EIEF

EINAUDI INSTITUTE FOR ECONOMICS AND FINANCE

EIEF Working Paper 06/14

September 2014

Endogenous Network Production Functions with Selectivity

by

William C. Horrace

(Syracuse University)

Xiaodong Liu

(University of Colorado at Boulder)

Eleonora Patacchini

(Cornell University, EIEF and CEPR)

Endogenous Network Production Functions with Selectivity

William C. Horrace[*] Xiaodong Liu[†] Eleonora Patacchini[‡]

Abstract

We consider a production function model that transforms worker inputs into outputs through peer effect networks. The distinguishing features of this production model are that the network is formal and observable through worker scheduling, and selection into the network is done by a manager. We discuss identification and suggest a variety of estimation techniques. In particular, we tackle endogenity issues arising from selection into groups and exposure to common group factors by employing a polychotomous Heckman-type selection correction. We illustrate our method using data from the Syracuse University Men's Basketball team, where at any point in time the coach selects a lineup and the players interact strategically to win games.

Key Words: Stochastic Frontier Model, Spatial Autoregressive Model, Peer Effects, Endogenous Network Formation, Selectivity.

JEL Codes: C31, C44, D24.

[*] Syracuse University, Email: whorrace@maxwell.syr.edu.

[†] University of Colorado at Boulder. E-mail: xiaodong.liu@colorado.edu.

[‡] Cornell University, EIEF and CEPR. E-mail: eleonora.patacchini@cornell.edu

1 Introduction

Endogeneity in production function estimation is not a new issue. Endogeneity of inputs can arise for a variety of reasons: input measurement error, simultaneity of unobservables and inputs, and endogeneity of "explanatory" outputs in multiple-output distance function analysis (to name a few). In service industries, these problems are exacerbated in obvious ways. However, one could imagine that the main challenge in estimating a service production function is the specification of the function itself. In particular, the way that labor is transformed into output may be unclear. Production in a service industry is typically not "serial" as it might be on a manufacturing assembly line, where productivity of worker A may only affect the productivity of worker B, who (in turn) only affects worker C.[1] Service industries may be characterized by teams of workers whose individual productivities are interrelated in complex ways and (in particular) through networks. Consider an architectural firm which simultaneously produces design plans for a variety of projects with teams of architects and draftsmen, who may work across multiple projects in a given workday. In this setting worker interrelatedness may be determined by networks established by a single manager, who assigns workers to teams based on both observable and unobservable characteristics of workers. This implies formal and measurable time-varying networks which may be endogenous due to selectivity.[2] Understanding network effects in production may be important for worker scheduling and design of worker incentive schemes.

The purpose of this paper is to specify an econometric model that incorporates peer effects on worker productivity (output).[3] That is, a worker's productivity is a function of the productivities of the co-workers on her team, where teams are assigned by managers. Individual team members interact through time-varying interaction schemes which serve as proxies for the managerial decision and which function as the mechanism for group formation and individual interrelatedness. In most econometric network models, selection into groups is as much an individual choice as is the behavior that stems from a given network structure.[4] In this setting endogeneity problems may arise if the

[1] This is not to suggest that a manufacturing process could not be more complicated, but the traditional assembly line process possesses this feature.

[2] There may also be informal networks, but they are not the focus here. Informal networks may arise through a principle-agent problem of imperfect montioring. A manager may order a worker to split her time evenly on the two projects, but she may not, in practice. An alternative way to conceptualize this phenomenon is that the formal network is measured with error.

[3] Peer effects have been indicated as one of the main empirical determinants of several important social phenomena (see Jackson and Zenou, 2013, part III, for a collection of recent studies).

[4] Some studies exploit random assignement. For example, in lab experiments or (infrequently) in field experiments a scientist or social planner determines groupings (see, e.g., Falk and Ichino, 2006, or Guryan et al., 2009).

model does not account for unobserved individual characteristics driving both network formation and behavior over networks. We consider the unique situation where a manager selects workers into teams (networks) to produce output, and we call this model a *Network Production Function Model*. In the model, network connections are captured by a binary adjacency matrix, where adjacency is specified as a binary indicator of team membership. The salient feature of this model is that team membership is perfectly observable.[5],[6] In this model, the manager's selection decisions depend on the combination of individual characteristics at the team level, rather than individual-level characteristics. Such team-level factors contribute to the so called "correlated effects" (Manski, 1993), which could be confounded with peer effects and lead to identification problems.

We use a polychotomous Heckman-type correction to address this problem in the context of production networks. In team projects, the probability of selecting a worker for the project is not independent across workers. We exploit this interdependency for the identification and estimation of peer effects in network production functions. This is the main contribution of the paper.

More specifically, we consider productivity of a single project, involving a two-stage process. First, the manager chooses a team (lineup) of m workers (m is predetermined) from a population of n workers to work on the project of interest. Residual workers are assigned to other projects.[7] Next, workers work on the project to produce output for a given time period. For the population of n workers, the $n \times n$ adjacency matrix across all projects is potentially endogenous. By focusing on a single project of interest, we have an $m \times m$ submatrix of the adjacency matrix which is exogenous conditional on selection into the specific project. Thus, the network endogeneity is reduced to a selectivity bias, which can be corrected using a fixed effect estimator or a polychotomous Heckman-type bias correction procedure due to Lee (1983) and Dahl (2002).[8]

The resulting selectivity bias term is an inverse mills ratio (in the case of the Lee's parametric estimate) or a single index (in the case of the Dahl's semi-parametric estimate), varies across lineups and time, and can be interpreted in two interesting ways. First, it can be thought of as a fixed effect

[5] Manski (1993) suggests that it is not possible to identify network effects if researchers do not know how networks or reference groups are formed by individulas in the network.

[6] It is also possible for adjacency to be measured as cumulative time that individuals worked together on a project. This would be directly measurable from time-cards, but we do not explore it here.

[7] We note that, in any period the $n - m$ residual workers are assigned to other projects, and lags of the output from these projects (as well as the project of interest) are treated as explanatory variables in the output and selection equations. In this sense our specification is not unlike the multiple-output distance function (Fare and Primont, 1990) where a single output is modeled as functions of the remaining outputs.

[8] It is also interesting to note that the word "lineup" evokes an image of workers standing in a line. Our notion of lineup allows us to abstract from the complicated endogenous network for all the workers to a simple, fixed and complete network of workers in a project.

that purges and quantifies the *correlated effects* of Manski (1993). That is, Manski noted that there may be unobserved effects, "wherein individuals in the same group tend to behave similarly because they have similar individual characteristics or face similar institutional environments."[9] In this case the group is the observed lineup, and the "institutional environment" is the manager's selection of the lineup into the project of interest. In this sense we use Heckman (1979) to solve Manski's correlated effects problem. In fact, in terms of estimation, we employ a fixed effect estimator in the style of Lee (2007) that differences out the correlated effect. Second, the selection bias term is loosely interpretable as managerial competence or efficiency. That is, all things being equal and averaging out luck, it is the manager's lineup selection that produces any unobserved team effect and, hence, variability of worker output. This is similar to the notion of inefficiency in the stochastic frontier literature (Aigner et al., 1977; and Meeusen and van den Broeck, 1977), so our selectivity bias term can be thought of as efficiency if it increases output and as inefficiency if it lowers it. Also, insofar as our bias term may be estimated from a first-stage selection equation, it is interpretable as x-efficiency in the stochastic frontier literature (Alvarez et al., 2006).[10]

Our empirical example is the network production function for college basketball. While this may only loosely represent a service industry production process, it is sufficient for the purpose of illustration. In this setting there are n players on a team engaged in two projects at any given period of time: five players interact to produce offense and defense, and $n - 5$ players sit on the bench to produce rest (which is inversely correlated with fatigue).[11] Our measure of active player productivity is *efficiency,* which aggregates time-averaged performance statistics on points, rebounds, blocks, steals, misses, assists, and other measures of offensive and defensive activity for each player. We include a measure of lagged fatigue as an explanatory variable to control for the productivity of benched players. Our data are all player substitutions during the regular 2011-2012 season of the Syracuse University men's college basketball team. We find statistically significant positive production spillovers across players in the same category (guards or forwards), but insignificant effects across players in different categories. When selectivity bias is taken into account, our estimate of peer effects in productivity is 0.0534. That is, a one unit increase in the average efficiency of the other active guards (forwards) induces a 0.0534 increase in the efficiency of an individual guard

[9] Manski (1993) page 533.

[10] More generally, it is interpretable as another source of heterogeneity. However, it is still interesting to speculate on the ways it may embody (in)efficiency.

[11] We take the managerial decisions and performance of the opposing team as exogenous. In this sense our notion of strategic equlibrium is only partial.

(forward) once selectivity bias taken into consideration.

The rest of the paper is organized as follows. The next section reviews the related literature, while highlighting the contribution of our paper. Section 3 introduces the econometric specification of a network production model, while Section 4 considers the specification and estimation of a network production model with selectivity. Section 5 provides an empirical example, using data from the 2011-12 Syracuse University Men's basketball team. Section 6 concludes.

2 Related Literature

Our paper lies at the intersection of different literatures. We briefly review them below, while highlighting our contribution.

2.1 Econometrics of network models

A number of papers have dealt with the identification and estimation of peer effects with network data (see Blume et al., 2011 for an excellent survey). There are three main methodological approaches.

