

where $\eta_1, \ldots, \eta_n$ are independent scalar random variables, the $u_{i,j}$ are random shocks, and h is a conformable function.[2] Link decisions are allowed to be endogenous because η_i and ε_i are allowed to be dependent. The shocks $u_{i,j}$ are independent of (η_i, η_j) and of $(\varepsilon_i, \varepsilon_j)$ The implication is that the network is exogenous conditional on η_i, i.e.,

$$\mathbb{E}(\varepsilon_i | \eta_i, \boldsymbol{A}, \boldsymbol{x}) = \mathbb{E}(\varepsilon_i | \eta_i).$$

Because $(\boldsymbol{A})_{i,j}$ depends on (η_i, η_j) all link decisions involving agent i correlate with ε_i. However, for all $i' \neq i$ and $j \neq i$, the link decision $(\boldsymbol{A})_{i',j}$ is independent of ε_i. Consequently,

$$\mathbb{E}(\varepsilon_i | \boldsymbol{A}_i, \boldsymbol{x}) = \mathbb{E}(\mathbb{E}(\varepsilon_i | \eta_i, \boldsymbol{A}, \boldsymbol{x}) | \boldsymbol{A}_i, \boldsymbol{x}) = \mathbb{E}(\mathbb{E}(\varepsilon_i | \eta_i)) = \mathbb{E}(\varepsilon_i) = 0,$$

meaning that our moment restriction in (1.1) holds in the model of Auerbach (2019) and Johnsson and Moon (2019). Furthermore, for all $i' \neq i$, $(\boldsymbol{A})_{i,j}$ and $(\boldsymbol{A})_{i',j}$ are dependent as they are both functions of η_j. Hence, there is predictive information about the former in the latter, and (1.2) is satisfied.

The stylized model in (1.3) can be generalized in a number of ways without jeopardising the validity of (1.1) and (1.2). The features that are embedded in it that are important for our purposes are that (i) η_i only affects link decisions involving agent i; (ii) the $\eta_1, \ldots, \eta_n$ are independent conditionally on the $x_1, \ldots, x_n$; and (iii) the decision to link depends on characteristics of both agents involved. These restrictions allow η_i to be replaced by a vector of agent-specific unobserved heterogeneity and permit link decisions to depend on a set of additional (observable or unobservable) variables. The function h could also be allowed to be pair-specific. These generalizations allow for excess heterogeneity and homophily of unrestricted form.

The specification in (1.3) is suitable when link formation is cooperative. A version of

[2] Auerbach (2019) and Johnsson and Moon (2019) also require that the $u_{i,j}$ be i.i.d. across agent pairs and impose either certain shape restrictions or a complete functional form on the function h to achieve identification. These restrictions are needed as the η_i need to be estimable (at a sufficiently fast rate), up to a strictly-monotone transformation. As this is not important for our developments we do not impose such restrictions here.

it for non-cooperative situations is

$$(\boldsymbol{A})_{i,j} = (\boldsymbol{A})_{j,i} = \begin{cases} 1 & \text{if } h(\eta_i) > u_{i,j} \text{ and } h(\eta_j) > u_{j,i} \\ 0 & \text{otherwise} \end{cases}, \tag{1.4}$$

where, now, $u_{i,j}$ and $u_{j,i}$ are pair-specific shocks. Clearly, such a specification also satisfies all requirements for (1.1) and (1.2) to hold. Again, (1.4) can be extended in a variety of ways without compromising this.

In (1.3) and (1.4) the source of endogeneity, η_i, has the interpretation of an unobserved agent-specific characteristic. The model of Arduini, Patacchini and Rainone (2015), in contrast, induces dependence via the random shocks, which can be seen as shocks to meeting probabilities between agents. A stripped-down representation of their model, which is for a directed network, is

$$(\boldsymbol{A})_{i,j} = \begin{cases} 1 & \text{if } h(\eta_i, \eta_j) > u_{i,j} \\ 0 & \text{otherwise} \end{cases}, \tag{1.5}$$

where, now, ε_i is dependent on the shocks $\boldsymbol{u}_i = (u_{i,1}, \ldots, u_{i,n})'$ but assumed independent of the heterogeneity terms η_i and η_j. If $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ are all assumed to be mutually independent we have

$$\mathbb{E}(\varepsilon_i | \boldsymbol{u}_i, \boldsymbol{A}, \boldsymbol{x}) = \mathbb{E}(\varepsilon_i | \boldsymbol{u}_i).$$

Under a set of additional restrictions this result can again be used as foundation for the construction of a control-function strategy; see Arduini, Patacchini and Rainone (2015).[3] However, the conditions given here already suffice for (1.1) and (1.2) to hold and, with it, for our instrumental-variable approach to go through.

