Web appendix for "Identification of Social Interactions through Partially Overlapping Peer Groups"

By Giacomo De Giorgi and Michele Pellizzari and Silvia Redaelli

Appendix A. Additional institutional details

The first set of data extracted from the university archives has been used in Garibaldi et al. (2007) and kindly passed on to us. Our dataset is an extensively updated version of the same sample of students with information on admission tests, teaching classes, course evaluations, labor market outcomes, exchange programs, etc. The currently available dataset covers all students enrolled at Bocconi since 1989.

Until the academic year 1999/2000, Bocconi offered four other degree programs in addition to the Business/Economics: one in "Economic and Social Sciences" (DES), one in "Economics of Financial Market Institutions" (CLEFIN), one in "Management of the Public Administration and International Institutions" (CLAPI) and one in "Law and Business Administration" (CLELI).[1] These degree programs differ both in their curricula and in the number of students admitted in each academic year.[2]

In their application forms, prospective students had to rank the five programs according to their preferences. Admission was based on a standardized entry test combined with high school performance. Applicants were then ranked according to these results and, starting from the top of the ranking, students were assigned to their preferred programs depending on availability. Specifically, a student was allocated to her first choice if there were still places available in that program; otherwise, if all places in her first choice had already been taken by students higher up in the ranking, the candidate was assigned to her second choice, and so on.

It is important to notice that in this mechanism, the student's stated preferences across the five programs do not influence the probability of being admitted, thereby excluding any strategic behavior in the reporting of preferences. This allows us to use this information to construct our indicator of ex-ante preferences. In particular, we consider students who indicated the DES degree - the more academically oriented version of Economics - as a first or a second choice as "determined" to do economics since the beginning of their studies.[3]

[1] Created in 1970, Business (Degree in Business Administration) and Economics (Degree in Economics) are the oldest degrees offered at Bocconi University. Four years later, they were joined by DES, a more quantitative and academic version of the Economics. All the other degrees (CLEFIN, CLAPI and CLELI) were introduced in 1990.

[2] Enrolment ceilings and admission tests were introduced in 1984.

[3] These are students who either had Business/Economics as a first choice and DES as a second or DES as first and Business/Economics as second, and who did not get a place in the DES.

Admitted candidates who decided not to register freed places for students further down in the ranking. However, only a few students (48 out of 753 in our cohort) who had been initially rejected took up a place freed by others, possibly because at the time of making these decisions most people had already obtained admission to another university and started to make arrangements for registration and accommodation.[4]

Eventually, the admission procedure in September 1998 led to 1,385 students (against a ceiling of 1,600) enrolled in the common Business/Economics track, followed by CLELI (239, against a ceiling of 350), CLEFIN (208, against a ceiling of 230), and CLAPI and DES (with, respectively, 132 and 91 against ceilings of 200 each). Once enrolled, Business/Economics students were not allowed to switch to any of the other degrees, while students enrolled in the CLELI, CLEFIN, CLAPI and DES programs could move to Business/Economics only after the first academic year.

In the academic year 1999/2000 Bocconi introduced a major reform of its structure (the so-called "Bocconi 2000" plan). In particular, the Business/Economics was abolished and students were forced to choose a specific degree upon entering the university with relatively limited chances to move across programs at later stages. Moreover, the information on the random allocation of students to classes has unfortunately been lost for the earlier cohorts of students and it is reliable only starting with the academic year 1998/1999. This forces us to use only the cohort of students enrolled in the Business/Economics program in the academic year 1998/1999.

Appendix B. Monte-Carlo Simulations

In this appendix we use simulation methods to investigate two slightly more technical issues that we have only marginally addressed in the main text. First, in Section B.1 we investigate the role of measurement error in the definition of the peer groups. Second, in Section B.2 we extend the simulation setting to analyze the role of heterogeneous correlated effects on the difference between the OLS and the IV estimators.

B.1. Measurement error in the definition of peers

We design a simple Monte-Carlo experiment where we allow for the groups to be mismeasured in a random fashion. To avoid confusion, in what follows we define the *friends* of a generic individual i as those students who truly affect i's outcome and *peers* those students who take classes together with i. Hence, measurement error arises because not all friends are peers and not all peers are friends.

