Discussion Paper Series – CRC TR 224

Discussion Paper No. 174
Project B 02

"No Man is an Island":
An Empirical Study on Team Formation and Performance

Alessandra Allocca*

May 2020

*Center for Doctoral Studies in Economics, University of Mannheim, Germany
(alessandra.allocca@gess.uni-mannheim.de)

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

"No Man is an Island":
An Empirical Study on Team Formation and Performance

Alessandra Allocca*

May 14, 2020
Click here for the latest version

Abstract

Many organizations rely on decentralized arrangements where employees choose their projects and teams. Most of the empirical literature on working collaborations instead focuses on teams that are exogenously formed. I develop a structural entry model with heterogeneous strategic interactions where agents decide whether to join a project. The decision depends on who else may potentially join the project, the project quality, as well as other individual and project characteristics. In turn, this decision affects the probability of project completion. I estimate the model using a novel dataset from an important scientific collaboration. I find that agents' decisions to select into projects highly depend on the pool of teammates and the size of the team whereas projects' quality is of lesser importance. Heterogeneity in agents' characteristics explains this selection, which needs to be accounted for to obtain unbiased estimates of teams' performance. With a counterfactual experiment, I show that moving from a decentralized to a centralized arrangement leads to fewer completed projects.

Keywords: Teamwork, Entry Game, Innovation, Personnel Economics.
JEL Codes: C57, C72, L2, M50, O32.

*Center for Doctoral Studies in Economics, University of Mannheim, Germany (alessandra.allocca@gess.uni-mannheim.de). I am deeply grateful to my advisors, Michelle Sovinsky, Laura Grigolon and Emanuele Tarantino. Special thanks to David Byrne, Bernhard Ganglmair, Phil Haile, Ariel Pakes, Saverio Simonelli, Konrad Stahl, and Hidenori Takahashi for their helpful comments. I benefited from discussions with Francesco Paolo Conteduca, Franco Esteban Cattaneo, Yihan Yan, Robert Aue, Cristina Bélles-Obrero, Ruben Hipp, Ekaterina Kazakova, Harim Kim, Niccolò Lomys, Yasemin Özdermir, Daniel Savelle, André Stenzel, Christoph Wolf, Annalisa Allocca, and with seminar participants and discussants at Mannheim, UCL, Bonn-Mannheim Ph.D. Workshop (Bonn), XXIX School in Economic Theory (Jerusalem), ENTER Jamboree (Tilburg), MACCI IO day (Mannheim), SAET Conference (Ischia), CRC Young Researchers Workshop (Mainz), JEI (Madrid). I thank Federico Ferrini for the opportunity of using the data. I gratefully acknowledge financial support from the German Research Foundation (DFG) through CRC TR 224 (Project B02).

No man is an Iland, intire of itselfe;
every man is a peece of the Continent,
a part of the maine [...].

John Donne - 1624

1 Introduction

Teamwork is a crucial element in determining the success of firms or research institutions. Over the centuries, the paradigm of working organization has shifted toward the execution of more specialized tasks, usually assigned to workers in a centralized fashion. At the same time, the organization of teams is evolving. Many companies and institutions now adopt decentralized approaches to production, such as open work-flows and *Agile* business practices.[1,2] As an example, Valve Corporation, one of the US leading companies in entertainment software and technology, states that open work-flows are a primary competitive advantage in recruiting and retention: "We've heard that other companies have people allocate a percentage of their time to self-directed projects. At Valve, that percentage is 100. Since Valve is flat, people don't join projects because they're told to. Instead, you'll decide what to work on after asking yourself the right questions." (Valve, *Handbook for new employees*, page 8). Scientific institutions, in which researchers often collaborate voluntarily, adopt similar arrangements (Guimera et al., 2005; Wuchty et al., 2007). This evidence naturally raises many questions: which elements drive the decision to join projects? How can we measure the performance of teams when the decision to join projects is endogenous? Is decentralization better for successful outcomes?

I address these questions with a unique dataset from an important collaborative experiment in science, Virgo. The ultimate goal of Virgo is the detection of gravitational waves, and the founders of the LIGO/Virgo (LIGO is the U.S. counterpart of Virgo) collaboration were awarded the 2017 Nobel Prize in Physics. Researchers involved in Virgo choose which project(s) to work on and use an on-line platform to report their activities. Within this framework, I disentangle the effects that complementarity/substitutability among researchers and projects' ex-ante heterogeneity (i.e. project ex-ante unobserved quality) have on the decision to join a project, controlling for researchers' and projects' exogenous characteristics. I show that to correctly assess the performance of an endogenously formed team one needs to take into account what drives the sorting of researchers into projects.

[1] *Let Employees Choose When, Where, and How to Work*, Harvard Business Review, N. Koloc, 2014.
[2] https://www.agilebusiness.org/page/About

The relevance of the analysis is twofold. First, the evaluation of workers' performance is a cornerstone of the literature in organizational economics. As workers' and organization's incentives are usually not aligned, conflicts of interest can generate inefficiencies. Since Holmstrom (1982), the literature on team production has focused on the analysis of optimal monitoring and incentives for workers, in terms of payment and career schemes.[3] My analysis contributes to this literature by documenting that a decentralized mechanism of project participation can create misalignment through another channel: the allocation of workers to projects based on individual preferences.

A large amount of public and private funds is allocated every year to scientific organizations, which are usually based on decentralized arrangements for project participation.[4,5] Similarly, within-firm workforce has also evolved toward a more flexible organization system. Hence, it is crucial to take into account individual preferences related to project participation, especially to understand if a decentralized mechanism can improve upon a centralized allocation of workers to projects, and to establish when a decentralized mechanism is desirable.

Second, endogenous project participation can bias the estimates of the parameters of interest (e.g. performance, productivity, and efficiency) if sorting is neglected. Workers may sort into projects for reasons that are not observable by the econometrician. The empirical setting of this paper provides a clean framework to address these issues.

To analyze the drivers of team formation, I develop and estimate an entry game with incomplete information *à la* Aguirregabiria and Mira (2007) where agents decide simultaneously whether to join a project. By revealed preference, an agent joins a project only if its payoff from doing so exceeds that from not joining. The former payoff depends on exogenous project characteristics, including a measure of ex-ante quality, the expectation on potential project-mates' actions, and a project-agent specific component. In this setting, strategic complementarities and substitutabilities may arise. Once agents make their entry decisions, the project is developed and ends with an outcome. The outcome is a function of different (observed and unobserved) characteristics, including information about other team members.

My main finding is that the pool of expected project-mates drives the decision to join a project while project quality is of lesser importance. The larger the pool, the lower the

[3] See (Bolton et al., 2005) and (Prendergast, 1999).

[4] For example, ERC, DFG, NRC, NSF, UK Research Councils grants.

[5] Recent empirical papers in innovation study the mechanisms behind collaborations and interactions in the innovation process (Akcigit et al., 2018) and in technical standards development (Ganglmair et al., 2018).

probability of joining a project, because of congestion or increasing coordination and communication costs (Becker and Murphy, 1992). Heterogeneity in researchers' characteristics plays an important role in explaining selection into projects. For example, senior researchers are more likely to join projects of expected larger size relative to junior researchers. I show that controlling for projects' ex-ante quality and endogenous project participation matters for obtaining unbiased estimates of teams' performance.

To assess the desirability of a decentralized mechanism, I consider a counterfactual centralized mechanism in which strategic interactions have no value. I find that the new allocation leads to excessive project participation and decreases the probability of project completion. Hence, a decentralized mechanism of task allocation within a firm can be more efficient because workers internalize costs and benefits of working together better than a centralized mechanism.

Starting from the seminal work of Lazear (1998), many papers have studied working collaborations.[6] Empirical works on peer effects analyze group interactions and how these affect productivity (Falk and Ichino, 2006; Mas and Moretti, 2009; Bandiera et al., 2010; Lindquist et al., 2015, among others). They show that co-workers can exert economically significant effects on their peers, via channels not explicitly created by the management system, such as social connections and network effects.

This paper contributes to the studies on working collaborations in several ways. First, a key challenge in this body of empirical work concerns the identification of the main determinants of workers' selection into teams. To address this challenge, I develop the methodologies used to study firm's entry decision into markets (see Aguirregabiria and Suzuki (2015) for a recent survey of the literature). A crucial difference between firms' and workers' decisions, however, is that the latter can gain from the presence of others. This distinction plays an important role in my structural model. Controlling for the determinants of selection proves to be crucial to correctly evaluate the performance of teams. Second, to estimate the model, it is important to observe individuals' decisions to enter a project, the project's characteristics, and its outcome. The data from the Virgo experiment are ideal to obtain this information. At the same time, with this unique data source, I can analyze the mechanisms behind knowledge creation in science. Third, using my estimates, I test the efficiency of the actual decentralized mechanism of researchers' allocation against a centralized mechanism to assess whether decentralization is a desirable design of teamwork within organizations. To my knowledge, this paper is the first to take a step toward understanding

[6]For instance, Hamilton et al. (2003) argue that teamwork is beneficial when there is specialization and knowledge transfer of information that may be valuable to other team members.

the determinants of decentralized team formation in working collaborations and to assess the efficiency of this mechanism.

The paper proceeds as follows: in the next section, I describe the institutional details and the data. I discuss the model in Section 3 and the empirical implementation in Section 4. Results from descriptive regressions and from the full model are presented in Section 5. I present the counterfactual in Section 6 and conclude in Section 7.

2 Institutional Details and Data Sources

I use unique data from a science experiment named Virgo[7], founded by the French National Center for Scientific Research (Centre National de la Recherche Scientifique – CNRS) and the Italian National Institute for Nuclear Physics (Istituto Nazionale di Fisica Nucleare – INFN)[8] in 1987 and completed in 2003. Virgo is operated in Italy, on the site of the European Gravitational Observatory (EGO), by an international collaboration consisting of about 200 people affiliated to 20 laboratories all over Europe. Virgo has two "sisters" in the United States, LIGO Livingston and LIGO Hanford. This joint collaboration has proven very successful and indeed the founders[9] were awarded the Nobel Prize in Physics in 2017.

Virgo consists of a giant laser interferometer. Interferometers work by merging two or more sources of light to create an interference pattern, which can be measured and analyzed. The interference patterns generated by the interferometers contain information about the object or phenomenon being studied. Virgo studies phenomena related to gravitational waves. The detection of gravitational waves, predicted by Albert Einstein's general relativity, has challenged physicists for over a century. During the 1970s, the discovery of the anomalies in the arrival times of radio pulses, due to a close neutron star, represented a crucial step toward the gravitational waves detection because it showed how catastrophic astronomical events can determine ripples in space-time. In 2015, the merger of two black holes radiated an amount of energy equivalent to $3.0 + -0.5$ solar mass in the form of gravitational waves. This event was recorded by LIGO. Subsequently, other events were recorded also by Virgo.

Building up the laser interferometer requires an incredible amount of resources and time:

[7] `http://www.virgo-gw.eu/`

[8] National Research Centers in France and Italy.

