WORKING PAPERS

N° 1348

July 2022

"Peer Effects and Endogenous Social Interactions"

Koen Jochmans

Toulouse School of Economics

PEER EFFECTS AND ENDOGENOUS SOCIAL INTERACTIONS

KOEN JOCHMANS*

TOULOUSE SCHOOL OF ECONOMICS, UNIVERSITY OF TOULOUSE CAPITOLE

This version: July 21, 2022

Abstract

This paper proposes a solution to the problem of the self-selection of peers in the linear-in-means model. We do not require to specify a model for how the selection of peers comes about. Rather, we exploit two restrictions that are inherent in many such specifications to construct conditional moment conditions. The restrictions in question 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 made between other pairs of individuals that are located sufficiently far away in the network. These conditions imply that instrumental variables can be constructed from leave-own-out networks.

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

JEL classification: C31, C36

---

*Address: Toulouse School of Economics, 1 esplanade de l'Université, 31080 Toulouse, France. E-mail: `koen.jochmans@tse-fr.eu`.

I am grateful to the European Research Council for support through grant n° 715787 (MiMo), and to the French Government and the ANR for support under the Investissements d' Avenir program, grant ANR-17-EURE-0010.

I am thankful to the editor, the associate editor, and two referees, as well as to Adam Rosen, Andrei Zeleneev, seminar participants at UCSD and UPF, and participants to the North American Summer Meeting of the Econometric Society in Montréal and the conference on Econometrics of Games, Matching and Networks in Toulouse for much-appreciated feedback.

The Stata module `ivnet` implements the estimator developed in this paper. I am most grateful to Vincenzo Verardi for help in the development of this module.

The first public version of this paper dates from August 18, 2020 and is available as `arXiv:2008.07886v1 [econ.EM]`.

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 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 not possible in many situations. In the context of the linear-in-means model of social interactions (Manski 1993, Bramoullé, Djebbari and Fortin 2009, and Blume, Brock, Durlauf and Jayaraman 2015), a recent literature has worked on approaches to deal with the self-selection problem. A review is provided by Bramoullé, Djebbari and Fortin (2020). The current paper is an addition to this growing body of work.

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, paving the way for inference. Arduini, Patacchini and Rainone (2015) and Johnsson and Moon (2021) (and also Auerbach 2022, albeit in a somewhat different context) weaken some of these requirements and propose two-step control-function approaches (Heckman and Robb, 1985). These methods, however, require data on a single large network that also needs to be sufficiently dense. As such they are not well suited for the conventional sampling paradigm where we observe many, possibly small, networks; typical examples would be schools, classrooms, or neighborhoods (also see Manski 1993, p. 537, for a discussion on the (in)compatibility of the linear-in-means model with different types of sampling schemes.) Furthermore, they are subject to the usual limitation of the control-function approach, which is its lack of robustness to misspecification of the link-formation process.

These limitations of the control-function approach can be sidestepped by taking an instrumental-variable route. Kelejian and Piras (2014) followed one such path, regressing link outcomes on exogenous variables that are presumed to drive the link decisions to cleanse them from endogenous factors. Applications of this idea are also documented in Santavirta and Sarzosa (2019) and Lee, Liu, Patacchini and Zenou (2021). While this is a general and simple technique, it of course presumes that such exogenous variables are available to the econometrician. Furthermore, the instruments 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 can be found in Lee, Liu, Patacchini and Zenou (2021), and we equally observed this in our own numerical experiments (not reported).

Here, instead, 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 when located sufficiently far away from one another in the network. 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 provide relevant conditional moment restrictions that pave the way for the construction of instrumental variables.

Conditions (i) and (ii) are sufficiently general to cover settings where networks are formed either cooperatively or non-cooperatively and allow for the possibility of transfers between agents. They can accommodate degree heterogeneity as well as (dis)assortative matching of unrestricted form (see, e.g., Newman 2010 for definitions of these concepts and discussion on their importance in network-formation models). The conditions are satisfied in quite general dyadic models of link formation, inclusing those of Johnsson and Moon

(2021) and Arduini, Patacchini and Rainone (2015). They do rule out (unrestricted forms of) interdependent link formation behavior such as that stemming from transitivity, where individuals are more likely to link if they have more connections in common. Such behavior poses substantial econometric complications that are difficult to handle with cross-sectional data alone. Interdependency is equally ruled-out in the control-function approach, as such models are typically incomplete.

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, instead, are faced with a situation in which both types of peer effect are endogenous. In the simplest case we construct instrumental variables as follows. For each individual we set up the subnetwork obtained on removing all link decisions in which this individual is involved. Under our conditions this leave-own-out network is exogenous and contains useful predictive information about the individual's 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 yield additional instrumental variables by involving characteristics of peers further away in the leave-own-out network. This is an intuitive modification of Bramoullé, Djebbari and Fortin (2009).

In the sequel we first introduce the linear-in-means model and present the conditional moment restrictions under which we will work. We next validate these restrictions in a general specification of link formation and also provide examples in which they will fail to hold. We then discuss the construction of instruments from leave-own-out networks. The performance of the resulting two-stage least-squares estimator is evaluated in a simulation experiment.

1 Setup

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 networks. 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 going from i to j in two steps, and so on.

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

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

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,

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

This 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). The full model is more complicated because, due to simultaneity, $\boldsymbol{Hy}$ would be endogenous even if peer selection were exogenous.

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 link decisions involving a given agent are not all independent of one another, conditional on the covariate.

A useful extension of the above is as follows. Suppose that agents can be partitioned into (known) groups; let $g(i)$ indicates the group to which agent i belongs and write $|g(i)|$ for its number of members. Then (1.1) and (1.2) can be replaced by the assumptions that

$$\mathbb{E}(\varepsilon_i|\boldsymbol{A}_{-g(i)},\boldsymbol{x}) = 0, \tag{1.3}$$

and

$$\mathbb{E}((\boldsymbol{A})_{i,j}|\boldsymbol{A}_{-g(i)},\boldsymbol{x}) \neq \mathbb{E}((\boldsymbol{A})_{i,j}|\,\boldsymbol{x}), \tag{1.4}$$

respectively. Here, $\boldsymbol{A}_{-g(i)}$ denotes the $(n-|g(i)|)\times(n-|g(i)|)$ adjacency matrix of the subnetwork obtained from $\boldsymbol{A}$ on deleting the rows and columns relating to all link decisions involving any of the members of the group $g(i)$. The conditions in (1.1) and (1.2) essentially restrict the self-selection problem to be caused by agent-specific unobservables that do not affect the decision of any other pair of agents to form a link between them. Conditions (1.3) and (1.4) would allow these drivers to be dependent within groups or even allow group-level unobservables to cause the endogeneity. The conditions are also compatible with interdependent link formation, provided that the interdependency is restricted to within-group interactions (see below).

2 Motivation

Network formation Before turning to our instrumental-variable approach we provide motivation and justification for the conditions in (1.1) and (1.2) in a general model of dyadic network formation.