The specific setting is the following: we simulate a simple world where an outcome y_i is a function of the (weighted) mean outcome and the (weighted) mean exogenous characteristics of one's friends, as well as individual traits (x_i), the sequence of class-specific unobserved confounder (U_i^g) and an iid error (ϵ_i):

$$y_i = \alpha + \beta E(y_{-i}|F_i) + \gamma E(x_{-i}|F_i) + \delta x_i + U_i^g + \epsilon_i.$$

[4] Note also that candidates in the lower tail of the distribution of the admission test were not offered any of these residual places.

where F_i is the indicator of the group of friends of student i. For simplicity, in the simulation we consider only a one-dimensional x_i. Consistently with our empirical analysis in the main text, $E(y_{-i}|F_i)$ and $E(x_{-i}|F_i)$ are computed weighting each friend by the number of courses taken together. Friends who are not peers are assigned the weight of the average peer.

The correlated effect U_i^g is generated as to reproduce the sum of the macro-shocks (u_i^g) that student i cumulates in the 7 compulsory classes of our application:

$$U_i^g = \sum_{g=1}^{7} u_i^g$$

where the simplest example of an unobservable u_i^g shock would be teacher quality. This unobservable U_i^g represent the correlated effect in the Manski (1993) wording and it is the source of endogeneity of $E(y_{-i}|F_i)$ that our IV strategy is designed to address.

Next, we generate the vector of the x_i's, the class shocks u^g and the idiosyncratic errors ϵ_i for a sample of n students. Specifically, we assume $x \sim N(\bar{x}, \sigma_x)$, $u^g \sim N(0, \sigma_u)$ and $\epsilon \sim N(0, \sigma_\epsilon)$. Given the linear recursive structure of the model, we can solve for the full vector of the outcomes y, once the groups of friends are defined.

Consistently with our application, we fix the number of courses to 7 and the number of classes for each course to 10, allowing for an uneven distribution of students in each class within each course.[5] Then, we allocate students to classes, and consequentially to peers, according to a totally random procedure and peers are defined, just like in the main text, as those students who take courses in the same classes.

Friends are defined by a probabilistic process that depends on the number of courses two students have taken together. Such process, and consequently the extent of measurement error, is regulated by three crucial parameters:

$$\begin{aligned} Pr(F_{ij} &= 1|meet_{ij} = 0) = \pi_0 \\ Pr(F_{ij} &= 1|meet_{ij} = 1) = \pi_1 \\ Pr(F_{ij} &= 1|meet_{ij} = 7) = \pi_7. \end{aligned}$$

where $F_{ij} = 1$ if i and j are friends and $meet_{ij}$ measures the number of courses i and j have taken together. Hence, π_0 is the probability of two students being friends, given that they never met in the classroom, π_1 is the corresponding probability for those students who met once, and finally π_7 gives the likelihood of friendship when the students met 7 times, i.e. the maximum in our simulated (and actual) data.

We assume, as it seems natural, that π increases (non-linearly) in the number of meetings: the more often two students meet the more likely they will be friends and interact. This structure allows a large degree of flexibility in the extent of mis-measurement we can generate in the simulations, while preserving a relatively simple structure. Each scenario will be defined by the triplet (π_0, π_1, π_7) and all the other values of $\pi_m = Pr(F_{ij} =$

[5] Namely, we introduce some small random variation in class size within each course.

$1|meet_{ij} = m)$ are computed under a simple linearity assumption: $\pi_m = a + bm$ for $m = 1, .., 7$.[6] For example, a scenario of no measurement error is one with $\pi_0 = 0$, $\pi_1 = \pi_7 = 1$, so that all peers would also be friends and vice versa. Increasing π_0 or reducing one or both π_1 and π_7 leads to more mis-measurement.