The above discussion shows that (1.1) and (1.2) are implied by many commonly-used specifications for network formation. The most important limitation of the requirements in (i)–(iii) is that they rule out situations where link decisions are interdependent. Transitivity, for example, where a pair of agents are more likely to be linked when they have more neighbors in common, calls for a simultaneous-equation model. Such a design would violate

[3] These restrictions include the need to consistently estimate the thresholds $h(\eta_i, \eta_j)$; Arduini, Patacchini and Rainone (2015) use a parametric logit specification (as in Graham 2017 and Jochmans 2018) to do so.

(i), as $(\boldsymbol{A})_{i,j}$ will generally depend on all $\eta_1, \ldots, \eta_n$ in such a case. Without access to panel data, as in Goldsmith-Pinkham and Imbens (2013), for example, dealing with such a design appears complicated.

2 Approach

Start with the baseline model. Here, the self-selection of ones' peers causes the spillover effect $\sum_{j=1}^{n}(\boldsymbol{H})_{i,j}x_j$ to be endogenous because the weights $(\boldsymbol{H})_{i,1}, \ldots, (\boldsymbol{H})_{i,n}$ correlate with the unobserved component ε_i. Because $\boldsymbol{H}$ is a row-normalized adjacency matrix, $(\boldsymbol{H})_{i,j}$ depends on all of $(\boldsymbol{A})_{i,1}, \ldots, (\boldsymbol{A})_{i,n}$. It does not depend on $(\boldsymbol{A})_{i',j'}$ for any of $i' \neq i$ and $j' \neq i$, however. By (1.1), the link decisions that do not involve agent i are exogenous. Furthermore, by (1.2), these $(\boldsymbol{A})_{i',j'}$ are not independent of $(\boldsymbol{H})_{i,j}$. This suggests the construction of instrumental variables by looking at linear combinations of $x_1, \ldots, x_n$, with weights coming from the leave-own-out network $\boldsymbol{A}_i$. A fruitful way of doing so is discussed next.

For each i' we define the $n \times n$ matrix

$$(\boldsymbol{H}_{i'})_{i,j} = \begin{cases} (\boldsymbol{A})_{i,j} \Big/ \sum_{j' \neq i'}(\boldsymbol{A})_{i,j'} & \text{if } i \neq i' \text{ and } j \neq i' \text{ and } \sum_{j' \neq i'}(\boldsymbol{A})_{i,j'} > 0 \\ 0 & \text{otherwise} \end{cases} .$$

This is the row-normalized version of the adjacency matrix $\boldsymbol{A}_{i'}$ introduced previously, only complemented with one additional zero row and one additional zero column. This augmentation is done for notational considerations, as it maintains the dimension of these matrices to $n \times n$. We stress that $\boldsymbol{H}_i$ is not obtained from setting to zero the ith row and column of $\boldsymbol{H}$. We can interpret $\boldsymbol{H}_i$ as the transition matrix on the network obtained on ruling-out links that involve agent i. From (1.1), the entries of this matrix are uncorrelated with ε_i. Furthermore, from (1.2), $(\boldsymbol{H}_i)_{i',j'}$ and $(\boldsymbol{H})_{i,j}$ are dependent for all (i,j) and (i',j'), conditional on the regressors.

The construction of these leave-own-out matrices is illustrated graphically in Figure 1. Plot (a) shows a directed wheel graph involving five agents, with Agent 1 at the center and the remaining agents in the periphery. Each arrow represents a link, with its weight