Let $\boldsymbol{v}_i$ and $\boldsymbol{u}_{i,j}$ be (infinite-dimensional) vectors at the agent level and the dyad level, respectively. Let $\boldsymbol{z}$ denote an (infinite-dimensional) vector that may contain variables that may depend on all agents and all dyads in the network; $\boldsymbol{z}$ may contain agent-specific variables as well as all elements of $\boldsymbol{x}$, for example. The formation of links is determined via

$$(\boldsymbol{A})_{i,j} = \begin{cases} 1 & \text{if } a(\boldsymbol{v}_i, \boldsymbol{v}_j, \boldsymbol{u}_{i,j}, \boldsymbol{z}) > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{2.5}$$

for some function a. This function is not assumed to be symmetric in the labels of the agents. This setup covers both directed and undirected networks and also accommodates both cooperative and non-cooperative link formation. Allowing the variables that enter a to be infinite dimensional yields a fully nonparametric specification that permits unrestricted heterogeneity (see, e.g., Hoderlein and Mammen 2007 for this type of formulation in a different context). The link function may thus differ across dyads and so, for example, how $\boldsymbol{z}$ affects $(\boldsymbol{A})_{i,j}$ may vary across dyads. Equation (2.5) covers general forms of assortative and disassortative matching between agents and allows for non-dyadic spillover effects through $\boldsymbol{z}$, which may include $\boldsymbol{x}$.

We complete the link-formation model in (2.5) with the following restrictions:

A.1 $\mathbb{E}(\varepsilon_i | \boldsymbol{v}_i, \boldsymbol{u}_{i,1}, \ldots, \boldsymbol{u}_{i,n}, \boldsymbol{A}, \boldsymbol{x}) = \mathbb{E}(\varepsilon_i | \boldsymbol{v}_i, \boldsymbol{u}_{i,1}, \ldots, \boldsymbol{u}_{i,n})$.

A.2 $(\boldsymbol{v}_i, \boldsymbol{v}_j, \boldsymbol{u}_{i,j}) \perp\!\!\!\perp (\boldsymbol{z}, \boldsymbol{x})$.

A.3 $\boldsymbol{v}_i \perp\!\!\!\perp \boldsymbol{v}_j | \boldsymbol{x}$ unless $i = j$.

A.4 $\boldsymbol{u}_{i,j} \perp\!\!\!\perp \boldsymbol{u}_{i',j'} | \boldsymbol{x}$ unless $i' = i$ or $j' = j$.

A.5 $\boldsymbol{v}_i \perp\!\!\!\perp \boldsymbol{u}_{i',j'} | \boldsymbol{x}$ unless $i' = i$ or $j' = i$.

A.1 states that the source of the endogeneity of network formation is the dependence of ε_i on $\boldsymbol{v}_i$ and $\boldsymbol{u}_{i,1}, \ldots, \boldsymbol{u}_{i,n}$. **A.2** imposes that these drivers are independent of the remaining covariates. **A.3** requires the variables $\boldsymbol{v}_i$ to be (conditionally) independent across agents while **A.4** demands that the variables $\boldsymbol{u}_{i,j}$ are independent between dyads that do not have an agent in common. This allows for $\boldsymbol{u}_{i,j}$, $\boldsymbol{u}_{i,j'}$, and $\boldsymbol{u}_{j'',i}$ to be dependent. **A.5** allows for $\boldsymbol{v}_i$ to depend on $\boldsymbol{u}_{i,j}$ and $\boldsymbol{u}_{j,i}$.

Proposition 1. *Let* **A.1** –**A.5** *hold. Then* (1.1) *and* (1.2) *are satisfied.*

Proof. By virtue of **A.1**, iterating expectations yields

$$\mathbb{E}(\varepsilon_i | \boldsymbol{A}_{-i}, \boldsymbol{x}) = \mathbb{E}(\mathbb{E}(\varepsilon_i | \boldsymbol{v}_i, \boldsymbol{u}_{i,1}, \ldots, \boldsymbol{u}_{i,n}, \boldsymbol{A}, \boldsymbol{x}) | \boldsymbol{A}_{-i}, \boldsymbol{x}) = \mathbb{E}(\mathbb{E}(\varepsilon_i | \boldsymbol{v}_i, \boldsymbol{u}_{i,1}, \ldots, \boldsymbol{u}_{i,n}) | \boldsymbol{A}_{-i}, \boldsymbol{x}).$$

From (2.5), $\boldsymbol{A}_{-i}$ is a function of $\boldsymbol{v}_{i'}$ for all $i' \neq i$, of $\boldsymbol{u}_{i',j}$ and $\boldsymbol{u}_{j,i'}$ for all $i' \neq i$ and $j \neq i$, and of $\boldsymbol{z}$. By **A.2**–**A.5** all of these variables are independent of $\boldsymbol{v}_i, \boldsymbol{u}_{i,1}, \ldots, \boldsymbol{u}_{i,n}$ conditional on $\boldsymbol{x}$. Further, **A.2** equally states that $\boldsymbol{v}_i, \boldsymbol{u}_{i,1}, \ldots, \boldsymbol{u}_{i,n}$ are independent of $\boldsymbol{x}$. We thus have that

$$\mathbb{E}(\mathbb{E}(\varepsilon_i | \boldsymbol{v}_i, \boldsymbol{u}_{i,1}, \ldots, \boldsymbol{u}_{i,n}) | \boldsymbol{A}_{-i}, \boldsymbol{x}) = \mathbb{E}(\mathbb{E}(\varepsilon_i | \boldsymbol{v}_i, \boldsymbol{u}_{i,1}, \ldots, \boldsymbol{u}_{i,n}) | \boldsymbol{x}) = \mathbb{E}(\mathbb{E}(\varepsilon_i | \boldsymbol{v}_i, \boldsymbol{u}_{i,1}, \ldots, \boldsymbol{u}_{i,n})),$$

which is equal to the unconditional mean $\mathbb{E}(\varepsilon_i) = 0$. This shows the first part of the proposition. Next, from (2.5), using the fact that $\boldsymbol{v}_j$ enters $(\boldsymbol{A})_{i',j}$ and $(\boldsymbol{A})_{j,i'}$ for all $i' \neq i$, the dependence between $\boldsymbol{u}_{i,j}$ and $\boldsymbol{u}_{i',j}, \boldsymbol{u}_{j,i'}$ for all $i' \neq i$, and the presence of the common term $\boldsymbol{z}$,