(i) The network is assumed exogenous. Identification relies on network topology and estimation is performed using 2SLS or GMM. The possible presence of unobserved factors responsible for network endogeneity is treated by network fixed effects (see, e.g., Lee, 2007; Bramoullé et al., 2009; Calvó-Armengol et al., 2009; Lee et al., 2010; Liu and Lee, 2010).

(ii) Self-selection of individuals into groups is explicitly taken into account. A selection equation based on individual decisions is added in approach (i) to treat possible network endogeneity. An individual-level selection correction term is then added in the outcome equation. This approach is considered in Liu et al. (2012).

(iii) Parametric modeling assumptions and Bayesian inferential methods are employed to integrate a network formation model with the model of behavior over the formed networks. The selection equation is based on individual decisions as in approach (ii). The network formation and the outcome equation are estimated jointly (see, e.g., Mele, 2013; Goldsmith-Pinkham and Imbens, 2013; Hsieh and Lee, 2013).

In our network production function model selection is done by a social planner (manager), rather than being the result of individual decisions. Hence, the possible network endogeneity can be treated by a group-level selection correction term. We show in this paper that the group-level selection correction term can either be treated as a group fixed effect as in approach (i) or be directly estimated as in approach (ii). Either approach is computationally simple, and thus we do not rely on Bayesian methods.

2.2 Network effects in productivity

There is a limited literature on networks in production processes. Guryan et al. (2009) consider performance of professional golfing pairs, but their parings are randomly assigned and the pairings are competitors not teammates. Bandiera et al. (2009) analyze the productivity of fruit pickers, but their networks are based on worker characteristics, and not on managerial formation of teams. Mas and Moretti (2009) consider peer effects in the performance of supermarket cashiers, but do not specifically employ teams or networks in their analysis. Hamilton et al. (2003) analyze the effect of teams on clothing manufacturing, but do not exploit team composition in a network analysis framework. In all these studies, when production networks or pairings are employed, they are assumed exogenous. Here, we specify a model where endogeneity is assured but replaced with team-level selectivity bias, which can be corrected using a fixed effect estimator or a polychotomous Heckman-type bias correction procedure.

2.3 Production function literature

Our focus is a single firm where the unit of observation is the worker who is observed over time. This is in contrast to the spatial production function work of Druska and Horrace (2004) or Glass et al. (2013), where the unit of observation is the firm, and exogenous networks are conceptualized as output/input spillovers across firms (or countries) measured as geographic distances or contiguity in a spatial estimation framework, and where consistency arguments are for large numbers of firms (or countries). In these papers it is not easy to conceptualize the network (spillover) mechanism or to argue that the adjacency matrix is the correct proxy for the mechanism.[12] In our case (the single firm) the network mechanism is clearly based on labor force peer-effects (e.g., Kandel and Lazear, 1992), and the adjacency matrix, based on the manager's assignments, would seem to be an excellent proxy for this mechanism.[13] The downside to our approach is that employee-level data (administrative data) may not be available to the econometrician. However, the methods considered herein could be used by managers, and the data available to them on employee and project characteristics would be quite detailed. Fortunately for us, the econometric model is also

[12] In their defense Druska and Horrace's distance and contiguity networks are a proxy for infrastructure (roads and bridges) on the island of Java. They find strong output spillovers across rice farms in the dry season and weak spillovers in the rainy season, when travel between villages on the island may be difficult.

[13] Manski (1993) argues that the spatial correlation model "makes sense in studies of small-group interactions, where the sample is composed of clusters of friends, co-workers, or household members... But it does not make sense in studies of neighborhood and other large-group effects, where the sample members are randomly chosen individuals."

suited for estimation of peer effects in sports teams, where all networks (the coach's decisions) are observed and where performance is directly measurable by the econometrician. Therefore, we illustrate our model using data from the Syracuse University Men's College Basketball team.

3 A General Network Production Model

Consider a firm with n workers and a manager that allocates workers to various projects (peer groups) in each time period $t = 1, .., T$. The number and composition of projects is unimportant to the econometric specification, but they may have implications for identification and estimation. When the manager allocates workers to projects she explicitly specifies an $n \times n$ adjacency matrix which determines the interrelatedness of the workers' productivity. Let the adjacency matrix be denoted by $A_t^o = [a_{ij,t}^o]$, where $a_{ij,t}^o = 1$ if workers i and j are assigned to the same project in period t and $a_{ij,t}^o = 0$ otherwise. We set $a_{ii,t}^o = 0$. Let the row-normalized A_t^o be $A_t = [a_{ij,t}]$, where $a_{ij,t} = a_{ij,t}^o / \sum_{k=1}^n a_{ik,t}^o$.[14] Then productivity of the worker i in period t is given by

$$y_{it} = \rho \sum_{j=1}^n a_{ij,t} y_{jt} + x_{it}\beta + u_{it}. \tag{1}$$

In this model, the dependent variable y_{it} is the productivity of worker i in period t. The term $\sum_{j=1}^n a_{ij,t} y_{jt}$ is the average productivity of worker i's co-workers assigned to the same project as i in period t, with its coefficient ρ capturing the peer effect. x_{it} is a $1 \times k_x$ vector of exogenous variables. u_{it} is the regression disturbance. In matrix form, (1) can be written as

$$Y_t = \rho A_t Y_t + X_t \beta + U_t, \tag{2}$$

where $Y_t = (y_{1t}, \cdots, y_{nt})'$, $X_t = (x_{1t}', \cdots, x_{nt}')'$, and $U_t = (u_{1t}, \cdots, u_{nt})'$.

If we assume that A_t is exogenous so that $\mathrm{E}(U_t | A_t, X_t) = 0$, then model (2) can be estimated using spatial panel data methods (see Lee and Yu, 2010 for a survey). However, it is reasonable to believe that the manager may have some information about U_t and her choices of how to allocate workers to projects may be correlated with U_t. If this is the case, then $\mathrm{E}(U_t | A_t, X_t) \neq 0$ and A_t is endogenous.

To find a remedy for the problem of endogenous adjacency matrix, we focus on the workers allocated to a specific project. Let d_{it} be an indicator variable such that $d_{it} = 1$ if worker is assigned

[14] For simplicity, we assume no worker is assigned to a project alone so that $\sum_{k=1}^n a_{ik,t}^o > 0$ for all i.

to the project in period t and $d_{it} = 0$ otherwise. Suppose m_t workers are allocated to the project. Then, for worker i assigned to the project (i.e. $d_{it} = 1$), (1) can be written as

$$y_{it} = \rho \frac{1}{m_t} \sum_{j=1, j \neq i}^{n} d_{jt} y_{jt} + x_{it}\beta + \mathrm{E}(u_{it}|D_t) + u_{it}^*, \tag{3}$$

where $D_t = (d_{1t}, \cdots, d_{nt})'$ and $u_{it}^* = u_{it} - \mathrm{E}(u_{it}|D_t)$. By construction, $\mathrm{E}(u_{it}^*|D_t) = 0$ and, thus, the weights d_{jt} in the peer effect regressor can be considered exogenous. We refer to $\mathrm{E}(u_{it}|D_t)$ as the selectivity bias.

Note, as m_t is often predetermined (e.g., in sports games, the number of active players m_t is fixed), d_{it} is not independent across i. Hence, in the our econometric model, instead of modeling the probability of a certain worker is assigned to a project (i.e. $\Pr(d_{it} = 1)$), we consider the probability of a set of workers (a lineup) is assigned to a project.

4 A Network Model with Selectivity

4.1 The econometric model

In time period t, the manager allocates a lineup of m_t workers from a set of n workers to a project.[15] Suppose there are q_t possible lineups, with a lineup denoted by L_s for $s = 1, \cdots, q_t$. Then, the manager allocates lineup L_s to the project in period t if and only if $d_{st}^* > \max_{r \neq s} d_{rt}^*$, where

$$d_{st}^* = \pi_{st} + \xi_{st}, \quad \text{for } s = 1, \cdots, q_t. \tag{4}$$

In (4), π_{st} is the deterministic component of d_{st}^* and ξ_{st} is a scalar random innovation with zero mean and unit variance. Let d_{st} be a dummy variable such that $d_{st} = 1$ if the lineup L_s is chosen to play in period t and $d_{st} = 0$ otherwise. Then, $d_{st} = 1$ if and only if $\epsilon_{st} < 0$ where $\epsilon_{st} = \max_{r \neq s} d_{rt}^* - d_{st}^*$.

The productivity of lineup L_s in period t is given by the following model

$$Y_{st} = \rho W_t Y_{st} + X_{st}\beta + U_{st}. \tag{5}$$

In (5), $Y_{st} = [y_{it}]_{i \in L_s}$ is an $m_t \times 1$ vector of observations on the dependent variable of the workers in lineup L_s in period t. W_t is a constant weighting matrix given by $W_t = \frac{1}{m_t - 1}(1_{m_t} 1_{m_t}' - I_{m_t})$. $W_t Y_{st}$ measures the average productivity of a worker's co-workers in lineup L_s in period t, with

[15] m_t is assumed to be predetermined.

its coefficient ρ capturing the peer effect. $X_{st} = [x_{it}]_{i \in L_s}$ is an $m_t \times k_x$ matrix of observations on k_x exogenous variables of the workers in lineup L_s in period t. U_{st} is an $m_t \times 1$ vector of regression disturbances such that $U_{st} \sim i.i.d.(0, \Sigma)$. We allow for possible correlation between U_{st} and $\xi_t = (\xi_{1t}, \cdots, \xi_{q_t,t})$ such that

$$\mathrm{E}(U_{st}|d_{st} = 1, \pi_t) = \lambda_s(\pi_t) 1_{m_t}, \tag{6}$$

where $\pi_t = (\pi_{1t}, \cdots, \pi_{q_t,t})$.

