Essays in Empirical Industrial Organization

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

Alessandra Allocca

Frühjahrssemester 2020

Abteilungssprecher: Prof. Volker Nocke, Ph.D.
Referent: Prof. Michelle Sovinsky, Ph.D.
Korreferent: Prof. Laura Grigolon, Ph.D.

Tag der mündlichen Prufung: 13. August 2020

A Paolo.

Contents

Acknowledgements

This dissertation is the result of a long life-changing journey that I made with the precious support of many travelers.

I am immensely indebted to my supervisor Michelle Sovinsky. Since I met Michelle on the first Empirical IO class, I immediately knew she would have been the perfect guide for this journey. I thank Michelle for her enchanting ability to see far where none is able to see, for showing me the real beauty of doing research, and for teaching me how to face the hardest challenges with her contagious smile. I am equally indebted to *Professor Dori* for her never-ending optimism, even in the darkest moments when my *Nemo* seemed lost forever.

I am deeply grateful to my advisors Laura Grigolon and Emanuele Tarantino. I thank Laura for her constant patience and care, and for the infinite support as a real friend. She has taken my hand and guided me through the intricate meanders of the academic world with the gentleness of a pure soul. The time spent together has been invaluable to me. I strongly believe the fate decided to connect me and Emanuele the very first day in Mannheim. I thank him for strongly believing in my research since the beginning, for being so amazingly stubborn when I was about to give up, for being there with his sharp advice anytime, unconditionally.

I would like to express my gratitude to Liana Jacobi for the insightful discussions that have enriched me as a researcher. Her strength and passion have been a true inspiration. I also thank Bernhard Ganglmair, whose enthusiasm and energy proved to me that innovation is not just a field of research but also a state of mind.

My gratitude goes to the Micro group of the University of Mannheim (the third floor) which provided me with an intellectually stimulating and supportive environment. In particular, I thank Volker Nocke, Martin Peitz, Helena Perrone, Nico Schutz, Hidenori Takahashi, and Thomas Tröger. It has been a great honor to be part of this group.

I thank the GESS-CDSE and the administration of the Economics Department of the University of Mannheim, in particular Marion Lehnert and Nadine Scherer. My life as a Ph.D. student has been easier with their constant support. I also thank the Collaborative Research Center Transregio 224 for having supported my research, and the EGO/Virgo collaboration for allowing me to use the data for the first chapter of this dissertation.

I began this long journey with Can, Lukas, Mirac, Matthew, Sebastian, and Franziska, with whom I shared fears and rewards. Along the way, I met many more extraordinary friends: Esteban, with whom I share a similar passionate approach to life; he was like a elder brother, a really smart one; Katia, who has always been able to spread her bright energy all over the place, and Niccolò, with whom I had some of the most vivid and histrionic discussions of my Ph.D.; I particularly thank him for showing me how extremely fascinating rigor in research can be. I also thank Cristina, Eleftheria and Harim, for the cheerful unforgettable moments spent together, and Andrè, Anton, Daniel, Enrico, Robert and Ruben, for constantly taking care of me.

I would also like to thank Veronica and Vincenzo, whose living example taught me to never give up on my dreams, and Laura (again) and Hendrik, for making me feel the warmth of a true family in Mannheim together with the little Marcel, who filled my darkest days with his overwhelming sweetness.

These years would have been even tougher without my beautiful ladies, Yasemin and Yihan. I thank them for playing the role of the eldest sisters even though I was supposed to be the most mature. Many times, knowing that I would have met with them in the office was my only source of joy. We grew up together and became an unbreakable trio, I was so lucky to have them by my side.