$$\begin{aligned}
\mathbb{E}((\boldsymbol{A})_{i,j} | \boldsymbol{A}_{-i}, \boldsymbol{x}) &= \mathbb{P}(a(\boldsymbol{v}_i, \boldsymbol{v}_j, \boldsymbol{u}_{i,j}, \boldsymbol{z}) > 0 | \boldsymbol{A}_{-i}, \boldsymbol{x}) \\
&= \mathbb{E}(\mathbb{P}(a(\boldsymbol{v}_i, \boldsymbol{v}_j, \boldsymbol{u}_{i,j}, \boldsymbol{z}) > 0 | \boldsymbol{v}_j, \boldsymbol{u}_{i',j} \text{ and } \boldsymbol{u}_{j,i'} \text{ for all } i' \neq i, \boldsymbol{z}, \boldsymbol{x}) | \boldsymbol{A}_{-i}, \boldsymbol{x}) \\
&\neq \mathbb{E}(\mathbb{P}(a(\boldsymbol{v}_i, \boldsymbol{v}_j, \boldsymbol{u}_{i,j}, \boldsymbol{z}) > 0 | \boldsymbol{v}_j, \boldsymbol{u}_{i',j} \text{ and } \boldsymbol{u}_{j,i'} \text{ for all } i' \neq i, \boldsymbol{z}, \boldsymbol{x}) | \boldsymbol{x}) \\
&= \mathbb{E}((\boldsymbol{A})_{i,j} | \boldsymbol{x})
\end{aligned}$$

follows easily. Hence, there is predictive power in $\boldsymbol{A}_{-i}$ about $(\boldsymbol{A})_{i,j}$ conditional on $\boldsymbol{x}$. This shows the second part of the proposition. The proof of the proposition is, thus, complete. $\square$

We remark that Proposition 1 does not require the presence of traditional instrumental variables, i.e., observables that enter a but are independent of ε_i. Rather, at the heart of Proposition 1 lies the observation that entries in a given column or row of the network adjacency matrix are usually correlated as link decisions involving a given agent are, at least partly, determined by the same factors. Whether these factors are observable in the data, or even known to be determinants of network formation by the econometrician, is irrelevant.

Equations (1.3) and (1.4) hold if (2.5) is complemented with the following restrictions:

B.1 $\mathbb{E}(\varepsilon_i | \boldsymbol{v}_i, \boldsymbol{u}_{i,1}, \ldots, \boldsymbol{u}_{i,n}, \boldsymbol{A}, \boldsymbol{x}) = \mathbb{E}(\varepsilon_i | \boldsymbol{v}_i, \boldsymbol{u}_{i,1}, \ldots, \boldsymbol{u}_{i,n})$.

B.2 $(\boldsymbol{v}_i, \boldsymbol{v}_j, \boldsymbol{u}_{i,j}) \perp\!\!\!\perp (\boldsymbol{z}, \boldsymbol{x})$.

B.3 $\boldsymbol{v}_i \perp\!\!\!\perp \boldsymbol{v}_j | \boldsymbol{x}$ unless $j \in g(i)$.

B.4 $\boldsymbol{u}_{i,j} \perp\!\!\!\perp \boldsymbol{u}_{i',j'} | \boldsymbol{x}$ unless $i' \in g(i)$ or $j' \in g(j)$.

B.5 $\boldsymbol{v}_i \perp\!\!\!\perp \boldsymbol{u}_{i',j'} | \boldsymbol{x}$ unless $i' \in g(i)$ or $j' \in g(i)$.

These conditions allow for group-level dependence in the drivers of peer selection.

Proposition 2. *Let* **B.1** –**B.5** *hold. Then* (1.3) *and* (1.4) *are satisfied.*

Proof. The proof follows the same steps as the proof of Proposition 1 and is omitted. $\square$

Failure of (1.1) Our conditional moment restrictions are not compatible with general interdependency in peer selection. To see this consider a simple example where agents have a taste for transitivity. Moreover,

$$(\boldsymbol{A})_{i,j} = \begin{cases} 1 & \text{if } v_i + v_j + \sum_{k=1}^{n} (\boldsymbol{A})_{j,k} \, (\boldsymbol{A})_{k,i} > u_{i,j} \\ 0 & \text{otherwise} \end{cases} . \tag{2.6}$$

Assume that scalar v_i and scalar $u_{i,j}$ are (i) independent across, respectively, agents and dyads; (ii) independent of each other; and (iii) independent of the regressors. Suppose that $\mathbb{E}(\varepsilon_i|v_i, \boldsymbol{A}, \boldsymbol{x}) = \mathbb{E}(\varepsilon_i|v_i)$ so that endogeneity is solely due to v_i. The specification in (2.6) implies that each $(\boldsymbol{A})_{i,j}$ depends on all of $v_1, \ldots, v_n$, and thus so do all elements of $\boldsymbol{A}_{-i}$. Therefore,

$$\mathbb{E}(\varepsilon|\boldsymbol{A}_{-i}, \boldsymbol{x}) = \mathbb{E}(\mathbb{E}(\varepsilon_i|v_i)|\boldsymbol{A}_{-i}, \boldsymbol{x}) \neq \mathbb{E}(\mathbb{E}(\varepsilon_i|v_i)) = 0,$$

and (1.1) fails. Furthermore, no entries of the adjacency matrix will satisfy the validity condition demanded of an instrument and so a weaker version of (1.1), such as (1.3), cannot be used to solve the problem.

On the other hand, (1.3) can be useful if the transitivity is restricted within groups, say

$$(\boldsymbol{A})_{i,j} = \begin{cases} 1 & \text{if } v_i + v_j + \sum_{k \in g(i)} (\boldsymbol{A})_{j,k} \, (\boldsymbol{A})_{k,i} > u_{i,j} \\ 0 & \text{otherwise} \end{cases} .$$

Indeed, in this case, $(\boldsymbol{A})_{i,j}$ depends on v_i, v_j, and $v_{i'} \in g(i)$. Therefore, $(\boldsymbol{A})_{i',j}$ for $i' \notin g(i)$ will be independent of v_i but will depend on v_j. Hence, this model is compatible with (1.3) and (1.4).

Failure of (1.2) The relevance of $\boldsymbol{A}_{-i}$ follows from the fact that the link decisions of a given agent are not independent of one another. Moreover, (1.2) exploits the dependence within a given row and column of the adjacency matrix. The condition fails in the basic random-graph model of Erdős and Rényi (1959, 1960), where links are formed according to

$$(\boldsymbol{A})_{i,j} = \begin{cases} 1 & \text{if } 0 > u_{i,j} \\ 0 & \text{otherwise} \end{cases} , \tag{2.7}$$

where $u_{i,j}$ is a random variable that is independent and identically distributed (i.i.d.) across dyads. This implies that $(\boldsymbol{A})_{i,j}$ is i.i.d. across dyads.

Of course, in the Erdős and Rényi (1959, 1960) model there is little margin for any self selection of the peer group to come about. An extension of the model in which endogeneity is present and (1.2) fails is the following simple directed model with one-way heterogeneity:

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

where v_i is i.i.d across agents and allowed to depend on ε_i, and $u_{i,j}$ is i.i.d. across dyads and independent of ε_i, as before. A symmetrized version that introduces heterogeneity on both sides is

$$(\boldsymbol{A})_{i,j} = \begin{cases} 1 & \text{if } v_i + v_j > u_{i,j} \\ 0 & \text{otherwise} \end{cases}. \tag{2.8}$$

This is essentially the model entertained by Johnsson and Moon (2021). Maintaining the same assumptions on the random variables as before, this setup does satisfy both (1.1) and (1.2).