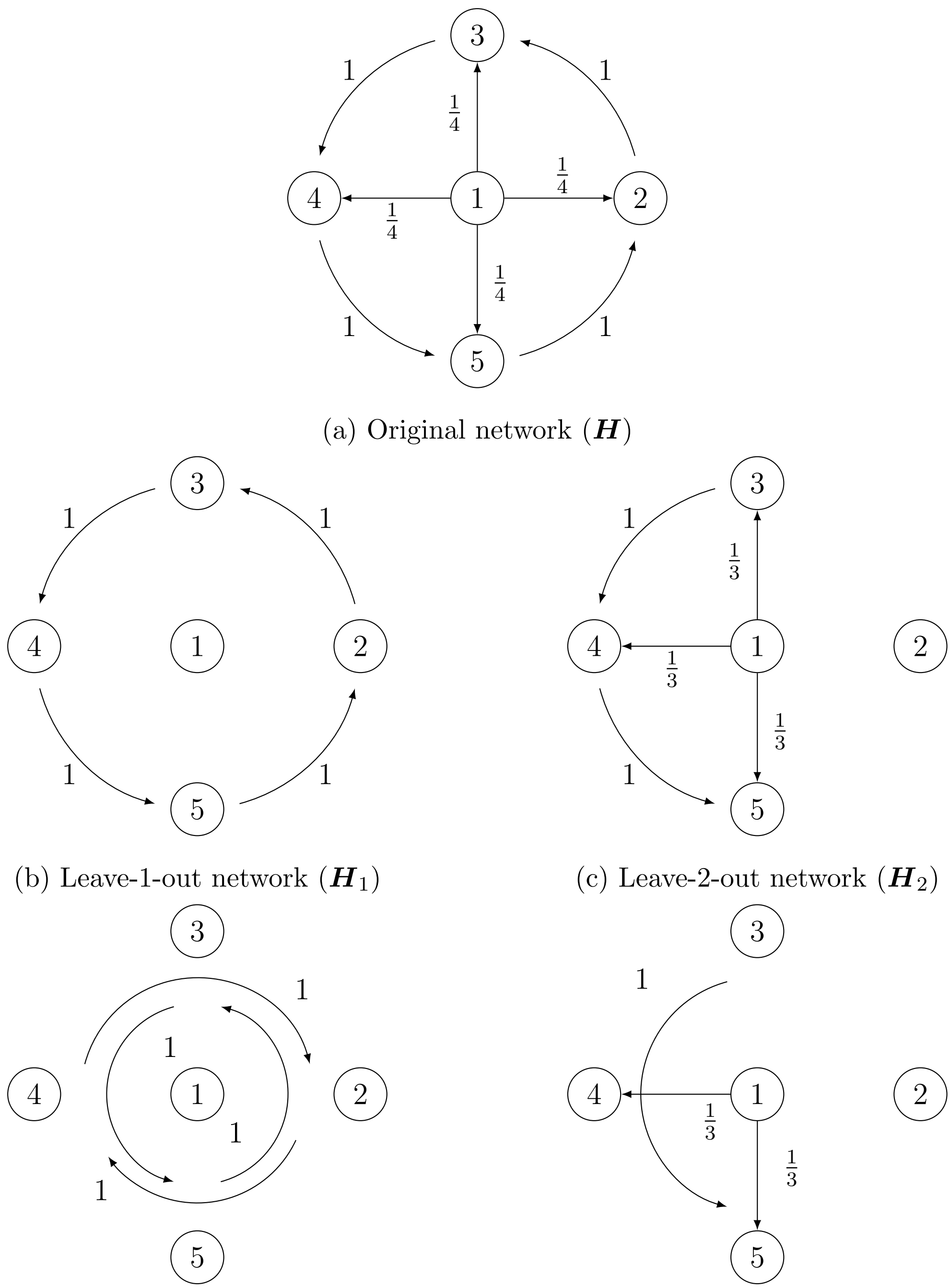
$(\boldsymbol{H})_{i,j}$ given alongside it. In the same way, Plots (b) and (c) give the subnetworks obtained on leaving out Agent 1 or Agent 2, respectively. Note that the weights provided for the links in these subnetworks—given by $(\boldsymbol{H}_1)_{i,j}$ and $(\boldsymbol{H}_2)_{i,j}$, respectively—will generally be different than those in the original network. Because of the symmetry of the problem, the leave-own out graphs for Agents 3,4, and 5 follow as in Plot (b), after a reshuffling of the indices of the agents in the periphery. Plots (d) and (e) give the subnetworks associated with the matrices $(\boldsymbol{H}_1^2)_{i,j}$ and $(\boldsymbol{H}_2^2)_{i,j}$, respectively. These will serve us later on and we will return to them at that point.

Recall that $(\boldsymbol{H})_{i,j}$ is the probability of arriving at agent j, from agent i, in a single step in the network defined by the original adjacency matrix $\boldsymbol{A}$. The entries of the $n \times n$ matrix

$$(\boldsymbol{Q}_1)_{i,j} = (n-1)^{-1} \textstyle\sum_{i' \neq i} (\boldsymbol{H}_i)_{i',j},$$

in contrast, give the probability of arriving at agent j in the network defined by $\boldsymbol{A}_i$, no matter the starting point, in a single step. The average $\sum_{j=1}^n (\boldsymbol{Q}_1)_{i,j} x_j$ is exogenous and will correlate with the spillover term, $\sum_{j=1}^n (\boldsymbol{H})_{i,j} x_j$. A simple intuition can be given by considering the example of network centrality: if agent j is involved in many links, it is likely that she will also be linked with agent i. We may then predict the link decision between agents i and j by looking at the linking behavior of agent j with all the other agents in the network. These choices are exogenous. A two-stage least-squares approach has just suggested itself for the baseline model: we instrument the endogenous spillover effect $\boldsymbol{Hx}$ by $\boldsymbol{Q}_1\boldsymbol{x}$. Other instruments than $\boldsymbol{Q}_1\boldsymbol{x}$ can equally be constructed under our exclusion restriction; some examples follow below. The current choice, however, has a natural extension to the full model.