My career as an economic student began at the University of Naples Federico II. I would like to express my gratitude to the members of the Economics Department (just to mention a few: Riccardo Martina, Marco Pagano, Marco Pagnozzi, and Saverio Simonelli) who have been extremely supportive throughout my Ph.D. and represented an inspiration to pursue a career in academia. At the University of Naples I also met many brilliant friends. I thank Alessandra, for our intimate confessions and for understanding me deeply inside, better than anyone else; Anna, for her delicate presence and for making me laugh in a lovely way about my moments of despair, and Antonio, who has proved with his adorable and acute optimism to be my biggest fan.

A special thank to my lifetime friends, the distance has never represented a limit for them: Ludovica, for painting my days with vivid shiny colors and for loving me unconditionally as I am; Pino, for having kept me alive and kicking throughout these years with his brilliant challenges and for his gifted touch of sensitivity, and Francesca, whose perseverance in life has been an example for me. I thank Francesco for encouraging me with his vibrant *mentalità* and for being the perfect shoulder to lean upon, and Evelina and Mirko, for their mindset of true globetrotters. We will use many more brushes together.

My family has been extremely important during these years. Being unable to say the last goodbye to Nonna Rosa was a heart-wrenching moment for me, but I am sure she is constantly looking down on me. I thank zio Michele and zia Giuseppina for their coming-home parties with delicious pizzas, zio Tonino and zia Angela for their constant affection, and my cousins Rosalba, Raffaele, Vincenzo and Giuseppe, Diamante and Lina.

I am deeply indebted to mamma e papà, the milestones of my life. I thank them for their immense patience, countless sacrifices and unconditional support. I am aware these years were especially hard for them, and this has made their dedication to me even more precious. I thank mamma for being my greatest and most powerful strength and papà for opening-up my eyes to the world and beyond. I thank the brightest star of my universe, Turi, whose passion and curiosity have been the spices of my intellectual path, and whose sweetness has been a safe harbor during the storms. Not by chance, the first chapter of this dissertation has been developed after a discussion with her.

Finally, I would like to thank my life companion Paolo. You are my biggest inspiration, Paolo, and I have never been able to take a step without you by my side. I have seen through your pupils, written with your hands, fought with your braveness, lived with your passionate heart. Without you, this journey would have not existed. You are simply perfect for me.

List of Figures

List of Tables

1. Introduction

The field of empirical industrial organization uses data to analyze the structure of industries in the economy by measuring the parameters that drive the behaviors of firms and consumers in these industries.

Part of the literature focuses on markets in which firms interact in an imperfectly competitive setting. Research in this field heavily relies on models with a game-theoretic foundation. Starting from the seminal work of Bresnahan and Reiss (1991b), many market structure models endogenize the number of firms entering a market. Not only industries but also other types of organizations operate through interactions among their members and can be analyzed with analogous game-theoretic models. The underlying hypothesis is that agents making a certain decision receive a non-negative payoff, conditional on the expectations or actions of other (potential and actual) agents acting in the same environment.

These considerations have been crucial in shaping the first two self-contained chapters of this dissertation. In the first chapter, I study the workers' decisions of joining teams within an important scientific experiment. In the second chapter, joint with Laura Grigolon, we provide empirical evidence of the link between common ownership and firms' decisions of entering markets in the Ontarian cancer drug industry.

The fact that decision-makers operating in a strategic environment have in expectation non-negative payoffs is parallel to revealed preference arguments at the basis of discrete choice models of consumer behavior. As in the market entry literature, consumers' choices are interpreted as revealing something about an underlying latent utility. By observing how consumers' decisions change, as their choice sets and market conditions change, one can gain insight into the underlying determinants of consumers' preferences. In the third chapter, joint with Liana Jacobi and Michelle Sovinsky, we analyze the potential complementaritarities in consumption of the so-called *sin goods* (marijuana, alcohol and tobacco) taking into account persistence in behavior.

For the development of this dissertation, I rely on rigorous descriptive analyses and the development and estimation of structural models. With these approaches, it is possible to give informed assessments to policy-makers and in the case of structural models to quantify the impact of feasible policy changes.