[9] "Pioneers Rainer Weiss and Kip S. Thorne, together with Barry C. Barish, the scientist and leader who brought the project to completion, ensured that four decades of effort led to gravitational waves finally being observed.".

this process is divided into intermediate steps, that I define as macro-projects. Macro-projects relate to the different phases of the development of the experiment: they could refer to the Infrastructure System of the Interferometer or to the Injection System, which takes care of the optics of the high power laser.[10] Therefore, different skills and knowledge are required depending on the actual task to perform. Macro-projects are then split into smaller tasks which do not compete with each other. I define them as projects.

The dataset spans more than 4 years from June 2012 to September 2016). June 2012 is the starting point of a new phase of the experiment (Advanced Virgo) which was completed in January 2017. Projects were set up in the *Technical Design Report* in April 2012. The report contains detailed descriptions of the projects and has been edited as a joint effort of the researchers working in Virgo at that time. Importantly, as the Report is compiled before Advance Virgo started, the projects are pre-determined and not designed to tailor specific researchers' characteristics.

In Virgo, the assignment of projects to researchers happens in a decentralized fashion: each member of the experiment voluntarily decides whether or not to join a certain project. The only exception holds for new entrants in the experiment; they are usually students or junior researchers who, during a few weeks at the start of their experience in Virgo, are exogenously allocated to projects. Researchers are paid a fixed wage by regulated contracts, in line with the respective national collective agreements,[11] so their salary does not depend on the performance. Moreover, because projects are relatively short lived, there are no long-term monetary or career incentives in joining a particular project.

The dataset used in the empirical analysis comprises several sources. I web-scrape information regarding the characteristics of the projects, the final outcomes of the projects and the projects' participants from the Logbook of Virgo. I complement my dataset by hand-collecting data on researchers' characteristics (nationality, gender, level of education, professional seniority) from several on-line sources, mainly personal websites, available *curricula* and LinkedIn profiles. These are discussed in turn.

2.1 The Logbook

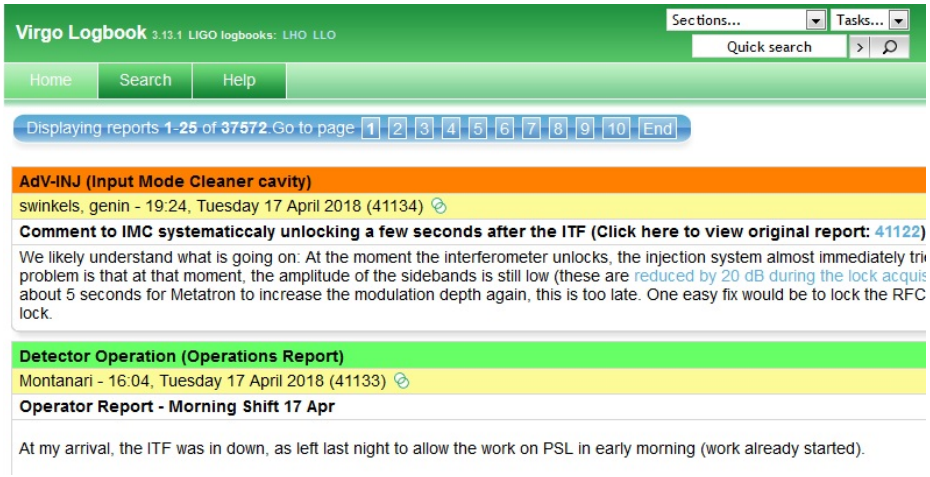
Researchers in Virgo communicate using an on-line platform: the Logbook,[12] which consists of web-pages held by project teams. With the advent of new electronic notebooks it has become possible to incorporate valuable information into enterprise-wide information

[10] A detailed description of the macro-projects is available upon request.
[11] The agreements may differ among Countries.
[12] Source: `https://tds.ego-gw.it/itf/osl_virgo/index.php`

management systems (McAlpine et al., 2006). The logbook allows researchers to record information on working projects and experiences such as results of measurements, tests, data taking, that describe the results of activities and tasks, are required for future activities or may be of value in the future. The Logbook is therefore a meeting platform for project-seeking researchers, who might also work from different locations, and hence it represents a communication platform with minimal search frictions (Hitsch et al., 2010). Moreover, researchers are obliged to report their work on the Logbook. This facilitates monitoring among researchers, as reports are observable and their content is verifiable. Because of these features, moral hazard or free riding are of limited concern. Furthermore, it is also unlikely that researchers coordinate beforehand about who is joining projects outside the on-line platform, as projects are short and coordination would result in observing long delays in the execution of the projects which are not observed in the data.

Each web-page of the Logbook consists of logs (or entries). A log represents a description or an update of a project; it is identified by the title of the macro-project and the project it refers to, the name of the author(s), the time and date, the (chronological) number, the main text and possibly images, comments or other files attached. A screenshot example of a Logbook page is given by figure 1. I provide two examples of projects in Appendix A.

Figure 1: Example of a Logbook page



This web-page consists of two logs belonging to different projects. For each log, the first row identifies the title of the macro-project; the second row identifies the name(s) of the project participants, together with time and day of the log; the third row identifies the project; the fourth part identifies the actual text of the project. In this example, the first is a project with two participants, the second is a single author project.

Table 1 shows descriptive statistics. The full dataset contains 16 macro-projects and

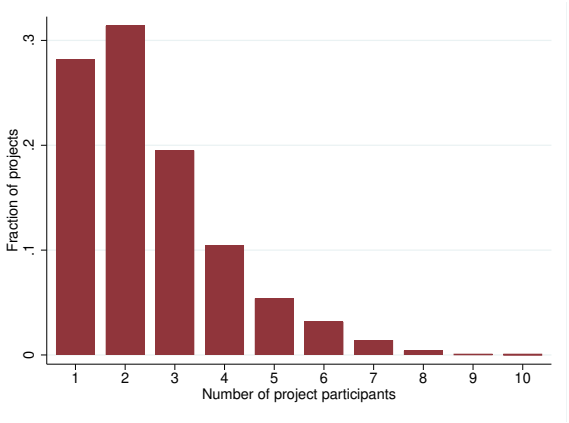
2, 243 projects. The average number of logs per project is 1.2: projects usually do not consist of multiple sequential rounds; this motivates the decision to model joining a project as a static one. In Appendix B I show the logs distributions per projects. Around 70% are team projects, the rest are solo projects. The maximum observed team size is 10 with an average of 3.09 participants per team.

	Mean	**St. Dev.**
Sample period	June 12 - Sept. 16	
No. of projects (obs)	2,243	
No. of macro-projects	16	
Logs per project	1.2	0.4
Team projects	71.8%	0.47
Team size	3.09	1.31
Max team size	10	
No. of projects with pre-determined teams	152	
No. of projects with external companies	71	

Table 1: Descriptive Statistics

Figure 2 shows the distribution of project participants. Projects with two participants are the most frequent, followed by solo projects.

Figure 2: Distribution of project participants



While project participation is mostly decentralized, in 152 projects (6% of the total number of projects) there are teams that are exogenously formed or pre-determined (formed off-line). For those teams, I do not observe the actual participants. Further, I defined these teams as "pre-determined teams". In very few projects (72), there is the participation of external companies; these companies supply Virgo with instruments and tools for the lab

experiments and help researchers to set up those instruments.

Table 2 reports the frequency of the macro-projects in terms of number of projects.

	Frequency
Macro-Project 1	305
Macro-Project 2	167
Macro-Project 3	75
Macro-Project 4	170
Macro-Project 5	463
Macro-Project 6	35
Macro-Project 7	18
Macro-Project 8	1
Macro-Project 9	135
Macro-Project 10	85
Macro-Project 11	320
Macro-Project 12	27
Macro-Project 13	59
Macro-Project 14	76
Macro-Project 15	115
Macro-Project 16	192
Total	2,243

Table 2: Frequency of macro-projects

In order to determine the final outcome of a project, I require a quantifiable measure of output. One possibility is to use publications that resulted from the projects. Unfortunately, this is not a viable option for two reasons. First, not all the projects end with a publication. Second, in Virgo the general rule is that publications that follow from a project must contain the names of all Virgo researchers in alphabetical order.[13] Therefore I require a less noisy measure. Fortunately, the Logbook comments represent an important data source for this scope. Depending on how complete a project is, it can end with different outcomes. I examine the text to determine a measure of outcome. In particular, I classify each text into one of two different categories.[14] When a project has more than one log, I only consider the latest one to measure project outcome. The categories are the following:

0. Describe a problem or a task proposing possible solutions (with no actual intervention); fix or understand a problem or perform a task temporarily/partially, do a measurement still in progress.

[13]By checking the research web-pages of some of the researchers in Virgo (for instance, on the platform `https://www.researchgate.net/`), it emerges that many publications have above $1,000$ authors.

[14]I perform robustness checks with three categories. For now, I implement the classification manually. Initially, I used tools from Supervised Machine Learning (in particular, classification methods) to determine measures of success. However, this classification proved less fruitful than manual classification because the jargon of the text is very detailed; therefore any set of *features* I gave as inputs to the classifiers was not improving the classification.

1. Fix or understand a problem, perform a task with success, complete or improve a measurement or survey.

In the sample, 11% of the projects are in class 0 and the rest in class 1. The classified texts are the measure of project outcome that I use in the empirical estimation. Some examples can be found in Appendix B.

2.2 Researchers' Characteristics

Virgo consists of 192 researchers. The pool of researchers that collaborate in Virgo is very heterogeneous. I hand-collect data on their demographic characteristics (nationality, gender, education, professional seniority, field of research) from several on-line sources, mainly personal websites, available *curricula* and LinkedIn profiles. Researchers can have very different backgrounds, work in different fields and have different nationalities. In order to coherently classify them in terms of professional seniority and education, I use the information available on the websites of the main European National Research Centers.[15] Figure 3 shows the table of conversion for professional seniority.[16]

Table of conversion professional seniority			
Academia	Research Institution (Italy)	Research Institution (France)	Technical Profession (no degree)
PhD		Engineer	
		Technologist	
		Ingénieur d'études	
Post-doc	Post-doctoral fellow		
Researcher/Assistant Prof	Researcher	Ingénieur de recherche	Technician
		Chargé de Recherche	(Technicien d'atelier) Assistant ingénieur
Associate Prof	First Researcher	First Engineer	First technician
Full Prof	Director of Research	Diriger des Recherches	
		Director technologist	

Table 3: Table of conversion professional seniority

Figure 4 provides descriptive statistics of researchers' demographics. 85% are male. Around 20% of the researchers in Virgo are juniors, whereas 80% are seniors.[17] Not all seniors have a Ph.D.; this is because seniors include technicians that do not hold a degree

[15] `http://www.differencebetween.net/miscellaneous/difference-between-technician-and-technologist`, `http://www.guide-des-salaires.com/fonction/technicien-datelier`,`http://www.cnrs.fr/en/join/engineer-technician-permanent.htm`,`https://cadres.apec.fr/Emploi/Marche-Emploi/Fiches-Apec/Fiches-metiers/Metiers-Par-Categories/Etudes-recherche-et-developpement/charge-de-recherche`,https://www.dgdr.cnrs.fr/drhchercheurs/concoursch/chercheur/carriere-en.htm

[16] When I am not able to find the professional position, I deduce it from the age, h-index or field of research. When two different levels of seniority are stated, I take the highest.

