An Estimable Model of Production Interactions in Endogenous Networks*

Giacomo De Giorgi

IEE/GSEM University of Geneva, BREAD, CEPR, IPA

Michele Pellizzari

IEE/GSEM University of Geneva, CEPR and IZA

Tomas Rodriguez

Universidad de los Andes

Abstract

We propose a model where forward-looking agents first decide to form links with each other and, then, engage in a production activity jointly with their linked peers. Exogenous linking opportunities facilitate the creation of network connections and the return to productive effort varies with the personal attributes of the connected agents. We apply our model to a purposely built dataset of college students containing information on the endogeneous networks of study partners linked with administrative records on the students' characteristics and academic performance. Identification relies upon the random assignment of students to classrooms, which generates exogenous opportunities for socialisation. Using the estimated structural parameters, we investigate the implications of two counterfactual experiments, one where students are streamed into classes by ability and one with single-sex classes.

Keywords: Networks, Production, Estimation, Counterfactual, Education
JEL Codes: D85, C15, C73, C63, E23, I23

*We would like to thank Bocconi administration for allowing us to develop this project. We are particularly thankful to Tito Boeri, Enrica Greggio, Marco Ottaviani, Bruno Pavesi, Sara Prati, Michelangela Verardi, Erika Zancan. Marco Jazzetta provided superb support with various technological aspects related to the project. We also benefited from excellent research assistance from Vikram Bahure and Isabel Melguizo López. We received valuable comments from seminar participants at the University of Geneva and the UCL-IFS workshop on "Causal Learning with Interactions". All errors remain our sole responsibility.
We acknowledge generous financial support from the Swiss National Science Foundation, project number 100018-165618.
Corresponding author: Michele Pellizzari, Institute of Economics and Econometrics, Geneva School of Economics and Management, University of Geneva, 40 Bd du Pont d'Arve, CH-1211, Geneva 4, Switzerland. Email: michele.pellizzari@unige.ch

1 Introduction

We present an estimable model of network formation and production interactions and provide a novel estimation method and counterfactual analysis for high dimensional and highly non-linear/non-differentiable problems. The simple idea of our model is that agents derive utility from two sources: (i) the enjoyment of being socially connected with other agents and (ii) the value of the output produced. Social interactions obviously affect the first of these sources of utility but they also indirectly affect the second, as the returns to production effort depend on the identities of one's peers.

As discussed in Jackson (2008), Jackson, Rogers and Zenou (2017) and Jackson (2019), it is essential to understand the role of economic networks and their structures for interpreting a large portion of human interactions and their outcomes. Hence, it is crucial to estimate the fundamental parameters of the problem so to be able to perform counterfactual analysis and policy simulations. While this logic appears rather linear, resolving the many problems and provide credible, yet workable, solutions to endogenous networks and mediated outcomes has proven a formidable task. It often is the case that models of peer effects treat the network formation stage as an identification nuisance (Manski (1993), Kline and Tamer (2020), and the extensive discussion in Graham and Áureo de Paula (2020)). Most studies in this area resort to credible sources of exogeneity for the identification of the social interaction parameters but pay little or no attention to the process of network formation (De Giorgi, Pellizzari and Redaelli, 2010; Sacerdote, 2001).

As shown in Carrell, Sacerdote and West (2013) sidestepping the network formation stage can lead to misleading policy conclusions. Kline and Tamer (2020) eloquently write:

> [...*Carrell et al. (2013) find evidence that [...] when they attempted to manipulate peer groups, individuals formed a network of connections within each manipulated peer group that lead to individuals avoiding interacting with individuals that the manipulated peer group had been designed for them to interact with...*]

These recent developments in network studies make it clear that, if one wants to provide in-

sightful policy prescriptions, one needs to jointly model network formation and production (or actions) and, then, estimate the fundamental parameters governing both processes. Achieving such a comprehensive understanding of social interactions is all the more important given the crucial role that networks play in so many domains: e.g. education (Calvó-Armengol, Patacchini and Zenou, 2009; Carrell, Sacerdote and West, 2013), insurance and credit (Angelucci and De Giorgi, 2009; Angelucci, De Giorgi and Rasul, 2018; Banerjee, Chandrasekhar, Duflo and Jackson, 2013; Cai, De Janvry and Sadoulet, 2015), criminal behavior (Bayer, Hjalmarsson and Pozen, 2009; Glaeser, Sacerdote and Scheinkman, 1996; Patacchini and Zenou, 2008), consumption (Ambrus, Mobius and Szeidl, 2014; De Giorgi, Frederiksen and Pistaferri, 2019), labor markets (Beaman, 2011).

One important premise behind our process of network formation is that agents have the opportunity to revise their linking decisions repeatedly. Following Jackson and Watts (2002), we model these revisions as a discrete time process. Each period a pair of randomly chosen agents is allowed to revise its linking decision, i.e. a linked pair can decide to split or two unlinked agents can decide to form a link. As in Bala and Goyal (2000) and Christakis, Fowler, Imbens and Kalyanaraman (2020), we assume a form of myopia: potential pairs of nodes make their linking decisions taking the rest of the network as given, as if their decisions would not influence those of others. This assumption is quite common in the literature, as it conveniently simplifies the solution of the models while remaining reasonably weak (Jackson, 2008; Jackson and Wolinsky, 1996).

While Mele (2017) describes the identification and estimation results for network architecture similar to ours, we expand his setting with a payoff structure that depends on endogenous and interlinked actions. In our specific application, such actions are the decisions to exert production effort whose returns depend on the attributes of one's linked agents or peers. Incorporating realistic peer effects in the definition of the payoffs from network formation is the ultimate goal of this paper, but it adds a substantive layer of complexity to the solution and estimation of the model (Badev, 2017; Battaglini, Patacchini and Rainone, 2019; De Paula, 2020).

We allow for both observable and unobservable heterogeneity, in particular agents are ex-ante

heterogenous on characteristics which make them more or less appealing to each other and more or less productive. These traits are observable to the econometrician but part of the pair-specific surplus remains observable only to the agents. Graham (2017) proposes a network formation model with a similar feature.

Given our modelling choices, we adopt a novel two-steps estimation procedure. First, following Jackson and Watts (2002), we argue that our network is pairwise stable and agents take it as given when making their effort decisions. These assumptions allow us to estimate the parameters of the production process separately from the others using GMM. In the second step, we proceed to estimate the parameters of the network formation process through a novel (quasi) maximum likelihood procedure. To keep the computation feasible, we use an iterative procedure where we first randomly split the full sample of pairs into an estimation and a validation set. Then, we draw a large number of samples of pairs of a given tractable size from the estimation set and for each such draws we maximize in the parameter space the likelihood of observing the network in our data. Since linking decisions are defined by a double inequality condition - both agents in the pair must be better off with the link than without it -, the resulting likelihood is highly non-linear and non-differentiable. Hence, we perform the maximisation using a derivative-free algorithm (simulated annealing). Eventually, we obtain many candidate parameter estimates, one for each drawn sample, and we select as our point estimate the one associated with the highest likelihood in the validation sample. We use the other estimates to construct empirical standard errors. Finally, we use these results to produce policy counterfactuals, following the approach of Jackson and Watts (2002).

We apply our model to data on college students for whom we observe both their network of study partners and their academic performance. We collected the data on study partners directly asking students to name the students with whom they prepared for exams, worked on problem sets or revised class material. In addition, we pair these networks with to the entire administrative records of students' socio-demographic and exams grades. In our setting, the students are randomly assigned to teaching classes and the randomisation is repeated every academic year generating exogenous variation in meeting opportunities that we can exploit for

identification purposes.

Our results suggest that students choose their study partners mostly with the objective of improving their academic performance rather than simply for the purpose of enjoying social interactions. This is consistent with the way we collected the data, given that we explicitly asked students to name their study partners. Nevertheless, students also show strong homophilic preferences, especially along the gender dimension. We also find evidence of strong complementarities in studying. Having more peers of high ability significantly improves own academic performance.

Using the estimated parameters we perform two policy experiments: one in which we allocate students to classes according to their level of ability and one in which we form single-sex classes. Both these policies are relevant in many real-world education settings but producing these counterfactual is notoriously difficult. As noted by De Paula (2020), one common feature of many network models is the multiplicity of equilibria, which imposes problems both for estimation and, particularly, for performing and presenting counterfactual analyses. The reason why we build our model within the evolutionary framework of Jackson and Watts (2002) is that it offers a compelling way for approaching these issues.

We find that tracking, i.e. forming classes by ability, increases average grades for higher ability students and lowers it for the less able ones (Duflo, Dupas and Kremer, 2011). When we organize classes by gender we show that the changes in the distribution of GPA are negligible, both in aggregate and by gender.

Our work is also related to a large literature which treats network formation as an iterative revision process along which agents make decisions myopically, including Mele (2017), Christakis et al. (2020) and Badev (2018). While our approach allows us to theoretically circumvent the multiplicity of equilibria, it suffers from the same pitfalls as other papers in the family which rely on simulations in order to approximate the long run distributions of Markov Chains induced by the network formation/revision process.Given our main objective of studying the academics outcomes mediated by the network and our two stage game of linking and effort decisions, this is the only method with which can approximate these long run distributions and,

thus, evaluate counterfactuals.

The reminder of the paper proceeds as follows: Section 2 discusses the theoretical set-up of our model; Section 3 details the estimation strategy; Section 4.1 describes the institutional context and the data; Section 4.3 provides the estimation results; Section 5 details our counterfactual policy exercises. Finally, Section 6 contains some concluding remarks.

2 Theoretical model

In this section, we formulate an estimable model of network formation and production. The crucial idea of our theory is that agents obtain utility from two sources: (i) the enjoyment of being socially connected with other agents and (ii) the output produced. Social interactions obviously affect the first of these sources of utility but they indirectly affect also the second, , as the returns to production effort depend on the identities of one's peers.

The basic premise behind the process of network formation is that the agents repeatedly obtain opportunities to revise their linking decisions. Following Jackson and Watts (2002), we model these revisions as a discrete time process. Each period a pair of agents is randomly chosen and is allowed to revise its linking decision, taking the rest of the network as given.[1] Upon being drawn, the agents in the pair assess the desirability of the link in question and decide whether to revise it, i.e. linking if they were not linked or separating if they were linked. Sequences of such purposeful link additions or deletions form "improving paths" in the space of all networks. This process is perturbed by "mutations" or "trembles", that happen with small probability ε and reverse the decision taken by the pair. Jackson and Watts (2002) show that this updating process defines an irreducible and aperiodic Markov chain over the space of all possible networks and therefore has a unique stationary distribution. As $\varepsilon \to 0$ this stationary distribution has support on the set of pairwise stable networks and on cycles formed by "improving paths" in the space of networks.