Chapter 1. In Chapter 1, I present an empirical structural model that quantifies the main drivers of endogenous team formation and team performance when the allocation of individuals to teams is decentralized.

Many companies currently adopt decentralized approaches to production. These arrangements are widespread in scientific institutions, as fellow researchers typically collaborate on a voluntary basis. The emergence of such arrangements poses several challenges to an economist. First, it is important to understand which elements drive the decision to join projects. Second, it becomes critical to develop tools to correctly measure the performance of teams when the decision to participate in projects is endogenous. These steps are fundamental to assess if decentralization is desirable to obtain successful outcomes with a larger probability.

To address these challenges, I use unique data from Virgo, an international collaborative experiment in science. Researchers involved in Virgo choose which project(s) to work on. For the analysis, I use the information on projects' characteristics, outcomes, and participants.

I develop and estimate an entry game with incomplete information where heterogeneous agents decide simultaneously whether to join a project (*à la* Aguirregabiria and Mira, 2007). The payoff of joining depends on exogenous project characteristics, including a measure of ex-ante quality, and the expectation on the actions of potential project-mates. Strategic complementarities and substitutabilities can arise in this setting as workers might find beneficial or detrimental to work with others. I measure project outcome in terms of probability of project completion.

I find that the pool of expected project-mates drives the decision to join a project while project quality is less important. The larger the pool, the lower the probability of joining a project, as a consequence of the congestion effect due to increasing coordination and communication costs. Heterogeneity in researchers' characteristics explains the selection into projects. I show that controlling for both projects' ex-ante quality and

endogenous project participation matters for obtaining unbiased estimates of teams' performance.

Finally, I consider a counterfactual centralized mechanism in which strategic interactions have no value. I find that this alternative allocation leads to excessive project participation and decreases the probability of project completion. Hence, adopting a decentralized mechanism of project allocation within a firm can be more efficient because workers internalize the costs and benefits of working with each other.

Chapter 2. In Chapter 2, joint with Laura Grigolon, we document the features of a highly innovative industry characterized by a concentrated ownership structure, the Ontarian cancer drug industry. The analysis has the objective of studying the effect of common ownership on the decision of generic producers to enter the market.

Common ownership, namely the practice for large institutional investors of owning stakes in competing firms, has raised the attention of antitrust scholars because the degree of common ownership grew in recent years. Some empirical studies show that it has a large effect on the strategic behavior of companies held by institutional shareholders. Common ownership linkages are a well-established feature of many industries, including the cancer drug industry, for which hospital and public drug program spending in Ontario is dramatically increasing over time. These factors make it an appealing setting to understand the consequences of the common ownership phenomenon.

We use unique data on the timing of cancer drug entry in the market (branded and generics) and collect information on patents, drug approvals, and drug indications. We complement our dataset by gathering ownership data mainly from 13F filings. With these data, we empirically assess the presence of common ownership and quantify which components mainly drive the link between common ownership and market entry. In particular, we show that investors' concentration plays an important role in defining common ownership in the years before the entry of a generic in the market.

Common ownership may have anticompetitive effects and be harmful to welfare. With the results of this paper, we make the first important step in identifying the target of eventual policy interventions to reduce this practice, for this industry as well as for other innovative industries characterized by a high level of concentration.

Chapter 3. In Chapter 3, joint with Liana Jacobi and Michelle Sovinsky, we analyze the potential complementarities in use when individuals choose to consume bundles containing marijuana, alcohol, or cigarettes (*sin goods*), taking into account persistence in consumption for these substances.

Two-thirds of Americans are in favor of marijuana legalization. This substance, however, might be consumed in combination with other substances, such as alcohol and tobacco. Moreover, past use of one of the substances might have consequences for the consumption of that substance and other *sin goods*, especially if one considers complementarity in consumption. Therefore, it is important to understand whether consuming marijuana affects the consumption of other substances and what changes when one considers the potentially addictive nature of these products.