A possible specification of U_{st} that leads to (6) is given by

$$U_{st} = \alpha_{st} 1_{m_t} + V_{st}, \tag{7}$$

where α_{st} is an i.i.d. time-varying lineup-specific error component with mean zero and variance σ_α^2, and V_{st} is an $m_t \times 1$ vector of i.i.d. random innovations with mean zero and variance σ_v^2. The error component α_{st} can be interpreted as a random shock in period t that may affect different lineups differently. Suppose the manager has some information about the realization of α_{st} but no information about that of V_{st} when she chooses a lineup. Then,

$$\mathrm{E}(U_{st}|d_{st} = 1, \pi_t) = \mathrm{E}(\alpha_{st}|\epsilon_{st} < 0, \pi_t) 1_{m_t} = \int\int_{-\infty}^{0} \frac{\alpha_{st} g_{st}(\alpha_{st}, \epsilon_{st}|\pi_t)}{\Pr(\epsilon_{st} < 0|\pi_t)} d\epsilon_{st} d\alpha_{st} 1_{m_t} = \lambda_s(\pi_t) 1_{m_t},$$

where $g_{st}(\alpha_{st}, \epsilon_{st}|\pi_t)$ is the conditional joint density of α_{st} and ϵ_{st}.[16]

Let $U_{st}^* = U_{st} - \lambda_s(\pi_t) 1_{m_t}$. (5) can be written as

$$Y_{st} = \rho W_t Y_{st} + X_{st}\beta + \lambda_s(\pi_t) 1_{m_t} + U_{st}^*. \tag{8}$$

The selectivity bias $\lambda_s(\pi_t)$ introduces a group correlated effect (Manski, 1993) to the model. As pointed out by Dahl (2002), semi-parametric estimation of ρ and β along with the unknown function $\lambda_s(\cdot)$ would face the "the curse of dimensionality" due to the presence of a large number of alternatives. To make the estimation feasible, restrictions need to be imposed on $\lambda_s(\cdot)$. In the following subsections, we consider three different approaches for estimation of (8).

[16] The specification given by (7) is merely an example to motivate the assumption (6). The validity of the proposed estimators does not rely on this specification.

4.2 The parametric selection correction approach

Let $F_{st}(\cdot|\pi_t)$ denote the conditional distribution function of $\epsilon_{st} \equiv \max_{r\neq s} d^*_{rt} - d^*_{st}$. Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal distribution and density respectively. Lee (1983) suggests using the transformation $J_{st}(\cdot) \equiv \Phi^{-1}(F_{st}(\cdot|\pi_t))$ to reduce the dimensionality of the selectivity bias. In terms of $J_{st}(\epsilon_{st})$, the selectivity bias is given by

$$\mathrm{E}(U_{st}|d_{st} = 1, \pi_t) = \mathrm{E}[U_{st}|J_{st}(\epsilon_{st}) < J_{st}(0), \pi_t].$$

Note, by construction, $J_{st}(\epsilon_{st})$ is a standard normal random variable and its marginal distribution does not depend on π_t. However, the joint distribution of U_{st} and $J_{st}(\epsilon_{st})$ may still depend on π_t. As pointed out by Dahl (2002) and Bourguignon et al. (2007), the following assumption is implicitly imposed in Lee (1983).

Assumption 1 *The joint distribution of U_{st} and $J_{st}(\epsilon_{st})$ does not depend on π_t.*

Assumption 1 implies that $\mathrm{E}[U_{st}|J_{st}(\epsilon_{st}) < J_{st}(0), \pi_t] = \mathrm{E}[U_{st}|J_{st}(\epsilon_{st}) < J_{st}(0)]$. Furthermore, to obtain an explicit functional form of the selectivity bias, we make the following assumption that is widely used in empirical studies.

Assumption 2 *U_{st} and $J_{st}(\epsilon_{st})$ are i.i.d. with a joint normal distribution given by*[17]

$$\begin{bmatrix} U_{st} \\ J_{st}(\epsilon_{st}) \end{bmatrix} \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma & \sigma_{12} 1_{m_t} \\ \sigma_{12} 1'_{m_t} & 1 \end{bmatrix}). \tag{9}$$

Given Assumption 2, the selectivity bias is given by

$$\mathrm{E}(U_{st}|d_{st} = 1, \pi_t) = -\sigma_{12} \frac{\phi(J_{st}(0))}{F_{st}(0|\pi_t)} 1_{m_t}. \tag{10}$$

Let $P_{st} = \Pr(d_{st} = 1|\pi_t)$ be the probability of choosing lineup L_s in period t given π_t. As $\mathrm{E}(U_{st}|d_{st} = 1, \pi_t) = \lambda_s(\pi_t) 1_{m_t}$, $J_{st}(0) = \Phi^{-1}(F_{st}(0|\pi_t))$ and $P_{st} = F_{st}(0|\pi_t)$, it follows from (10) that

$$\lambda_s(\pi_t) = -\sigma_{12} \frac{\phi(\Phi^{-1}(P_{st}))}{P_{st}}. \tag{11}$$

[17] The likelihood function of the model based on the joint normal distribution (9) is given in Appendix A.

The transformation using $J_{st}(\cdot)$ greatly reduces the dimensionality of the multiple index function $\lambda_s(\pi_t)$ because it allows $\lambda_s(\pi_t)$ to depend on π_t only through P_{st} with a single unknown parameter σ_{12}. Substitution of (11) into (8) gives

$$Y_{st} = \rho W_t Y_{st} + X_{st}\beta - \sigma_{12}\frac{\phi(\Phi^{-1}(P_{st}))}{P_{st}} 1_{m_t} + U_{st}^{*}. \tag{12}$$

For the network model, Lee's approach can be implemented as follows.

Step 1: Let $\pi_{st} = z_{st}\gamma$, where z_{it} is a $1 \times k_z$ vector of exogenous variables. Then, γ can be estimated by maximizing the likelihood function

$$\ln L = \sum_{t=1}^{T}\sum_{s=1}^{q_t} d_{st} \ln P_{st}. \tag{13}$$

It proves convenient to assume that ξ_{st} is independently and identically Gumbel distributed so that $P_{st} = \exp(z_{st}\gamma)/\sum_{r=1}^{q_t}\exp(z_{rt}\gamma)$ (McFadden, 1974). Then, γ can be estimated by a conditional logit estimator $\hat{\gamma}$.

Step 2: With the predicted probabilities $\hat{P}_{st} = \exp(z_{st}\hat{\gamma})/\sum_{r=1}^{q_t}\exp(z_{rt}\hat{\gamma})$ obtained in the first step, we consider the feasible counterpart of (12)

$$Y_{st} = \rho W_t Y_{st} + X_{st}\beta - \sigma_{12}\frac{\phi(\Phi^{-1}(\hat{P}_{st}))}{\hat{P}_{st}} 1_{m_t} + U_{st}^{**}, \tag{14}$$

and estimate $(\rho, \beta', \sigma_{12})'$ by the two-stage least squares (2SLS) estimator with linearly independent columns in $W_t X_{st}$ as instruments for $W_t Y_{st}$. The correct asymptotic covariance matrix of the 2SLS estimator can be derived in a similar way as in Lee et al. (1980) with appropriate modifications.

4.3 The semi-parametric selection correction approach

Dahl (2002) proposes an alternative selection correction approach based on the index sufficiency assumption that the joint distribution of U_{st} and ϵ_{st} depends on π_t only through $P_{st} = \Pr(d_{st} = 1|\pi_t)$. Based on this idea, we impose the following assumption to reduce the dimensionality of the selectivity bias.

Assumption 3 *$\lambda_s(\pi_t) = \mu(P_{st})$.*

Assumption 3 implies that the multiple index selectivity bias $\mathrm{E}(U_{st}|d_{st}=1,\pi_t)$ depends on π_t only through P_{st}, and, thus, equation (8) becomes

$$Y_{st}=\rho W_t Y_{st}+X_{st}\beta+\mu(P_{st})1_{m_t}+U_{st}^{*}. \tag{15}$$

For the parametric approach, Assumption 2 implies that the functional form of $\mu(\cdot)$ is given by $\mu(P_{st})=-\sigma_{12}\phi(\Phi^{-1}(P_{st}))/P_{st}$. For the semi-parametric approach, we approximate $\mu(P_{st})$ by series expansions (see, Andrews, 1991; Newey, 1997) without imposing functional form assumptions on $\mu(\cdot)$.

Thus, the semi-parametric selection correction approach can be implemented in a similar two-step procedure as the parametric approach.

Step 1: We obtain the predicted probabilities $\hat{P}_{st}$ from, say, a conditional conditional logit regression.