To visualize the instrumental variable just constructed it may be useful to think about the matrix $\boldsymbol{Q}_1$ as a transition matrix, thus again inducing a network. Plot (a) in Figure 2 shows this network in our directed wheel-graph example. It is substantially different from the original network (as given in Plot (a) of Figure 1 above). While in the original network, given by $\boldsymbol{H}$, the agents in the periphery where connected only by a counter-clockwise circle of links, $\boldsymbol{Q}_1$ also features an additional clockwise circle, as well as new direct links between



(a) Original network ($\boldsymbol{H}$)

(b) Leave-1-out network ($\boldsymbol{H}_1$)

(c) Leave-2-out network ($\boldsymbol{H}_2$)

(d) Leave-1-out network ($\boldsymbol{H}_1^2$)

(e) Leave-2-out network ($\boldsymbol{H}_2^2$)

Figure 1: Transition matrix $\boldsymbol{H}$ for the directed wheel-graph network [(a)] together with transition matrices for its leave-own-out subnetworks $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ [(b) and (c)] and their iterations $\boldsymbol{H}_1^2$ and $\boldsymbol{H}_2^2$ [(d) and (e)].

agents at opposite ends of the circle. The weights assigned to these link are also generally different.

In the full linear-in-means model,

$$\boldsymbol{y} = \alpha \boldsymbol{\iota}_n + \delta \boldsymbol{H}\boldsymbol{y} + \beta \boldsymbol{x} + \gamma \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\varepsilon},$$

the presence of $\boldsymbol{H}\boldsymbol{y}$ as a regressor would induce an endogeneity problem even if $\boldsymbol{H}$ were exogenous. If $-1 < \delta < 1$, and if all the agents in the network are linked to at least one other agent,

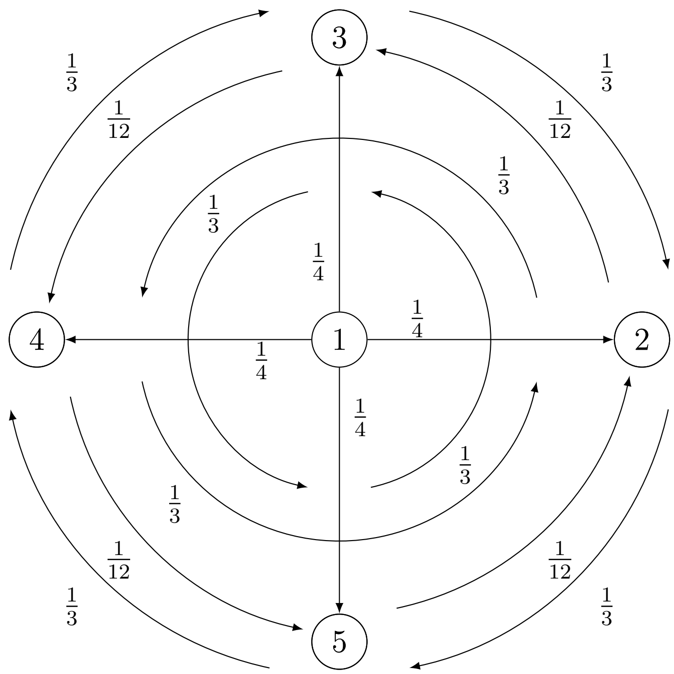
$$\boldsymbol{H}\boldsymbol{y} = \mu\, \boldsymbol{\iota}_n + \beta \boldsymbol{H}\boldsymbol{x} + \lambda \sum_{s=0}^{\infty} \delta^s \boldsymbol{H}^{s+2} \boldsymbol{x} + \sum_{s=0}^{\infty} \delta^s \boldsymbol{H}^{s+1} \boldsymbol{\varepsilon}, \tag{2.1}$$

where we write $\mu = \alpha/(1-\delta)$ and $\lambda = \delta\beta + \gamma$. The argument of Bramoullé, Djebbari and Fortin (2009) and De Giorgi, Pellizzari and Redaelli (2010) is that $\boldsymbol{H}^2\boldsymbol{x}$, $\boldsymbol{H}^3\boldsymbol{x}$, and so on can be used as instrumental variables for $\boldsymbol{H}\boldsymbol{y}$ when the network is exogenous, provided that $\lambda \neq 0$. The validity of these variables as instruments breaks down when the network is endogenous.

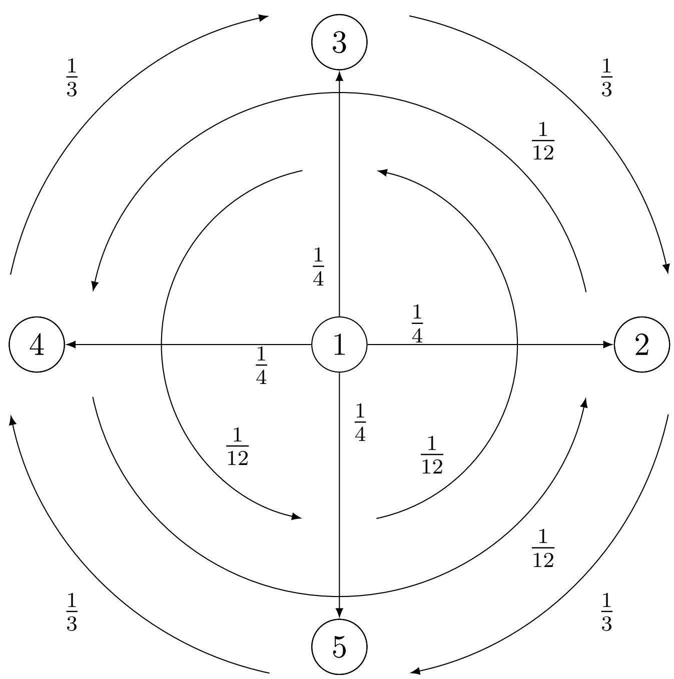
On inspecting the expansion in (2.1) a natural extension to our approach in the baseline model presents itself. Because $\boldsymbol{H}_i$ is a matrix of transition probabilities, it can be iterated on, in the same way as $\boldsymbol{H}$, to yield probabilities of arriving at each agent when taking multiple steps through the network. In full analogy to $\boldsymbol{Q}_1$, the entries of the $n \times n$ matrix