3 Instrumental variables

In the remainder of the paper we work under (1.1) and (1.2). What follows can be adapted to (1.3) and (1.4) in a straightforward manner.

Baseline In

$$y_i = \alpha + \beta x_i + \gamma \left(\sum_{j=1}^{n} (\boldsymbol{H})_{i,j} \, x_j \right) + \varepsilon_i, \tag{3.1}$$

the spillover term can be endogenous because $(\boldsymbol{H})_{i,j} = {}^{(\boldsymbol{A})_{i,j}}/_{\sum_{j'=1}^{n} (\boldsymbol{A})_{i,j'}}$ is a function of $(\boldsymbol{A})_{i,1}, \ldots, (\boldsymbol{A})_{i,n}$ and all of these variables are allowed to covary with ε_i. By (1.1) the latter is, however, mean independent (given $\boldsymbol{x}$) of $(\boldsymbol{A})_{i',j}$ for all $i' \neq i$ and $j \neq i$. Hence, the link decisions that do not involve agent i—i.e., $(\boldsymbol{A})_{i',j}$ for $i' \neq i$ and $j \neq i$—are exogenous. Furthermore, by (1.2), these $(\boldsymbol{A})_{i',j}$ contain predictive power on $(\boldsymbol{H})_{i,j}$. In (2.5), for example, they covary with $(\boldsymbol{A})_{i,i'}$ and $(\boldsymbol{A})_{i,j}$ and, because $\boldsymbol{H}$ is a row-normalized adjacency matrix, every link decision in the leave-own-out network $\boldsymbol{A}_{-i}$ will be a valid and relevant instrument for $(\boldsymbol{H})_{i,j}$.

The above observation suggests the construction of instrumental variables as linear combinations of $x_1, \ldots, x_n$ with weights coming from $\boldsymbol{A}_{-i}$. There is no unique way of doing so. Here we discuss one way to proceed that will equally have a natural extension to the full model. To describe it, we first introduce, for each of the n agents i, the $n \times n$ matrices

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

This is the row-normalized version of the adjacency matrix $\boldsymbol{A}_{-i}$ introduced previously, only complemented with one additional row of zeros and one additional column of zeros, each of these at location i. This augmentation is done for notational considerations, as it maintains the dimension of these matrices to $n \times n$, the size of the matrix $\boldsymbol{H}$ pertaining to the full network. $\boldsymbol{H}_{-i}$ corresponds to the transition matrix on the network obtained on ruling-out any links that involve agent i.

The construction of these leave-own-out matrices is illustrated graphically in Figure 1. The left plot 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, the middle and right plots give the subnetworks obtained on leaving out Agent 1 and Agent 2, respectively. Because of the symmetry of the problem, the leave-own out graphs for Agents 3,4, and 5 are the same as in the right plot up to a rotation of the indices of the agents in the periphery. They are, thus, not given separately.

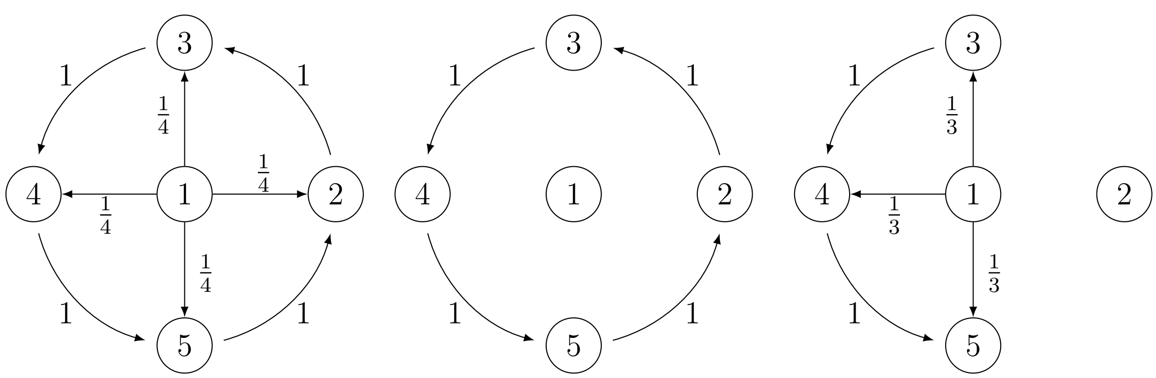

Figure 1: Transition matrix for a directed wheel graph on five agents (left) together with transition matrices for the implied leave-one-out subnetworks (middle and right).

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. It may be useful to think about the matrix $\boldsymbol{Q}_1$ as a transition matrix, thus again inducing a network. The right plot in Figure 2 shows this network in our wheel-graph example; the original network is repeated in the left plot. The former differs considerably from the latter. While in the original network the agents in the periphery where connected only by a counter-clockwise circle of links, the induced network also features an additional clockwise circle, as well as new direct links between agents at opposite ends of the circle. The weights assigned to these link are also generally different.

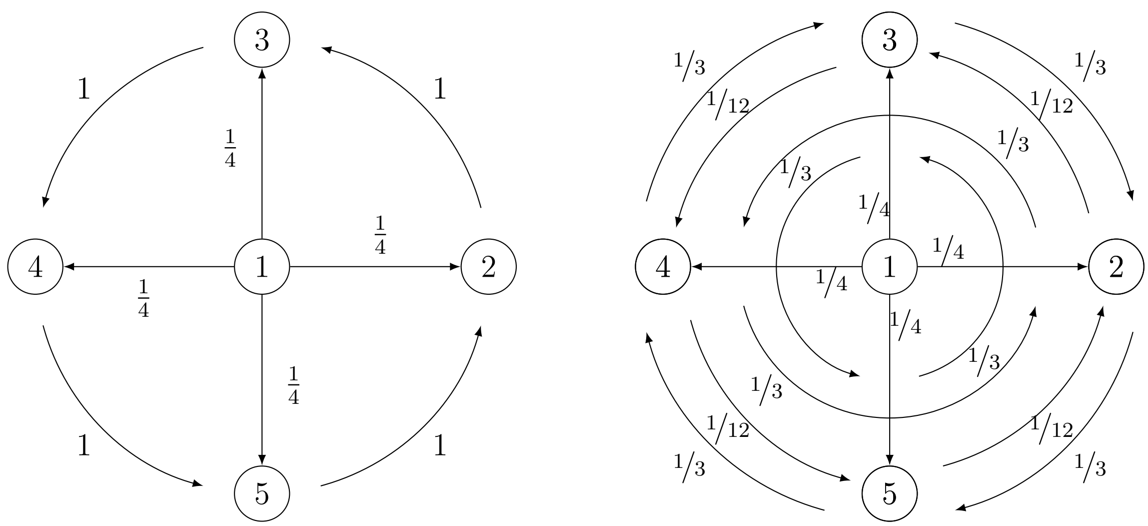

Figure 2: Transition matrix $\boldsymbol{H}$ for a directed wheel graph on five agents (left) together with the transition matrix $\boldsymbol{Q}_1$ for the implied network (right).