Step 2: We replace $\mu(P_{st})$ in (15) by its (feasible) series approximation $\sum_{k=1}^{K}\kappa_k b_k(\hat{P}_{st})$, where the functions $b_k(\cdot)$ are referred to as the basis functions,[18] and estimate $(\rho,\beta')'$ together with the series expansion coefficients κ_k by the 2SLS estimator with linearly independent columns in W_tX_{st} as instruments for W_tY_{st}.[19]

4.4 The fixed-effect approach

From a different perspective, the selectivity bias $\lambda_s(\pi_t)$ in (8) can be considered as a time-varying lineup-specific fixed effect. To avoid estimating the unknown function $\lambda_s(\cdot)$, we can apply a within transformation to eliminate this term from (8).

Suppose $X_{st}=[X_{1,st},1_{m_t}x_{2,st}]$, where $X_{1,st}$ is an $m_t\times k_1$ matrix of observations on k_1 individual-varying exogenous variables and $x_{2,st}$ is a $1\times k_2$ vector of individual-invariant exogenous variables $(k_1+k_2=k_x)$. Then, equation (8) can be written as

$$Y_{st}=\rho W_t Y_{st}+X_{1,st}\beta_1+1_{m_t}x_{2,st}\beta_2+\lambda_s(\pi_t)1_{m_t}+U_{st}^{*}. \tag{16}$$

Let $Q_t=I_{m_t}-\frac{1}{m_t}1_{m_t}1_{m_t}'$ denote the within-transformation projector. Then, as $Q_t1_{m_t}=0$ and

[18] Dahl (2002) find similar results in his application using either polynomial or Fourier series as basis functions.

[19] For consistency and asymptotic normality, the number of basis functions should increase with the sample size (see, Andrews, 1991; Newey, 1997). In practice, the number of basis functions is chosen by the researcher.

$Q_t U_{st}^* = Q_t U_{st}$, pre-multiplication of (8) by Q_t gives

$$Q_t Y_{st} = \rho Q_t W_t Y_{st} + Q_t X_{1,st}\beta_1 + Q_t U_{st}. \tag{17}$$

Then, ρ and β_1 can be estimated from the within model (17) by the conditional maximum likelihood (CML) approach in Lee (2007).[20]

The fixed-effect approach does not impose any restrictions on $\lambda_s(\pi_t)$. However, given the special structure of the weighting matrix W_t, the workers in the chosen lineup form a complete network. The within transformation may cause an identification problem similar to the one studied in Lee (2007). This can be seen from the reduced form equation of (17). Suppose $|\rho| < 1$, then it follows from (5) that

$$Y_{st} = (I_{m_t} - \rho W_t)^{-1} X_{1,st}\beta_1 + (I_{m_t} - \rho W_t)^{-1} U_{st}. \tag{18}$$

For $W_t = \frac{1}{m_t - 1}(1_{m_t} 1'_{m_t} - I_{m_t})$, we have $Q_t(I_{m_t} - \rho W_t)^{-1} = \frac{m_t - 1}{m_t - 1 + \rho} Q_t$. Therefore, pre-multiplication of (18) by Q_t gives

$$Q_t Y_{st} = \frac{m_t - 1}{m_t - 1 + \rho} Q_t X_{1,st}\beta_1 + \frac{m_t - 1}{m_t - 1 + \rho} Q_t U_{st}. \tag{19}$$

From (19), we can see that the within model (17) can be identified if m_t varies over t. On the other hand, if $m_t = m$ for all t, then the peer effect coefficient ρ cannot be identified from β_1 after the within transformation.

To identify the peer effect when $m_t = m$ for all t, we need to introduce some exclusion restrictions. One possibility is to introduce heterogenous peer effects. Let $W_{1s}^o = [w_{ij,1s}^o]$ be an adjacency matrix with $w_{ij,1s}^o = 1$ if the ith and jth workers in the lineup s are of the same type and $w_{ij,1s}^o = 0$ otherwise. Let $W_{2s}^o = [w_{ij,2s}^o]$ be an adjacency matrix with $w_{ij,2s}^o = 1$ if the ith and jth workers in the lineup s are of different types and $w_{ij,2s}^o = 0$. By construction, $\frac{1}{m-1}(W_{1s}^o + W_{2s}^o) = W \equiv \frac{1}{m-1}(1_m 1'_m - I_m)$. Let W_{1s} and W_{2s} be row-normalized W_{1s}^o and W_{2s}^o respectively, such that $W_{1s} 1_m = W_{2s} 1_m = 1_m$.[21]

[20] The CML estimator is consistent and asymptotically normal as in Lee (2007) as the sample size $\sum_{t=1}^{T} m_t$ goes to infinity.

[21] Sometimes, W_{s1}^0 (or W_{s2}^0) may have a row of zeros. For example, if worker i has no co-worker of the same type in a lineup, then $w_{ij,1s}^o = 0$ for all j. Then, the corresponding row of W_{s1} (or W_{s2}) is also zero. As a result, $W_{s1} 1_{m_t} \neq 1_{m_t}$ (or $W_{s2} 1_{m_t} \neq 1_{m_t}$), and the likelihood function cannot be derived for the transformed dependent variable $Q_t Y_{st}$ (see Liu and Lee, 2010). In this case, the model after within transformation given by (21) can be estimated by the GMM approach in Liu and Lee (2010).

Then, (16) can be generalized to a model with heterogenous peer effects given by

$$Y_{st} = \rho_1 W_{1s} Y_{st} + \rho_2 W_{2s} Y_{st} + X_{1,st}\beta_1 + 1_{m_t} x_{2,st}\beta_2 + \lambda_s(\pi_t)1_{m_t} + U^*_{st}, \tag{20}$$

where ρ_1 captures the same-type peer effect and ρ_2 captures the cross-type peer effect. Premultiplying (20) by $Q = I_m - \frac{1}{m}1_m 1'_m$, we have

$$QY_{st} = \rho_1 QW_{1s} Y_{st} + \rho_2 QW_{2s} Y_{st} + QX_{1,st}\beta_1 + QU_{st}. \tag{21}$$

As $(I_m - \rho_1 W_{1s} - \rho_2 W_{2s})^{-1} = \rho_1 W_{1s}(I_m - \rho_1 W_{1s} - \rho_2 W_{2s})^{-1} + \rho_2 W_{2s}(I_m - \rho_1 W_{1s} - \rho_2 W_{2s})^{-1} + I_m$, it follows from the reduced form equation of (20) that

$$\begin{aligned} QY_{st} &= Q(I_m - \rho_1 W_{1s} - \rho_2 W_{2s})^{-1} X_{1,st}\beta_1 + Q(I_{m_1} - \rho_1 W_{1s} - \rho_2 W_{2s})^{-1} U^*_{st} \qquad (22) \\ &= \rho_1 QW_{1s}(I_m - \rho_1 W_{1s} - \rho_2 W_{2s})^{-1} X_{1,st}\beta_1 + \rho_2 QW_{2s}(I_m - \rho_1 W_{1s} - \rho_2 W_{2s})^{-1} X_{1,st}\beta_1 \\ &\quad + QX_{1,st}\beta_1 + Q(I_{m_1} - \rho_1 W_{1s} - \rho_2 W_{2s})^{-1} U^*_{st} \\ &= \rho_1 \mathrm{E}(QW_{1s} Y_{st}|X_{st}, d_{st}) + \rho_2 \mathrm{E}(QW_{2s} Y_{st}|X_{st}, d_{st}) + QX_{1,st}\beta_1 \\ &\quad + Q(I_{m_1} - \rho_1 W_{1s} - \rho_2 W_{2s})^{-1} U^*_{st}. \end{aligned}$$

Therefore, $(\rho_1, \rho_2, \beta'_1)$ can be separately identified if $\mathrm{E}(QW_{1s} Y_{st}|X_{st}, d_{st})$, $\mathrm{E}(QW_{2s} Y_{st}|X_{st}, d_{st})$, and columns in $QX_{1,st}$ are linearly independent for some t.

To better understand this identification condition, we consider a special case that $\rho_1 = \rho_2 = 0$ in the data generating process. In this case, it follows from the reduced form equation of (20) that $\mathrm{E}(QW_{1s} Y_{st}|X_{st}, d_{st}) = QW_{1s} X_{1,st}\beta_1$ and $\mathrm{E}(QW_{2s} Y_{st}|X_{st}, d_{st}) = QW_{2s} X_{1,st}\beta_1$. Thus, a necessary condition for $\mathrm{E}(QW_{1s} Y_{st}|X_{st}, d_{st})$, $\mathrm{E}(QW_{2s} Y_{st}|X_{st}, d_{st})$, and $QX_{1,st}$ to be linearly independent is that QW_{1s}, QW_{2s} and Q are linearly independent. Although $QW^0_{1s} + QW^0_{2s} = (m-1)QW = -Q$, for the row normalized adjacency matrices W_{1s} and W_{2s}, QW_{1s}, QW_{2s} and Q can still be linearly independent in general. Therefore, model (21) can be identified. The CML estimator in Lee et al. (2010) can be easily generalized to estimate (21).

To summarize, for model (16), the fixed-effect approach can be implemented by the following steps.

Step 1: We estimate the within equation (17) by the CML estimator in Lee (2007).

Step 2: We obtain the predicted probabilities $\hat{P}_{st}$ from, say, a conditional logit regression.