$$(\boldsymbol{Q}_2)_{i,j} = (n-1)^{-1} \textstyle\sum_{i' \neq i} \sum_{j'=1}^{n} (\boldsymbol{H}_i)_{i',j'}\, (\boldsymbol{H}_i)_{j',j},$$

give the probability of arriving at agent j in the network induced by $\boldsymbol{A}_i$, no matter the starting point, in two steps. Under our moment condition in (1.1) these weights are, again, exogenous. This, then, allows to instrument the endogenous right-hand side variables, $\boldsymbol{H}\boldsymbol{x}$ and $\boldsymbol{H}\boldsymbol{y}$, by the exogenous variables $\boldsymbol{Q}_1\boldsymbol{x}$ and $\boldsymbol{Q}_2\boldsymbol{x}$. In light of the above the interpretation of this is immediate. Like in the exogenous case, we require $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ to be sufficiently different. Contrary to Bramoullé, Djebbari and Fortin (2009), it is more difficult here to give simple primitive conditions for instrument relevance, however. In their case (2.1) implies a linear reduced form from which such conditions can be derived. This does not apply here.



(a) Network induced by $\boldsymbol{Q}_1$



(b) Network induced by $\boldsymbol{Q}_2$

Figure 2: Networks induced by the instrumentation matrices $\boldsymbol{Q}_1$ [(a)] and $\boldsymbol{Q}_2$ [(b)] for the directed wheel-graph network

The subnetworks induced by iterating on $\boldsymbol{H}_i$ in our wheel-graph example are given in Plots (d) and (e) of Figure 1 while Plot (a) in Figure 2, in turn, gives the network induced by $\boldsymbol{Q}_2$. Again, iterating steps results in networks that are quite different from their one-step counterparts—in Plots (b) and (c) in Figure 1 and Plot (a) in Figure 2, respectively—with some existing links being severed, other links being given a different weight, and new links being created.

Because the parameters in our model are overidentified we can consider additional instruments. One natural way to do so is by taking additional steps through the network. Letting

$$(\boldsymbol{Q}_s)_{i,j} = (n-1)^{-1} \sum_{i' \neq i} \sum_{j_1=1}^{n} \cdots \sum_{j_{s-1}=1}^{n} (\boldsymbol{H}_i)_{i',j_1} (\boldsymbol{H}_i)_{j_1,j_2} \cdots (\boldsymbol{H}_i)_{j_{s-1},j},$$

for any integer s it is apparent that $\boldsymbol{Q}_s\boldsymbol{x}$ is a valid instrumental variable. Instruments so constructed play a role analogous to $\boldsymbol{H}^s\boldsymbol{x}$ in the approach of Bramoullé, Djebbari and Fortin (2009). Of course, like there, as s increases the transition matrix $\boldsymbol{H}_i^s$ will tend to its steady-state distribution, so that higher iterations will provide increasingly less (additional) information. We also note that other instruments are equally possible. For example, noting that

$$\sum_{j=1}^{n} (\boldsymbol{Q}_s)_{i,j} \boldsymbol{x}_j = \boldsymbol{\iota}_n' \boldsymbol{H}_i^s \boldsymbol{x}/(n-1)$$

is an average of the vector $\boldsymbol{H}_i^s\boldsymbol{x}$, it would be natural to consider second moments like $\boldsymbol{x}'\boldsymbol{H}_i'\boldsymbol{H}_i\boldsymbol{x}$ and $\boldsymbol{x}'\boldsymbol{H}_i'\boldsymbol{H}_i^2\boldsymbol{x}$, and so on.