Under (1.1) and (1.2) the average

$$\textstyle\sum_{j=1}^{n} (\boldsymbol{Q}_1)_{i,j}\, x_j$$

is exogenous in (3.1) and correlates with the problematic spillover term $\sum_{j=1}^{n} (\boldsymbol{H})_{i,j}\, x_j$. A way forward is thus to complement the model in (3.1) with the first-stage regression specification

$$\left(\sum_{j=1}^{n}(\boldsymbol{H})_{i,j}\, x_j\right) = \pi_0 + \pi_1 x_i + \pi_2 \left(\sum_{j=1}^{n}(\boldsymbol{Q}_1)_{i,j}\, x_j\right) + \epsilon_i, \tag{3.2}$$

and estimate the parameters of the baseline model by a standard two-stage least-squares procedure.

The regression coefficient π_2 in (3.2) is a complex function of features of the network such as the probability of observing a link between any two agents, a triangle between any three agents, and similar larger-dimensional sub-network configurations. Even in the relatively simple model of Johnsson and Moon (2021) this involves integrals of nonlinear functions up to n dimensions. It does, therefore, not seem feasible to derive closed-form expressions for a general network size. For small networks generated from (2.8) expressions can be derived. While they are difficult to interpret directly, they can be combined with low-dimensional parametrizations to study the dependence of the regression coefficient on the underlying model.

One such parametrization has $v_i \sim \text{Exponential}(\lambda_v)$ and $u_{i,j} \sim \text{Exponential}(\lambda_u)$. The upper two plots in Figure 3 provide the value of π_2 as a function of the scale parameters for $n \in \{3, 4, \}$ when the regressor x_i is zero mean and independent across agents. The plots show that, for any value of λ_v, π_2 increases with λ_u. The larger λ_u the more likely it is that $u_{i,j}$ takes on small values and, hence, the more likely that $(\boldsymbol{A})_{i,j} = 1$ for given values of v_i and v_j. Larger values of λ_u thus yield denser networks, all else equal, which increases instrument strength. The limit case $\lambda_u \to \infty$ corresponds to the degenerate situation where unit mass is being placed at zero and links form with probability one. The comparative statics for λ_v are different. Larger values of λ_v mean less variability in v_i, and so less degree heterogeneity across agents. This translates into an instrument with smaller predictive power. Here the limit case $\lambda_v \to \infty$ yields an empty network.

Another parametrization is one where $v_i \sim \text{Normal}(\mu, \sigma^2)$ and $u_{i,j} \sim \text{Normal}(0, 1)$. The lower two plots in Figure 3 give π_2 as a function of μ and σ. The plots reveal the same comparative statics as in the exponential parametrization. Larger values of μ yield denser networks while larger values of σ yield more between-agent heterogeneity in link behavior.

Both these features bring with them predictive power.

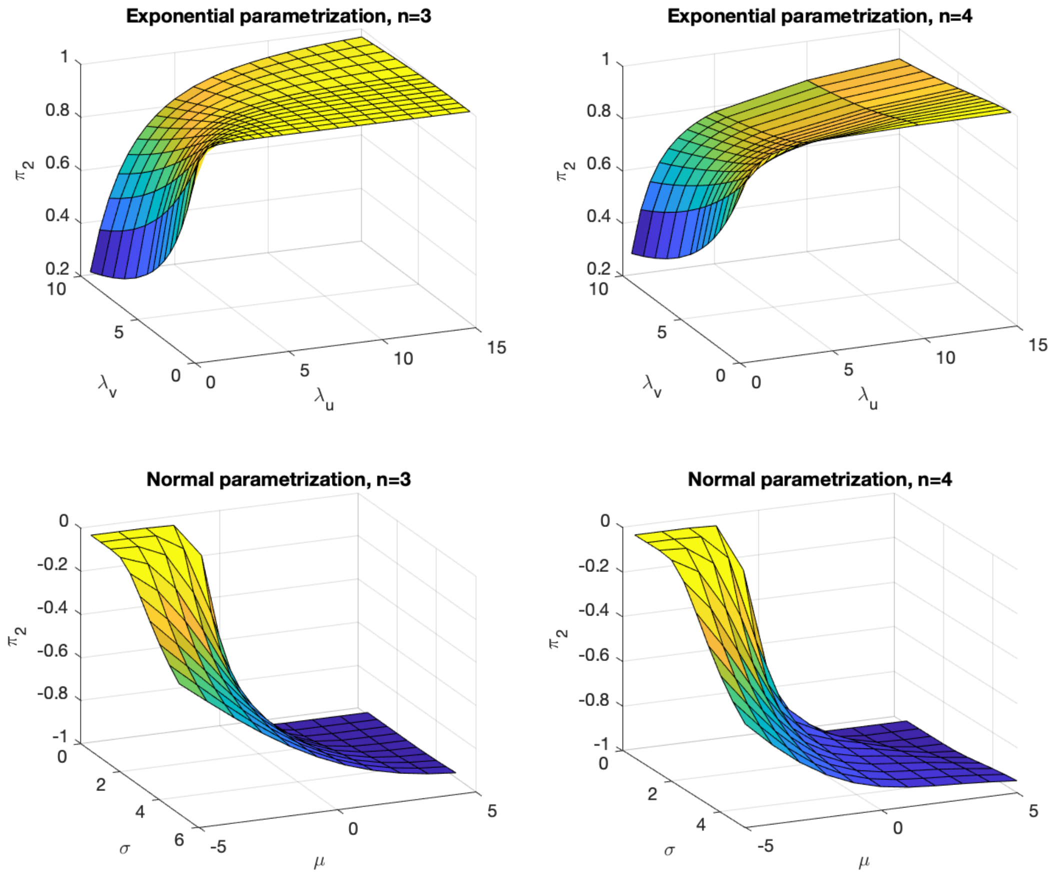

Figure 3: Theoretical calculations of the value of the regression coefficient π_2 in (3.2) under model (2.8) with the regressor zero mean and independent across agents. Upper plots: $v_i \sim \text{Exponential}(\lambda_v)$ and $u_{i,j} \sim \text{Exponential}(\lambda_u)$. Lower plots: $v_i \sim \text{Normal}(\mu, \sigma^2)$ and $u_{i,j} \sim \text{Normal}(0, 1)$. Left plots: $n = 3$. Right plots: $n = 4$.

To see how instrument strength evolves with the size of the network we must resort to simulation. Figure 4 plots the value of π_2 as a function of n for $\lambda_v \in \{1, 5, 10\}$ and $\lambda_u = 1/2$ in the exponential parametrization, as obtained by simulating a large number of i.i.d. networks from this parametrization and computing the regression slope in question. While larger values of λ_v yield a smaller coefficient, its magnitude increases with n in each case. The plot also shows that π_2 is a concave function of n, and so the increase in the magnitude is a decreasing function of n.