Throughout the simulation we set the number of observations to 1150 (as in the actual data) and the parameters β, γ, δ and σ_ϵ to the following reasonable values:

Parameter	Value
n	1150
β	0.9
γ	0.2
δ	0.5
σ_ϵ	0.2

Given that our main interest in the simulation is on measurement error and its relation to the endogenous macro-shock effect, we let π_0, π_1, and σ_u vary so as to construct a large set of scenarios. For convenience, we fix π_7 to a constant equal to 0.95. We then estimate the model from the simulated data by OLS and IV, where the IV's for student i, as in the paper, are the x's of the excluded peers. Given measurement error, these may not necessarily be also excluded friends. For each of the simulated scenarios we replicate the data 100 times, estimating the model at each replication. In Table B.1 we present results averaged over the 100 replications.

In the upper panel of Table B.1, we consider a scenario where the class shocks u_i^g have a relatively large variance and, thus, the OLS estimates are more heavily affected by endogeneity bias. Overall, the simulation shows that the IV strategy we introduce in the paper performs extremely well in all the scenarios, irrespective of the degree of mis-measurement. While the OLS always overestimates the true β by over 10 percent, the IV is never biased by more than 4 percent. Similar results are reported in the lower panel where we consider a scenario with $\sigma_u = 0.1$ (low endogeneity), although the OLS and the IV estimators are now a lot more similar.

A final caveat should be borne in mind when comparing our empirical results with the simulation. In fact, while in our analysis we consider a binary outcome and employ a linear probability model, the simulated results are generated by a simple linear-in-means model. In principle, it is possible to simulate a model that more closely mimics the one we use to produce our main results. However, solving such model recursively would be a lot more complicated (mainly because it may feature multiple equilibria) and the role of measurement error might be confounded by the specific functional form assumptions. Thus, we prefer to present simulation results based on a simpler model in order to focus attention on the role of mis-measurement per se.

[6] This linearity assumption could also be modified, although in unreported results we find that it affects results only very marginally.

B.2. Heterogeneous correlated effects

Apart from those few students who take all courses in the same classes, the correlated effects of any two peers would be different. This is enough to make it impossible to draw unambiguous predictions about the sign of the bias of the OLS estimator. Nevertheless, the results in Table B.1 show that in the various specifications that we considered the OLS is larger than the IV. In this section we complicate the structure of the Mote-Carlo experiment to show that allowing the class shocks to have heterogeneous effects across students easily leads to the opposite result (Iv larger than OLS).

In particular, we maintain exactly the same structure of the simulated data as in Section B.1 and we only change the specification of the correlated effect U_i^g, which is now the *weighted* sum of the 7 unobservable class shocks, with weights that are individual specific:

$$U_i^g = \sum_{g=1}^{7} \theta_i^g \times u_i^g$$

In this specification we allow, for example, a good teacher in economics to have a different effect on different students. Technically, while we maintain the same distributional assumptions for the u_i^g's, we draw a vector of 7 individual weights $theta_i^g$'s for each student from a normal with mean zero and variance σ_θ so that that the same class shock u_i^g can have effects of different sign depending on the student.

In Figure B.1 we plot the difference between the IV and the OLS estimators under different assumptions about σ_θ (on the horizontal axis) and σ_u. Results show that, along the entire range of variation in the degree of heterogeneity (σ_θ) the difference between the IV and the OLS estimators is almost equally likely to be positive or negative. If anything, there is a slight tendency to a larger frequency of positive differences as the degree of heterogeneity increases.