The condition that $\lambda \neq 0$ is crucial to the approach when the network is exogenous. It requires that $\boldsymbol{x}$ affects $\boldsymbol{y}$, either directly or through $\boldsymbol{H}\boldsymbol{x}$, and also that endogenous and exogenous peer effects do not exactly cancel each other out in the reduced form. Moreover, when $\lambda = 0$ we would have

$$\mathbb{E}(\boldsymbol{H}\boldsymbol{y}|\boldsymbol{H}, \boldsymbol{x}) = \mu\, \boldsymbol{\iota}_n + \beta \boldsymbol{H}\boldsymbol{x}$$

when link formation is exogenous; the $\boldsymbol{H}^s\boldsymbol{x}$, for all $s > 1$, no longer contain predictive information about $\boldsymbol{H}\boldsymbol{y}$, conditional on $\boldsymbol{H}\boldsymbol{x}$. The situation is different when link formation is endogenous. Indeed, here, there will generally still be information on $\boldsymbol{H}\boldsymbol{y}$ in $\boldsymbol{Q}_s\boldsymbol{x}$ coming from the fact that $\mathbb{E}(\boldsymbol{x}'\boldsymbol{Q}_s'\boldsymbol{H}^{p+1}\boldsymbol{\varepsilon}) \neq 0$, for any pair of integers p, s, in this case. This is so because, while the entries of $\boldsymbol{H}_i\boldsymbol{x}$ do not correlate with ε_i, they do correlate with all ε_j for $j \neq i$ when link decisions are endogenous. The implication is that endogeneity of the network can yield identification in settings where no exogenous variables are present in the linear-in-means model.

3 Inference

We now consider a collection of G independent networks, of size $n_1, \ldots, n_G$, respectively, and rechristen $n = \sum_{g=1}^{G} n_g$, the total number of observations. Each of these networks comes with its associated adjacency matrix, $\boldsymbol{A}_g$—and, thus, its row-normalized version $\boldsymbol{H}_g$—as well as with the variables $\boldsymbol{y}_g = (y_{g,1}, \ldots, y_{g,n_g})'$ and $\boldsymbol{x}_g = (x_{g,1}, \ldots, x_{g,n_g})'$, which follow the linear-in-means model. We may write

$$\boldsymbol{y}_g = \boldsymbol{X}_g\boldsymbol{\vartheta} + \boldsymbol{\varepsilon}_g,$$

where we combine all regressors in $\boldsymbol{X}_g = (\boldsymbol{\iota}_{n_g}, \boldsymbol{H}_g\boldsymbol{y}_g, \boldsymbol{x}_g, \boldsymbol{H}_g\boldsymbol{x}_g)$ and collect all parameters to estimate in $\boldsymbol{\vartheta} = (\alpha, \delta, \beta, \gamma)'$. The estimator of $\boldsymbol{\vartheta}$ that we consider here is the two-stage least-squares estimator using the instruments introduced above.

On denoting the instrument matrix as $\boldsymbol{Z}_g$ the estimator can be written in its usual form

$$\boldsymbol{\vartheta}_n = (\textstyle\sum_g \boldsymbol{X}_g'\boldsymbol{Z}_g(\sum_g \boldsymbol{Z}_g'\boldsymbol{Z}_g)^{-1}\sum_g \boldsymbol{Z}_g'\boldsymbol{X}_g)^{-1}(\sum_g \boldsymbol{X}_g'\boldsymbol{Z}_g(\sum_g \boldsymbol{Z}_g'\boldsymbol{Z}_g)^{-1}\sum_g \boldsymbol{Z}_g'\boldsymbol{y}_g).$$

Its large-sample properties can be deduced from the work of Hansen and Lee (2019) on clustered data sets. Their results are particularly well-suited for the problem at hand. In particular, they allow for arbitrary dependence within each network, for networks to grow in size, and do not require data to be identically distributed. The following four conditions need to be imposed:

(i) For some $2 \leq r < \infty$,

$$\frac{\left(\sum_g n_g^r\right)^{2/r}}{n} \leq c < \infty, \qquad \max_g \frac{n_g^2}{n} \overset{n\uparrow\infty}{\longrightarrow} 0,$$

where c is an arbitrary constant.

(ii) For some s with $r < s$, $\sup_{g,i} \mathbb{E}(|y_{g,i}|^{2s}) < \infty$ and $\sup_{g,i} \mathbb{E}(|x_{g,i}|^{2s}) < \infty$.

(iii) The matrices

$$\frac{\sum_g \mathbb{E}(\boldsymbol{Z}_g' \boldsymbol{\varepsilon}_g \boldsymbol{\varepsilon}_g' \boldsymbol{Z}_g)}{n} \qquad \text{and} \qquad \frac{\sum_g \mathbb{E}(\boldsymbol{Z}_g' \boldsymbol{Z}_g)}{n}$$

have minimum eigenvalue bounded away from zero.

(iv) The matrix

$$\frac{\sum_g \mathbb{E}(\boldsymbol{Z}_g' \boldsymbol{X}_g)}{n}$$

has maximal column rank

These four conditions are intuitive. A discussion on (i)–(ii) is in Hansen and Lee (2019). Conditions (iii)-(iv) are nothing else than the usual rank conditions that are needed in any instrumental-variable problem.