We develop and estimate a dynamic model of multi-substance use allowing for persistence in behavior. For the empirical analysis, we uniquely combine data from two primary sources. The first are individual-level panel data from the Panel Study of Income Dynamics (PSID) survey, which contains information on demographics and consumption behaviors of young adolescents in the US. The second source are pricing data for marijuana, alcohol and cigarettes collected from administrative tax data and transaction data.

Our parameter estimates show that it is important to account for correlation across unobservables and persistence in behavior when analyzing the decision of using the *sin goods* in combination. Moreover, we find that the past use of a substance influences not only its current use but also the decision of using the substance together with other substances. Our results provide insightful information on the long-run effect of marijuana legalization for the concurrent and future consumption of potentially substitutable products.

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

> No man is an Iland, intire of itselfe;
> every man is a peece of the
> Continent, a part of the maine [...].
>
> John Donne - 1624

2.1. 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 workflows 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 workflows 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

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

which researchers often collaborate voluntarily, adopt similar arrangements (Guimera, Uzzi, Spiro, and Amaral, 2005; Wuchty, Jones, and Uzzi, 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.

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

[3] See (Bolton, Dewatripont, 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, Caicedo, Miguelez, Stantcheva, and Sterzi, 2018) and in technical standards development (Ganglmair, Simcoe, and Tarantino, 2018).

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 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 (Bandiera, Barankay, and Rasul, 2010; Falk and Ichino, 2006; Lindquist, Sauermann, and Zenou, 2015; Mas and Moretti, 2009, 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 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.

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

2.2. 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: 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

[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.".

[10] A detailed description of the macro-projects is available upon request.

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.

2.2.1. The Logbook

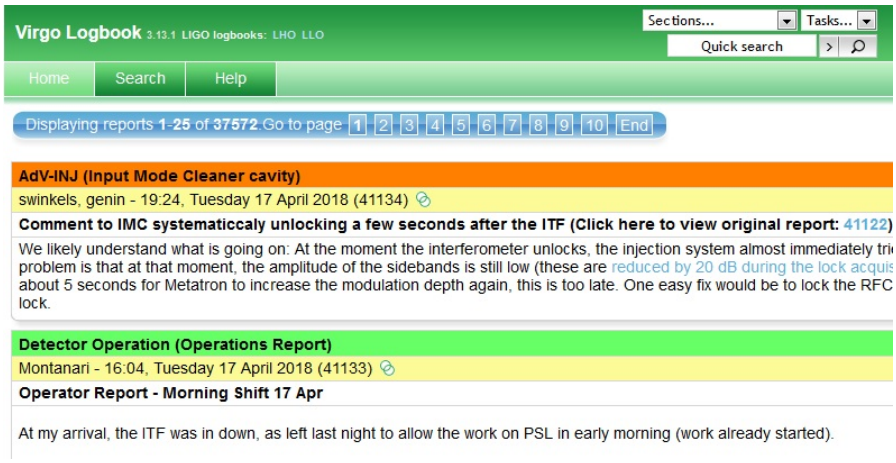
Researchers in Virgo communicate using an on-line platform: the Logbook,[12] which consists of web-pages held by project teams. With the advent of new electronic notebooks it has become possible to incorporate valuable information into enterprise-wide information management systems (McAlpine, Hicks, Huet, and Culley, 2006). The log-

[11] The agreements may differ among Countries.
[12] Source: `https://tds.ego-gw.it/itf/osl_virgo/index.php`

book 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, Hortaçsu, and Ariely, 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 89 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 2.1. I provide two examples of projects in Appendix A.

Figure 2.1.: Example of a Logbook Page



Notes: 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 2.1 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 2.1.: Descriptive Statistics

Figure 2.2 shows the distribution of project participants. Projects with two participants are the most frequent, followed by solo projects.