Step 3: Let $\hat{r}_{st} = \frac{1}{m} 1_m'(Y_{st} - \hat{\rho} W_t Y_{st} - X_{1,st}\hat{\beta}_1)$, where $\hat{\rho}$ and $\hat{\beta}_1$ are the first-step estimates. We consider the regression

$$\hat{r}_{st} = x_{2,st}\beta_2 + \mu(\hat{P}_{st}) + \zeta_{st}, \tag{23}$$

where the selectivity bias $\mu(\hat{P}_{st})$ is either given by $-\sigma_{12}\phi(\Phi^{-1}(\hat{P}_{st}))/\hat{P}_{st}$ in the parametric approach or approximated by $\sum_{k=1}^{K} \kappa_k b_k(\hat{P}_{st})$ in the semi-parametric approach, and ζ_{st} is the error term. We estimate β_2 together with the unknown parameters in $\mu(\hat{P}_{st})$ by the OLS estimator.

4.5 Comparison of the estimation approaches

Like other Heckman-type two-step selection bias correction procedures, the two approaches proposed in Sections 4.2 and 4.3 have the advantage of computational simplicity. However, both approaches impose strong restrictions on the selectivity bias $\lambda_s(\pi_t)$ to reduce its dimensionality.[22] Furthermore, because of the endogeneity of the peer effect regressor, the model needs to estimated by the 2SLS estimator that relies on the existence of valid and relevant instruments. This may be quite challenging in empirical applications. In our empirical example, for instance, the valid instruments are quite weak, although we experimented with several sets of instruments. Therefore, the 2SLS estimates may not be reliable.

On the other hand, the fixed effect approach proposed in Section 4.4 does not impose any restrictions on the selectivity bias $\lambda_s(\pi_t)$. After we eliminate the selectivity bias using the within transformation, we can use the CML or GMM estimator to estimate the peer effect. The CML and GMM exploit both linear and quadratic moment conditions, and, thus, may outperform the 2SLS estimator that only uses linear moment conditions, when the linear moment conditions are weak (see, Lee et al., 2010; Liu and Lee, 2010). However, as shown in Section 4.4, the within transformation makes the identification of the peer effect more challenging because the workers in the chosen lineup form a complete network. In particular, we show that the within equation is not identified if m_t does not vary over time. In this case identification can be achieved by imposing exclusion restrictions through heterogenous peer effects.

[22] See Assumptions 1 and 2 for the parametric approach and Assumption 3 for the semi-parametric approach.

5 An Empirical Illustration

As an empirical illustration, we estimate a network production function for a basketball team, where a coach selects lineups of players over the course of a game. As the valid instruments turn out to be quite weak in the empirical example, with the first stage F statistic lower than 5 (J. and Yogo, 2005), estimators that leverages 2SLS may not be reliable for this data. Hence, we use the fixed effect estimation approach. As the number of active players is constant over time (i.e., $m_t = m$), we split players into two types, guards and forwards, to identify the peer effects. We detail the application of the fixed effect estimator for the specification considered in the empirical example in Appendix B.

5.1 Data

Our data are for the Syracuse University Men's Basketball team over the 2011-2012 season. The team played 33 games during the regular season (we exclude March Madness games). We define a time period as the time interval between two consecutive substitutions.[23] We removed overtime periods from the data, since the manager's allocation strategy may be different in overtime. We removed time periods of less than 30 seconds, since there might not be enough observations on players' productivities in those extremely short periods. We thus observe 79 different lineups (of 5 active players) over 448 time periods, in total 2,240 observations.[24]

There are two outputs in a basketball game: the production of offense/defense (some measure related to the "on court" productivity of active players) and rest (players sitting on the bench).[25] We take the opposing team's strategy as exogenous, using only a measure of the team's Rating Percentage Index (RPI) from the previous year which we describe below.

5.2 Variable definition

The dependent variable Y_{st} of equation (20) is measured using using the efficiency statistic EFF_{it}:

$$EFF_{it} = (PT_{it} + REB_{it} + AST_{it} + STL_{it} + BLK_{it} - MFG_{it} - MFT_{it} - TO_{it})/Mins_{it}$$

[23] Our time periods have irregular length.

[24] An important problem, which is common to most existing empirical studies, is a possible misspecification of the network structure. The main threats are sampling issues due to the fact that only a subset of connections are observed (see, e.g., Chandrasekhar and Lewis, 2011; Lin, 2013; Liu et al., 2013). In our case, the coach selects lineups to produce output, so that networks are accurately measured.

[25] It could be argued that there a multiple offensive outputs (points, rebounds, assists, etc.) and multiple defensive outputs (steals, blocks, rebounds, etc.). However our purpose is to illustrate the econometric contribution, and not to perform a comprehensive empirical analysis.

where PT_{it} is points, REB_{it} is rebounds, AST_{it} is assists, STL_{it} is steals, BLK_{it} is blocks, MFG_{it} is missed field goals, MFT_{it} is missed free throws, TO_{it} is turn overs, and $Mins_{it}$ is minutes played for player i in period t.[26] These are period-by-period statistics and not season-long aggregates. Over the course of the entire season and across players the average efficiency is 0.37 with a standard deviation of 1.07, a minimum of -3.75, and a maximum of 8.28. This is not calculated when a player is on the bench.

The individual-varying exogenous variables in the main equation (the $X_{1,st}$'s) are $Experience_{it}$ and $Fatigue_{it}$. $Experience_{it}$ is minutes played from the start of the game to the end of period $t-1$. It has and average of 9.91 minutes, a standard deviation of 7.81, a minimum of 0, and a maximum of 37.58 minutes. For active player i in period $t-1$, $Fatigue_{it}$ is minutes *continuously* played until the end of period $t-1$; for inactive players in period $t-1$, fatigue is 0. The average fatigue across the entire season is 3.78 minutes with a standard deviation of 5.09 minutes. The high variance is due to the fact that there are players who almost always continuously play and those who almost never play.

The exogenous variables that do not vary over i in the main equation (the $x_{2,st}$'s) are the opposing team's Rating Percentage Index (RPI_t), $Home_t$, a dummy variable equal to 1 if the game is played in the Syracuse University Carrier Done (two-thirds of the games were played at home in the 2011-2012 season), and $2nd\text{-}Half_t$, a dummy variable equal to 1 if the current period is in the second half of the game. The rating percentage index is one of the systems used to rank NCAA teams and is based on a teams wins, losses and its strength of schedule. This system has been in use in college basketball since 1981 to aid in the selecting and seeding of teams appearing in the 68-team men's tournament (March Madness). The index is based on a team's winning percentage, its opponents' winning percentage, and the winning percentage of those opponents' opponents. For the teams in our data the average RPI from the 2010-11 season is 0.55 with a standard deviation of 0.08.

The exogenous variables in the selection equation are lineup-level aggregations of variables from the main equation. $Lineup\text{-}efficiency_{st}$ is the total efficiency score of the lineup s from the start of the game until the end of period $t-1$. It has an average of 1.63 and a standard deviation of 1.22. $Lineup\text{-}experience_{st}$ is the total minutes played by the lineup at the end of period $t-1$. It has an average of 49.57 minutes and a standard deviation of 34.03 minutes. $Lineup\text{-}fatigue_{st}$ is

[26] This assumes equal weights for each individual productive activities. Other weighting schemes could be considered, but a similar efficiency measure is employed by the National Basketball Association to rank player productivity, so we use it as a matter of convenience.

the total minutes continuously played by the lineup at the end of period $t-1$. It has an average of 18.92 minutes and a standard deviation of 14.06 minutes. $Lineup\text{-}fouls_{st}$ is the total fouls by the lineup at the end of period $t-1$. It has an average of 3.13 fouls and a standard deviation of 2.77 fouls. $One\text{-}substitution_t$ is a dummy variable equal to 1, if one player was substituted to achieve the lineup at time t. It has an average of 0.68 and a standard deviation of 0.46. $Two\text{-}substitution_t$ is a dummy variable equal to 1, if two players were substituted to achieve the lineup at time t. It has an average of 0.22 and a standard deviation of 0.42. The omitted category is three or more players were substituted. Variable definitions and descriptive statistics are summarized in Appendix C (Table C.1).

5.3 Estimation results

5.3.1 Results without selectivity bias correction

Let us start by presenting the ML estimation results without accounting for selectivity bias (Lee, 2004). As the number of time periods ($T = 448$) is much larger than the number of players in the Syracuse University Men's Basketball team ($n = 19$), we can use player dummies to control for unobserved player-specific characteristics. Results are contained in Table 1.

[insert Table 1 here]

Model 1 considers the benchmark outcome equation (5) with homogenous peer effects.[27] The estimation results are reported in column 1 of Table 1. In line with expectations, it appears that player's experience is positively correlated with his productivity (0.0154 efficiency units per minute played), and the effect of fatigue is negative (-0.0083 efficiency units per minute continuously played), although it is not statistically significant. The quality of the opposing team plays a strong role in decreasing player's productivity (statistically significant -1.1677), and the second-half of a game seems to be less productive that the first half (significant -0.2159). Peer effects in productivity appear positive and statistically significant. In terms of magnitude, an unit increase in the average efficiency of the teammates induces a 0.0841 increase in the efficiency of the individual player.