By Hansen and Lee (2019, Theorems 8 and 9), the estimator $\boldsymbol{\vartheta}_n$ is consistent as $n \to \infty$ and has a normal limit distribution. The robust estimator of its asymptotic variance equals

$$\boldsymbol{V}_n = (\boldsymbol{S}_n' \boldsymbol{W}_n \boldsymbol{S}_n)^{-1} (\boldsymbol{S}_n' \boldsymbol{W}_n \boldsymbol{\Omega}_n \boldsymbol{W}_n \boldsymbol{S}_n)(\boldsymbol{S}_n' \boldsymbol{W}_n \boldsymbol{S}_n)^{-1},$$

where we use the shorthand notation

$$\boldsymbol{S}_n = \textstyle\sum_g \boldsymbol{Z}_g' \boldsymbol{X}_g, \qquad \boldsymbol{W}_n = (\textstyle\sum_g \boldsymbol{Z}_g' \boldsymbol{Z}_g)^{-1}, \qquad \boldsymbol{\Omega}_n = \textstyle\sum_g \boldsymbol{Z}_g' \hat{\boldsymbol{\varepsilon}}_g \hat{\boldsymbol{\varepsilon}}_g' \boldsymbol{Z}_g,$$

and $\hat{\boldsymbol{\varepsilon}}_g = \boldsymbol{y}_g - \boldsymbol{X}_g \boldsymbol{\vartheta}_n$ are the residuals from the two-stage least-squares procedure. We then have

$$\boldsymbol{V}_n^{-1/2} (\boldsymbol{\vartheta}_n - \boldsymbol{\vartheta}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{I}_4)$$

as $n \to \infty$.

4 Simulations

The procedure was evaluated in a Monte Carlo experiment. We generated networks via the link formation process

$$(\boldsymbol{A})_{i,j} = (\boldsymbol{A})_{j,i} = \begin{cases} 1 & \text{if } \eta_i + \eta_j > c \\ 0 & \text{otherwise} \end{cases},$$

where the η_i are independent standard-normal variates and we set $c = -\sqrt{2}\Phi^{-1}(.25)$, for Φ the standard-normal distribution function. In this way, the unconditional link-formation probability is .25. We then drew $x_i \sim N(1,1)$ and generated outcomes from the full model, inducing endogeneity in link formation by generating

$$\varepsilon_i = \varphi(\eta_i) + u_i, \qquad u_i \sim N(0,1),$$

for different choices of the function φ. The parameters were set as $\alpha = 0$, $\beta = 1$, $\gamma = .5$ and $\delta = .5$. Data were generated for 250 groups, each consisting of 25 agents. Results are presented for the estimator of Bramoullé, Djebbari and Fortin (2009) (TSLS-X) and for our proposal (TSLS-E). The former instruments $\boldsymbol{Hx}$ by itself and $\boldsymbol{Hy}$ by $\boldsymbol{H}^2\boldsymbol{x}, \ldots, \boldsymbol{H}^4\boldsymbol{x}$. This approach is valid when $\varphi(\eta)$ does not depend on η. The latter instruments $\boldsymbol{Hx}$ by $\boldsymbol{Q}_1\boldsymbol{x}$ and $\boldsymbol{Hy}$ by $\boldsymbol{Q}_2\boldsymbol{x}, \ldots, \boldsymbol{Q}_4\boldsymbol{x}$. We use two overidentifying moments to discipline the sampling distribution of the estimators—ensuring that their first two moments exist—so that we can meaningfully report on their bias and standard deviation (see, e.g., Mariano 1972).

Table 1 contains the bias and standard deviation of the estimators, the mean and standard deviation of the implied t-statistics, as well as the empirical rejection frequency of two-sided t-tests (at the 5% significance level). We do not report results for the estimator of the intercept. Four different specifications for φ were considered: (i) a constant, (ii) a linear function, (iii) an exponential function, and (iv) a sine function. All results were obtained over 5,000 Monte Carlo replications and all the variables were redrawn in each iteration.