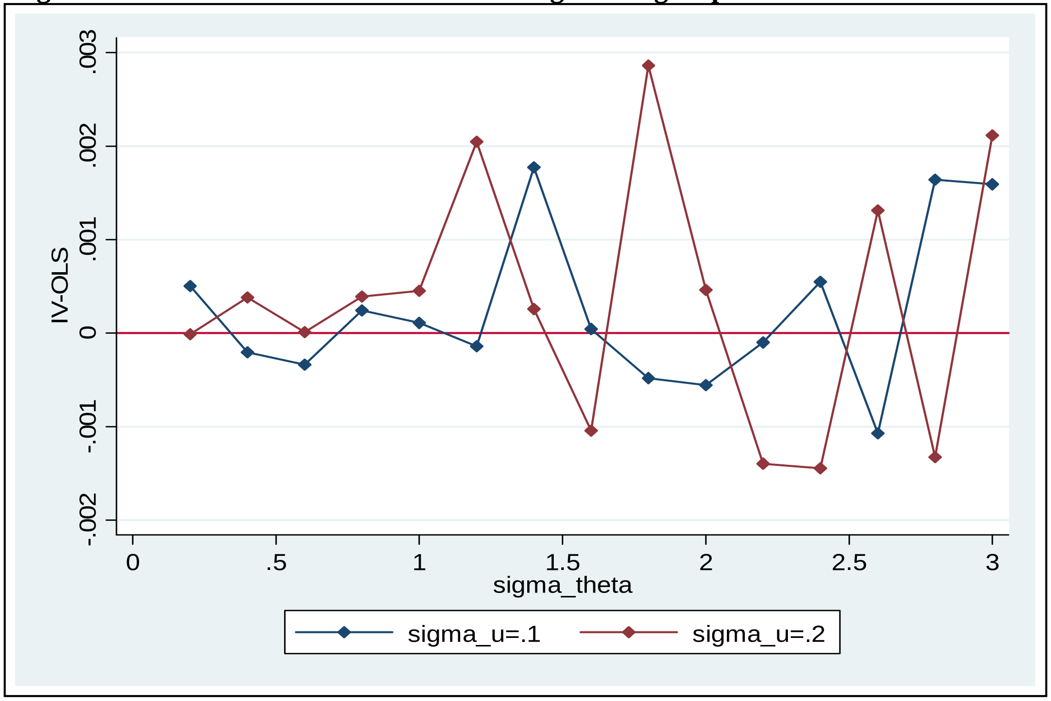
Given that in our main results the IV estimator is larger than the OLS by a sizeable 8-9 times, it is worth noticing that the limited variation in the endogenous variable $E(y|G_i)$ exacerbates the bias in the OLS estimate. To clarify this point, consider a simple linear model with just one regressor: $y = x\beta + \varepsilon$, where x is endogenous and a valid instrument z is available. In this simple case, the OLS estimator can be written as: $\widehat{\beta}_{OLS} = \beta + \frac{Cov(x,\varepsilon)}{Var(x)}$. In the particular case of the linear probability model, it is easy to show that, for given $Var(E(x|y))$, the bias is larger the smaller the variance of the endogenous variable.

Table B.1: Simulation results

Panel 1: High endogeneity (s_u=0.2)

p_0	p_1	% of peers who are friends	% of friends who are not peers	% of peers who are not friends	Ratio OLS/true parameter	Ratio IV/true parameter
no measurement error		1.00	0.00	0.00	1.132 (0.010)	1.010 (0.010)
0	0.3	0.34	0.00	0.66	1.106 (0.014)	0.983 (0.014)
0	0.7	0.72	0.00	0.28	1.120 (0.014)	0.996 (0.015)
0.05	0.3	0.73	0.08	0.27	1.131 (0.014)	1.016 (0.014)
0.05	0.7	0.73	0.08	0.27	1.131 (0.013)	1.016 (0.013)
0.1	0.3	0.41	0.16	0.59	1.146 (0.011)	1.039 (0.011)
0.1	0.7	0.74	0.16	0.26	1.134 (0.011)	1.014 (0.013)
Panel 2: Low endogeneity (s_u=0.1)						
no measurement error		1.00	0.00	0.00	1.033 (0.006)	1.001 (0.006)
0	0.3	0.34	0.00	0.66	1.013 (0.007)	0.992 (0.008)
0	0.7	0.72	0.00	0.28	1.022 (0.007)	0.994 (0.008)
0.05	0.3	0.73	0.08	0.27	1.032 (0.006)	1.004 (0.007)
0.05	0.7	0.73	0.08	0.27	1.032 (0.006)	1.004 (0.006)
0.1	0.3	0.41	0.16	0.59	1.065 (0.007)	1.043 (0.007)
0.1	0.7	0.74	0.16	0.26	1.049 (0.006)	1.021 (0.006)

Figure B.1: IV-OLS difference under heterogeneous group shocks.