Model 2 of Table 1 considers heterogenous peer effects. We split players into two types, guards and forwards (no differentiation of centers from forwards), and distinguish between peer effects arising from "same-type" teammates and peer effects arising from "cross-type" teammates. The estimation results are reported in column 2 of Table 1. It appears that the peer effects are mostly

[27] We assume normality of the error distribution so that $U_{st} \sim i.i.d.N(0, \sigma^2 I_m)$.

due to interactions between players of the same type. The same-type peer effect is 0.0638 (significant) and the cross-type peer effect is 0.0345 (insignificant). It appears that once we condition on observed and unobserved player characteristics, there are no endogenous effects at work between players of different types.

Model 3 of Table 1 is the restricted heterogenous peer effects where we only consider the same-type peer effect. The estimation results reported in column 3 of Table 1 remain roughly unchanged from Model 2.

5.3.2 Results with selectivity bias correction

As explained in Section 4.4 and detailed in Appendix B, the fixed-effect approach can be implemented in three steps. First, we use a within transformation to eliminate selectivity bias and estimate the transformed outcome equation by the CML approach (detailed in Section B.2). Covariates that do not vary at the individual level (RPI, $Home$ and $2nd\text{-}Half$) are eliminated by the within group transformation. As the number of active players is constant over time (i.e. $m_t = m$), the transformed outcome equation is not identified for Model 1. Hence, we have to exploit heterogenous peer effects to achieve identification. The fixed-effect CML estimation results are reported in Table 2.

[insert Table 2 here]

With both same-type and cross-type peer effects in Model 2, the peer effects are not significant due to multicollinearity of those two effects in our data. When we only consider the same-type peer effect in Model 3, the peer effect is positive and statistically significant, but lower in magnitude than the corresponding estimate in Table 1 without selectivity bias correction. In line with the estimates in Table 1, a player's experience is positively associated with her performance. The effect of fatigue is negative and becomes statistically significant once selectivity bias is corrected. Furthermore, the likelihood ratio test (test statistic is 0.96) fails to reject the restriction that cross-type peer effect is zero at conventional significance levels.

[insert Table 3 here]

Table 3 reports the second step conditional logit estimation of the selection equation.[28] The estimates reveal the factors that are important when the coach selects the lineup. In particular, the

[28] To reduce the total number of alternatives, we restrict the set of possible lineups to the lineups that actually employed by the coach in a game.

past productivity, fatigue and number of fouls of the players in a lineup play important roles in the coach's lineup choices.

[insert Table 4 here]

Table 4 reports the third step, where the effects of the individual-invariant regressors are recovered and the selectivity bias is estimated. The estimation procedure is detailed in Section B.3. For the parametric approach, the joint normality assumption (9) implies the selectivity bias has a specific functional form (11) with a single unknown parameter σ_{12}. For the semiparametric approach, the selectivity bias is approximated by a series expansion. For the parametric approach, the estimate of σ_{12} is insignificant. When parametric restrictions are removed, the Wald test suggests the coefficients of the series expansion are jointly significant and hence selection does play a role in the outcome equation. The estimated effects of the individual-invariant regressors are in line with the estimates in Table 1 except the coefficient of the home game dummy is now positive.

6 Conclusion

This paper makes contributions to both the network and production function literatures. The proposed network production function mitigates traditional problems in the identification and estimation of peer effects, including endogenous network formation and network topology misspecification. In our proposed model, the network is (and peer groups are) well-defined, and selection into groups is not an individual choice but the decision of a manager (social planner) who has historical information on the observable and unobservable characteristics of the workers. This allows selection into a single project to be at the team-level, and allows the network structure to be fixed by the manager (predetermined for the workers), who selects teams (lineups) into the set structure. The selection process can be modelled in a Heckman-type framework (Heckman, 1979). Being at the team level, the selection correction term captures the "correlated effects" of Manski (1993). Thus, our approach tackles in a single step the selection and the corrected effects problems in the network literature. The solution comes at a cost of the need for administrative data on each worker's history which may not be readily available.

Regarding the production function literature, our analysis considers issues related to the estimation of managerial efficiency (the managerial selection bias correction term), the determinants of efficiency through the selection equation, and multi-output (project) distance functions.

Our empirical example suggests that peer effects exist among players in a basketball game and that a selectivity bias correction matters.

References

Aigner, D., Lovell, C. and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models, *Journal of Econometrics* **6**: 21–37.

Alvarez, A., Amsler, C., Orea, L. and Schmidt, P. (2006). Interpreting and testing the scaling property in models where inefficiency depends on firms characteristics, *Journal of Productivity Analysis* **25**: 201–212.

Andrews, D. W. K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models, *Econometrica* **59**: 307–345.

Bandiera, O., Barankay, I. and Rasul, I. (2009). Social connections and incentives in the workplace: evidence from personnel data, *Econometrica* **77**: 1047–1094.

Blume, L. E., Brock, W. A., Durlauf, S. N. and Ioannides, Y. M. (2011). Identification of social interactions, *in* J. Benhabib, A. Bisin and M. O. Jackson (eds), *Handbook of Social Economics*, Vol. 1B, North-Holland, pp. 855–966.

Bourguignon, F., Fournier, M. and Gurgand, M. (2007). Selection bias corrections based on the multinomial logit model: Monte carlo comparisons, *JOURNAL OF ECONOMIC SURVEYS* **21**: 174–205.

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

Calvó-Armengol, A., Patacchini, E. and Zenou, Y. (2009). Peer effects and social networks in education, *The Review of Economic Studies* **76**: 1239–1267.

Chandrasekhar, A. and Lewis, R. (2011). Econometrics of sampled networks. Working paper, MIT.

Dahl, G. (2002). Mobility and the returns to education: testing a roy model with multiple markets, *Econometrica* **70**: 2367–2420.

Druska, V. and Horrace, W. C. (2004). Generalized moments estimation for spatial panel data: Indonesian rice farming, *American Journal of Agricultural Economics* **86**: 185–198.

Falk, A. and Ichino, A. (2006). Clean evidence on peer effects, *Journal of Labor Economics* **24**: 39–58.

Fare, R. and Primont, D. (1990). A distance function approach to multioutput technologies, *Southern Economic Journal* **56**: 879–891.

Glass, A., Kenjegalieva, K. and Sickles, R. (2013). A spatial autoregressive production frontier model for panel data: with an application to european countries. Working Paper, Rice Univeristy.

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

Guryan, J., Kroft, K. and Notowidigdo, M. (2009). Peer effects in the workplace: evidence from random groupings in professional golf, *American Economic Journal: Applied Economics* **1**: 34–68.

Hamilton, B. H., Nickerson, J. and Owan, H. (2003). Team incentives and worker heterogeneity: an empirical analysis of the imact of teams on productivity and participation, *Journal of Politicial Economy* **111**: 465–497.

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

Hsieh, C. S. and Lee, L. F. (2013). A social interaction model with endogenous friendship formation and selectivity. Working Paper, Ohio State Univeristy.

J., S. and Yogo, M. (2005). Testing for weak instruments in linear iv regression, *in* D. W. K. Andrews and J. Stock (eds), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, Cambridge University Press, pp. 80–108.

Jackson, M. O. and Zenou, Y. (eds) (2013). *Economic Analyses of Social Networks*, Edward Elgar Publishing.

Kandel, E. and Lazear, E. (1992). Peer pressure and partnerships, *Journal of Political Economy* **100**: 801–817.

Lee, L. F. (1983). Generalized econometric models with selectivity, *Econometrica* **51**: 507–512.

Lee, L. F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial econometric models, *Econometrica* **72**: 1899–1926.

Lee, L. F. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects, *Journal of Econometrics* **140**: 333–374.

Lee, L. F., Liu, X. and Lin, X. (2010). Specification and estimation of social interaction models with network structures, *The Econometrics Journal* **13**: 145–176.

Lee, L. F., Maddala, G. S. and Trost, R. P. (1980). Asymptotic covariance matrices of two-stage probit and two-stage tobit methods for simultaneous equations models with selectivity, *Econometrica* **48**: 491–503.

Lee, L. F. and Yu, J. (2010). Some recent developments in spatial panel data models, *Regional Science and Urban Economics* **40**: 255–271.

Lin, X. (2013). Estimation of a local-aggregate network model with sampled networks, *Economics Letters* **118**: 243–246.

Liu, X. and Lee, L. F. (2010). Gmm estimation of social interaction models with centrality, *Journal of Econometrics* **159**: 99–115.

Liu, X., Patacchini, E. and Rainone, E. (2013). The allocation of time in sleep: a social network model with sampled data. CPR working paper No. 162, Center for Policy Research, Syracuse University.

Liu, X., Patacchini, E., Zenou, Y. and Lee, L. F. (2012). Criminal networks: Who is the key player? CEPR working paper No. 8185.

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

Mas, A. and Moretti, E. (2009). Peers at work, *American Economic Review* **99**: 112–145.

McFadden, D. L. (1974). Conditional logit analysis of qualitative choice behavior, *in* P. Zarembka (ed.), *Frontiers in Econometrics*, New York: Academic Press, pp. 105–142.

Meeusen, W. and van den Broeck, J. (1977). Efficiency estimation from cobb-douglas production functions with composed error, *International Economic Review* **18**: 435–444.

Mele, A. (2013). A structural model of segregation in social networks. working paper, Johns Hopkins University.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators, *Journal of Econometrics* **79**: 147–168.