Our empirical approach is grounded on the assumption that the networks that we observe are

[1] One can think of the process as starting when the agents have their first socialization opportunities. However, theoretically, the starting point of the process does not matter.

draws from the long run distribution of such a process. Under the additional assumption that the true parameters of our models are such that we can rule out cycles formed by improving paths, we can estimate the parameters of the model with maximum-likelihood. We can then approximate the long run distribution of different outcomes of interest for some counterfactual scenarios by simulating the associated Markov chains.

The underlying network formation process is, therefore, just like in Mele (2017) but in his model preferences are such that the long run distribution corresponds to an exponential random graph model. Our most important innovation consists in a more realistic and useful definition of the payoffs to creating network connections, which in our model depend not only of fixed parameters but also on the outcome of the endogenous production process. Hence, the approach to estimation and counterfactual simulation of Mele (2017) cannot be used in our setting and we have, therefore, developed a new method.

In the remaining of this section, we define the elements of the model and the solution concept. We discuss the estimation approach in the next section (Section 3) and the simulations of the counterfactual scenarios in Section 5, after the presentation of the data we use for the empirical exercise (4.1) and the parameter estimates (4.3).

2.1 Utility

We consider agents that are ex-ante heterogeneous along two dichotomous dimensions: gender (F) and ability (Q). Although our data contain a continuous proxy of ability (the score in the entry test at university), we prefer to discretize it into above or below the sample median and we label these groups as high or low ability. This simplification allow us to preserve empirical tractability, while still capturing a crucial feature of the data.

Given this heterogeneity, four possible types of pairs exist between any two generic agents i

and j (l_{ij}):

$$l_{ij} = \begin{cases} 0 \quad if \quad F_i \neq F_j \ \& \ Q_i \neq Q_j \\ 1 \quad if \quad F_i = F_j \ \& \ Q_i = Q_j \\ 2 \quad if \quad F_i = F_j \ \& \ Q_i \neq Q_j \\ 3 \quad if \quad F_i \neq F_j \ \& \ Q_i = Q_j \end{cases} \tag{1}$$

The utility of a generic agent i is defined as the weighted sum of two components, one related to the enjoyment of social connections and one to the net value of production output:[2]

$$U_i(G|X,T,H,\Omega) = \phi[S(G_i|X)] + (1-\phi)[Y(e_i, G_i|X, \tau_{v_i}, \eta_i) - C(e_i|X_i)] - K(G_i|M_i, \Omega) \tag{2}$$

where G is the (non-normalised) adjacency matrix of the network (and G_i is its ith row vector) and X is the matrix of individual characteristics observable to both the agents and the econometrician ($X_i = \{F_i, Q_i\}$ is the ith row vecotr of X). $S(G_i|X)$ is the utility value of social connections, $Y(e_i, G_i|X, T, H)$ is output, which is a function of individual effort e_i and the agent's network connections G_i, which are both endogenously determined in the model. In addition, output also depends on exogenous factors, i.e. the characteristics of the individual and her connections (X) and a series of shocks (τ_i and η_i, which are elements of T and H, respectively), which are discussed later. $C(e_i|X_i)$ is the cost of exerting production effort and $K(G_i|M_i, \Omega)$ is the cost to agent i of creating the connections in network G, given exogenous linking opportunities M and idiosyncratic pair-specific factors (Ω), that are observable to the agents and not to the econometrician. We now discuss each of the functions in equation 2 more in details.

The utility value of social connections is:

$$S(G_i|X) = \left[\sum_{j \neq i} \mathbb{1}(l_{ij} = 1)\nu + \mathbb{1}(l_{ij} = 2)\nu_{\neg Q} + \mathbb{1}(l_{ij} = 3)\nu_{\neg F} + \mathbb{1}(l_{ij} = 0) \right]. \tag{3}$$

where the ν parameters measure the utility values derived from the different types of links:

[2] In our setting, the production process consists of studying for academic exams and the output is measured with exam grades but the model is general and can be applied to a variety of other contexts.

- ν is the utility of fully homophilic connections;
- $\nu_{\neg Q}$ is the utility of connections with same F and different Q;
- $\nu_{\neg F}$ is the utility of connections with same Q and different F;
- the utility of fully heterophilic connections is normalised to 0.

$C(e_i|X_i)$ is the cost of study effort e_i, which we define as follows:

$$C(e_i|X_i) = \frac{1}{2}\frac{e_i^2}{(1+\delta Q_i)} \tag{4}$$

with $\delta > 0$.

$K(G_i|M_i, \Omega)$ is the cost of creating all the links involving agent i but, given how we model the process of network formation, it is more useful to start from the definition of the cost of a single link. Without loss of generality, consider a network G where i and j are indeed linked and define the cost to i of creating such a link as:

$$K(G_i|M_i, \Omega) - K(G_{-ij}|M_i, \Omega) = \theta_0 + \theta_1 m_{ij} + \theta_2 d_i + \omega_{ij} \tag{5}$$

where G_{-ij} refers to network G absent link ij. m_{ij} is the $\{ij\}$th element of M and we interpret it as some exogenous factor facilitating or impeding the formation of a link between i and j.[3] d_i is the total number of i's connections (i''s degree). We expect $\theta_2 \leq 0$ if there are scale effects in link creation and creating $\theta_2 \geq 0$ if creating many connections might be tiring. The sign of θ_1 depends on the specific definition of M. If, as in our application, higher values of m_{ij} should lower the cost of link formation, then we expect θ_1 to be negative. The term ω_{ij}, which is the $\{ij\}$th element of Ω, captures pair-specific idiosyncratic factors that make the link more or less costly to establish. For example, i and j might particularly like or dislike each other,

[3] In our empirical implementation M is the result of the class assignment process. The generic m_{ij} is equal to the number of times agents i and j have been assigned to the same classroom. The higher m_{ij} the lower the cost of creating a link between the two agents.

thus affecting the cost of interacting with each other. We assume that ω_{ij} is observable to the agents at the time of making the linking decision but the econometrician does not observe it.
We think of equation 5 as describing the psychological cost of starting a conversation with a stranger or the time and effort required to organise meetings.

2.2 Production

Output is produced via a production function that depends on both individual effort and ability:

$$Y(e_i, G_i|X, \tau_i, \eta_i) = \mu_e(G_i|X)e_i + \tau_i + \eta_i \tag{6}$$

where $\mu_e(G_i|X)$ is the marginal product of effort and depends on the number of low and high ability peers:

$$\mu_e(G_i|X) = \beta_0 + \beta_1 \sum_{j \neq i} g_{ij}Q_j + \beta_2 \left[\sum_{j \neq i} g_{ij}Q_j\right]^2 + \beta_3 \sum_{j \neq i} g_{ij}(1-Q_j) + \beta_4 \left[\sum_{j \neq i} g_{ij}\,(1-Q_j)\right]^2 \tag{7}$$

We allow some non-linearity in the relationship between the number peers (separately by type) and the returns to individual effort. The intuition behind this formulation is that, for example, when the number of high ability peers is very high, the benefits of interacting with skilled individuals might be lower when there are many of them.
τ_i is a set of fixed effects for the different contexts in which interactions take place. In our setting, for example, these will be the classes to which individual i is assigned to throughout his bachelor program but in other contexts they could be the schools or the production sites where output is produced.
Finally, η_i is a random shock to the individual productive performance (luck, attention, etc.) that is realised only once all the linking and effort decisions are made. We impose no distributional assumptions on either τ_i nor η_i.

2.3 Timing

The timing of the model is the following:

At time zero the exogenous meeting opportunities M are generated. Then, the pair level idiosyncratic shocks Ω are realized: some pairs might particularly like or dislike each others. In a sense, M and Ω are very similar objects: they both affect the cost of forming links between any pair of agents. The only substantial difference is that M is observable to the econometrician whereas Ω is not.

Next, agents make linking decisions. We model network formation as a dynamic process in which all pairs of agents sequentially get the possibility to form or remove the links among them. This phase can be interpreted as an intensive socialization period in which pairs of agents have abundant occasions to to revise their linking decisions. A crucial innovation of our approach is that such decisions are forward looking: when deciding whether to form or break a link agents take into account the implications of such decision on the output that they will be able to produce, above and beyond the direct enjoyment of the social interaction.

Finally, once the network is formed, agents exert production effort and output is produced, conditional on the random shocks H and T, giving rise to a non-degenerate distribution of Y for ex-ante identical agents.

2.4 Solution

The model is solved backwards. First, we solve for the optimal effort and output for a given network structure. Then, we model optimal network formation following Jackson and Watts (2002).

2.4.1 Optimal effort and production

For a given network architecture, agents choose effort optimally by maximising expected output:

$$max_{e_i} \; E_H \left[Y(e_i, G_i | X, \tau_i, \eta_i) - C(e_i | X_i) \right] \tag{8}$$

Expectations are taken over the distributions of H, the only unobservable unknown to the agents at the time of the effort choice.

Optimal effort is:

$$e_i^* = (1 + \delta Q_i) \mu_e(G_i | X) \tag{9}$$

Maximised output Y is:

$$Y_i^* = \tau_0 + (1 + \delta Q_i) \left(\mu_e(G_i | X) \right)^2 + \tau_i + \eta_i \tag{10}$$

2.4.2 Optimal linking

In the framework put forth in Jackson and Watts (2002), beginning from any network, pairs of agents are randomly chosen and allowed to decide whether to form a link or not (in case they do not have a link among them) or whether to preserve a link or remove it (in case they do have a link among them). If both agents benefit from forming or preserving a link, then they do so, and otherwise they do not form, or remove the link. Agents are assumed to act myopically when they make these decisions, namely they do not internalize the fact that they are part of an endless link revision process and that their decision might influence those of other pairs. Link revisions are decided taking the rest of the network as fixed.

Without loss of generality, consider the case of a pair of agents i and j that are not connected in the given network G. When called on to revise their status, they will decide to form the link if they both expect it to improve their levels of utility:

$$E_H \left[U_i(G_{(+ij)} | X, T, H, \Omega) - U_i(G | X, T, H, \Omega) \right] \geq 0 \tag{11}$$

$$E_H \left[U_j(G_{(+ij)} | X, T, H, \Omega) - U_j(G | X, T, H, \Omega) \right] \geq 0 \tag{12}$$

where $G_{(+ij)}$ is network G augmented with link ij. The expectation is taken over the distribution of H, the only observable that is unknown to the agents at the time of the linking decision. The link will actually be created only if the above conditions are met and there is no mutation, which happens with probability $1-\varepsilon$. With probability ε, a mutation or tremble occurs and the link is not created, despite conditions 11 and 12 being jointly satisfied.