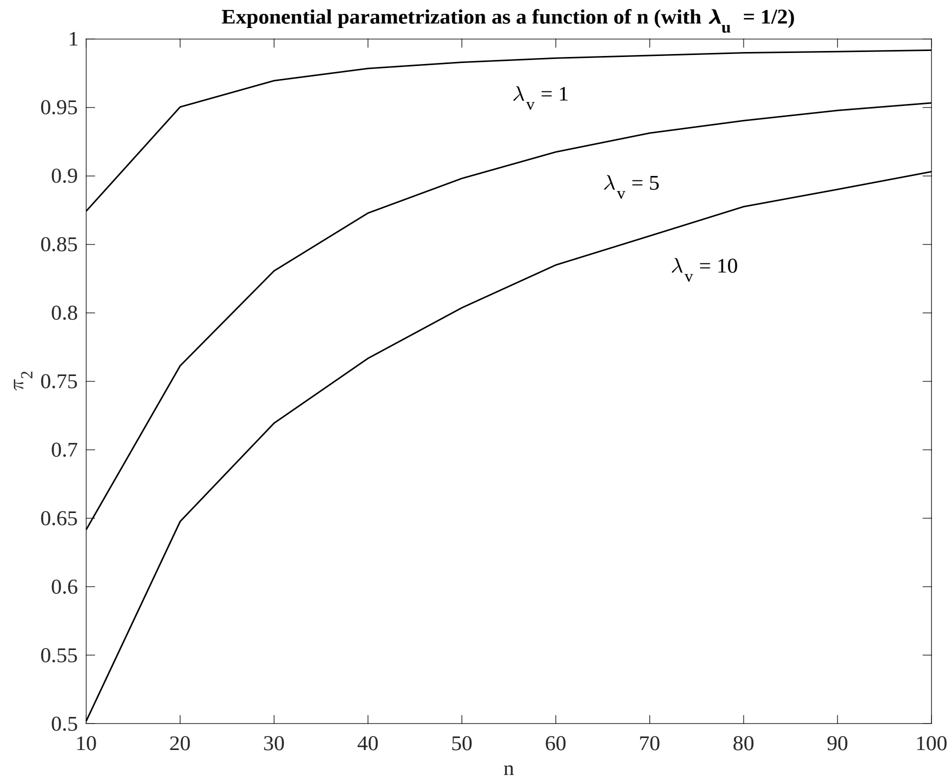

Figure 4: Simulated value of the regression coefficient π_2 in (3.2) under model (2.8) as a function of n. The regressor is zero mean and independent across agents, $v_i \sim \text{Exponential}(\lambda_v)$, and $u_{i,j} \sim \text{Exponential}(\lambda_u)$.

Full model 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,

$$\boldsymbol{H}\boldsymbol{y} = \frac{\alpha}{1-\delta} \boldsymbol{\iota}_n + \beta \boldsymbol{H}\boldsymbol{x} + (\delta\beta + \gamma) \sum_{s=0}^{\infty} \delta^s \boldsymbol{H}^{s+2} \boldsymbol{x} + \sum_{s=0}^{\infty} \delta^s \boldsymbol{H}^{s+1} \boldsymbol{\varepsilon}. \tag{3.3}$$

The argument of Bramoullé, Djebbari and Fortin (2009) is that $\boldsymbol{H}^2\boldsymbol{x}$, $\boldsymbol{H}^3\boldsymbol{x}$, and so on can be used as instrumental variables for $\boldsymbol{Hy}$ when the network is exogenous, provided that $\delta\beta + \gamma \neq 0$. This argument cannot be used in our setting as the validity of these variables as instruments breaks down when peer selection is endogenous. However, on inspecting the expansion in (3.3) 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 our wheel-graph example, the subnetworks induced by iterating on $\boldsymbol{H}_i$ once to get $\boldsymbol{H}_i^2 = \boldsymbol{H}_i\boldsymbol{H}_i$ are given in Figure 5. This suggests a generalization of our approach in the baseline model to the full model that mimics the core idea underlying the instrument construction in Bramoullé, Djebbari and Fortin (2009) for the exogenous case.

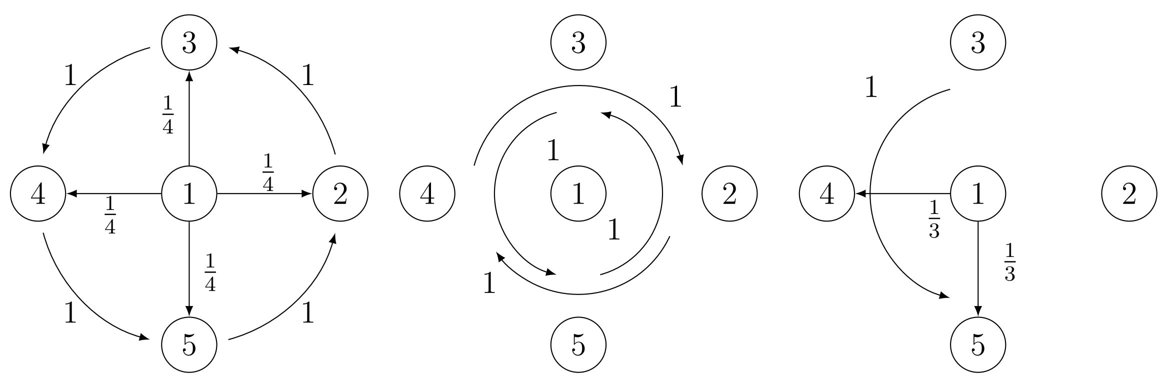

Figure 5: Transition matrix for a directed wheel graph on five agents (left) together with the iterated transition matrices for the implied leave-one-out subnetworks (middle and right).