[17] Senior level 1 is the equivalent of Associate Professor in Academia; senior level 2 is comparable to the definition of Full Professor.

or engineers. Not surprisingly, more than 60% are specialized in Physics. For 18 researchers (around 17% of the total number) I was not able to find information online (most likely they are technicians or seniors that do not have an online identity; for some of them, I only observe their nickname, therefore I am not able to go back to their original names); they appear in only 93 projects.

	Frequency
Professional seniority	
Juniors	20%
Seniors level 1	63%
Seniors level 2	17%
Field of Research	
Physics	61%
Engineering	24%
Others	15%
Other demographics	
Males	85%
Italians	58%
With Ph.D.	36%
No. of researchers	*174*

Table 4: Descriptive statistics for researchers

As I will discuss in Section 4, entry models with heterogeneous strategic interactions are computationally intense. Therefore, I exploit the information on researchers' demographic characteristics to reduce the burden of computation of the empirical model. In particular, I assign each researcher exclusively to a certain type, which is as a combination of two characteristics: field of specialization and professional seniority. I simplify further by aggregating Seniors level 1 and Seniors level 2 together in the category "Seniors", and Engineers and researchers specialized in fields other than Physics in the category "Other fields". Following this specification, an example of type is: specialized in Physics, Junior researcher. Table 5 shows the distribution of researchers by types; table 6 shows the number of projects for each type.

	# of Researchers
Non-Physics Seniors	65
Non-Physics Juniors	2
Physics Seniors	74
Physics Juniors	33
Non classified	18
Total	*192*

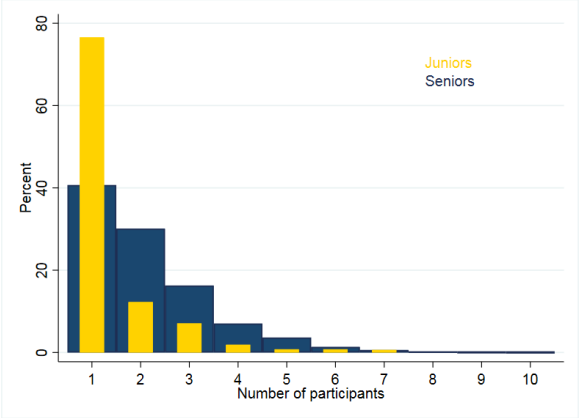
Table 5: Number of researchers by type

	# of Projects
Non-Physics Seniors	1,058
Non-Physics Juniors	18
Physics Seniors	1,648
Physics Juniors	485
Non classified	93

Table 6: Number of projects by type

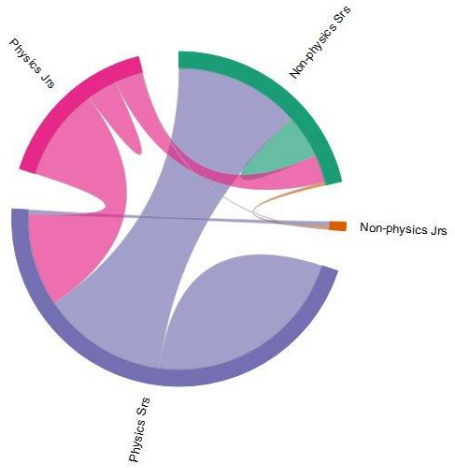
Figure 3 shows the distributions of project participation with individuals in the same level of seniority for Juniors (light yellow bars) and Seniors (dark blue bars). The frequency with which a junior works with one or more juniors is around 20% and it is visibly lower than the frequency of a senior working with one or more seniors, around 60%.

Figure 3: Homophily by professional seniority



In Appendix B, I show the distributions separately for Non-Physics Seniors and Physics Seniors. They have similar patterns. From the previous figure, it emerges that juniors do not frequently work with other juniors, but this does not necessarily imply that they work alone. Figure 4 gives a more comprehensive illustration of the bilateral project connections among researchers' types.

Figure 4: Chord diagram of bilateral project connections by types



The length of the purple arch corresponds to the total number of projects with at least one Physics Senior (1,648). The length of the pink arch corresponds to the total number of projects with at least one Physic Junior (485). Likewise for the other types: the green arch is for Non Physics Seniors (1,058) and the orange arch for Non-Physics Juniors (18). Non-Physics Juniors work in very few projects.

The pink flow that links Physics Juniors and Physics Seniors represents the projects in which the two types collaborate. Same holds for the pink flow that links Physics Juniors and Non-Physics Seniors. The pink flow that turns back into the pink part represents the projects in which Physics Juniors collaborate with other individuals of the same type. One can easily see that Physics Juniors are working more frequently with Seniors (both Physics and Non-Physics) than with Juniors from the same field. Moreover, a big portion of Physics Seniors collaborate with Non-Physics Seniors, as suggested by the purple flow that links the two types. The evidence suggests that the allocation of researchers to projects is non-random. I show that these paths hold also when controlling for project characteristics. I exploit this variation for identifying the main determinants of project participation in the structural model.

3 Model

In this section, I present a structural model to quantify the determinants of projects' outcomes controlling for the endogenous drivers of working collaborations. For every project, each researcher type observes the set of exogenous characteristics and the set of potential

entrants, and her own idiosyncratic shock. She decides whether to join a working project by comparing post-entry single period payoffs.[18] In the last stage, a project ends and the outcome realizes. The model is composed of two parts. First, I present the structural model of project participation as an entry game of incomplete information. Then, I define the outcome equation; this expresses the outcome of a project as a function of different factors, including the number of researchers determined in the first stage.

3.1 Game of Project Participation

I model the decision to join a project as an entry game with incomplete information, following the literature on estimating games of incomplete information (e.g. Seim (2006)). The model is static and agents make their decisions simultaneously. The payoff form joining a project is positive, while the payoff from not joining is normalized to zero.[19] For every project, a type decides whether to join a working project by comparing single-period payoffs.

3.1.1 Payoff Function

Consider a set of projects $\mathcal{J} = \{1, ..., J\}$ indexed by j, where each project belongs to a macro-project. A researcher is defined uniquely by her type g, with $g = 1, ..., G$. An agent-type wants to join a project because of the other agent-types she might work with or because the project has desirable features (for instance, high ex-ante quality). The latent benefit of joining a project can capture short-term reputational concerns, willingness to learn or intrinsic motivation. The payoff associated with the decision of joining a project depends on the following factors. First, it depends on the exogenous characteristics of the project, including the macro-project the project belongs to and its ex-ante quality. Second, it depends on the strategic interactions among participants, namely who else is potentially joining the project. Finally, the decision hinges also on a stochastic component, that gives information about the ability of the type to work on the project. The payoff of researcher of type i associated to joining project j takes the following form:

$$\pi_{ij} = \alpha_i X_i + \eta' D_j + \sum_{g=1}^{G} \delta_{ig} N_{gj} + \tilde{q_j} + \epsilon_{ij} \tag{1}$$

[18] I abstract from any dynamic consideration in this paper for several reasons. First of all, the projects are relatively small and last a short amount of time. Therefore long-term reputation concerns that lie under any dynamic decision are rather negligible. Second, a dynamic setting will have the disadvantage of ignoring the intermediate steps in terms of project final outcomes. Repeated interactions among players may obviously affect the decision of joining certain projects.

[19] This assumption is standard in the literature of entry games.

X_i is the vector of type characteristics (dummies for the four types: Physics Seniors, Physics Juniors, Non-Physics Seniors, Non-Physics Juniors), D_j is comprehensive of macro-project dummies and other project exogenous characteristics (number of pre-determined teams, number of external firms) of project j; N_{gj} denotes the number of researchers of type g in project j;[20] $\tilde{q_j}$ is the ex-ante project quality and it is unobserved by the econometrician; ϵ_{ij} is the researcher type and project-specific shock. Each researcher observes her own project-individual-specific shock, but only knows the distribution of the others' errors; therefore, the described entry game is a game of incomplete information.

The vector of parameters to estimate is given by $\theta_1 = (\alpha, \eta, \delta)$. In particular, δs capture the strategic substitutability/complementarity with respect to the teammates. Because of imperfect information about her teammate payoff, i can only form expectation of their optimal choices. Based on the expected teammate distribution across projects, each researcher type chooses whether to join a project by maximizing her payoffs given her own type. In particular, type i joins project j if:

$$\mathbb{E}[\pi_{ij}] = \alpha_i X_i + \eta' D_j + \sum_{g=1}^{G} \delta_{ig}\, \mathbb{E}[N_{gj}] + \tilde{q_j} + \epsilon_{ij} \geqslant 0. \tag{2}$$

3.1.2 Equilibrium

Define as p^*_{ij} the equilibrium probability of entering project j for i. Then, the following has to hold:

$$p^*_{ij} = \Phi\left(\alpha_i X + \eta' D + \sum_{g=1}^{G} \delta_{ig} p^*_{gj} + \tilde{q_j}\right) \tag{3}$$

for all i and j. $\Phi(\bullet)$ is assumed to be a continuous CDF. Researcher type i's vector of equilibrium conjectures over all projects is then defined by the set of J equation probabilities. The system (3) defines the equilibrium conjectures as a fixed point of the mapping from i's conjecture of her teammates strategies into her teammates conjectures of the i's strategy. The existence is given by Brouwer's Fixed Point Theorem.[21]

[20] For sake of simplicity and in line with some of the literature on entry games of incomplete information, I assume that the number of others researchers enters the payoff linearly.

[21] The agents' own conjectures enter the probability simplex and are continuous in others' expected behaviour.

3.2 Project Outcome

Researchers produce scientific projects of varying outcomes. The final outcome of a project can be expressed as a function of different "inputs", which include the number of researchers who endogenously participate to the project (this is in spirit close to Akcigit et al. (2018)) and other project exogenous characteristics. Moreover, a project might be better because of some idiosyncratic heterogeneous ex-ante components. I include these components in what I define as the ex-ante project quality.

There exist $\mathcal{J}$ and $\mathcal{G}$ researchers' types. Each project j ends with a certain outcome. The variable $outcome_j$ takes on a value of 0 or 1, depending on the project classification (see 2.1). It underlies a continuous variable $outcome^*$, which is a latent variable for degree of project completion. N_{gj} denotes the number of researchers of type g in project j.[22] The regression to estimate is the following:

$$outcome_j^* = \tau + \sum_{g=1}^{G} \beta_g N_{gj} + \lambda' C_j + q_j. \tag{4}$$

C_j is a vector of control variables, which include macro-project dummies, and other types of controls (monthly dummies, number of pre-determined teams, number of external firms). The vector of parameters to estimate is given by $\theta_2 = (\tau, \beta, \lambda)$. The coefficient β_g measures how an additional researcher of type g affects the final project outcome, and it can be considered as a proxy for performance. In other words, if the coefficient for a particular type is positive, this means that adding a researcher of that type increases the probability of the project completion. The term q_j is the unexplained component of the outcome and measures the ex-ante unobserved project quality.