The case of pairs of agents who are linked and get the possibility to revise their status is solved symmetrically in the same manner.

3 Estimation

The estimation procedure that we develop is grounded on the assumption that the final network which we observe is drawn from the distribution resulting from the network formation and updating process. Jackson and Watts (2002) show that such distribution is stationary and, , as $\varepsilon \to 0$, its support falls within the set of pairwise stable networks (excluding cycles by assumption). Given these assumptions, the observed network is very likely to be pairwise stable.[4]

The estimation procedure has two steps. First, we estimate the parameters of the production function with GMM. Second, we estimate the remaining parameters, mostly related to the network formation process, with a novel maximum likelihood procedure.

3.1 Estimation of the production process

Given that agents make effort decisions taking the network as given, the production function can be estimated separately from the parameters governing the network formation process. Excluding the contextual effects T, whose numerosity depends on the specific context (in our application T is a vector of 25 class fixed effects), the production function 10 has five parameters: $\{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \delta\}$.

[4] In particular, we assume that the parameters of the model are such that there are no "cycles" in the graph of networks defined by improving myopic updates. Hence, in the limit as $\varepsilon \to 0$ the Markov chain must eventually reach a pairwise stable network. The chain never reaches an absorbing state but spends disproportionately large amounts of time in pairwise stable networks.

These parameters can be easily estimated using the following orthogonality conditions:

$$\begin{aligned}
0 &= E\big[Y_i - (1+\delta Q_i)\mu_e(G)^2 - \tau_i\big] \\
0 &= E\big[Q_i\big(Y_i - (1+\delta Q_i)\mu_e(G)^2 - \tau_i\big)\big] \\
0 &= E\big[\big(\textstyle\sum_{j\neq i} f(m_{ij}|X)Q_j\big)\big(Y_i - (1+\delta Q_i)\mu_e(G)^2 - \tau_i\big)\big] \\
0 &= E\big[\big(\textstyle\sum_{j\neq i} f(m_{ij}|X)Q_j\big)^2\big(Y_i - (1+\delta Q_i)\mu_e(G)^2 - \tau_i\big)\big] \\
0 &= E\big[\big(\textstyle\sum_{j\neq i} f(m_{ij}|X)(1-Q_j)\big)\big(Y_i - (1+\delta Q_i)\mu_e(G)^2 - \tau_i\big)\big] \\
0 &= E\big[\big(\textstyle\sum_{j\neq i} f(m_{ij}|X)(1-Q_j)\big)^2\big(Y_i - (1+\delta Q_i)\mu_e(G)^2 - \tau_i\big)\big]
\end{aligned} \tag{13}$$

where $f(m_{ij})$ is the predicted probability that i and j are connected conditional on m_{ij} (and controls), which we estimate based on the data frequencies. In essence, this GMM instruments the shares of high and low ability peers (linear and squared) with their predicted counterparts. The predictions are based on the exogenous meeting opportunities (and controls). Adding the assumption that η_i has mean zero and that it is orthogonal to individual ability Q_i (the first two conditions in the above list), we have a total of 6 conditions for 5 parameters. Adding the τ's would simply require adding one condition for each of them, stating their orthogonality with η_i: $0 = E\left[\tau_i\left(Y_i - (1+\delta Q_i)\mu_e(G)^2 - \tau_i\right)\right]$.

Notice that, given our assumptions, the use of instruments is not strictly necessary for the identification of the production function. The only unobservable in the production process is η_i, which is unknown to the agents at the time of making the linking decisions. Hence, the shares of high and low ability peers are exogenous in equation 10. Nevertheless, using predicted peers as instruments makes the estimation strategy robust to the presence of other unobservables that might potentially generate endogeneity issues.

3.2 Estimation of the network formation process

Once we have estimated the parameters of the production function, we proceed with the others. First of all, we impose a distributional assumption on Ω, namely, we assume that the ω_{ij} are all independently and identically distributed according to a standard normal $N(0,1)$. Next, we also

fix the value of θ_0 to 1. Given the absence of an obvious metrics for the cost of forming social connections, this assumption fixes the expected cost of establishing one's first link ($d_i = 0$) in the absence of any exogenous facilitation or impediment ($m_{ij} = 0$) to unity. We will, then, be able to interpret the other parameters of the model in the light of this normalisation. Eventually, we are left with the following 6 additional parameters to estimate: $\Lambda = \{\phi, \nu, \nu_{\neg Q}, \nu_{\neg F}\}$, which we label Λ.[5]

With these additional assumptions and normalizations, we can write the likelihood of observing the status of any given pair of agents. For example, the probability that agent i and j are linked to each other is:

$$\begin{aligned} P_1(\Lambda|X, M, T) &= P\big(E_H\big[U_i(G_{(+ij)}|X, M, T, H, \Omega) - U_i(G|X, M, T, H, \Omega)\big] \geq 0, \\ &\quad E_H\big[U_j(G_{(+ij)}|X, M, T, H, \Omega) - U_j(G|X, M, T, H, \Omega)\big] \geq 0\big) \end{aligned} \quad (14)$$

The two events in this probability depend on the parameters Λ and on the specific draws of ω_{ij}, which we assumed to be distributed according to a standard normal. The distribution of the production shocks H are irrelevant in equation 14 because they are unknown at the time of making the linking decisions. Hence, $P_1(\Lambda|X, M, T)$ is the probability that ω_{ij} takes a value such that both events in the equation are verified, conditional on a given set of parameters and on observable variables.

The probability of observing non-linked pairs is simply the complement to one of $P_1(\Lambda|X, M, T)$:

$$\begin{aligned} P_0(\Lambda|X, M, T) &= 1 - P\big(E_H\big[U_i(G_{(+ij)}|X, M, T, H, \Omega) - U_i(G|X, M, T, H, \Omega)\big] \geq 0 \\ &\quad E_H\big[U_j(G_{(+ij)}|X, M, T, H, \Omega)\big] \geq 0\big) - U_j(G|X, M, T, H, \Omega)\big] \geq 0\big) \end{aligned} \quad (15)$$

Using equations 14 and 15 we can write the full likelihood of any observed network as a function of the parameters Λ. However, if the purpose of such computation is the maximization of the resulting likelihood in search for parameter estimates, there are two serious computational problems. First, in many applications the sample size in the space of agent pairs risks being

[5]Recall that we have already also normalised the utility value of fully heterophilic connections to 0.

enormous. For example, in our application we have a rather small population of 577 agents but the total number of pairs raises to 166'176. Second, both equations 14 and 15 are based on double inequality conditions, thus they are highly non-linear end non-differentiable in the parameter vector. Hence, we cannot use standard optimization tools.

To overcome these problems and produce estimates of the parameters Λ (and their standard errors), we adopt the following procedure.

- First, we randomly split the total sample of pairs in two parts, one that we will use for estimation and one for validation of the model estimates. In our application, we reserve a random 75% of the pairs for estimation and the remaining 25% for validation. We stratify the samples on meeting opportunities to avoid having too many pairs that never met.

- Then, from the subsample of pairs reserved for estimation we randomly draw many samples of some manageable size. The pairs are drawn evenly from the populations of linked and non-linked pairs. In our application we draw 500 samples of 732 pairs.[6] Also at this stage, we stratify the random samples on meeting opportunities.

- Next, for each random sample we maximise the likelihood using simulated annealing. We thus produce many maximum likelihood parameter estimates.[7]

- Finally, for each vector of estimated parameters, we compute the likelihood of the model using the 25% of the original sample that was not used for estimation and set aside for validation. We can then rank the parameter vectors according to this out-of-sample likelihood and we select as our point estimate the vector associated with the highest likelihood. The other vectors are used to compute empirical standard errors.

[6] This sample size results from sampling 80% of the linked pairs and an equal number of non-linked pairs.

[7] We impose some lower and upper bounds to the parameters in order to limit the parameter space and make the simulated annealing maximization procedure computationally manageable. We also disregard solutions with one or more parameter values at boundaries of the parameter space. With upper and lower bounds on each of the 6 parameters, this is a rather strict requirement and ,in our application, we drop 268 results out of the 500 samples.

4 Empirical application

We apply and estimate the model described in the previous sections to network data about students at Bocconi university in Milan, Italy. The data allow us to reconstruct the full network of study partners, i.e. students who meet in groups to work on problem sets, exam preparation, etc. We also have access to detailed background information from the administrative archives of the university, allowing us to characterise the nature of each pair of students in terms of homophily.

4.1 Data and Setting

Our data cover the universe of students in the 2011 entry cohort of the BA in management, the largest program offered by Bocconi at that time. The duration of the program is three years, so the students started their first year in September 2011 and were supposed to obtain their degrees in the summer 2014.

In order to elicit the network of study partners, we purposely created an online survey, which we administered to all students in this cohort via the university web portal at the end of each of the three academic years of their program.

The survey was launched every year in June/July, once the summer exam session was completed and students were logging into the university web portal to see their grades. Upon connecting, a pop-up window invited them to fill the survey. The pop-up window would continue to appear at every log in until the survey was completed or when we took it off line (about one month after it was launched). The survey, available in Appendix A, essentially only asked two questions: (i) the share of lectures attended (in person) during the academic year and (ii) the names of study partners, both males and females up to 20 (10 for each gender). As students typed the names of their study mates into the survey tool, the system searched for them in the administrative archive of the university and attached to each nominated person a student number.

With the student numbers of all respondents and nominated students we recovered all the information regarding these individuals from the administrative archives of the university, including

their grades in all the courses and their background characteristics, such as gender, nationality, place of origin, high school grades. Importantly, we also know their scores in the entry test they were all required to take as part of the admission process.[8]

The academic environment of Bocconi university is that of a standard selective university in Europe or North America and data on its students have already been used for a number of studies (De Giorgi and Pellizzari, 2014; De Giorgi, Pellizzari and Redaelli, 2010; De Giorgi, Pellizzari and Woolston, 2012; Garibaldi, Giavazzi, Ichino and Rettore, 2012). One feature of the teaching logistics at Bocconi University is particularly important for our study, namely students are randomly assigned to different classes and, contrary to what happens in many other institutions, the randomization is repeated over time. More precisely, when students enter their first year, they are randomly divided into groups and they attend the compulsory courses of their program within such groups. The random allocation is then repeated at the beginning of each subsequent academic year, so that students attend the compulsory classes of each academic year with a different group of random classmates. Being randomly assigned to the same teaching group and physical classroom obviously affects socialization opportunities and the probability of studying together increases substantially with the number of times any two students have been assigned to the same class in the past. Using the terminology of our model in Section 2, the random allocation to classes generates exogenous variation in the cost of forming links and we use it as a measure of m_{ij}. More specifically, we define m_{ij} in our application as the number of times students i and j have been assigned to the same class.