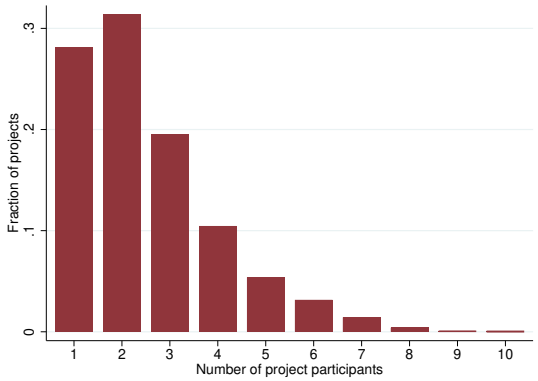


Figure 2.2.: Distribution of Project Participants

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 2.2 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 2.2.: 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.

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 C.

2.2.2. 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 online sources, mainly personal websites, available *curricula* and LinkedIn profiles. Re-

[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.

searchers 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 2.3 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 2.3.: Table of Conversion Professional Seniority

Figure 2.4 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 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.

[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,\unskip\protect\penalty\@M\vrulewidth\z@height\z@depth\dp ,

[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.

	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 2.4.: 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 2.5 shows the distribution of researchers by types; table 2.6 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 2.5.: 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 2.6.: Number of Projects by Type

Figure 2.3 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%.

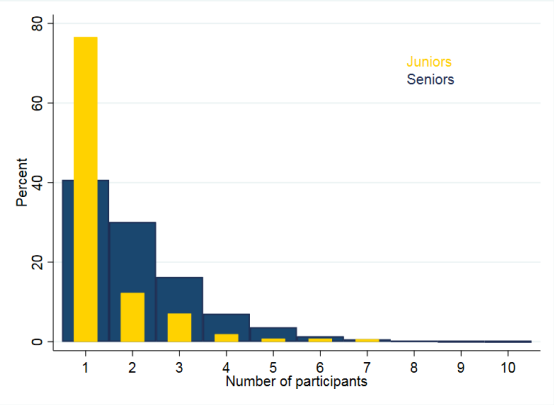
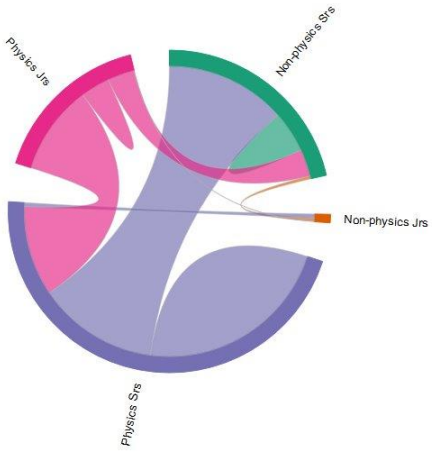


Figure 2.3.: Homophily by Professional Seniority

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 2.4 gives a more comprehensive illustration of the bilateral project connections among researchers' types.

Figure 2.4.: Chord Diagram of Bilateral Project Connections by Types



Notes: 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.

2.3. 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.

2.3.1. 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.

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,

[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.

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:

$$\pi_{ij} = \alpha_i X_i + \eta' D_j + \sum_{g=1}^{G} \delta_{ig} N_{gj} + \tilde{q_j} + \epsilon_{ij} \tag{2.1}$$

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:

$$\mathbb{E}[\pi_{ij}] = \alpha_i X_i + \eta' D_j + \sum_{g=1}^{G} \delta_{ig}\, \mathbb{E}[N_{gj}] + \tilde{q_j} + \epsilon_{ij} \geqslant 0. \tag{2.2}$$

[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.

Equilibrium

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

$$p^*_{ij} = \Phi\left(\alpha_i X + \eta' D + \sum_{g=1}^{G} \delta_{ig} p^*_{gj} + \tilde{q_j}\right) \tag{2.3}$$

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]

2.3.2. 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:

$$outcome^*_j = \tau + \sum_{g=1}^{G} \beta_g N_{gj} + \lambda' C_j + q_j. \tag{2.4}$$

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

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

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.