Appendix C: Additional results

Table C.1: Common courses

	Semester	**Area**
Management I	1st	Business
Mathematics	1st	Quantitative
Private Law	1st	Law
Accounting	2nd	Business
Economics I	2nd	Economics
Public Law	2nd	Law
Economics II	3rd	Economics
Management II	3rd	Business
Statistics	3rd	Quantitative

Table C.2: Characteristics of courses and lecturing classes

	Semester	Number of classes	Characteristics	Average	(s.d.)	Min	Max
Management I	I	10	Enrolled students	140.40	(14.92)	130	169
			Student questionnaires	80.70	(13.70)	62	109
			Average attendance[a] (%)	85.67	(1.12)	84.08	87.24
			Congestion[b] (1 to 5)	3.33	(0.15)	3.16	3.61
Mathematics	I	10	Enrolled students	140.80	(16.91)	125	164
			Student questionnaires	102.80	(63.86)	28	253
			Average attendance[a] (%)	83.89	(1.53)	81.39	86.51
			Congestion[b] (1 to 5)	3.77	(0.52)	3.00	4.57
Private Law	I	4	Enrolled students	351.75	(164.14)	189	510
			Student questionnaires	70.00	(27.02)	38	104
			Average attendance[a] (%)	79.73	(4.52)	74.91	83.89
			Congestion[b] (1 to 5)	3.07	(0.13)	2.95	3.23
Accounting	II	10	Enrolled students	142.80	(47.75)	109	258
			Student questionnaires	100.30	(61.17)	54	215
			Average attendance[a] (%)	84.80	(1.25)	82.26	86.58
			Congestion[b] (1 to 5)	3.46	(0.48)	3.02	4.40
Economics I	II	6	Enrolled students	216.50	(92.67)	85	316
			Student questionnaires	136.83	(103.78)	24	317
			Average attendance[a] (%)	84.92	(1.23)	83.56	86.84
			Congestion[b] (1 to 5)	3.63	(0.72)	2.83	4.82
Public Law	II	4	Enrolled students	351.75	(147.84)	217	528
			Student questionnaires	41.00	(20.12)	15	64
			Average attendance[a] (%)	82.72	(2.54)	79.45	85.62
			Congestion[b] (1 to 5)	2.89	(0.16)	2.67	3.03
Economics II	III	6	Enrolled students	222.83	(99.20)	156	381
			Student questionnaires	109.17	(52.42)	19	176
			Average attendance[a] (%)	83.87	(1.97)	81.42	86.80
			Congestion[b] (1 to 5)	2.96	(0.47)	2.47	3.72
Management II	III	8	Enrolled students	184.25	(104.07)	123	382
			Student questionnaires	80.75	(25.94)	56	125
			Average attendance[a] (%)	84.38	(0.63)	83.38	85.27
			Congestion[b] (1 to 5)	2.14	(0.25)	1.76	2.51
Statistics	III	8	Enrolled students	272.25	(90.00)	142	404
			Student questionnaires	140.75	(58.91)	35	203
			Average attendance[a] (%)	85.66	(1.04)	83.31	86.53
			Congestion[b] (1 to 5)	3.27	(0.93)	2.09	4.46

a. Self reported by the students.

b. Congestion is defined from students evaluations as the average answer given to the following question: *"For your learning, the number of students attending your class has been: insufficient (1), too low (2), ideal (3), too high (4), excessive (5)"*.

Table C.3: Correlation of individual and peers'/excluded peers' characteristics

Dependent variable:	**1=Determined economics**		**Admission test score**		**High school final grade**	
Fraction of peers determined to economics	0.070 (0.112)	-	-	-	-	-
Fraction of excluded peers determined to economics	-	-0.109 (0.435)	-	-	-	-
Peers' average admission test score	-	-	0.103 (0.101)	-	-	-
Excluded peers' average admission test score	-	-	-	-0.477 (0.347)	-	-
Peers' average high school grade	-	-	-	-	-0.159 (0.101)	-
Excluded peers' average high school grade	-	-	-	-	-	-0.455 (0.428)
Additional controls	Admission test score, high school final grade.		Determined to economics, high school final grade		Determined to economics, admission test score.	
Observations	1,141	1,141	1,141	1,141	1,141	1,141

All regressions also include the following controls: gender, high school type dummies, household income, highest income bracket, non resident dummy.
Standard errors in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%