The exogenous variation in socialization opportunities induced by the random allocation process is crucial for the identification of our model and it is one of the unique features of our data. To the best of our knowledge, there exist no other dataset in which it is possible to observe a actual (endogenous) network of agents in a setting where the costs of forming links is subject to some substantive exogenous variation.

For consistency with the static nature of our theoretical model, we aggregate outcomes and

[8]We replicated the survey for all undergraduate students at Bocconi university at the end of the academic years 2011/2012-2012/2013-2013/2014 and 2014/2015. Hence, we have more data than those used in this paper, covering multiple academic programs and cohorts. We decide to restrict our analysis to only one cohort and one academic program for the sake of both simplicity and computational convenience.

network connections for each student over the three years of their program. Hence, we define the outcome Y_i as GPA in all compulsory courses taken in any academic year and we define the network as the union of (undirected) named peers in any survey.[9] In other words, we consider two students as linked to each other if either one or the other named the other one in any survey, at the end of year one, two or three.

As the random reshuffling of class composition happens at the beginning of the each year any pair i, j can be randomly assigned to the same classroom up to 3 times, i.e. $m_{ij} \in [0, 1, 2, 3]$.

4.2 Descriptive statistics

In the administrative data of the university there are 996 students who enrolled in the first year of the BA in management in 2011/2012. 396 of them took all the three online surveys administered at the end of each year in their program (June/July 2012, 2013 and 2014) and 692 of them were nominated as study peers by at least one survey respondent. Given some overlap between these two groups, 790 students either responded to the survey or were nominated by some respondent. Eliminating dropouts and students with missing information on some of our key variables of interest, our working sample eventually comprises 577 students. Compared tot he full population, these students are more likely to be females (50% against about 40% in the original population) and slightly less skilled (their average score in the entry test is about 25% of a standard deviation below the average of the non-selected students and similarly for GPA). Table 1 shows some basic descriptive statistics about our working sample.

We measure the output of the study process with the grade point average over compulsory courses. We exclude languages and computer use, which, despite being compulsory, can be waived for foreign students and students holding specific certificates (e.g. TOEFL, ICDL, ECDL). The students in our sample take 24 such courses. Exam grades range from 0 to 30, with 18 being the minimum pass grade.[10] The average GPA in our sample is 25.4, with a

[9] We do not consider elective course to avoid problems of endogenous selection in courses of different difficulty. Moreover, elective courses are rarely taught in multiple classes. Eventually, we compute GPA over a set of 24 compulsory courses.

[10] This metrics comes from the historical tradition of having exams marked by commissions of 3 professors,

Table 1: Descriptive statistics about students

	Mean	Std	Min	Max
GPA[a]	25.413	2.051	20.442	29.669
Entry test	48.143	9.635	29.3	77
Q=1[b]	0.501	0.5	0.000	1.000
F=1[c]	0.497	0.500	0.000	1.000
Number of peers	2.118	2.396	0.000	16.000
Number of F=1 peers	1.061	1.485	0.000	10.000
Number of Q=1 peers	1.01	1.359	0.000	10.000
Obs.		577		

[a] All compulsory exams (excluding languages, computer use and seminars).

[b] 1=top 50% of the distribution of the entry test.

[c] 1=female.

minimum of 20 and a maximum of 29.6.

Admission to the university is selective. Students are ranked based on a linear combination of an entry test and high school results and the fixed set of available places in each program are assigned starting from the highest ranked student.[11] The entry test comprises 70 questions in four domains: reading comprehension, logical reasoning, mathematics and numerical reasoning. Each correct answer gives one point, wrong answers give -0.2 and there are bonuses for answering correctly most questions in a domain. We use the score in the entry test as our main measure of ability.[12] Students in our data score on average 48.1, with a minimum of 29.3 and a maximum of 77. The distribution is rather symmetric, with a median of 48.3, very similar to the mean. For consistency with the theoretical framework, we discretize ability into just two categories: high ability is defined as scoring above the median ($Q = 1$) and low ability as scoring below the median ($Q = 0$).

The sample displays a rather balanced gender distribution with 49.7% female students. On average students are linked to 2.1 peers in the network of study partners. About 9% of survey respondents do not nominate any peer and the maximum number of connections is 16, suggesting that the limit of 20 maximum nominations is not binding in our setting. Consistently

each giving a grade on the 0-10 scale and 6 begin the minimum passing grade. Today most exams are marked by one single professor by the metrics has remained.

[11] The entry test is weighted 55%, high school results (a combination of grades in core courses of the last 2 years) 43% and language certificates, if present, 2%.

[12] Our main findings are confirmed when replicating results using high school grades.

with the overall descriptive statistics, students on average have 1 female mate and 1 high ability mate.

The 577 students in our sample form 166,176 pairs. Based on the nominations in the surveys, we can classify such pairs as either being connected, if either one or both students nominated the other, or not connected. Table 2 reports some descriptive statistics about the student pairs, separately by their connection status.

Table 2: Descriptive statistics about student pairs

	Non-linked pairs	Linked pairs	All pairs
Both Female	0.247	0.3	0.247
	(0.431)	(0.458)	(0.431)
Both Male	0.252	0.298	0.252
	(0.434)	(0.458)	(0.434)
Both Q=1[a]	0.25	0.247	0.25
	(0.433)	(0.432)	(0.433)
Both Q=0[a]	0.249	0.293	0.249
	(0.480)	(0.479)	(0.480)
Same class $(m_{ij})^b$	0.329	1.074	0.332
	(0.541)	(0.722)	(0.544)
Link probability[c]	0.003	0.106	0.004
	(0.015)	(0.237)	(0.022)
Number of pairs	166,176	611	165,565

[a] 1=top 20% of the distribution of the entry test.

[b] Number of times assigned to the same class.

[c] Predicted probability of being linked based on m_{ij} (and standard set of covariates).

Some degree of homophily seems to be present in our data along the gender dimension. About 30% of the linked pairs are of the same gender, compared to approximately 25% of the non linked pairs. Regarding ability, there does not seem to be much assortative mating for the high ability students and some for the low ability ones.

Begin assigned to the same teaching class is strongly associated with being connected, supporting our intuition that the random allocation into classes generates variation in the cost of forming links. Linked students have been on average once in the same class compared to about 0.3 for the non linked pairs.

Another way to see the role of the class assignment process is presented in Table 3. The share

Table 3: Connections and class assignment

Number of times in the same class (m_{ij})	Share of linked pairs
0	0.0011
1	0.0073
2	0.0251
3	0.0677

of pairs that are connected increases substantially with the number of times the students in the pair have been assigned to the same class: from 0.1% of all pairs who have never been in the same class to 6.7% of pairs who have been in the same class in all three years of their program. Figure 1 shows the degree distribution of the network of study partners (i.e. the number of connections) together with some descriptive statistics about the network. About 20% of the students have no connection at all, 32.7% have only one study partner and 21.1% have two. Only 10% of the students have more than 5 connections.

Variable	Avg.	S.d.
Degree	2.11	2.4
Clustering (local)	0.22	0.34
Clustering (global)	0.14	-
Support	0.33	0.47
Path length	7.2	2.5

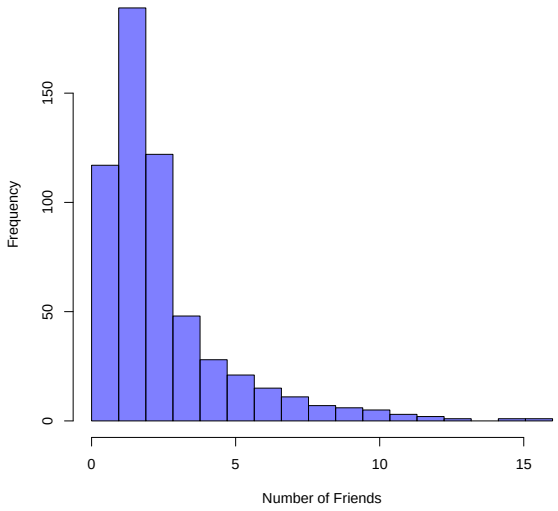
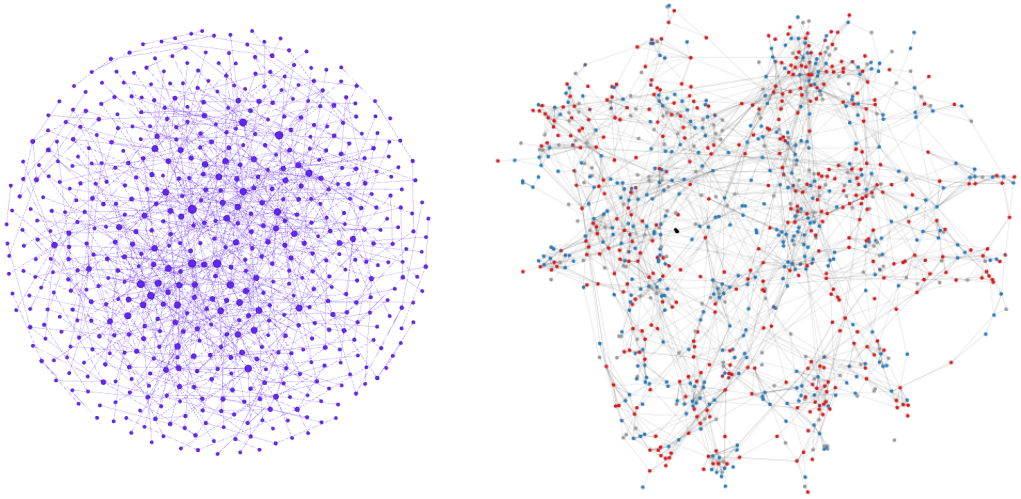


Figure 1: Network descriptive statistics

Figure 2a shows the entire network (2a) and gives some indication of clustering around GPA. When coloring the nodes (2b, blue for below median GPA and red for above median, it is evident that GPA tends to be clustered amongst connected students, which we take as suggestive

of the relevance of network effects.



(a) Network of study partners

(b) Clustering of GPA between Friends

Figure 2: Network and clustering

4.3 Results

In this section, we present the estimates of the parameters of our model applied to the student data described above. In the next section we will then use these estimates and the model to produce some counterfactual experiments that exemplify the usefulness of our framework.
We start with the parameters of the production function that are estimated via GMM, as discussed in Section 3.1. The estimates of these parameters are reported in Table 4. Our preferred estimates are those reported in the second column, where we use predicted connections based on the random class assignment to instrument the characteristics of actual connections. In other words, the estimates in the second column of Table 4 are produced using the exact moments described in equation 13. For comparison, in the first column we show the same estimates produced without using instruments, namely using the share of actual (instead of predicted) peers of high and low quality as instruments for themselves. Results are broadly comparable.