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 conditions in (1.1) and (1.2) these weights are, again, useful to construct instrumental variables. Moreover, we may instrument the two endogenous right-hand side variables, $\boldsymbol{Hx}$ and $\boldsymbol{Hy}$, by the two exogenous variables

$\boldsymbol{Q}_1\boldsymbol{x}$ and $\boldsymbol{Q}_2\boldsymbol{x}$. Figure 6 gives the network induced by $\boldsymbol{Q}_2$ in our wheel-graph example. Like in the exogenous case, we require $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ to be sufficiently different. Equally like in the case where the network is exogenous, this is more difficult in denser networks. It is more difficult to give simple primitive conditions for instrument relevance than in the exogenous-network case, however. Indeed, in Bramoullé, Djebbari and Fortin (2009), (3.3) implies a linear reduced form from which such conditions can be derived. This is not the case here.[1]

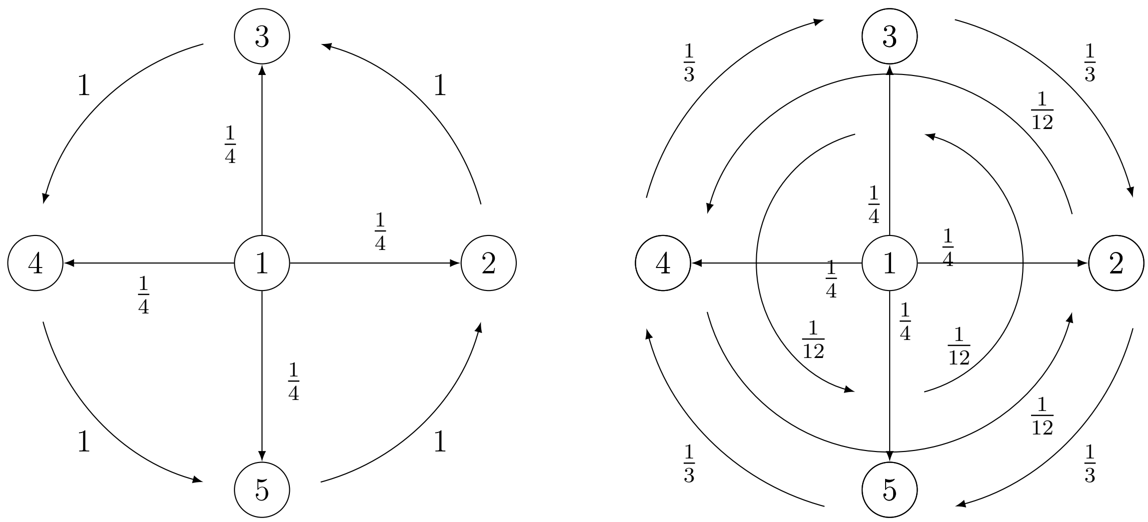

Figure 6: Transition matrix $\boldsymbol{H}$ for a directed wheel graph on five agents (left) together with the transition matrix $\boldsymbol{Q}_2$ for the implied network (right).

It is possible to construct additional instruments by taking additional steps through the leave-own-out networks. This yields $\boldsymbol{Q}_s\boldsymbol{x}$ for

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

[1]The condition that $\delta\beta+\gamma\neq 0$ is key when the network is exogenous. It requires that $\boldsymbol{x}$ affects $\boldsymbol{y}$ and that endogenous and exogenous peer effects do not exactly offset each other. Otherwise, we have that $\mathbb{E}(\boldsymbol{H}\boldsymbol{y}|\boldsymbol{H},\boldsymbol{x}) = \mu\,\boldsymbol{\iota}_n + \beta\boldsymbol{H}\boldsymbol{x}$, and so $\boldsymbol{H}^s\boldsymbol{x}$ (for any $s>1$) no longer contains predictive information about $\boldsymbol{H}\boldsymbol{y}$ conditional on $\boldsymbol{H}\boldsymbol{x}$. On the other hand, when link formation is endogenous there will generally still be information on $\boldsymbol{H}\boldsymbol{y}$ in $\boldsymbol{Q}_1\boldsymbol{x}$ and $\boldsymbol{Q}_2\boldsymbol{x}$ even if $\delta\beta+\gamma=0$. This information comes from the fact that $\mathbb{E}(\boldsymbol{x}'\boldsymbol{Q}_1'\boldsymbol{H}^{p+1}\boldsymbol{\varepsilon})\neq 0$ and $\mathbb{E}(\boldsymbol{x}'\boldsymbol{Q}_2'\boldsymbol{H}^{p+1}\boldsymbol{\varepsilon})\neq 0$ because the entries of $\boldsymbol{H}_{-i}\,\boldsymbol{x}$ are functions of variables that depend on ε_j for $j\neq i$.

and any integer s. These instruments play a role analogous to $\boldsymbol{H}^s\boldsymbol{x}$ in the approach of Bramoullé, Djebbari and Fortin (2009). Again, as s increases the transition matrix $\boldsymbol{H}^s_{-i}$ will tend to its steady-state distribution, so that higher iterations will provide increasingly less (additional) predictive power. The speed of convergence to the steady-state distribution is faster in denser networks.

4 Simulations

Our 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 i.i.d. standard-normal variates and we set $c = -\sqrt{2}\Phi^{-1}(p)$, for $p \in \{1/4, 1/2\}$ and Φ the standard-normal distribution function. The marginal probability of forming a link is p. 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 two sample sizes. The first has 250 i.i.d. networks of size 25. The second has 25 i.i.d. networks of size 250.

We present simulation results for the estimator of Bramoullé, Djebbari and Fortin (2009) (TSLS-X) and for our proposal (TSLS-E). The former instruments $\boldsymbol{H}\boldsymbol{x}$ by itself and $\boldsymbol{H}\boldsymbol{y}$ by $\boldsymbol{H}^2\boldsymbol{x}, \ldots, \boldsymbol{H}^4\boldsymbol{x}$. The latter instruments $\boldsymbol{H}\boldsymbol{x}$ and $\boldsymbol{H}\boldsymbol{y}$ by $\boldsymbol{Q}_1\boldsymbol{x}$, $\boldsymbol{Q}_2\boldsymbol{x}, \ldots, \boldsymbol{Q}_4\boldsymbol{x}$. In both cases 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 concerns the design where $p = 1/4$ and we have 250 networks of size 25. It contains the bias and standard deviation of the estimators, the mean and standard deviation

Table 1: Simulation results: 250 networks of size 25, $p = 1/4$

	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	size	bias	std	mean	std	size
					$\varphi(\eta) = 0$					
β	-0.0002	0.0130	-0.0167	1.0120	0.0492	-0.0009	0.0142	-0.0760	0.9928	0.0500
γ	0.0003	0.0505	0.0125	1.0033	0.0504	-0.0095	0.1322	-0.2229	1.0099	0.0532
δ	-0.0002	0.0223	-0.0047	1.0152	0.0524	0.0050	0.0590	0.2415	1.0108	0.0592
					$\varphi(\eta) = \eta$					
β	-0.0401	0.0166	-2.4547	1.0320	0.6864	0.0001	0.0187	-0.0022	1.0012	0.0516
γ	-0.4254	0.1373	-3.0446	0.8915	0.8924	0.0017	0.1289	-0.0589	0.9959	0.0516
δ	0.3536	0.0334	10.5145	0.8746	1.0000	-0.0013	0.0458	0.0635	0.9884	0.0448
					$\varphi(\eta) = \exp(3\Phi(\eta))$					
β	-0.0258	0.0637	-0.4226	1.0060	0.0684	-0.0005	0.0668	-0.0112	1.0015	0.0468
γ	-0.2610	0.5285	-0.4514	0.9856	0.0756	0.0009	0.3060	-0.0329	1.0154	0.0532
δ	0.2192	0.0284	7.6166	0.6550	1.0000	-0.0027	0.0399	0.0540	1.0202	0.0588
					$\varphi(\eta) = \sin(3\Phi(\eta))$					
β	-0.0120	0.0130	-0.9229	1.0046	0.1572	-0.0005	0.0146	-0.0477	0.9958	0.0484
γ	-0.1248	0.0513	-2.4664	1.0012	0.6904	-0.0055	0.1294	-0.1864	1.0180	0.0576
δ	0.1049	0.0170	6.2143	1.0519	1.0000	0.0033	0.0560	0.2065	1.0118	0.0556

of the implied t-statistics, as well as the empirical rejection frequency of two-sided t-tests (at the 5% significance level) under the null. 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.