2.4. 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 outcome 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.

2.4.1. 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, Hong, Krainer, and Nekipelov, 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 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]

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

[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

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:

$$Q_J(\theta, \boldsymbol{p}) = \frac{1}{J}\frac{1}{G}\sum_{j=1}^{J}\sum_{g=1}^{G} \log \Psi_i(d_{ij}|X, C, \tilde{q}; \boldsymbol{p}, \theta_1) \tag{2.5}$$

2.4.2. 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 term q_j, from which I construct the generalized residuals $\tilde{q_j}$ following Gourieroux, Monfort, Renault, and Trognon (1987):

$$\tilde{q_j} = E[q_j|X, \hat{\theta}_2], \tag{2.6}$$

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.

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.

2.4.3. 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

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

$$\ln L = \sum_{j=1}^{J} \log \Psi(outcome_j | N, C; \theta_2). \tag{2.7}$$

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

$$\tilde{q}_j = E[q_j | outcome_j = 0, N, C, \hat{\theta}_2] = \frac{-\phi(\hat{\tau} - \hat{\beta} N_j - \hat{\lambda}' C)}{1 - \Phi(\hat{\tau} - \hat{\beta} N_j - \hat{\lambda}' C)} \tag{2.8}$$

For $outcome_j = 1$:

$$\tilde{q}_j = E[q_j | outcome_j = 1, N, C, \hat{\theta}_2] = \frac{\phi(\hat{\tau} - \hat{\beta} N_j - \hat{\lambda}' C)}{\Phi(\hat{\tau} - \hat{\beta} N_j - \hat{\lambda}' C)} \tag{2.9}$$

a function of the 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:

$$f(d, outcome) = \prod_{j=1}^{J}\prod_{i=1}^{I} Pr(d_{ij}|X, C, \tilde{q}; P, \theta_1) \times \prod_{j=1}^{J} Pr(outcome_j|N, C; \theta_2). \quad (2.10)$$

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:

$$LL(\theta) = \frac{1}{J}\frac{1}{G}\sum_{j=1}^{J}\sum_{i=1}^{I} \log \Psi_i(d_j|X, C, \tilde{q}; \boldsymbol{p}, \theta_1) + \frac{1}{J}\sum_{j=1}^{J} \log \Psi(outcome_j|N, C; \theta_2) \quad (2.11)$$

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

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

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 ($\delta' s = 0$).

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

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.

2.5. 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.

2.5.1. 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 2.7 reports the results from preliminary OLS regressions of project outcome, where the dependent variable can take values 0 or 1 ("not completed" or "completed",

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

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 (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 2.7 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 2.7.: Project Outcome Results: OLS

Table 2.8 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 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.

[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.

	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 2.8.: Project Outcome Results: Logit

Researchers with different characteristics might affect differently the probability of project completion. Table 2.9 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 2.8.

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 2.9.: Project Outcome Results: Binary Outcome

Table 2.10 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 (2.1) 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 (specifi-

[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.

cation (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.

	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 2.10.: 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.

2.5.2. 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 2.11 column (1) correspond to specification (6) of Table 2.8 from the previous section and show the reduced-form results for the probability of

project completion. Column (2) corresponds to specification (3) of Table 2.10 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

[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

labour and the increased coordination costs (Adams, Black, Clemmons, and Stephan, 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. The results presented in table 2.12 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 2.9; 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 2.11. 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 2.11.

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

of Bitnet facilitates increased research collaboration between US universities and the specialization of research tasks.