Table C.4: First-stage regressions for line 2 in Table 7 in the main text

Dependent variable: fraction of peers choosing economics	**Restricted peers** (with exogenous effects)	**Restricted peers** (without exogenous effects)	**All peers**
	[1]	[2]	[3]
Instruments: excluded peers' mean characteristics:			
Admission test[2]	-1.270	-1.270	-0.580***
	(0.773)	(0.773)	(0.155)
Admission test squared	0.009*	0.009*	0.004***
	(0.006)	(0.006)	(0.001)
Fraction of non-resident[3]	-1.085	-1.085	-
	(0.802)	(0.802)	
Fraction of non-resident squared	0.782	0.782	-
	(0.656)	(0.656)	
Fraction of determined economics	-0.400***	-0.400***	-0.134***
	(0.105)	(0.105)	(0.018)
Individual characteristics			
Admission test[2]	-0.008*	-0.008*	0.000
	(0.005)	(0.005)	(0.001)
Admission test squared	0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)
High school final grade[4]	0.017	0.017	0.004
	(0.028)	(0.028)	(0.004)
1=determined economics	-0.009	-0.009	0.001
	(0.007)	(0.007)	(0.001)
1=female	0.006	0.006	0.000
	(0.005)	(0.005)	(0.001)
Log household income[5]	-0.000	-0.000	-0.000
	(0.002)	(0.002)	(0.000)
1=highest income bracket[5]	-0.006	-0.006	-0.002
	(0.019)	(0.019)	(0.004)
1=non resident[3]	0.006	0.006	0.000
	(0.007)	(0.007)	(0.001)
High school type dummies	Yes	Yes	Yes
Region of residence dummies	Yes	Yes	Yes
Nr. Obs.	1,141	1,141	1,141

1. Excluded instruments: averages of admission test, high school final grade, determined to do economics in the group of excluded peers who are not in one's peer group.
2. Normalised between 0 and 100. Average in the sample = 69.10
3. Resident outside the province of Milan.
4. Normalised between 0 and 100 (pass = 60). Average in the sample = 86.3
5. If a student declares that household income falls in the highest income bracket no further information is collected therefore household income is coded to 1 for households in the last bracket and an ad-hoc dummy controls for this group.

Robust standard errors in parentheses: p-values in square brackets.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table C.5: Estimated exogenous peer effects

Dependent variable: fraction of peers choosing economics	**Table 6, column 1**	**Table 6, column 2**	**Table 7, row 2 column 2**
	[1]	[2]	
Average (weighted) peers' characteristics:			
Admission test score[1]	-0.002	-0.003	0.028
	(0.005)	(0.006)	(0.245)
Admission test score - squared	-	-	-0.000
			(0.002)
High school final grade[2]	0.092	-0.724	-0.801
	(0.428)	(0.627)	(0.715)
% of females	0.057	0.094	0.094
	(0.088)	(0.095)	(0.095)
Log household income[3]	0.002	0.023	0.023
	(0.033)	(0.038)	(0.039)
% highest income bracket students[3]	-0.045	0.246	0.260
	(0.346)	(0.419)	(0.428)
% of students determined to economics	-0.086	-0.148	-0.153
	(0.102)	(0.111)	(0.117)
% of non-milanese students[4]	-0.016	0.043	0.264
	(0.089)	(0.097)	(0.554)
% of non-milanese students - squared	-	-	-0.241
			(0.456)
% of students with technical high school degree[5]	-0.064	-0.381	0.045
	(0.207)	(0.285)	(0.101)
% of students with foreign high school degree[5]	0.023	-0.035	-0.408
	(0.076)	(0.085)	(0.315)

1. Normalised between 0 and 100. Average in the sample = 69.10

2. Normalised between 0 and 100 (pass = 60). Average in the sample = 86.3

3. If a student declares that household income falls in the highest income bracket no further information is collected therefore household income is coded to 1 for households in the last bracket and an ad-hoc dummy controls for this group.

4. Resident outside the province of Milan.

5. The reference group is students with a classical or scientific degree (lyceums).

Robust standard errors in parentheses: p-values in square brackets.

* significant at 10%; ** significant at 5%; *** significant at 1%