4 Empirical Implementation

I estimate the model discussed in section 3 to quantify the main determinants of endogenous project participation, and the project outcome equation to assess the performance of teams and solo researchers. A complication to be tackled in the estimation is the fact that the ex-ante project quality, which is unobserved by the econometrician, affects the decision to join a project. Moreover, q_j influences the final outcome in two ways. First, it enters the outcome directly as a residual; second, it affects the project outcome indirectly through the number of project participants. I could potentially compute the residuals from the out-

[22] In an alternative specification, I allow the outcome to be a quadratic function of N_{gj}.

come equation and use them to estimate the parameters of the structural model. Because of selection, the measure of the residual from the outcome equation is likely to be biased. Hence, I estimate the two empirical components jointly. I express the residuals in terms of the outcome parameters and I use them to estimate the model of project participation.

First, I present the estimation procedure for the game of project participation. Then, I discuss the estimation for the project outcome. Finally, I describe the procedure for the joint estimation.

4.1 Game of Project Participation

I estimate a static game of project participation as an entry game with incomplete information. I abstract from any dynamic consideration in my setting, as discussed in 3. I assume that a player is privately informed about her own idiosyncratic shock and knows only the distribution of other players' shocks. The assumption is realistic if one thinks that a player has a different fit for/to each project, and the fit is not perfectly known by the others. Moreover, the models of incomplete information have an advantage in terms of computational burden.

Entry games with strategic interactions are likely to lead to multiple equilibria, especially in the presence of strategic complementarities. Solutions to this multiplicity problem have been proposed by, among others, Bjorn and Vuong (1984), Bresnahan and Reiss (1991b), Bresnahan and Reiss (1991a) and Berry (1992). Papers on moment inequalities (Ciliberto and Tamer (2009)) allow for general forms of heterogeneity across players providing a methodology for set identification without making equilibrium selection assumptions. However, bounds for the estimated coefficients are likely to give very little information on the type of strategic interactions among players if their ranges are too broad. This is not well suited in this setting given that one of the mail goals is to measure the degree of complementarity and substitutability among researcher types. Alternatively, Schaumans and Verboven (2008), for example, impose assumptions on the sign of the strategic parameter, but in this framework any assumption would appear to be *ad hoc*. Part of the recent literature deals with the multiplicity issue by using a two-step estimation procedure (Aguirregabiria and Mira, 2002, 2007; Bajari et al., 2010), without imposing any further assumptions on the strategic parameter. The method eliminates the need to solve the fixed-point problem when evaluating the corresponding (pseudo) likelihood function that is implied by the structural choice probabilities.[23] I adapt the two-step method to my static framework and, differently

[23](Aguirregabiria and Mira, 2007) have proposed a recursive extension of the two-step pseudo-likelihood estimator.

from the standard literature on entry games, I allow for strategic complementarity and substitutability. In the first step, I estimate the probabilities of entry conditional on project observables.[24] In the second step, I find the structural parameters that are most consistent with the observed data and these estimated equilibrium probabilities. A key assumption for the consistency of this approach is that, in the data, two projects feature the same equilibrium conditional on observables.[25]

4.1.1 Pseudo Log-likelihood Function

Let d_{ij} be the choice of researcher type i of joining or not project j. Moreover, let $\Psi_i = \Phi(\alpha_i X_i + \eta' C_j + \sum_{g=1}^{G} \delta_{ig} p_{gj} + \tilde{q})$, where Ψ_i follows a logistic distribution. The Pseudo-Likelihood Function is the following:

$$Q_J(\theta, \boldsymbol{p}) = \frac{1}{J}\frac{1}{G}\sum_{j=1}^{J}\sum_{g=1}^{G} \log \Psi_i(d_{ij}|X, C, \tilde{q}; \boldsymbol{p}, \theta_1) \tag{5}$$

4.2 Outcome Equation

I estimate the outcome equation using a discrete choice model. For $outcome^*$ being the latent continuous variable for degree of project completion, then

$$outcome = \begin{cases} 0 & \text{if } outcome^* < \tau \\ 1 & \text{otherwise} \end{cases}$$

I assume that the error terms are iid logistically distributed across observations and I set the location and scale parameter equal to 0 and 1, respectively.

The unobserved ex-ante quality is given by the residual of the logit regression. In the case of latent models (probit, logit, ordered probit, etc.) it is not possible to calculate the residuals directly, since the latent dependent variable $outcome^*$ is not observed. I do have an estimate of the conditional distribution of $outcome^*$ conditioned on the observable variables (vector X), based on the specification and the maximum likelihood parameter estimates. From this, I can obtain an estimate of the conditional distribution of the error

[24] Ideally, one should estimate these probabilities non-parametrically. However, the two-step procedure is embedded in a joint maximum likelihood estimation, therefore I estimate the first step parametrically to increase the speed of the estimation.

[25] Several authors have introduced extensions to allow for multiplicity of equilibria when two markets have the same observable characteristics. De Paula and Tang (2012), for instance, propose a test for the signs of state-dependent interaction effects that does not require parametric specifications of players' payoffs, the distributions of their private signals, or the equilibrium selection mechanism.

term q_j, from which I construct the generalized residuals $\tilde{q_j}$ following Gourieroux et al. (1987):

$$\tilde{q_j} = E[q_j|X, \hat{\theta}_2], \tag{6}$$

where $\hat{\theta}_2 = (\hat{\tau}, \hat{\beta}, \hat{\lambda})$ is obtained by maximum likelihood.[26] The residual captures all the unobserved factors that enter the ex-ante project quality. Researchers are likely to sort into projects because of this component. Therefore, sorting creates a problem of endogeneity that biases the results of the estimation.

4.3 Joint Estimation

I need to estimate the probability of completion of a project and the entry probability jointly to overcome the endogeneity issue, with the caveat that some covariates affect contemporaneously the two of them.

I follow Seim (2006), who estimates a model of entry with endogenous product-type choices by computing the joint equilibrium prediction for the location probabilities and the equilibrium number of entrants in a market. I compute the joint prediction for the probability of project completion and the equilibrium number of project participants. In Seim (2006), however, the location decision does not depend on the market-level unobservable, which influences only the probability of entry. Therefore, she is able to obtain the market-level unobservable so that the predicted number of entrants coincides with the observed number in each market. Then, she uses the market-level unobservable to compute the location-choice probabilities.

In this setting, the project-level unobservable q_j affects both the decision to join a project and the project outcome, directly and indirectly through N_j. Therefore, to account for this issue, I express the generalised residual $\tilde{q_j}$ (equations (9) and (10)) as a function of the

[26] Gourieroux et al. (1987) show that the score vector can be expressed in terms of generalised errors. Define the loglikelihood:

$$\ln L = \sum_{j=1}^{J} \log \Psi(outcome_j|N, C; \theta_2). \tag{7}$$

The first order derivative (score function) with respect to the constant (Greene (2003)) produces the generalized residual. For $outcome_j = 0$:

$$\tilde{q_j} = E[q_j|outcome_j = 0, N, C, \hat{\theta}_2] = \frac{-\phi(\hat{\tau} - \hat{\beta}N_j - \hat{\lambda}'C)}{1 - \Phi(\hat{\tau} - \hat{\beta}N_j - \hat{\lambda}'C)} \tag{8}$$

For $outcome_j = 1$:

$$\tilde{q_j} = E[q_j|outcome_j = 1, N, C, \hat{\theta}_2] = \frac{\phi(\hat{\tau} - \hat{\beta}N_j - \hat{\lambda}'C)}{\Phi(\hat{\tau} - \hat{\beta}N_j - \hat{\lambda}'C)} \tag{9}$$

outcome variables and I substitute it into the payoff function. By doing that, I estimate the equilibrium parameters of the model of project participation taking into account the project ex-ante quality (unobserved by the econometrician) and I correctly solve for the endogeneity in the outcome equation. For defined d_{gj} (action of type g for project j), the joint pseudo-likelihood is:

$$f(d, outcome) = \prod_{j=1}^{J}\prod_{i=1}^{I} Pr(d_{ij}|X, C, \tilde{q}; P, \theta_1) \times \prod_{j=1}^{J} Pr(outcome_j|N, C; \theta_2). \quad (10)$$

Equation (10) consists of two parts. The first part computes the likelihood of observing project participation choices conditional on the project-level unobservable $\tilde{q}$. Recall that $\tilde{q}$ is the random factor that affects also the probability of observing a particular outcome realization. Therefore, to derive the unconditional likelihood, the first component of the joint pseudo-likelihood is multiplied by the probability of observing an certain outcome such that predicted and actual probability of project completion are equal. Because of simultaneity, I derive the unconditional likelihood by expressing $\tilde{q}$ as a function of the outcome parameters and regressors and substitute it into the payoff function for the model of project participation. I assume that the error terms of the model of project participation and the outcome equation follow a logistic distribution.[27] The joint pseudo-loglikelihood is the following:

$$LL(\theta) = \frac{1}{J}\frac{1}{G}\sum_{j=1}^{J}\sum_{i=1}^{I} \log \Psi_i(d_j|X, C, \tilde{q}; \boldsymbol{p}, \theta_1) + \frac{1}{J}\sum_{j=1}^{J} \log \Psi(outcome_j|N, C; \theta_2) \quad (11)$$

4.3.1 Two-Step Procedure

In line with the estimation procedure for the model of project participation described in 4.1, I perform the joint estimation in two steps.

1. I maximize the joint loglikelihood without the vector $\boldsymbol{p}$ and obtain the reduced-form estimates of the equilibrium probabilities of entry, together with the estimates of $\theta_1^{Step1}, \theta_2^{Step1}$. In this step, I account for the correlation between the project outcome and the model project participation through $\tilde{q}_j$, but not for the endogenous entry as I do not consider the strategic interactions ($\delta's = 0$).

[27] I restrict the variance covariance matrix of the joint distribution of the error terms to be an identity matrix.

2. With the probabilities predicted in the first step, I construct the joint pseudo-loglikelihood function (11)[28] and obtain the final estimates for $\hat{\theta}_1, \hat{\theta}_2$.

4.4 Identification

The outcome equation is at the project level, whereas the payoff function is at the individual-project level. Some variables are included only in the payoff and do not impact the outcome directly. Indeed, type-specific characteristics affect only the decision of joining a project. The term N_{gj} contained in the outcome equation is the post-equilibrium total number of researchers in a project. The term $E[N_{gj}]$ in the payoff function represents the expectation of the number of potential entrants in a project for each researcher before she takes the decision to join/not join. The two terms are highly correlated. Simulation results show that the identification of the strategic coefficients (δ's) requires variation in the predicted entry probabilities from stage 1 across types. I observe the same set of researchers working both on solo and team projects, where teams are heterogeneous and can have different sizes. The identification strategy of the structural parameters exploit this heterogeneity in team memberships in the data.