Appendices

A The Likelihood Function under Joint Normality

If $U_{st} \sim N(0, \Sigma)$, the density function of Y_{st} is

$$f(Y_{st}) = (2\pi)^{-m_t/2}|\Sigma|^{-1/2}|I - \rho W_t| \exp(-\frac{1}{2}(Y_{st} - \rho W_t Y_{st} - X_{st}\beta)'\Sigma^{-1}(Y_{st} - \rho W_t Y_{st} - X_{st}\beta)).$$

Furthermore, if U_{st} and $J_{st}(\epsilon_{st})$ are i.i.d. with a joint normal distribution given by (9), the conditional distribution of $J_{st}(\epsilon_{st})$ given Y_{st} is

$$J_{st}(\epsilon_{st})|Y_{st} \sim N(\sigma_{12}1_m'\Sigma^{-1}(Y_{st} - \rho W_t Y_{st} - X_{st}\beta), 1 - \sigma_{12}^2 1_m'\Sigma^{-1}1_m).$$

Then, the log-likelihood function of equations (4) and (5) is given by

$$\begin{aligned} \ln L &= \sum_{t=1}^{T}\sum_{s=1}^{p_t} d_{st}[\ln f(J_{st}(\epsilon_{st})|Y_{st}) + \ln f(Y_{st})] \\ &= \sum_{t=1}^{T}\sum_{s=1}^{p_t} d_{st}[\ln \Phi(\frac{J_{st}(z_{st}\gamma) - \sigma_{12}1_m'\Sigma^{-1}(Y_{st} - \rho W_t Y_{st} - X_{st}\beta)}{\sqrt{1 - \sigma_{12}^2 1_m'\Sigma^{-1}1_m}}) \\ &\quad -\frac{m_t}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| + \ln|I_{m_t} - \rho W_t| - \frac{1}{2}(Y_{st} - \rho W_t Y_{st} - X_{st}\beta)'\Sigma^{-1}(Y_{st} - \rho W_t Y_{st} - X_{st}\beta)]. \end{aligned}$$

B The Empirical Model and the Fixed Effect Estimator

In this appendix, we detail the fixed effect estimator for the specification considered in the empirical example.

B.1 The empirical model

In the empirical application, we assume the manager chooses lineup s in period t (i.e., $d_{st} = 1$), if $d_{st}^* > \max_{r \neq s} d_{rt}^*$, where

$$d_{st}^* = z_{st}\gamma + \xi_{st}, \quad \text{for } s = 1, \cdots, q_t.$$

We assume ξ_{st} is independently and identically Gumbel distributed so that $P_{st} = \Pr(d_{st} = 1) = \exp(z_{st}\gamma)/\sum_{r=1}^{q_t}\exp(z_{rt}\gamma)$.

The outcome equation of the chosen lineup s in period t is given by

$$Y_{st} = \rho_1 W_{1s} Y_{st} + \rho_2 W_{2s} Y_{st} + X_{1,st}\beta_1 + 1_m x_{2,st}\beta_2 + U_{st}, \tag{24}$$

where $U_{st} = \alpha_{st} 1_m + V_{st}$ with $\alpha_{st} \sim N(0, \sigma_\alpha^2)$ and $V_{st} \sim N(0, \sigma_v^2 I_m)$. We assume, when the manager chooses a lineup, she has no information about the realization of individual random innovations V_{st} but may has some information about the random shock α_{st}. Thus,

$$\mathrm{E}(U_{st}|d_{st} = 1, \pi_t) = \mathrm{E}(\alpha_{st}|d_{st} = 1, \pi_t)1_m = \lambda_s(\pi_t)1_m,$$

where $\pi_t = (\pi_{1t}, \cdots, \pi_{q_t,t})$ and $\pi_{st} = z_{st}\gamma$. Then, the selection bias corrected outcome equation is

$$Y_{st} = \rho_1 W_{1s} Y_{st} + \rho_2 W_{2s} Y_{st} + X_{1,st}\beta_1 + 1_m x_{2,st}\beta_2 + \lambda_s(\pi_t)1_m + U_{st}^*,$$

where $U_{st}^* = U_{st} - \lambda_s(\pi_t)1_m$. By construction, $\mathrm{E}(U_{st}^*|d_{st} = 1, \pi_t) = 0$.

B.2 Estimation of the peer effect

To estimate the peer effect coefficients (ρ_1, ρ_2), we first eliminate the selectivity bias using a within transformation. Premultiplying (24) by $Q = I_m - \frac{1}{m}1_m 1_m'$, we have

$$QY_{st} = \rho_1 QW_{1s}Y_{st} + \rho_2 QW_{2s}Y_{st} + QX_{1,st}\beta_1 + QV_{st}. \tag{25}$$

To estimate (25), we generalize the CML approach in Lee, Liu and Lin (2010). The transformed disturbances QV_{st} in (25) are linearly dependent because its variance matrix $\sigma^2 Q$ is singular. Following Lee, Liu and Lin (2010), we consider an equivalent but more effective transformation. Let the orthonormal matrix of Q be $[P, 1_m/\sqrt{m}]$. The columns in P are eigenvectors of Q corresponding to the eigenvalue one, such that $P'1_m = 0$, $P'P = I_{m-1}$ and $PP' = Q$. Therefore, premultiplying (24) by P' gives

$$P'Y_{st} = \rho_1 P'W_{1s}Y_{st} + \rho_2 P'W_{2s}Y_{st} + P'X_{1,st}\beta_1 + P'V_{st}. \tag{26}$$

Let $\bar{Y}_{st} = P'Y_{st}$, $\bar{X}_{1,st} = P'X_{1,st}$, $\bar{V}_{st} = P'V_{st}$, $\bar{W}_{1s} = P'W_{1s}P$, and $\bar{W}_{2s} = P'W_{2s}P$. As

$P'W_{1s} = \bar{W}_{1s}P'$ and $P'W_{2s} = \bar{W}_{2s}P'$, (26) can be rewritten as

$$\bar{Y}_{st} = \rho_1 \bar{W}_{1s}\bar{Y}_{st} + \rho_2 \bar{W}_{2s}\bar{Y}_{st} + \bar{X}_{1,st}\beta_1 + \bar{V}_{st}, \tag{27}$$

where $\bar{V}_{st} \sim N(0, \sigma_v^2 I_{m-1})$. Hence, $(\rho_1, \rho_2, \beta_1', \sigma_v^2)$ can be estimated by maximizing the conditional likelihood function is given by

$$\begin{aligned} \ln L &= \sum_{t=1}^{T}\sum_{s=1}^{q_t} d_{st}[-\frac{m-1}{2}\ln(2\pi\sigma_v^2) + \ln|I_{m-1} - \rho_1\bar{W}_{1s} - \rho_2\bar{W}_{2s}| \\ &\quad -\frac{1}{2\sigma_v^2}(\bar{Y}_{st} - \rho_1\bar{W}_{1s}\bar{Y}_{st} - \rho_2\bar{W}_{2s}\bar{Y}_{st} - \bar{X}_{1,st}\beta_1)'(\bar{Y}_{st} - \rho_1\bar{W}_{1s}\bar{Y}_{st} - \rho_2\bar{W}_{2s}\bar{Y}_{st} - \bar{X}_{1,st}\beta_1)]. \end{aligned}$$

B.3 Estimation of the selectivity bias

Let $r_{st} = \frac{1}{m}1_m'(Y_{st} - \rho_1 W_{1s}Y_{st} - \rho_2 W_{2s}Y_{st} - X_{1,st}\beta_1)$. Then,

$$r_{st} = x_{2,st}\beta_2 + \lambda_s(\pi_t) + \zeta_{st}, \tag{28}$$

where $\zeta_{st} = -\lambda_s(\pi_t) + \alpha_{st} + \frac{1}{m}1_m' V_{st}$. Then, β_2 and unknown parameters in $\lambda_s(\pi_t)$ can be estimated from (28) with r_{st} replaced by $\hat{r}_{st} = \frac{1}{m}1_m'(Y_{st} - \hat{\rho}_1 W_{1s}Y_{st} - \hat{\rho}_2 W_{2s}Y_{st} - X_{1,st}\hat{\beta}_1)$, where $(\hat{\rho}_1, \hat{\rho}_2, \hat{\beta}_1')$ are the CML estimates. In this appendix, we give the asymptotic covariance of the OLS estimator for the parametric selection-bias correction approach. The asymptotic covariance of the semiparametric estimator can be derived in a similar way with appropriate modifications (see Dahl, 2002, footnote 24).