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 2.11 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 2.8	**Project participation (2) Spec. (3) Table 2.10**	**Joint Two-Stage Pseudo Likelihood**	
			Quality, no endogenous participation (3)	**Quality & endogenous participation (4)†**
# of project participants	0.556*** (0.055)		0.137*** (0.062)	-0.037 (0.07)
# of pre-determined teams	2.32*** (0.51)		3.52*** (0.512)	3.55** (0.52)
Type characteristics				
Non-Physics Junior		-0.35*** (0.056)	-0.41*** (0.055)	0.98*** (0.105)
Non-Physics Senior		-4.31*** (0.241)	-4.37*** (0.24)	-2.96*** (0.258)
Physics Junior		1.63*** (0.061)	1.62*** (0.062)	2.95*** (0.106)
Physics Senior		0.75*** (0.071)	0.71*** (0.072)	2.12*** (0.117)
Project characteristics				
# predetermined teams		-0.88*** (0.074)	-0.84*** (0.074)	-0.86*** (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 2.11.: Full Model

	Project completion (1) Spec. (3) Table 2.9	Project participation (2) Spec. (3) Table 2.10	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 2.12.: Full Model With Type-specific Coefficients.

2.6. 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]

2.6.1. 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.

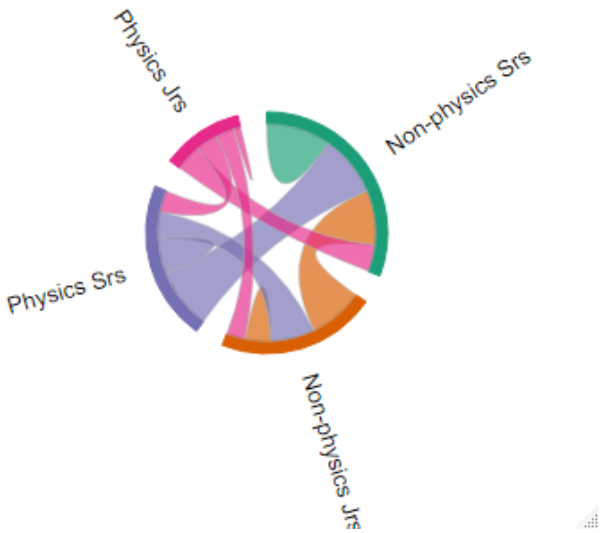


Figure 2.5.: Chord Diagram for Bilateral Project Connections: Counterfactual

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

Table 2.13.: Number of Projects by Type: Counterfactual

Figure 2.5 shows the results in terms of bilateral project connections. Compared to figure 2.4 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.

2.6.2. Project Outcome

To measure the effect of the alternative mechanism of project allocation on project outcome I use the estimated coefficients of Table 2.12 column (3); these results cor-

respond 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, Wuchty, and Uzzi, 2008). According to Banal-Estañol, Macho-Stadler, and Pérez-Castrillo (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.

2.7. 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

[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.

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 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.

A. Examples of Projects

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

Figure A.1.: Project 1

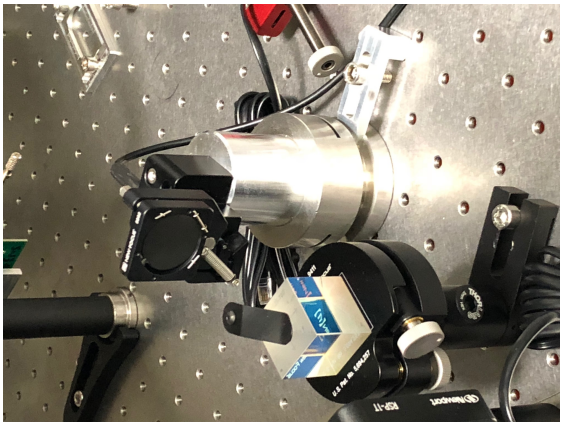
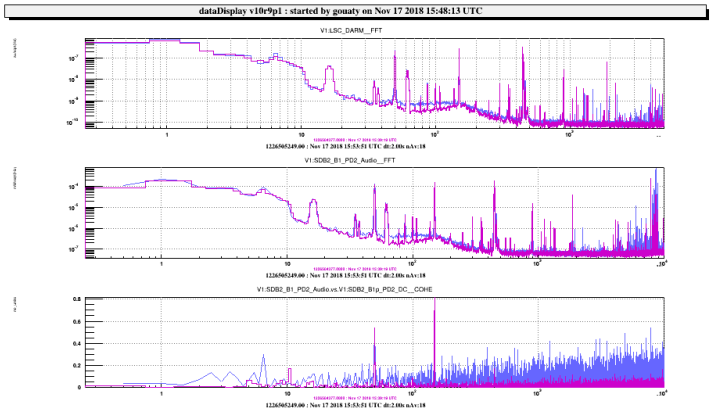