5 Results

In this section, I discuss a number of reduced-form preliminary results. Then, I address the estimation of the full model that comprises researchers' participation choices and project outcomes.

5.1 Preliminary Analysis

In this subsection, I show that free-riding is not a concern in this setting. Then, I present reduced-form results in support of the full structural model.

Table 7 reports the results from preliminary OLS regressions of project outcome, where the dependent variable can take values 0 or 1 ("not completed" or "completed", depending on the classification explained in section 2.1). This implementation is useful to understand whether there is a non-linear relation between the outcome and the number of project participants and to estimate the optimal threshold in the non-linear case. Specification (1) includes as covariates the total number of researchers (linear and square); specification (2) includes the previous covariates and the number of pre-determined teams, while specification

[28] For the second step, I initialize the loglikelihood at $\theta_1^{Step1}, \theta_2^{Step1}$.

(3) has the same covariates of (1) and in addition the number of external firms. Specification (4) includes all the covariates previously specified and macro-project dummies. Each specification of table 7 shows that the total number of project participants affects project outcomes positively and significantly. The square term through all the specifications has a negative and significant coefficient; this suggests a concave relationship. Several rationales can underlie this hump-shaped relationship. First, decreasing returns to scale in team production function. In particular, the marginal contribution of an additional researcher of a given type can be decreasing as the improvement on the pre-existing stock of skills already present in the project can shrink. Alternatively, free-riding in teams can imply that, as the number of researchers increases, some researchers can exploit the work of the other teammates. If free-riding plays an important role in this framework, one should expect to see that many projects present a number of researchers exceeding the optimal one, that is determined by the x-coordinate of the vertex of the parabola implied by the estimates of the outcome equation. In all specifications, the vertex of the parabola is significant for projects with more than 5 people. Only a small fraction of projects (around 5%, corresponding to 112 projects) operate with more than 5 participants. Therefore, free-riding does not seem to play a role in this setting.

	OLS **(1)**	**OLS** **(2)**	**OLS** **(3)**	**OLS** **(4)**
# of project participants	0.0559***	0.059***	0.0535***	0.055***
	(0.00955)	(0.015)	(0.015)	(0.015)
# of project participants^2	-0.0049**	-0.0054***	-0.0046**	-0.0053**
	(0.000962)	(0.002)	(0.002)	(0.002)
# of pre-determined teams		0.077***		0.057***
		(0.012)		(0.016)
# of external firms			0.051***	0.011
			(0.02)	(0.02)
Macro-Project dummies	No	No	No	Yes
Threshold	5.7***	5.51***	5.8***	5.2***
	(0.93)	(0.82)	(1.03)	(0.74)

Number of obs: 2,243. All regressions include the constant term. Standard errors in parenthesis.*p<0.10, **p<0.05, ***p<0.01.

Table 7: Project outcome results: OLS

Table 8 displays the results from logit regressions for the model of project outcome (equation (4)).[29] The covariates include the total number of researchers (all the specifications) and the number of researchers squared (specification (3)), the number of pre-determined

[29]I also perform other robustness checks using 3 categories for the outcome. Results from the ordered categorical model are very similar. Same holds for the results of the probit regressions.

teams (all the specifications), the number of external firms (all the specifications except for (1)), a dummy for whether a project is a comment to another project (specification (4)), time dummies (specification (5)) and macro-project dummies (specification (6)). Holding other things fixed, as the number of researchers in a project increases, the project is more likely to be completed. The number of pre-determined teams affects positively and significantly the project outcome, while the number of external firms and the dummy for comments are not significant. The results are similar across the different specifications. I use specification (6) in the full structural model, as this is the one with the best fit according to the AIC selection test.

	Binary Outcome (1)	**Binary Outcome (2)**	**Binary Outcome (3)**	**Binary Outcome (4)**	**Binary Outcome (5)**	**Binary Outcome (6)**
# of project participants	0.93***	0.92***	0.58***	0.5***	0.35***	0.55***
	(0.14)	(0.03)	(0.19)	(0.1)	(0.06)	(0.05)
# of pre-determined teams	2.02***	2***	1.5***	1.5***	1.5***	2.32***
	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)
# of external firms		0.9	0.97	0.9	0.86	0.63
		(0.73)	(0.72)	(0.75)	(0.75)	(0.77)
# of project participants^2			-0.04			
			(0.03)			
Dummy for comments				-0.1		
				(0.3)		
Macro-Project dummies	No	No	No	No	No	Yes
Time dummies	No	No	No	No	Yes	No
LL at convergence	-693	-691	-690	-691	-663	-655

Number of obs: 2,243. All the specifications include a constant. Standard errors in parenthesis. *p<0.10, **p<0.05, ***p<0.01.

Table 8: Project outcome results: Logit

Researchers with different characteristics might affect differently the probability of project completion. Table 9 shows reduced-form results from logit specifications of the outcome equation where the number project participants is in terms of researchers' types. The more the project participants for each type the higher the probability of project completion. These results hold also when controlling for the number of pre-determined teams (specification (2) and (3)) and macro-project dummies (specification (3)). Notice that the effect on the probability of project completion is not randomly distributed across types. I use specification (3) in the full structural model for consistency as this includes all the covariates of specification (6) in table 8.

In these regressions, I do not control for selection, therefore it is not possible to give an economic interpretation to the results as researchers might select into projects with a better ex-ante quality.

	Binary Outcome (1)	**Binary Outcome (2)**	**Binary Outcome (3)**
# of Non-Physics Juniors	-0.751	-0.781	-0.363
	(0.805)	(0.809)	(0.786)
# of Non-Physics Seniors	1.483***	1.394***	0.744***
	(0.099)	(0.098)	(0.116)
# of Physics Juniors	1.039***	0.968***	0.665***
	(0.154)	(0.152)	(0.147)
# of Physics Seniors	0.780***	0.770***	0.402***
	(0.050)	(0.050)	(0.070)
# of pre-determined teams		1.890***	0.624
		(0.509)	(0.565)
Macro-Project dummies	No	No	Yes

Number of obs: 2,243. All regressions include the constant term. Standard errors in parenthesis.*p<0.10, **p<0.05, ***p<0.01.

Table 9: Project Outcome Results: Binary Outcome

Table 10 presents the results from the discrete choice model of project participation that does not include the strategic interactions and the unobserved project-level component (ex-ante quality). In other words, I estimate equation (1) with $\delta_{ig} = 0$ and without controlling for q_j. I assume that the set of potential entrants is random across projects, and has cardinality equal to 10, as 10 is the maximum number of individuals I observe in a project.[30] The dependent variable is equal to 1 if a researcher joins a project and 0 otherwise. As shown by equation (1), the latent payoff of project participation is a function of type and project characteristics. In specification (1), I only control for types' characteristics by including dummies for Non-Physics Juniors and Seniors, and Physics Juniors and Seniors. The reference group is given by non-classified researchers. In this specification, Physics Juniors are more likely to join a project whereas the other types are less likely to do so. Results remain unchanged when controlling for the number of pre-determined teams (specification (2)). In this case, the higher the number of pre-determined teams, the lower the probability of joining (pre-determined teams might be a proxy for project complexity). When controlling for macro-project dummies (specification (3)), results change. In particular, Non-Physics Juniors and Seniors are less likely to join a project whereas people specialized in Physics are more likely to join a project. This can be due to the fact that most of the projects are in the field of physics. I use specification (3) in the full structural model as controlling for macro-project dummies seems to have an impact on the probability of joining a project.

[30] I perform robustness checks also with sets of 8 and 9 random potential entrants. I will perform robustness checks with different sets of potential entrants. One idea would be to determine the set of potential entrants for each project by looking at the empirical distribution of types that joint projects with similar characteristics before.

	Project participation (1)	**Project participation (2)**	**Project participation (3)**
Non-Physics Junior	-1.53***	-1.51***	-0.35***
	(0.025)	(0.027)	(0.056)
Non-Physics Senior	-5.45***	-5.55***	-4.31***
	(0.23)	(0.23)	(0.241)
Physics Junior	0.367***	0.38***	1.63***
	(0.028)	(0.028)	(0.061)
Physics Senior	-0.38***	-0.37***	0.75***
	(0.028)	(0.05)	(0.071)
# pre-determined teams		-0.24***	-0.88***
		(0.06)	(0.074)
Macro-project dummies	No	No	Yes

Number of projects: 2,243. Number of potential participants for each project: 10. All regressions include a constant. Standard errors in parenthesis.*p<0.10, **p<0.05, ***p<0.01.

Table 10: Model of project participation without strategic interactions

The results presented in this section are likely to suffer from endogeneity: researchers can sort into specific projects based on the project ex-ante quality, that is unobserved to the econometrician and correlated with some of the covariates (i.e. the number of project participants as well as macro-project dummies). In the next section, I show the results from the simultaneous estimation of the full structural model in which I account for selection and endogenous participation. Other reduced-form results are discussed in Appendix D.

5.2 Full Structural Model

In this subsection, I present the results from the joint estimation of the full structural model. First, I show the results from a simpler specification in which I do not account for heterogeneity in types. Then, I show the results when I estimate type-specific coefficients.

The results of Table 11 column (1) correspond to specification (6) of Table 8 from the previous section and show the reduced-form results for the probability of project completion. Column (2) corresponds to specification (3) of Table 10 from the previous section and shows reduced-form results from the discrete choice model of project participation without the strategic interactions and without controlling for ex-ante project quality.

Recall the reasons to joint a project: it could be because the project is more likely to end with completion (this is reflected in the quality term) or because an individual cares about who else is joining (this is reflected in the strategic component). In column (3) I allow only for correlation in project quality both in the outcome and in the project participation model by estimating the joint maximum likelihood. Allowing for correlation does not have a big impact on project participation but it changes the estimates of the outcome equation. The

magnitude of the parameter for the number of project participants shows that in column (1) I was overestimating the effect on the probability of project completion, hence I was overestimating the performance of teams.

In column (4), I control for the correlation in project ex-ante quality and endogenous participation (i.e. who an individual gets to work with) by estimating the game of project participation jointly with the outcome equation. First of all, notice that the number of project participants in the outcome equation is a proxy for entry. Indeed, once I control for endogenous entry in the joint estimation, the effect in the outcome equation goes away. More interestingly, there is evidence of selection into team size. When I control for endogenous selection on quality and team size, the coefficient for juniors non-physics turns out to be positive: non-physics juniors are more likely to enter compared to what I find in column (3). This can be explained for instance by the fact that junior researchers want to joint projects to learn from others and to gain experience. At the same time, seniors non-physics are still less likely to enter: they don't seem to obtain any gain from working with the others. The strategic coefficient for the number of expected entrants is negative: researchers dislike working with groups that are too large. This can be explained by the higher costs of coordination and communication that a researcher has to bear when working with larger groups. The existence of coordination costs that increase with team size has been shown to be an important obstacle for collaborative work (Becker and Murphy, 1992).[31] Optimal team size hinges on the trade-off between the benefits of specialization and division of labour and the increased coordination costs (Adams et al., 2005); in this setting, the second component seems to play a bigger role.