These estimates imply that changes to the set of study partners have very meaningful implica-

Table 4: Production function estimates

	Instruments	
	No[a]	Yes[b]
β_0	4.986	4.989
	(0.017)	(0.054)
β_1	0.009	0.049
	(0.012)	(0.041)
β_2	0.007	-0.003
	(0.012)	(0.063)
β_3	0.00005	-0.008
	(0.002)	(0.007)
β_4	-0.002	-0.001
	(0.002)	(0.012)
δ	0.034	0.036
	(0.007)	(0.007)
Observations	577	577

The dependent variable is the students' GPA in all compulsory courses.
[a] GMM estimates based on moments similar to those listed in equations 13 but where the predicted shares of peers of high and low quality are replaced by the actual shares.
[b] GMM estimates based on the moments listed in equations 13.

tions for student performance. For example, consider a low ability student with two connections, a high ability and a low ability one, adding one more study partner of high ability would increase the GPA of the student by 12% of a standard deviation, whereas adding one more low ability peer would decrease GPA by 2.3% of a standard deviation. Performing the same exercise on a high ability student produces effects of very similar magnitudes.

Table 5 reports the estimates of the other parameters of the model, estimated following the procedure described in Section 3.2.

Most of these parameters are quite unique and, to the best of our knowledge, there exists no other study that has tried to estimate them. For example, ϕ measures the weight students put on the pure utility benefits of socialisation when they form links. Our best estimate of this parameter is 0.16, suggesting that the students in our sample forms connections mostly for the benefits of the production interactions that they generate. Unfortunately, this parameter is estimated very imprecisely, perhaps suggesting that there is large heterogeneity across agents.

Table 5: Network formation estimates

Parameter	Estimate (std.err.)[a]
ϕ	0.160
	(0.236)
ν	0.932
	(0.207)
$\nu_{\neg Q}$	0.917
	(0.215)
$\nu_{\neg F}$	0.547
	(0.188)
θ_1	-0.076
	(0.024)
θ_2	-0.324
	(0.032)
Estimation[b]	
Log-likelihood	-246.847
Observations	732
Validation[c]	
Log-likelihood	-104.693
Observations	41,546

[a] The estimation procedure is described in details in Section 3.2.

[b] Log-likelihood computed using the estimated coefficients on the sample of pairs used for the estimation. The number of observations refers to the number of pairs in the sample used for estimation.

[c] Log-likelihood computed using the estimated coefficients on the sample of pairs that was not used used for the estimation (of any set of coefficients, both those reported in the table and those used to produce the standard errors). The number of observations refers to the number of pairs in the sample used for validation (i.e. pairs that were never used for estimation).

The νs are the socialisation benefits of different link types and should be interpreted keeping in mind that we have normalised the benefit of fully heterophilic connections to zero. Hence, our data seem to be consistent with a good degree of preference for homophily, with fully homophilic links yielding a benefit of over 0.93. Interestingly, homophily along the gender dimension appears to be substantially more important than along the ability dimension. The benefit generated by linking with someone of one's same gender but different ability is 0.92, very similar to fully homophilic links, whereas connections of same ability but different gender seem to be much less valuable with an estimated contribution to utility of 0.55. These parameters are estimated quite precisely.

As expected, θ_1 is negative, suggesting that the cost of forming links decreases with the number of times students are assigned to the same class. Recalling that we normalised to zero the cost of forming one's first link in the absence of any meeting, this is a non negligible effect. Interestingly, the cost of making connections decreases with degree, suggesting some sort of economies of scale in network formation, which are perhaps justified in the context of study partners. Once students organise a session to meet and work together, then having one more peer joining is minimal and the fixed cost of the logistics has been paid already. There might be non linearities in the relationship between the cost of linking and network degree but, with the relative small number of connections that we see in our data, they might be difficult to identify.

5 Counterfactual policies

In this section we use our model and its estimated parameters to generate two counterfactual scenarios that would arise as the result of different assignment policies.

The first policy we consider is one that many school systems or individual schools adopt to some degree, namely assigning students to teaching classes based on their ability. Given the very discrete measure of ability that we have adopted for our empirical exercise, we simulate a very stark version of this policy, one where all the high ability students are assigned to the same class and all low ability students are assigned to the same class, but it would be relatively

easy to modify the definition of the ability groups in the model and replicate the exercise for various versions of this policy.

The second policy that we consider is also one that is and has been adopted by many educational institutions, namely separating students by gender, with classes made of either only male or female students.

Technically, we estimate by simulation the long run distribution of some statistics of interest under the observed policy, and the long run distribution of the statistics under the counterfactual policy. Namely, for both the observed random class assignment policy and for any given counterfactual policy of interest, we take a small value of ε and for each of a large number of draws of the unobservables in the model, we sample from the unique stationary distribution of the resulting Markov chain. Concretely:

1. We draw a vector of the unobservables of the model (Ω and H), which are consistent with the observed network being pairwise stable. We refer to such a draw of unobservables as a "frame".

2. Fixing a value of ε we simulate the Markov chain for a large number of periods and take the realized empirical distribution of the statistic of interest as an approximation of the stationary one. In particular we focus on the individual GPAs, on the individual degree and on the assortativity of the network as a whole, according to gender and to ability.

We repeat this process for a large number of frames, and estimate the expected counterfactual distribution of the statistics of interest as the average long run distribution over all the frames.

Assortativity is a measure of homophily, and captures the extent to which agents tend to link with other agents of the same kind along a given a dimension relative to what we would expect if they formed their relations disregarding the dimension in question (Newman, 2003). In the case of a binary characteristic $c \in \{0, 1\}$ the formula is:

$$q_c = \frac{(c_{00} + c_{11}) - a_1 a_0}{1 - a_1 a_0}$$

In this expression c_{00} is the fraction of the total number of links in the graph occurring among

agents for whom $c = 0$, c_{11} is the fraction of the total number of links for whom $c = 1$, a_0 is the fraction of edges involving at least one agent for whom $c = 0$ and a_1 is the fraction of edges involving at least one agent for whom $c = 1$. The numerator is thus the difference between the fraction of links that occur among agents of the same kind and the fraction that we would expect if links were built disregarding characteristic c. This quantity is normalized by the value which this difference would take if there were no links between the two kinds of agents in the network. It follows that the coefficient is positive only if there is a greater fraction of links among agents of the same kind than what we would expect if links occurred at random, and can take a maximum value of 1.

5.1 Forming classes by ability

Allocating the students into classes according to their ability affects the network structure because the estimated parameters of the model imply that greater exposure of any two given students reduces their linking cost, thereby leading to a substantial change in the network structure and thus, potentially, to a change in the distribution of expected student performances.

Figure 3 shows the distribution of network level assortativity according to ability and gender under the observed allocation of students to classes and under the counterfactual allocation. The distributions for the assortativities under the actual observed allocations serve as benchmarks and are obtained by simulating the Markov chain using the same number of frames and number of simulations as we use to produce the counterfactuals.

Figure 4 shows the corresponding degree distributions. As expected, given that greater exposure reduces the cost of forming links, in this counterfactual exercise the distribution of assortativity by ability shifts to the right. On the other hand the policy does not have an impact on assortativity by gender with a straightforward qualitative interpretation.

As can be seen in Figure 4, the counterfactual assignment does not lead to a qualitative change in the degree distribution. That is, in expectation randomly chosen agents will tend to have the same number of study peers as under the observed assignment policy.

Figures 5 and 6 show the actual GPA distribution of our students in comparison to the coun-

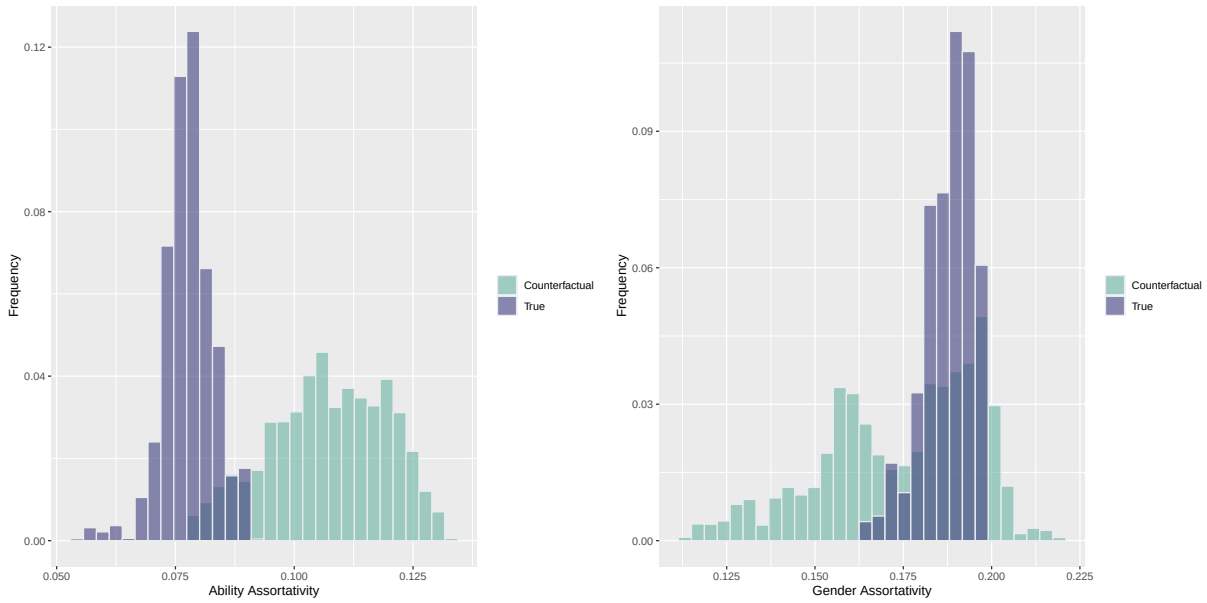


Figure 3: Sorting by ability: true and counterfactual distributions of assortativity by ability (Left) and gender (Right)