Table 1: Simulation results

	TSLS-X					TSLS-E				
	$\boldsymbol{\vartheta}_n - \boldsymbol{\vartheta}$		$\boldsymbol{V}_n^{-1/2}(\boldsymbol{\vartheta}_n - \boldsymbol{\vartheta})$			$\boldsymbol{\vartheta}_n - \boldsymbol{\vartheta}$		$\boldsymbol{V}_n^{-1/2}(\boldsymbol{\vartheta}_n - \boldsymbol{\vartheta})$		
	bias	std	mean	std	rate	bias	std	mean	std	rate
					$\varphi(\eta) = 0$					
β	0.0000	0.0130	0.0005	1.0320	0.0578	-0.0002	0.0134	-0.0165	1.0286	0.0550
γ	-0.0002	0.0414	-0.0077	1.0084	0.0542	-0.0002	0.1638	-0.2083	1.0287	0.0600
δ	0.0000	0.0203	0.0241	1.0047	0.0508	-0.0001	0.0610	0.2144	1.0235	0.0594
					$\varphi(\eta) = \eta$					
β	-0.0081	0.0181	-0.4580	1.0244	0.0816	0.0002	0.0180	0.0134	1.0049	0.0532
γ	-0.0833	0.1336	-1.2585	2.0672	0.4930	-0.0003	0.1406	-0.1060	1.0261	0.0572
δ	-0.0833	0.0651	7.7740	3.1595	0.9596	-0.0004	0.0459	0.1075	1.0224	0.0548
					$\varphi(\eta) = \exp(3\Phi(\eta))$					
β	-0.0073	0.0661	-0.1140	1.0071	0.0538	0.0007	0.0669	0.0083	1.0009	0.0524
γ	0.1873	0.4531	0.7775	1.8904	0.2462	-0.0079	0.5682	-0.0597	1.0186	0.0534
δ	0.1515	0.0611	8.1131	3.5276	0.9460	0.0013	0.1326	0.1113	1.0358	0.0626
					$\varphi(\eta) = \sin(3\Phi(\eta))$					
β	-0.0017	0.0130	-0.1336	0.9951	0.0496	-0.0003	0.0133	-0.0209	0.9899	0.0468
γ	-0.0315	0.0436	-0.7777	1.0663	0.1406	-0.0139	0.1552	-0.2240	1.0319	0.0652
δ	0.0379	0.0200	2.4044	1.3197	0.6308	0.0055	0.0572	0.2364	1.0230	0.0624

The estimator of Bramoullé, Djebbari and Fortin (2009) does well when link formation is exogenous. Otherwise, the coefficient estimates are biased, except for those of β. The latter observation can be explained by the fact that link formation is independent of the covariates and that the covariate is independent across agents in our simulation design. The (estimated) standard error (not reported) also tends to substantially underestimate the true variability in the point estimates. This is apparent from inspection of the sample variance of the t-statistics, which is much larger than unity. Together with the presence of bias, this implies that the sampling distribution of the t-statistics is far from a standard-normal distribution. Consequently, the t-test displays large overrejection rates. Using instruments constructed from the leave-own-out networks delivers estimators that are virtually unbiased for all the designs in Table 1. The associated t-statistics have a mean that is close to zero and a standard deviation that is close to unity. Furthermore, the empirical rejection frequencies are close to their nominal size of 5%, and this for all parameters and for all designs. Hence, the normal approximation does well for TSLS-E.

The findings discussed here were confirmed in a larger set of Monte Carlo designs, where the η_i were drawn from asymmetric distributions and network formation also depends on covariates. These results were similar in spirit to those reported here and, hence, are not discussed further here.

References

Angelucci, M. and G. De Giorgi (2009). Indirect effects of an aid program: How do cash injections affect ineligibles' consumption? *American Economic Review 99*, 486–508.

Arduini, T., E. Patacchini, and E. Rainone (2015). Parametric and semiparametric IV estimation of network models with selectivity. Mimeo.

Auerbach, E. (2019). Identification and estimation of models with endogenous network formation. Mimeo.

Blume, L. E., W. A. Brock, S. N. Durlauf, and R. Jayaraman (2015). Linear social interactions models. *Journal of Political Economy 123*, 444–496.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009). Identification of peer effects through social networks. *Journal of Econometrics 150*, 41–55.

Bramoullé, Y., H. Djebbari, and B. Fortin (2020). Peer effects in networks: A survey. *Annual Review of Economics 12*, 603–629.

De Giorgi, G., M. Pellizzari, and S. Redaelli (2010). Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics 2*, 241–275.

Goldsmith-Pinkham, P. and G. W. Imbens (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics 31*, 253—264.

Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica 85*, 1033–1063.

Hansen, B. E. and S. Lee (2019). Asymptotics theory for clustered samples. *Journal of Econometrics 210*, 268–290.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*, 153–161.

Heckman, J. J. and R. Robb (1985). Alternative methods for evaluating the impact of interventions. *Journal of Econometrics 30*, 239–267.

Hsieh, C.-S. and L. F. Lee (2016). A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics 31*, 301–319.

Jochmans, K. (2018). Semiparametric analysis of network formation. *Journal of Business & Economic Statistics 36*, 705–713.

Johnsson, I. and H. R. Moon (2019). Estimation of peer effects in endogenous social networks: Control function approach. Forthcoming in *Review of Economics and Statistics*.

Kelejian, H. H. and G. Piras (2014). Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes. *Regional Science and Urban Economics 46*, 140–149.

Lee, L. F., X. Liu, E. Patacchini, and Y. Zenou (2020). Who is the key player? A network analysis of juvenile delinquency. Forthcoming in *Journal of Business & Economic Statistics*.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies 60*, 531–542.

Mariano, R. S. (1972). The existence of moments of the ordinary least squares and two-stage least squares estimators. *Econometrica 40*, 643–652.

Miguel, E. and M. Kremer (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica 72*, 159–217.

Qu, X. and L. F. Lee (2015). Estimating a spatial autoregressive model with an endogenous spatial weight matrix. *Journal of Econometrics 184*, 209–232.

Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics 116*, 681–704.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association 101*, 1398–1407.