To conclude, the main finding is that prospective collaboration with the others mostly drives endogenous project participation. *Ceteris paribus*, the larger the number of project-mates, the lower the probability of joining a working project, because of congestion and increasing coordination and communication costs. Quality impacts project participation but not as much as endogenous entry (captured by team size). Finally, as selection into project is non random, controlling for quality and endogenous project participation matters for obtaining unbiased estimates of team performance.

Selection might depend on the characteristics of the researchers' types. For instance, a Junior researcher might attach more value to joining a project with a higher ex-ante quality than a Senior researcher because the first cares more about her reputation than the latter.

[31] It has been proven that lowering coordination costs can increase the returns to collaborative work. Agrawal and Goldfarb (2008) for instance show that a decrease in collaboration costs through the adoption of Bitnet facilitates increased research collaboration between US universities and the specialization of research tasks.

The results presented in table 12 allow me to explore the effect of heterogeneity on project participation and on the probability of project completion.

The results of column (1) correspond to specification (3) of Table 9; here the outcome depends on the number of project participants of each types. The other covariates are the same as the ones from column (1) of table 11. The more the project participants for each type the higher the probability of project completion. However, by comparing these results with those in column (1) of the previous table, one can see that the effect of the number of project participants on the outcome is not randomly distributed across types. Once again, it is not possible to give an economic interpretation to the results as these do not control for selection.

Column (2) presents the results from estimating the model of project participation without strategic interaction and without controlling for project ex-ante quality. The results are the same as column (2) of table 11.

In column (3), I control for quality both in the outcome equation and in the model of project participation. The coefficients for the number of project participants by types change with respect to column (1). This shows again that there is selection into quality that has to be taken into account when estimating the performance of teams.

In column (4) I look at the combined effect of quality and endogenous project participation on the outcome and on the probability of joining. As expected, the coefficients in the outcome are not significant: the number of project participants by types is a proxy for endogenous participation. When controlling for the further effect of endogenous entry, Non-Physics Juniors are more likely to enter relative to column (3), whereas Physics Juniors and Seniors are less likely to enter (again, relative to column (3)). Additionally, the larger the expected pool of project-mates, the higher the probability of participating for Seniors (both Non-Physics and Physics); vice versa for Juniors.

Table 11 showed that, on average, the larger the pool of participants, the lower the probability of joining a working project. Now, I find that this effect differs across types; this suggests that heterogeneity in researchers' characteristics plays an important role in explaining selection into projects. Controlling for endogenous participation, the probability of joining a project is lower for seniors relative to juniors. More importantly, for seniors, the larger the pool of expected participants, the higher the probability of joining. For juniors this result is flipped. The intuition is the following: if it is true that juniors suffer from implicit costs of coordination and congestion associated with larger groups, senior researchers instead benefit from working with larger groups in expectations, as perhaps they have more expertise in organizing and handling them.

	Project completion (1) Spec. (6) Table 8	**Project participation (2) Spec. (3) Table 10**	**Joint Two-Stage Pseudo Likelihood**	
			Quality, no endogenous participation (3)	**Quality & endogenous participation (4)†**
# of project participants	0.556***		0.137***	-0.037
	(0.055)		(0.062)	(0.07)
# of pre-determined teams	2.32***		3.52***	3.55**
	(0.51)		(0.512)	(0.52)
Type characteristics				
Non-Physics Junior		-0.35***	-0.41***	0.98***
		(0.056)	(0.055)	(0.105)
Non-Physics Senior		-4.31***	-4.37***	-2.96***
		(0.241)	(0.24)	(0.258)
Physics Junior		1.63***	1.62***	2.95***
		(0.061)	(0.062)	(0.106)
Physics Senior		0.75***	0.71***	2.12***
		(0.071)	(0.072)	(0.117)
Project characteristics				
# predetermined teams		-0.88***	-0.84***	-0.86***
		(0.074)	(0.074)	(0.075)
# of expected entrants		-	-	-1.44***
				(0.096)

Number of projects: 2,243. Number of potential participants for each project: 10. All regressions include macro-project dummies and a constant.
†Bootstrap standard errors in parenthesis.

Table 11: Full model

	Project completion (1) Spec. (3) Table 9	**Project participation (2) Spec. (3) Table 10**	**Joint Two-Stage Pseudo Likelihood**	
			Quality, no endogenous participation (3)	**Quality & endogenous participation (4)†**
# of Non-Physics Juniors	0.363 (0.78)		-1.45 (0.449)	-4.27 (3.9)
# of Non-Physics Seniors	0.745*** (0.115)		0.51*** 0.063	-3.67 2.6
# of Physics Juniors	0.665*** (0.146)		0.108** (0.093)	-2.87 (2.1)
# of Physics Seniors	0.401*** (0.07)		0.02 (0.052)	-2.52 (1.98)
# of pre-determined teams	0.62 (0.56)		0.62 (0.55)	0.62 (0.55)
Type characteristics				
Non-Physics Junior		-0.35*** (0.056)	-0.36*** (0.056)	0.99*** (0.106)
Non-Physics Senior		-4.31*** (0.241)	-4.3*** (0.24)	-2.17*** (0.24)
Physics Junior		1.63*** (0.061)	1.67*** (0.061)	-0.9*** (0.101)
Physics Senior		0.75*** (0.071)	0.78*** (0.072)	-1.38*** (0.09)
Project characteristics				
# pre-determined teams		-0.88*** (0.074)	-0.9*** (0.074)	-1.2*** (0.074)
# of expected entrants for Non-Physics Junior		-	-	-0.81*** (0.4)
# of expected entrants for Non-Physics Senior		-	-	2.02*** (0.2)
# of expected entrants for Physics Junior		-	-	-0.83*** (0.08)
# of expected entrants for Physics Senior		-	-	1.16*** (0.09)

Number of projects: 2,243. Number of potential participants for each project: 10. **All regressions include macro-project dummies and a constant**.
†Standard errors in parenthesis.

Table 12: Full model with type-specific coefficients.

6 Counterfactual

I use the results from the structural model to investigate the effect of alternative allocation mechanisms on project participation and project outcome.

In the previous section, I provided evidence of endogenous selection into projects. I have shown that this selection depends mainly on the expected pool of potential project-mates. Moreover, I found that researcher types influence the probability of project completion in an heterogeneous way. A straightforward experiment is a counterfactual scenario where project participants do not take into consideration who else is joining a project when deciding whether or not to join. In other words, let's assume that a manager allows for voluntarily project participation based only on project characteristics (including the project ex-ante quality), without revealing any further information on how the others are selecting into projects.[32]

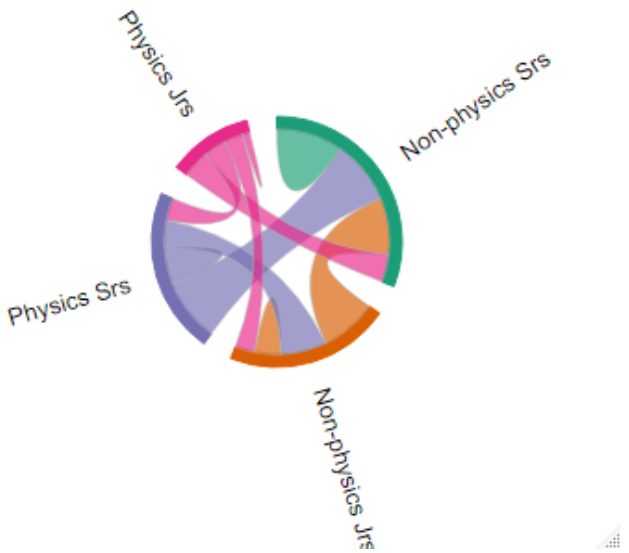
6.1 Project Participation

While the estimation of the structural game of project participation does not require solving for an equilibrium, the implementation of counterfactual experiments typically involves the computation of an equilibrium, or at least an approximation (Aguirregabiria, 2012). The multiplicity of equilibria follows from the presence of strategic complementarity. When I shut down the strategic component of the payoff function, I am able to abstract from this issue in simulating the optimal behavior of the agents.

For the counterfactual experiment, I use the predicted project ex-ante quality and the estimated parameters from the structural model, and I set the strategic interactions parameters (δ's in equation (1)) to zero. I find that, under this counterfactual, individuals join more projects, and this result is stronger for juniors (physics and non-physics). Recall that juniors are less likely to participate the higher the number of expected project-mates. When juniors cannot form these expectations, they do not internalize the eventual costs of coordination and communication deriving from working with larger teams. As a consequence, they are more prone to join a project.

[32] Another implicit assumption is that communication among researchers is not allowed.

Figure 5: Chord diagram for bilateral project connections: counterfactual



	# of Projects
Non-Physics Seniors	1,537
Non-Physics Juniors	1,078
Physics Seniors	1,194
Physics Juniors	520

Table 13: Number of projects by type: counterfactual

Figure 5 shows the results in terms of bilateral project connections. Compared to figure 4 in section 2, under this counterfactual experiment there is more diversity in collaborations as the connections among types are more frequent. For instance, the pink flows that connect Junior Physics to the other types have similar dimensions: this means that Junior Physics cooperate almost equally with all the others.

When the decision of joining a project does not depend on who else is joining, there is more participation and more variety in teams. The next step is to assess the effect of this reallocation on project outcomes.

6.2 Project Outcome

To measure the effect of the alternative mechanism of project allocation on project outcome I use the estimated coefficients of Table 12 column (3); these results correspond

to the first step of the joint estimation, in which I correct for selection into quality.[33]

I find that the counterfactual percentage of completed projects is 6% lower than that observed in the data. Therefore, shutting down the strategic interaction in the decision to join a working project leads to excessive participation, which affects negatively the probability of project completion; team variety does not alleviate this effect.

I show that under this hypothetical scenario more team diversity is achieved. Diversity is often claimed to be a crucial condition for radical innovation (Nelson and Winter, 1982; Singh and Fleming, 2010). It has been shown that structurally diverse teams are more likely to produce breakthroughs (Guimera et al., 2005; Jones et al., 2008). According to Banal-Estañol et al. (2019), teams that exhibit greater diversity in knowledge and skills, education, and/or scientific ability are generally more likely to be successful. In contrast, I find that under this counterfactual scenario higher efficiency of project outcome is not achieved due to the fact that there is excessive project participation. This seems to suggest that it is more efficient to let researchers decide their teams also based on who else is potentially joining, as they optimally internalize the costs and benefits of working with others.

The results from this counterfactual experiment provide a concrete measure of the important role played by the strategic motives in the allocation of researchers to projects and for the probability of project completion.