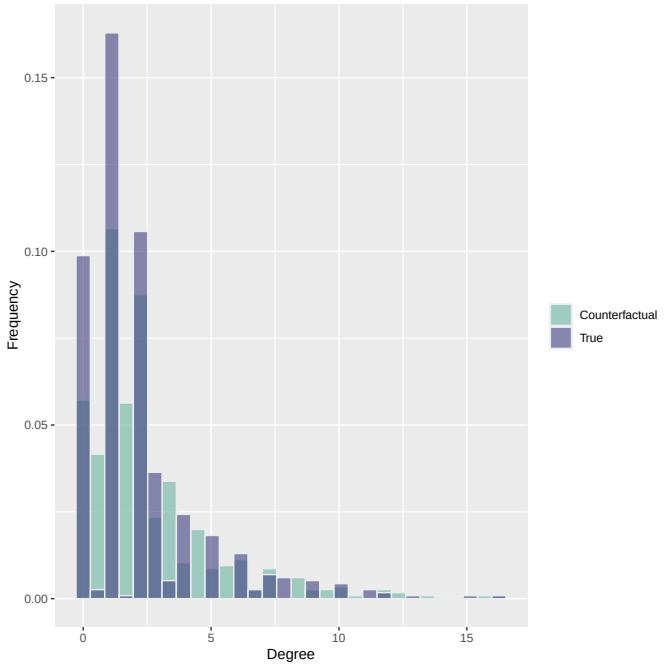


Figure 4: Sorting by ability: true and counterfactual expected degree distributions

terfactual distribution. The means of both distributions are very similar. The mean of the true distribution is 25.45 and the mean of the counterfactual distribution is 25.46. However, Figure 6, which reports the distributions separately for the two ability groups, suggests that students with scores above the median benefit slightly by the assortative allocation, whereas students with scores below the median perform slightly worse. In particular, in the counterfactual, the mean outcome of the high ability students is 26.05, approximately 1% of a standard deviation above the true mean of 25.89. On the other hand, the mean outcome of the low ability students is 24.86, approximately 1% standard deviation below the observed mean of 25.02.

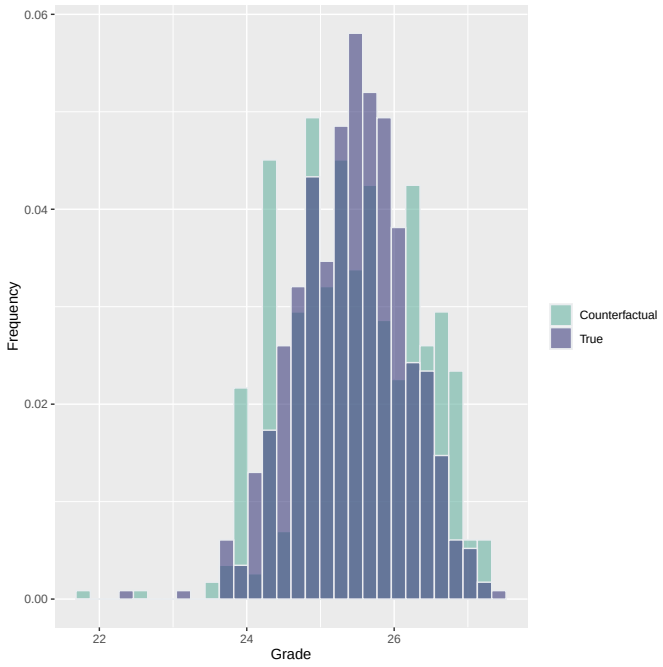


Figure 5: Sorting by ability: true and counterfactual distributions of GPA

5.2 Forming classes by gender

As in the case of our first counterfactual exercise, allocating the students into classes according to their gender alters the distribution of networks that might arise because the cost of building a link among two students decreases with their exposure to each other, and exposure is largely determined by class assignments.

As shown in Figure 7 the distribution of the assortativity coefficient by gender under the coun-

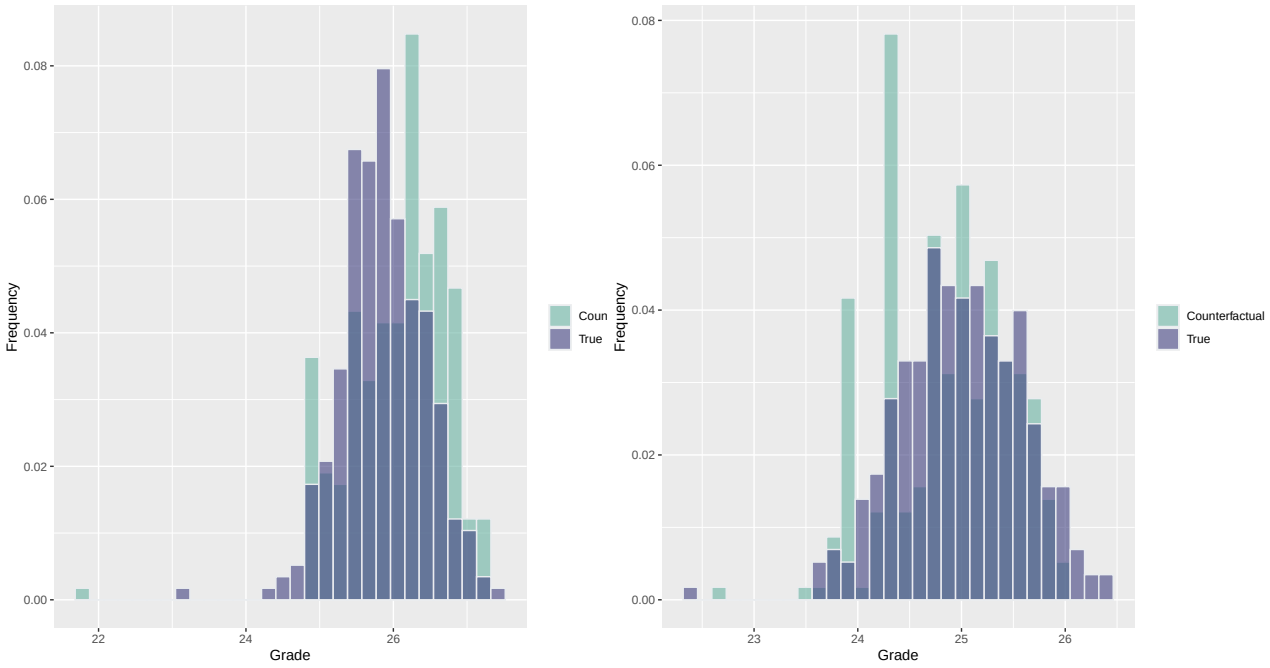


Figure 6: Sorting by ability: true and counterfactual distributions of GPA for high (Left) and low (Right) ability students

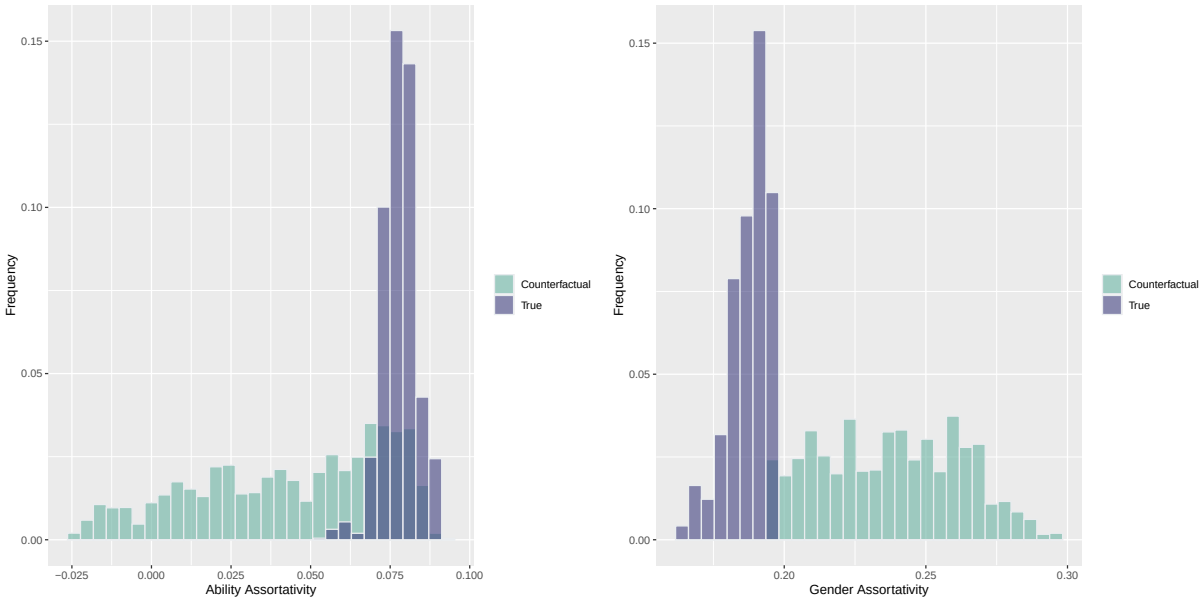


Figure 7: Sorting by gender: true and counterfactual distributions of assortativity by ability (Left) and gender (Right)

terfactual policy shifts to the right relative to the distribution that we obtain under the observed assignment policy. At the same time, as can be seen in Figure 8 the policy has no effect over the expected degree distribution of the network.

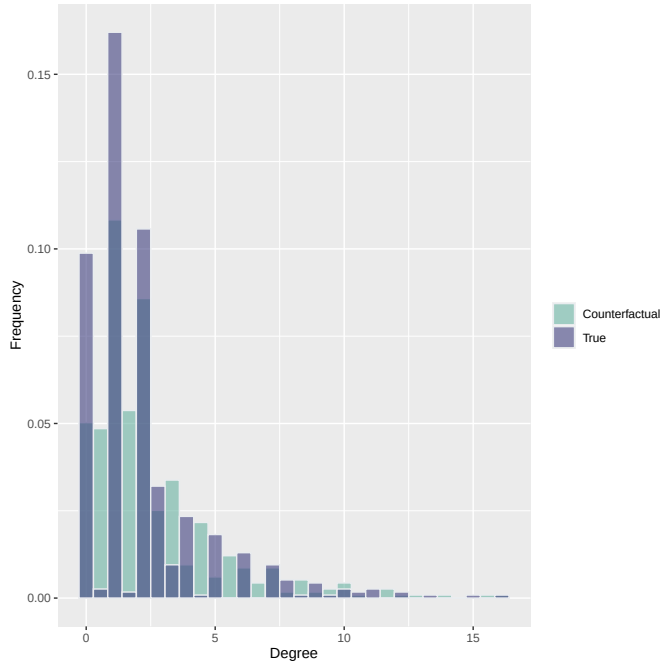


Figure 8: Sorting by gender: true and counterfactual expected degree distributions

As shown in Figures 9 this policy does not generate differences in the distribution of GPA, neither overall nor by gender.