The estimator of Bramoullé, Djebbari and Fortin (2009) does well when link formation is exogenous. Otherwise, the coefficient estimators are biased, except for the one 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 presence of bias implies that the sampling distribution of the t-statistic is not centered

Table 2: Simulation results: 250 networks of size 25, $p = 1/2$

	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	size	bias	std	mean	std	size
					$\varphi(\eta) = 0$					
β	-0.0002	0.0130	-0.0149	1.0004	0.0528	-0.0018	0.0157	-0.1735	1.0045	0.0552
γ	-0.0002	0.0689	-0.0025	1.0155	0.0532	-0.0354	0.2307	-0.4632	1.0788	0.0976
δ	-0.0001	0.0316	0.0226	1.0144	0.0524	0.0145	0.0909	0.4784	1.0810	0.0992
					$\varphi(\eta) = \eta$					
β	-0.0379	0.0170	-2.2743	1.0322	0.6224	-0.0014	0.0216	-0.0898	1.0229	0.0552
γ	-0.4161	0.2333	-1.7507	0.9198	0.4152	-0.0307	0.3037	-0.3206	1.0684	0.0792
δ	0.3689	0.0735	5.0615	0.8707	1.0000	0.0103	0.1116	0.3232	1.0662	0.0796
					$\varphi(\eta) = \exp(3\Phi(\eta))$					
β	-0.0213	0.0649	-0.3803	1.0194	0.0728	-0.0039	0.0698	-0.0407	0.9770	0.0472
γ	-0.2459	0.7542	-0.2456	0.9888	0.0544	-0.1334	0.8262	-0.4093	1.0226	0.0776
δ	0.2227	0.0740	2.9705	0.8600	0.8788	0.0321	0.2145	0.4561	1.0256	0.0928
					$\varphi(\eta) = \sin(3\Phi(\eta))$					
β	-0.0107	0.0133	-0.8023	0.9981	0.1292	-0.0017	0.0162	-0.1650	0.9957	0.0540
γ	-0.1191	0.0724	-1.7126	1.0493	0.3992	-0.0401	0.2243	-0.4686	1.0505	0.0904
δ	0.1044	0.0263	4.1521	1.1528	0.9796	0.0163	0.0893	0.4836	1.0621	0.0948

at zero. 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.

In Table 2 we have results for when $p = 1/2$, and so the networks are more dense. The chief impact of this design change is that both estimators become less precise. This is in line with our discussion from above. TSLS-E also suffers from somewhat more bias than before. This results in some size distortion, with overrejections of as much as five

Table 3: Simulation results: 25 networks of size 250, $p = 1/4$

	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	size	bias	std	mean	std	size
					$\varphi(\eta) = 0$					
β	-0.0000	0.0125	-0.0034	0.9934	0.0460	-0.0006	0.0128	-0.0460	0.9904	0.0430
γ	-0.0009	0.0814	-0.0088	1.0005	0.0540	-0.0825	0.4106	-0.4205	0.9458	0.0615
δ.	-0.0003	0.0399	-0.0040	1.0009	0.0540	0.0458	0.2029	0.4736	0.9533	0.0645
					$\varphi(\eta) = \eta$					
β	-0.0050	0.0189	-0.2689	0.9846	0.0595	-0.0001	0.0183	-0.0065	1.0101	0.0515
γ	-0.4665	0.6514	-3.0298	4.9107	0.7665	-0.0392	0.3156	-0.2120	0.9949	0.0515
δ	0.4458	0.1439	13.6949	3.0706	0.9985	0.0175	0.1396	0.2284	0.9795	0.0520
					$\varphi(\eta) = \exp(3\Phi(\eta))$					
β	-0.0022	0.0682	-0.0558	0.9628	0.0415	-0.0010	0.0664	-0.0155	0.9512	0.0400
γ	-0.0618	1.9939	0.4965	7.1590	0.7850	-0.0641	0.7538	-0.0818	0.9592	0.0430
δ	0.1791	0.0965	12.0535	5.6056	0.9575	0.0178	0.1801	0.1163	0.9575	0.0390
					$\varphi(\eta) = \sin(3\Phi(\eta))$					
β	-0.0022	0.0128	-0.1691	0.9800	0.0490	-0.0006	0.0130	-0.0431	0.9733	0.0435
γ	-0.1986	0.1339	-2.7439	1.8327	0.6745	-0.0786	0.2691	-0.4333	0.9736	0.0645
δ.	0.1817	0.0403	6.5818	1.5210	0.9990	0.0431	0.1365	0.4767	0.9739	0.0760

percentage points.

Table 3, finally, contains results for the case where we have few large networks, and $p = 1/4$. Both TSLS-X and TSLS-E are notably more variable in such a setting. Both due to the presence of bias and underestimation of the sampling variability, the normal approximation performs poorly for TSLS-X. For TSLS-E, on the other hand, it continues to be quite accurate, leading to little discrepancy between theoretical and actual size in hypothesis testing.

Conclusion

This paper has introduced an instrumental-variable approach to deal with self selection of peers in the linear in means model. Identification is achieved through conditional moment conditions that are implied by a large class of network-formation models. A particularly simple approach that builds on leave-own-out networks has been introduced. This leads to two-stage least-squares estimators that are straightforward to implement and carry a close resemblance to the estimator of Bramoullé, Djebbari and Fortin (2009). The latter is arguably the default estimator under the assumption of exogenous construction of peer groups.

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. (2022). Identification and estimation of a partially linear regression model using network data. *Econometrica 90*, 347–365.

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.

Erdős, P. and A. Rényi (1959). On random graphs. *Publicationes Mathematicae 6*, 290–297.

Erdős, P. and A. Rényi (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5*, 17–61.

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

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

Hoderlein, S. and E. Mammen (2007). Identification of marginal effects in nonseparable models without monotonicity. *Econometrica 75*, 1513–1518.

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.

Johnsson, I. and H. R. Moon (2021). Estimation of peer effects in endogenous social networks: Control function approach. *Review of Economics and Statistics 103*, 328–345.

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 (2021). Who is the key player? A network analysis of juvenile delinquency. *Journal of Business & Economic Statistics 39*, 849–857.

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.

Newman, M. (2010). *Networks*. Oxford University Press.

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

Santavirta, T. and M. Sarzosa (2019). Effects of distruptive peers in endogenous social networks. Mimeo.

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.