7 Conclusion

This paper develops and estimates an empirical structural model that quantifies the main drivers of endogenous team formation and team performance when the allocation of individuals to working projects is decentralized. The empirical analysis relies on novel data from an important scientific experiment, which represents an ideal setting to study the decentralized allocation of individuals to projects. In particular, the decision to join a project is mostly driven by two forces: on the one hand, a researcher can sort into a project because of the prospect of collaborating with other project-mates; on the other hand, a researcher can join a project because of its ex-ante better quality. Controlling for individual and project characteristics, I disentangle the role played by these two determinants on the

[33]Ideally, I should use the results from the joint structural model in which I control for quality and endogenous participation, but the coefficients of the outcome equation as expected turn out to be non-significant, and I cannot make any inference from the results. By using the estimates from column (3) I do not control for the indirect effect of quality (namely, the selection based on the expected number of project-mates whose decision in turn depends on the project ex-ante quality). However, I am still able to quantify at least partially the efficiency gain or loss when moving from a decentralized to an alternative mechanism of project participation.

researcher's decision to join a project. I also show how ignoring either force can lead to biased estimates in a decentralized framework of allocation of workers to projects and, hence, to incorrect conclusions regarding team performance. My main finding is that prospective collaboration is the most important driver of whether to join a particular project. *Ceteris paribus*, the larger the number of project-mates, the lower the probability of joining a working project on average, as a consequence of the congestion or increasing coordination and communication costs. However, heterogeneity in researchers' characteristics plays an important role in explaining selection into projects: for example, senior workers are more likely join projects of expected larger size (as measured by the number of project-mates) relative to junior workers. Finally, to assess the role of strategic collaboration, I consider a counterfactual centralized mechanism in which this channel is shut down. When doing so, I show how this leads to excessive entry and generates inefficiency in terms of project outcomes. My results suggest that adopting a decentralized mechanism of task allocation within a firm can be more efficient because workers internalize the costs and benefits of working with other project-mates.

So far, the empirical literature has focused on working collaborations characterized by exogenous team formation. However, this paper suggests that analyzing the effect of endogenous team formation is crucial to study the problem of efficient allocation of resources within a working organization. This aspect has become increasingly relevant since many institutions are moving from a centralized allocations of workers to projects to a (partly) decentralized one. This paper provides insights on the economic consequences of decentralization for an efficient allocation of resources. Ignoring the factors that drive endogenous team formation may result in incorrect conclusions regarding the efficiency of decentralized mechanisms of project participation. An interesting follow-up would be to test different allocation algorithms in order to find the one that achieves the highest possible outcome in terms of efficiency.

The estimation procedure I have proposed in this paper could be adapted to study the mechanisms behind endogenous alliances and partnerships. One example in industrial organization is R&D joint ventures. One could potentially analyze the consequences of policy restrictions targeted to joint ventures participants.

This paper leaves some aspects for future investigation. One concern is that there can be constraints affecting the decision to participate to a project, such as time or availability constraints. I control for them by including type-specific dummies in the model of project participation. However, if these constraints are researcher's specific or time variant, then this can represent an issue because I do not explicitly model these constraints in the decision

of joining working projects. Likewise, I do not consider potential spillovers among projects: when a researcher works on a project that is not successful, she can possibly adjust her expectation regarding the outcome of a correlated project. Moreover, I also assume that the researcher's investment of time and expertise is strictly project-specific, where in a real world setting some knowledge and skills can be transferable across projects. These are important topics for future research. Despite these assumptions, the paper moves a first step forward the analysis of endogenous team formation by proposing a tractable framework and using a novel source of data. Exploring the above additional questions can shed a light on our comprehension of team formation and allows us to understand why no man is an island.

References

Adams, J. D., Black, G. C., Clemmons, J. R., and Stephan, P. E. (2005). Scientific teams and institutional collaborations: Evidence from us universities, 1981–1999. *Research policy*, 34(3):259–285.

Agrawal, A. and Goldfarb, A. (2008). Restructuring research: Communication costs and the democratization of university innovation. *American Economic Review*, 98(4):1578–90.

Aguirregabiria, V. (2012). A method for implementing counterfactual experiments in models with multiple equilibria. *Economics Letters*, 114(2):190–194.

Aguirregabiria, V. and Mira, P. (2002). Swapping the nested fixed point algorithm: A class of estimators for discrete markov decision models. *Econometrica*, 70(4):1519–1543.

Aguirregabiria, V. and Mira, P. (2007). Sequential estimation of dynamic discrete games. *Econometrica*, 75(1):1–53.

Aguirregabiria, V. and Suzuki, J. (2015). Empirical games of market entry and spatial competition in retail industries.

Akcigit, U., Caicedo, S., Miguelez, E., Stantcheva, S., and Sterzi, V. (2018). Dancing with the stars: innovation through interactions. Technical report, National Bureau of Economic Research.

Bajari, P., Hong, H., Krainer, J., and Nekipelov, D. (2010). Estimating static models of strategic interactions. *Journal of Business & Economic Statistics*, 28(4):469–482.

Banal-Estañol, A., Macho-Stadler, I., and Pérez-Castrillo, D. (2019). Evaluation in research funding agencies: Are structurally diverse teams biased against? *Research Policy*, 48(7):1823–1840.

Bandiera, O., Barankay, I., and Rasul, I. (2010). Social incentives in the workplace. *The Review of Economic Studies*, 77(2):417–458.

Becker, G. S. and Murphy, K. M. (1992). The division of labor, coordination costs, and knowledge. *The Quarterly Journal of Economics*, 107(4):1137–1160.

Berry, S. T. (1992). Estimation of a model of entry in the airline industry. *Econometrica: Journal of the Econometric Society*, pages 889–917.

Bjorn, P. A. and Vuong, Q. H. (1984). Simultaneous equations models for dummy endogenous variables: a game theoretic formulation with an application to labor force participation.

Bolton, P., Dewatripont, M., et al. (2005). *Contract theory*.

Bresnahan, T. F. and Reiss, P. C. (1991a). Empirical models of discrete games. *Journal of Econometrics*, 48(1-2):57–81.

Bresnahan, T. F. and Reiss, P. C. (1991b). Entry and competition in concentrated markets. *Journal of political economy*, 99(5):977–1009.

Ciliberto, F. and Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6):1791–1828.

De Paula, A. and Tang, X. (2012). Inference of signs of interaction effects in simultaneous games with incomplete information. *Econometrica*, 80(1):143–172.

Falk, A. and Ichino, A. (2006). Clean evidence on peer effects. *Journal of Labor Economics*, 24(1):39–57.

Ganglmair, B., Simcoe, T., and Tarantino, E. (2018). Learning when to quit: An empirical model of experimentation. Technical report, National Bureau of Economic Research.

Gourieroux, C., Monfort, A., Renault, E., and Trognon, A. (1987). Generalised residuals. *Journal of econometrics*, 34(1-2):5–32.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Guimera, R., Uzzi, B., Spiro, J., and Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702.

Hamilton, B. H., Nickerson, J. A., and Owan, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of political Economy*, 111(3):465–497.

Hitsch, G. J., Hortaçsu, A., and Ariely, D. (2010). What makes you click?—mate preferences in online dating. *Quantitative marketing and Economics*, 8(4):393–427.

Holmstrom, B. (1982). Moral hazard in teams. *The Bell Journal of Economics*, pages 324–340.

Jones, B. F., Wuchty, S., and Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *science*, 322(5905):1259–1262.

Lazear, E. P. (1998). *Personnel economics for managers*. Wiley New York.

Lindquist, M. J., Sauermann, J., and Zenou, Y. (2015). Network effects on worker productivity.

Mas, A. and Moretti, E. (2009). Peers at work. *The American Economic Review*, 99(1):112–145.

McAlpine, H., Hicks, B. J., Huet, G., and Culley, S. J. (2006). An investigation into the use and content of the engineer's logbook. *Design Studies*, 27(4):481–504.

Nelson, R. R. and Winter, S. G. (1982). The schumpeterian tradeoff revisited. *The American Economic Review*, 72(1):114–132.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of economic literature*, 37(1):7–63.

Schaumans, C. and Verboven, F. (2008). Entry and regulation: evidence from health care professions. *The Rand journal of economics*, 39(4):949–972.

Seim, K. (2006). An empirical model of firm entry with endogenous product-type choices. *The RAND Journal of Economics*, 37(3):619–640.

Singh, J. and Fleming, L. (2010). Lone inventors as sources of breakthroughs: Myth or reality? *Management science*, 56(1):41–56.

Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039.

A Examples of Projects

In project 1 (figure 6), researchers align two mirrors on a lab desk so that a laser can pass through the lens. In project 2 (figure 7), researchers analyze data collected from a measurement experiment.

Figure 6: Project 1

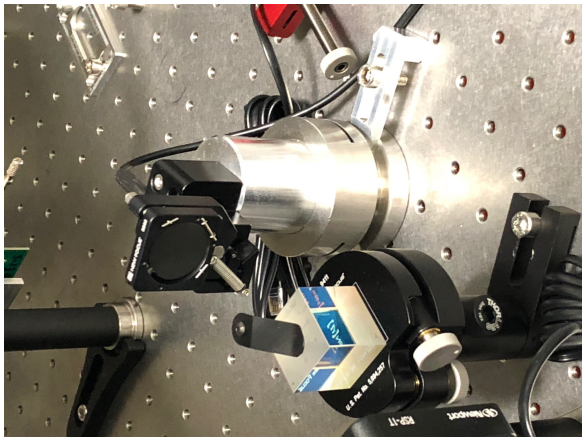
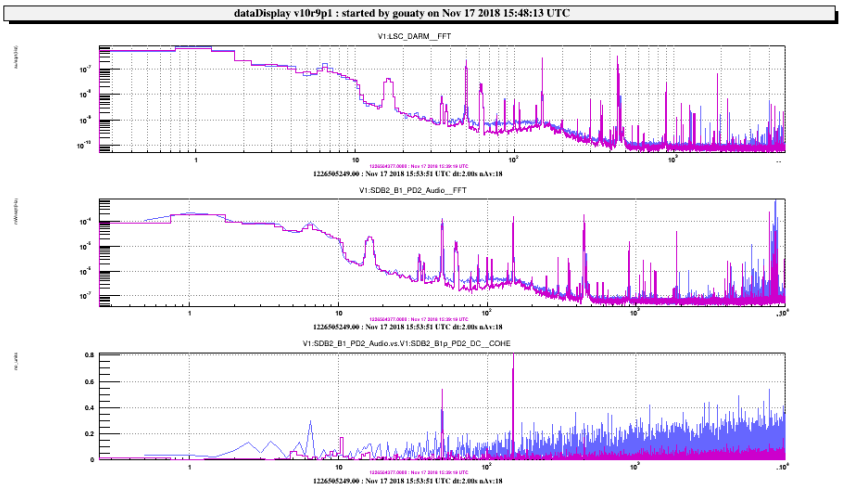