6 Concluding Remarks

In this paper we propose a model of network formation and production interactions, which mimics the structure of many instances of human interaction.

The crucial feature of our model compared to the existing literature is the definition of the payoffs to network formation in a much more useful form, namely as a combination of the direct utility benefits of socialisation and, importantly, also of the output of a production process where linked agents interact. We believe that this is an important innovation in the direction taken by Badev (2017) and Battaglini, Patacchini and Rainone (2019). The work of Carrell,

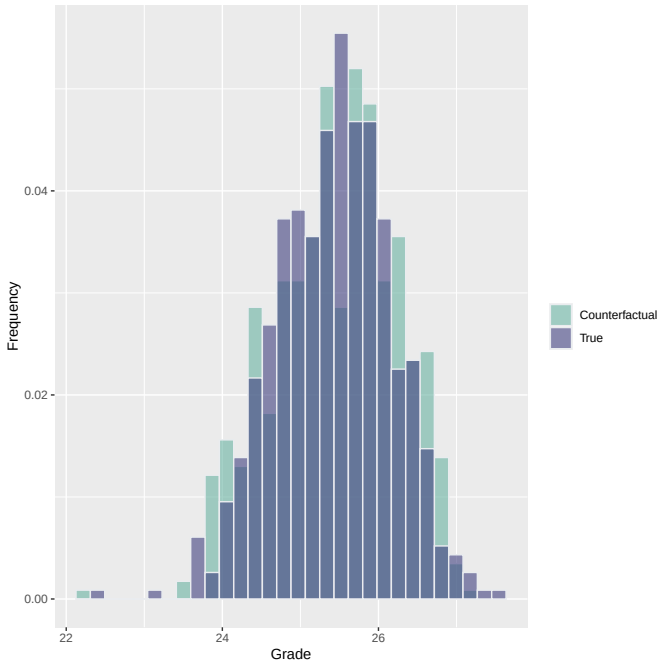


Figure 9: Sorting by gender: true and counterfactual distributions of GPA

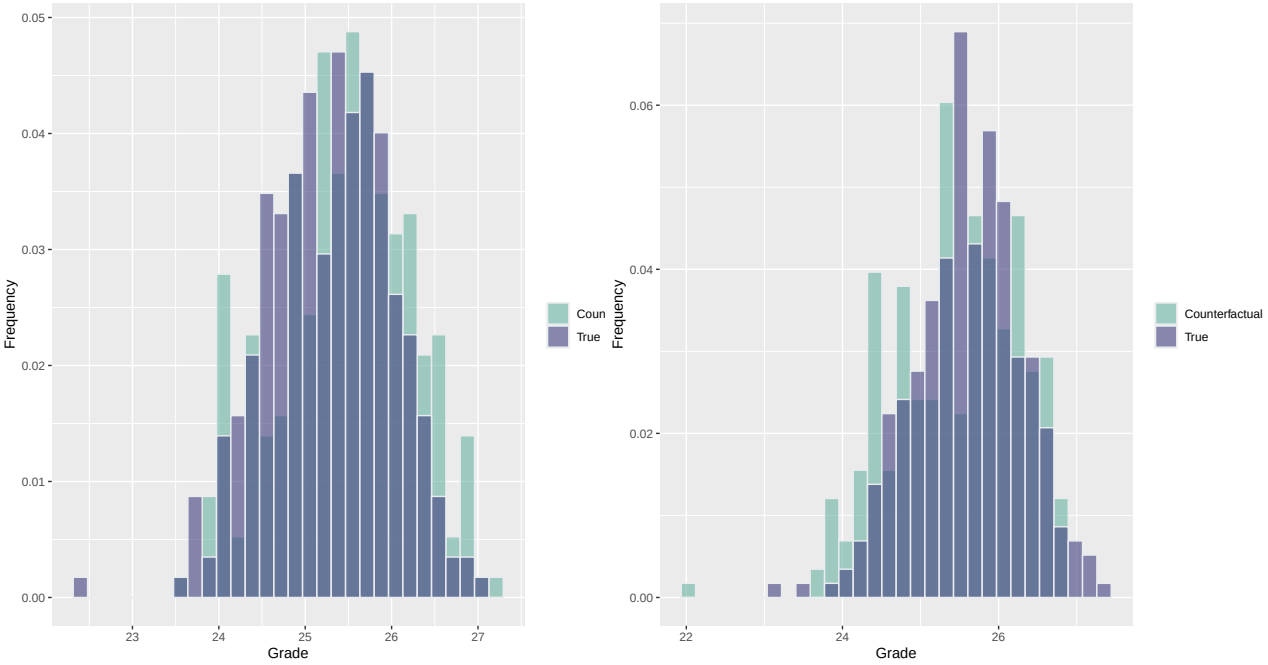


Figure 10: Sorting by gender: true and counterfactual distributions of GPA for female (Left) and male (Right) students

Sacerdote and West (2013) and Kline and Tamer (2020) among others, clearly indicates that a deep understanding the role of networks in human behaviour can only be achieved by jointly modelling the process of network formation and the activities that take place within the network. This is exactly what our model attempts to generate.

Armed with our theoretical framework and our estimation procedure, one can provide insightful policy prescriptions for a broad range of processes featuring interactions among agents connected via complex networks of non-trivial sizes. For example, in our empirical application we discuss the outcomes of college class assignments in terms of ability and gender and find that, while tracking has a positive effect on high-ability students, it also has a negative impact on low ability ones, therefore increasing human capital inequalities. Single-sex classes provide no gains for either gender or average GPA.

Our work also highlights a number of interesting avenues for future research. For example, we simplified our production process to depend only on the exogenous characteristics of the linked peers and it might be useful to expand it to include also endogenous effects, such as peers' effort. However, such an extension would generate a cascade of feedback effects that could span the entire network and make the complexity of the problem intractable. Imposing assumptions, such as those on degree and depth in de Paula, Richards-Shubik and Tamer (2018), might be a solution to limit the cascade which is worth exploring. Another obvious direction of future work is related to the technicalities of the counterfactual simulations. With a large number of agents, it is known that the mixing times of the Markov Chain resulting from the link revision process are very slow, implying that our simulation procedure becomes quite unreliable. Improving the simulation procedure would represent an important advancement to the literature and it would allow producing better policy advice.

References

Ambrus, Attila, Markus Mobius and Adam Szeidl. 2014. "Consumption Risk-Sharing in Social Networks." *American Economic Review* 104(1):149–82.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.104.1.149*

Angelucci, Manuela and Giacomo De Giorgi. 2009. "Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption?" *American Economic Review* 99(1):486–508.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.99.1.486*

Angelucci, Manuela, Giacomo De Giorgi and Imran Rasul. 2018. "Consumption and Investment in Resource Pooling Family Networks." *The Economic Journal* 128(615):2613–2651.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12534*

Badev, Anton. 2017. "Discrete Games in Endogenous Networks: Equilibria and Policy.".

Badev, Anton. 2018. "Nash equilibria on (un) stable networks." *arXiv preprint arXiv:1901.00373* .

Bala, Venkatesh and Sanjeev Goyal. 2000. "A Noncooperative Model of Network Formation." *Econometrica* 68(5):1181–1229.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00155*

Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo and Matthew O. Jackson. 2013. "The Diffusion of Microfinance." *Science* 341(6144).
URL: *https://science.sciencemag.org/content/341/6144/1236498*

Battaglini, Marco, Eleonora Patacchini and Edoardo Rainone. 2019. Endogenous Social Connections in Legislatures. Working Paper 25988 National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w25988*

Bayer, Patrick, Randi Hjalmarsson and David Pozen. 2009. "Building Criminal Capital behind Bars: Peer Effects in Juvenile Corrections*." *The Quarterly Journal of Economics*

124(1):105–147.

URL: *https://doi.org/10.1162/qjec.2009.124.1.105*

Beaman, Lori A. 2011. "Social Networks and the Dynamics of Labour Market Outcomes: Evidence from Refugees Resettled in the U.S." *The Review of Economic Studies* 79(1):128–161.

URL: *https://doi.org/10.1093/restud/rdr017*

Cai, Jing, Alain De Janvry and Elisabeth Sadoulet. 2015. "Social Networks and the Decision to Insure." *American Economic Journal: Applied Economics* 7(2):81–108.

URL: *https://www.aeaweb.org/articles?id=10.1257/app.20130442*

Calvó-Armengol, Antoni, Eleonora Patacchini and Yves Zenou. 2009. "Peer Effects and Social Networks in Education." *Review of Economic Studies* 76(4):1239–1267.

URL: *https://EconPapers.repec.org/RePEc:oup:restud:v:76:y:2009:i:4:p:1239-1267*

Carrell, Scott E., Bruce I. Sacerdote and James E. West. 2013. "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation." *Econometrica* 81(3):855–882.

URL: *https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA10168*

Christakis, Nicholas, James Fowler, Guido W. Imbens and Karthik Kalyanaraman. 2020. Chapter 6 - An empirical model for strategic network formation. In *The Econometric Analysis of Network Data*, ed. Bryan Graham and Áureo de Paula. Academic Press pp. 123 – 148.

URL: *http://www.sciencedirect.com/science/article/pii/B9780128117712000122*

De Giorgi, Giacomo, Anders Frederiksen and Luigi Pistaferri. 2019. "Consumption Network Effects." *The Review of Economic Studies* 87(1):130–163.

URL: *https://doi.org/10.1093/restud/rdz026*

De Giorgi, Giacomo and Michele Pellizzari. 2014. "Understanding Social Interactions: Evidence from the Classroom." *The Economic Journal* 124(579):917–953.

URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12083*

De Giorgi, Giacomo, Michele Pellizzari and Silvia Redaelli. 2010. "Identification of Social Interactions through Partially Overlapping Peer Groups." *American Economic Journal: Applied Economics* 2(2):241–75.
URL: *https://www.aeaweb.org/articles?id=10.1257/app.2.2.241*

De Giorgi, Giacomo, Michele Pellizzari and William Gui Woolston. 2012. "CLASS SIZE AND CLASS HETEROGENEITY." *Journal of the European Economic Association* 10(4):795–830.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1542-4774.2012.01073.x*

De Paula. 2020. Chapter 3 - Strategic network formation. In *The Econometric Analysis of Network Data*, ed. Bryan Graham and Áureo de Paula. Academic Press pp. 41 – 61.
URL: *http://www.sciencedirect.com/science/article/pii/B9780128117712000092*

de Paula, Áureo, Seth Richards-Shubik and Elie Tamer. 2018. "Identifying Preferences in Networks With Bounded Degree." *Econometrica* 86(1):263–288.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA13564*

Duflo, Esther, Pascaline Dupas and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101(5):1739–74.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.101.5.1739*

Garibaldi, Pietro, Francesco Giavazzi, Andrea Ichino and Enrico Rettore. 2012. "College Cost and Time to Complete a Degree: Evidence from Tuition Discontinuities." *The Review of Economics and Statistics* 94(3):699–711.