Under the joint normality assumption

$$\begin{bmatrix} \alpha_{st} \\ J_{st}(\epsilon_{st}) \end{bmatrix} \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\alpha^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{bmatrix}),$$

we have $\lambda_s(\pi_t) = \mathrm{E}(\alpha_{st}|d_{st} = 1, \pi_t) = \mathrm{E}(\alpha_{st}|J_{st}(\epsilon_{st}) < J_{st}(0)) = -\sigma_{12}\varphi(J_{st}(0)) = -\sigma_{12}\varphi(\Phi^{-1}(P_{st}))$, where $\varphi(\cdot) = \phi(\cdot)/\Phi(\cdot)$. Hence, (28) can be written as

$$r_{st} = x_{2,st}\beta_2 - \sigma_{12}\varphi(\Phi^{-1}(P_{st})) + \zeta_{st}.$$

Let $\varphi_{st} = \varphi(\Phi^{-1}(P_{st}))$ and $\hat{\varphi}_{st} = \varphi(\Phi^{-1}(\hat{P}_{st}))$, where $\hat{P}_{st} = \exp(z_{st}\hat{\gamma})/\sum_{r=1}^{q_t}\exp(z_{rt}\hat{\gamma})$ and $\hat{\gamma}$ is the conditional logit estimator. The (infeasible) OLS estimator of $\delta = (\beta_2', \sigma_{12})'$ is given by

$$\tilde{\delta} = (\sum_{t=1}^{T} \hat{h}_t' \hat{h}_t)^{-1} \sum_{t=1}^{T} \hat{h}_t' \sum_{s=1}^{q_t} d_{st} r_{st} = \delta + (\sum_{t=1}^{T} \hat{h}_t' \hat{h}_t)^{-1} \sum_{t=1}^{T} \hat{h}_t' \sum_{s=1}^{q_t} d_{st} [\zeta_{st} + \sigma_{12}(\hat{\varphi}_{st} - \varphi_{st})],$$

where $\hat{h}_t = \sum_{s=1}^{q_t} d_{st}(x_{2,st}, -\hat{\varphi}_{st})$. Let $A_{st} = \Phi^{-1}(P_{st})\varphi_{st} + \varphi_{st}^2$ and

$$\Gamma_t = \sum_{s=1}^{q_t} d_{st} \frac{\partial \varphi(\Phi^{-1}(P_{st}))}{\partial \gamma'} = -\sum_{s=1}^{q_t} d_{st} A_{st} \frac{1}{\phi(\Phi^{-1}(P_{st}))} (P_{st} - P_{st}^2) z_{st}.$$

Let $A = \text{diag}\{\sum_{s=1}^{q_t} d_{st} A_{st}\}_{t=1,\cdots,T}$, $H = (h_1', \cdots, h_T')'$, and $\Gamma = (\Gamma_1', \cdots, \Gamma_T')'$. We have $\sqrt{T}(\tilde{\delta} - \delta) \xrightarrow{d} N(0, \text{plim}(\frac{1}{T} H'H)^{-1} \Omega (\frac{1}{T} H'H)^{-1})$, where

$$\Omega = \frac{1}{T} H' (\sigma_\zeta^2 I_T - \sigma_{12}^2 A + \sigma_{12}^2 \Gamma \Sigma_\gamma \Gamma') H,$$

with $\Sigma_\gamma = [\sum_{t=1}^{T} \sum_{s=1}^{q_t} P_{st}(z_{st} - \sum_{s=1}^{q_t} P_{rt} z_{rt})'(z_{st} - \sum_{s=1}^{q_t} P_{rt} z_{rt})]^{-1}$.[29] Furthermore, under certain regularity conditions, we can show that the feasible OLS estimator

$$\hat{\delta} = (\sum_{t=1}^{T} \hat{h}_t' \hat{h}_t)^{-1} \sum_{t=1}^{T} \hat{h}_t' \sum_{s=1}^{q_t} d_{st} \hat{r}_{st}$$

is asymptotically equivalent to $\tilde{\delta}$.

C Data Description

[insert Table C.1 here]

[29] As $\text{Var}(\zeta_{st} | d_{st} = 1) = \sigma_\zeta^2 - \sigma_{12}^2 A_{st}$, σ_ζ^2 can be estimated by $\hat{\sigma}_\zeta^2 = \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{q_t} d_{st} \{\hat{\zeta}_{st}^2 - \hat{\sigma}_{12}^2 [\Phi^{-1}(\hat{P}_{st}) \hat{\varphi}_{st} + \hat{\varphi}_{st}^2]\}$, where $\hat{\zeta}_{st}$ is the OLS estimation residual.

Table C.1: Description of Data

	Label	Variable Definition	Mean	SD	Min	Max
y	Efficiency	Current period efficiency score that is given by (points+rebounds+assists+steals+blocks-misses-turnovers)/minutes.	0.37	1.07	-3.75	8.28
x_1	Experience	Minutes played from the start of the game till the end of period t-1.	9.91	7.81	0	37.58
	Fatigue	Minutes continuously played at the end of period t-1.	3.78	5.09	0	37.58
x_2	RPI	The previous year RPI of the opposing team.	0.55	0.08	0.38	0.66
	Home	A dummy variable taking value one if it is a home game.	0.67	0.47	0	1
	2nd Half	A dummy variable taking value one if it is the second half.	0.41	0.49	0	1
z	Lineup efficiency	The total efficiency score of the players in the lineup from the start of the game till the end of period t-1.	1.63	1.22	-2.65	5.04
	Lineup experience	The total minutes played by the players in the lineup from the start of the game till the end of period t-1.	49.57	34.03	0	152.78
	Lineup fatigue	The total minutes continuously played by the players in the lineup at the end of period t-1.	18.92	14.06	0	91.70
	Lineup fouls	The total number of fouls of the players in the lineup at the end of period t-1.	3.13	2.77	0	15
	One-substitution	A dummy variable taking value one if it takes one substitution from the lineup in period t-1 to reach this lineup.	0.68	0.46	0	1
	Two-substitution	A dummy variable taking value one if it takes two substitutions from the lineup in period t-1 to reach this lineup.	0.22	0.42	0	1

Number of observations: 2240; number of periods: 448

Table 1: ML Estimation of the Outcome Equation without Selectivity Bias Correction

Dep. Var.: Player Efficiency	Model 1	Model 2	Model 3
Peer effects	0.0841***		
	(0.0279)		
Same-type peer effects		0.0638***	0.0651***
		(0.0175)	(0.0175)
Cross-type peer effects		0.0345	
		(0.0216)	
Experience	0.0154***	0.0154***	0.0156***
	(0.0053)	(0.0053)	(0.0053)
Fatigue	-0.0083	-0.0084	-0.0083
	(0.0059)	(0.0058)	(0.0058)
RPI	-1.1677***	-1.1538***	-1.1914***
	(0.3043)	(0.3035)	(0.3028)
Home	-0.0030	-0.0034	-0.0031
	(0.0490)	(0.0489)	(0.0489)
2nd Half	-0.2159***	-0.2142***	-0.2197***
	(0.0683)	(0.0681)	(0.0680)
Player dummies	Yes	Yes	Yes
Log likelihood	-3294.83	-3291.17	-3292.47
Sample size	2240	2240	2240

Model 1: the outcome equation with homogenous peer effects

Model 2: the outcome equation with both same-type and cross-type peer effects

Model 3: the outcome equation with only cross-type peer effects

Standard errors in parentheses.

Statistical significance: ***p<0.01 ; **p<0.05 ; *p<0.1.

Table 2: Fixed Effect ML Estimation of the Outcome Equation

Dep. Var.: Player Efficiency	Model 2	Model 3
Same-type peer effects	0.1432	0.0534***
	(0.1000)	(0.0220)
Cross-type peer effects	0.1532	
	(0.1655)	
Experience	0.0337***	0.0316***
	(0.0098)	(0.0091)
Fatigue	-0.0198***	-0.0186***
	(0.0075)	(0.0070)
Player dummies	Yes	Yes
Log likelihood	-2590.55	-2591.03
Sample size	2240	2240

Model 2: the outcome equation with both same-type and cross-type peer effects

Model 3: the outcome equation with only cross-type peer effects

Standard errors in parentheses.

Statistical significance: ***p<0.01 ; **p<0.05 ; *p<0.1.

Table 3: Conditional Logit Estimation of the Selection Equation

Dep. Var.: Probability of Lineup Selection	
Lineup efficiency	0.1565*
	(0.0829)
Lineup experience	-0.0268***
	(0.0071)
Lineup fatigue	-0.0766***
	(0.0129)
Lineup fouls	-0.1199***
	(0.0512)
One-substitution	4.4993***
	(0.2712)
Two-substitution	2.2187***
	(0.2439)
Player dummies	Yes
Log likelihood	-755.94
Sample size	448

Standard errors in parentheses.

Statistical significance: ***p<0.01 ; **p<0.05 ; *p<0.1.

Table 4: OLS Estimation of Individual-Invariant Regressors in the Outcome Equation

Dep. Var.: Lineup-Averaged	Model 2		Model 3	
Estimation Residuals from Table 2	Parametric	Series	Parametric	Series
RPI	-1.0734***	-1.0157***	-1.3081***	-1.2369***
	(0.2339)	(0.2331)	(0.3100)	(0.3094)
Home	0.0080	0.0151	0.0067	0.0157
	(0.0388)	(0.0386)	(0.0515)	(0.0512)
2nd Half	-0.3526***	-0.3557***	-0.3570***	-0.3597***
	(0.0360)	(0.0362)	(0.0477)	(0.0480)
σ12	-0.0325		-0.0234	
	(0.0324)		(0.0429)	
Wald test for selectivity bias		11.1764**		9.1486*
		[0.0247]		[0.0575]
Sample size	448	448	448	448

Model 2: the outcome equation with both same-type and cross-type peer effects

Model 3: the outcome equation with only cross-type peer effects

Standard errors in parentheses; p values in brackets

Statistical significance: ***p<0.01 ; **p<0.05 ; *p<0.1.