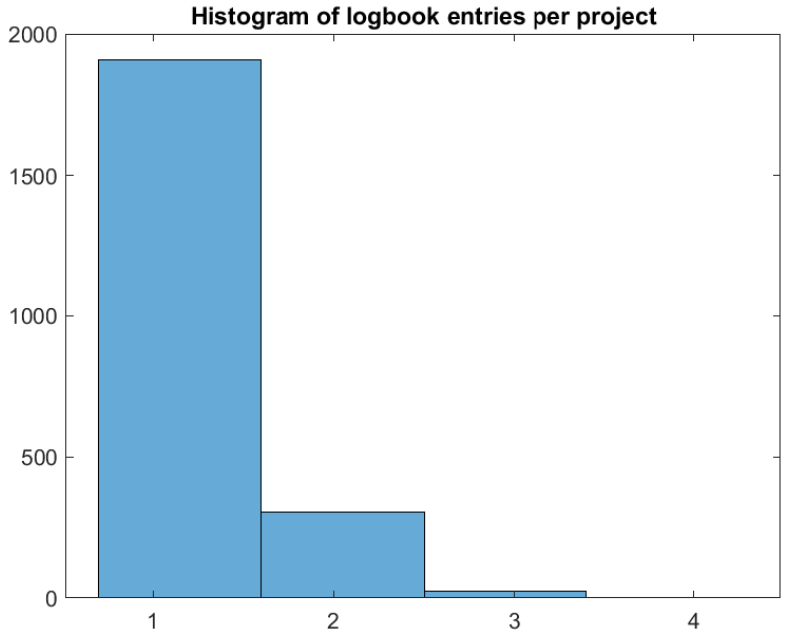
Figure 7: Project 2



B Other Descriptive Statistics

The following histogram shows the distribution of Logbook entries per project. Most of the projects have only one entry.

Figure 8: Logbook Entries Distribution



The following plots shows the distributions of project participation for Non-Physics Seniors (figure 9) and Physics Seniors (figure 10). The two distributions look very similar.

Figure 9: Distribution of Non-Physics Seniors

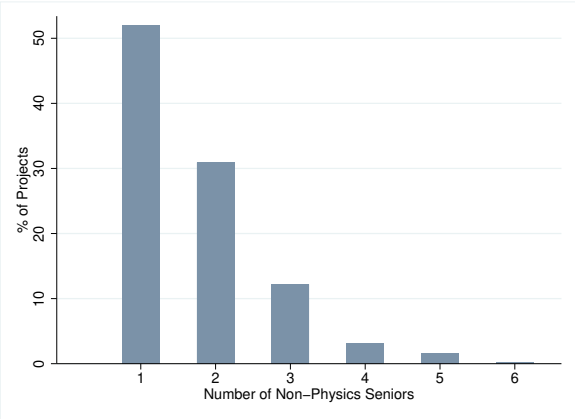
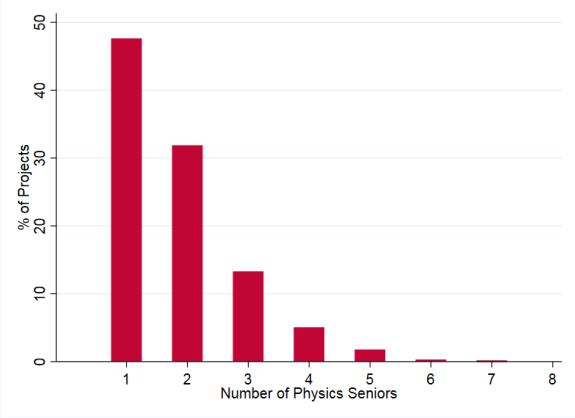


Figure 10: Distribution of Physics Seniors



C Examples of Outcome Classifications

Project with classification 0: "Looking at $NARM\_LOCK\_state$ it seems that the lock could hold until around 10 UTC this morning. From that time a series of relocks attempts (with lock periods of different duration) has triggered until around 14:20 UTC were the lock could not be achieved anymore [...]."

Project with classification 1: "As foreseen after the completion of Long Towers scaffolding [...] also the DET Tower has been equipped with a Frigerio Style scaffolding. The installation could be completed, yesterday, in a single day [...]."

	Predicted probabilities	**Predicted probabilities**	**Predicted probabilities**	**Predicted probabilities**
		In absence of Physics Seniors	In absence of Physics Juniors	In absence of Non-Physics Juniors
Zero Non-Physics Senior	0.528***	0.423***	0.488***	0.53***
	(0.0085)	(0.017)	(0.01)	(0.0085)
One Non-Physics Senior	0.245***	0.29***	0.25***	0.24***
	(0.008)	(0.016)	(0.01)	(0.008)
Two Non-Physics Seniors	0.145***	0.19***	0.16***	0.144***
	(0.007)	(0.014)	(0.0084)	(0.007)
Three Non-Physics Seniors	0.057***	0.0566***	0.069***	0.057***
	(0.0047)	(0.007)	(0.006)	(0.004)
Four Non-Physics Seniors	0.014***	0.0167***	0.017***	0.014***
	(0.0025)	(0.0043)	(0.003)	(0.0025)
Five Non-Physics Seniors	0.0075***	0.0063***	0.008***	0.0075***
	(0.0018)	(0.0022)	(0.002)	(0.0018)
Six Non-Physics Seniors	0.0008	0.0003	0.001	0.0008
	(0.0006)	(0.0004)	(0.0008)	(0.0006)

Predicted margins calculated from multinomial regressions. All the regressions include macro-project dummies and a constant.
Standard errors in parenthesis.

Table 14: Predicted probabilities of project participation for Non-Physics Seniors

D Other Reduced-form Results

Table 14 and 15 contain additional reduced-form evidence that shows clear paths in the connections among different types even controlling for projects' characteristics.

In particular, table 14 shows the predicted probabilities from a multinomial logit regression where the dependent variable takes value 0 if there are no Non-Physics Senior in a project, 1 if one Non-Physics Senior joins, 2 if two Non-Physics Seniors join and so on and so forth. All the regressions include macro-tasks dummies and a constant. Column (1) includes as regressors the number of project participants of all other types. In column (2), I predict the probabilities of being in a project in absence of Physics Seniors. One can see that the probability of not having Non-Physics Seniors decreases from 53% in column (1) to 42% in column (2): this means that in absence of Seniors in Physics it is more likely that there will be one or more Non-Physics Senior in the project. Hence, the finding suggests substitutability among seniors. Column (3) shows that in absence of Physics Juniors it is more likely that a Non-Physics Senior is in a project, as the predicted probability of not having Non-Physics Seniors changes to 49%, but the effect is milder than before. The results of column (4) do not change from those in column (1) as there are few Non-Physics Juniors in the sample.

In table 15, I show the predicted probabilities of being in a project for Physics Seniors, in the same spirit of the previous table. The average predicted probabilities in column (1) are higher than those from the previous table: it is more likely that one or more Physics Seniors are in a project relative to Non-Physics Seniors. Again, the results in column (2) suggest substitutability among seniors: indeed, for a Physics Senior, the probability of being in a project in absence of a Non-Physics Senior is 81%, which is higher than the average probability reported in column (1) (74%). The predicted probabilities in column (3) do not differ from those of column (1), meaning that there is no reduced-form evidence for complementarity/substitutability between researchers in Physics. Again, the results of column (4) do not change from those in column (1) as there are few Non-Physics Juniors in the sample.

	Predicted probabilities	**Predicted probabilities**	**Predicted probabilities**	**Predicted probabilities**
		In absence of Non-Physics Seniors	In absence of Physics Juniors	In absence of Non-Physics Juniors
Zero Physics Senior	0.265***	0.19***	0.265***	0.265***
	(0.007)	(0.01)	(0.008)	(0.007)
One Physics Senior	0.35***	0.39	0.35***	0.35***
	(0.009)	(0.34)	(0.01)	(0.009)
Two Physics Seniors	0.23***	0.25	0.23***	0.23***
	(0.008)	(0.26)	(0.009)	(0.008)
Three Physics Seniors	0.09***	0.10	0.09***	0.09***
	(0.006)	(0.54)	(0.006)	(0.006)
Four Physics Seniors	0.037***	0.03	0.033***	0.03***
	(0.003)	(0.3)	(0.004)	(0.003)
Five Physics Seniors	0.013***	0.01***	0.01***	0.013***
	(0.0023)	(0.0025)	(0.002)	(0.0023)
Six Physics Seniors	0.002**	0.002**	0.004**	0.002**
	(0.0009)	(0.001)	(0.001)	(0.0009)
Seven Physics Seniors	0.0009	0.001	0.001	0.0009
	(0.0006)	(1.46)	(0.001)	(0.0006)

Predicted margins calculated from multinomial regressions. All the regressions include macro-project dummies and a constant.
Standard errors in parenthesis.

Table 15: Predicted probabilities of project participation for Physics Seniors

E Robustness Checks

An alternative specification for the outcome equation is the following:

$$outcome_j^* = \tau + \sum_{g=1}^{G} \beta_g \mathbf{1}_{gj} + \sum_{g=1}^{G} \sum_{i=1}^{G\setminus\{g\}} \psi_{gi} \mathbf{1}_{gj} \mathbf{1}_{ij} + \lambda' C_j + q_j \tag{12}$$

In this specification, the types enter as dummies. Moreover, I introduce interaction terms to capture spillovers and complementarities across different researchers' types. $\mathbf{1}_{ij}$ is a dummy equal to 1 if at least one researcher of type $i \neq g$ is in project and 0 otherwise. Results from this specification are presented in table 16. Adding the interaction terms to capture spillovers across types does not leave enough explanatory power for identification for none of the specifications.

	Binary Outcome (1)	Binary Outcome (2)	Binary Outcome (3)
Dummy non-physics senior	0.89	0.79	0.69
	(1.35)	(1.35)	(1.33)
Dummy non-physics junior	0.31	0.29	0.45
	(0.78)	(0.78)	(0.81)
Interaction non-physics senior/non-physics junior	0.01	0.05	0.08
	(1.61)	(1.6)	(1.61)
Dummy physics senior	0.12	-0.06	-0.36
	(1.36)	(1.36)	(1.33)
Interaction non-physics senior/physics senior	0.09	0.19	0.23
	(1.34)	(1.34)	(1.31)
Dummy physics junior	1.66	1.62	1.45
	(1.16)	(1.16)	(1.13)
Interaction non-physics senior/physics junior	-0.77	-0.77	-0.73
	(0.54)	(0.54)	(0.55)
Interaction non-physics junior/physics junior	-2.84*	-2.86**	-2.94*
	(1.7)	(1.68)	(1.7)
Interaction physics senior/physics junior	-0.95	-0.91	-0.96
	(1.14)	(1.14)	(1.1)
# pre-determined teams	1.21**		0.52
	(0.48)		(0.49)
Macro-project dummies	No	No	Yes

Number of obs: 2,243. St. errors in parentheses. All regressions include the constant. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 16: Project Outcome with Interactions