Glaeser, Edward L., Bruce Sacerdote and José A. Scheinkman. 1996. "Crime and Social Interactions*." *The Quarterly Journal of Economics* 111(2):507–548.
URL: *https://doi.org/10.2307/2946686*

Graham, Bryan S. 2017. "An Econometric Model of Network Formation With Degree Hetero-

geneity." *Econometrica* 85(4):1033–1063.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA12679*

Graham, Bryan and Áureo de Paula. 2020. *The Econometric Analysis of Network Data*. Academic Press.

Jackson, Matthew O. 2008. *Social and Economic Networks*. Princeton University Press.
URL: *http://www.jstor.org/stable/j.ctvcm4gh1*

Jackson, Matthew O. 2019. *The Human Network*. Pantheon.
URL: *https://www.penguinrandomhouse.com/books/541370/the-human-network-by-matthew-o-jackson/*

Jackson, Matthew O and Alison Watts. 2002. "The evolution of social and economic networks." *Journal of economic theory* 106(2):265–295.

Jackson, Matthew O. and Asher Wolinsky. 1996. "A Strategic Model of Social and Economic Networks." *Journal of Economic Theory* 71(1):44 – 74.
URL: *http://www.sciencedirect.com/science/article/pii/S0022053196901088*

Jackson, Matthew O., Brian W. Rogers and Yves Zenou. 2017. "The Economic Consequences of Social-Network Structure." *Journal of Economic Literature* 55(1):49–95.
URL: *https://www.aeaweb.org/articles?id=10.1257/jel.20150694*

Kline, Brendan and Elie Tamer. 2020. Chapter 7 - Econometric analysis of models with social interactions. In *The Econometric Analysis of Network Data*, ed. Bryan Graham and Áureo de Paula. Academic Press pp. 149 – 181.
URL: *http://www.sciencedirect.com/science/article/pii/B9780128117712000134*

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *The Review of Economic Studies* 60(3):531–542.
URL: *http://www.jstor.org/stable/2298123*

Mele, Angelo. 2017. "A Structural Model of Dense Network Formation." *Econometrica* 85(3):825–850.

URL: *https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA10400*

Newman, Mark EJ. 2003. "Mixing patterns in networks." *Physical review E* 67(2):026126.

Patacchini, Eleonora and Yves Zenou. 2008. "The strength of weak ties in crime." *European Economic Review* 52(2):209 – 236.

URL: *http://www.sciencedirect.com/science/article/pii/S0014292107001274*

Sacerdote, Bruce. 2001. "Peer Effects with Random Assignment: Results for Dartmouth Roommates*." *The Quarterly Journal of Economics* 116(2):681–704.

URL: *https://doi.org/10.1162/00335530151144131*

Appendix A: The online survey

First Page (pop-up window): Introduction

Dear student,

we ask you to take the time (approximately 10 minutes) to answer a few short questions for a research project that analyses the process of human capital formation of our students. All the data will be anonymized and will be used for the unique and exclusive purpose of scientific research (no commercial use) according to the current "code of conduct for the treatment of personal data for statistical and scientific purposes" (available at this link).
The research project is coordinated by prof. Tito Boeri (Bocconi University) and prof. Michele Pellizzari (University of Geneva) and a more detailed description is available following this link. Only the researchers officially listed in the project will have access to the data.

Thank you!

Underneath the text, there are two buttons:

1. I accept to participate
2. Not now. Please, remind me next time.

For those choosing 2, the introduction pop-up window is presented again next time they log in until we take the survey off line.

To those choosing 1, the next page is presented.

Second page: Attendance

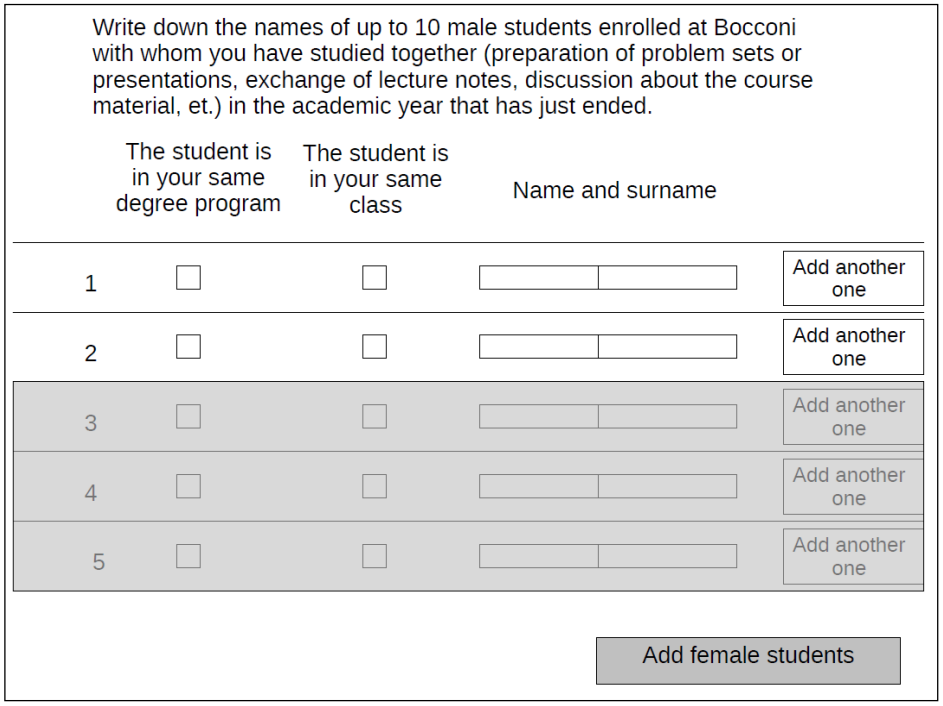
What percentage of the lectures of this semester did you attend (on average across all the courses)?

Less than 50% – between 51% and 70% – between 71% and 90% – more than 90%

Third page: Study mates

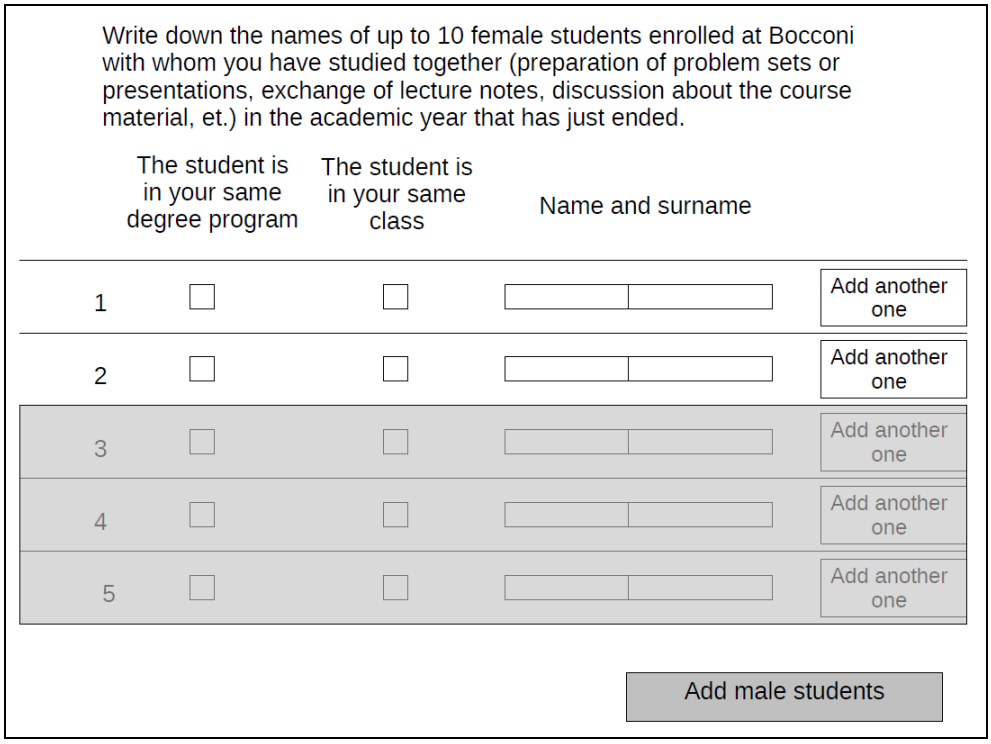
Students with surname starting with letters A to M are first asked about male study mates, the other students are first asked about female study mates.

The template to indicate study mates is the following (for males):



When clicking on "Add another one", the next gray line becomes white and writable.

Similarly for female study mates:



Every time the student clicks on "Add another one", the system checks if the name that was indicated appears in the archive of enrolled students using the information on whether the student was in the same class and program. Different outcomes are possible.

1. *If there is a unique match, the system records the indicated student (and his/her student number) and moves on to the next name.*
2. *If there is no match, the system searches again disregarding he information about the class and the program. If a unique match is found, the system records the indicated student (and his/her student number) and moves on to the next name.*
3. *If there are multiple matches.*
   a) *If there are fewer than 10 matches, the systems presents to the respondent all the possible matches in a drop-down menu, indicating together with the students' names, also their program, class and year of enrollment. The respondent chooses one name, which is recorded by the system and moves on to the next.*
   b) *If there are more than 10 matches.*
      i. *If the respondent did not indicate whether the student was in his/her same class and program, the system asks to add these pieces of information and try searching again. The respondent can, however, disregard this suggestion and click on "Add another one" to move on. In this case, the system records the indicated name and surname without an associated student number.*
      ii. *If the respondent did provide the information about same class/program, the system records the indicated name and surname without an associated student number and moves on to the next name.*

For all students, regardless of whether they are first asked about male or female mates, in the second page (i.e. when they are asked the second gender), the button "Add another one" is replaced by a button names "Completed".

Fourth page: No study mates
(Only for respondents who indicated no student, either male or female)

It is always possible for the respondent to click directly on the buttons in the bottom right corners of the two previous pages, thus indicating no study mate. To these students, the system presents the following question:

You did not list any student, male or female, can you tell us why?
1. You did not study with anyone, you always studied on your own.
2. You studied with students who are not enrolled at Bocconi.
3. You did not want to name your study mates as you thought it would be inappropriate.

Fifth page: Thank you

Thank you for participating!
Thank you for taking the time to answer our questions!