Figure A.2.: Project 2



B. Other Descriptive Statistics

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

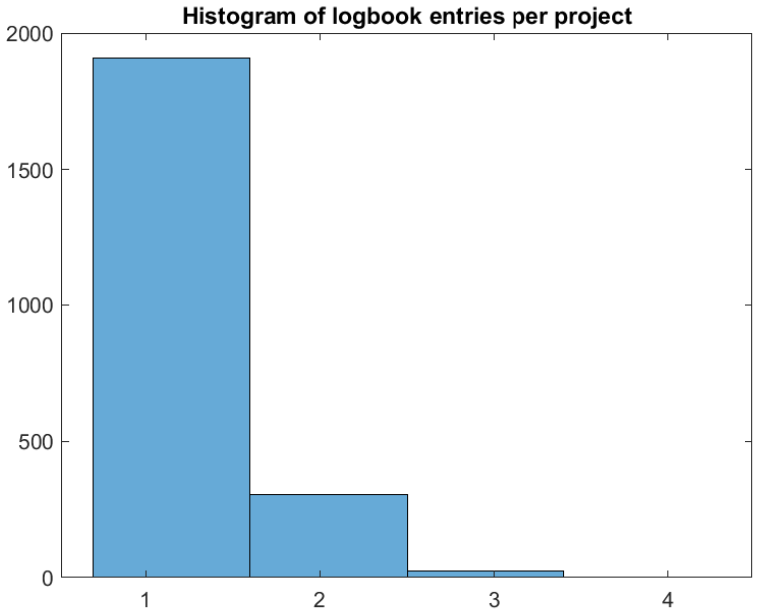


Figure B.1.: Logbook Entries Distribution

The following plots shows the distributions of project participation for Non-Physics Seniors (figure fig. B.2) and Physics Seniors (figure fig. B.3). The two distributions look very similar.

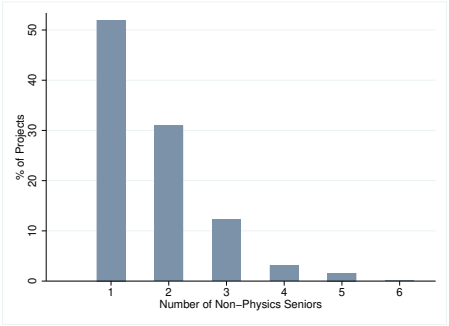


Figure B.2.: Distribution of Non-Physics Seniors

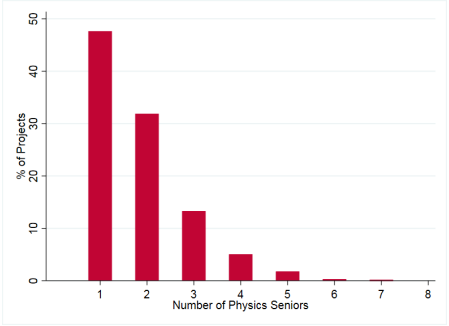


Figure B.3.: Distribution of Physics Seniors

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 [...]."

D. Other Reduced-form Results

Table D.1 and D.2 contain additional reduced-form evidence that shows clear paths in the connections among different types even controlling for projects' characteristics.

In particular, table D.1 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 D.2, 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 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 D.1.: Predicted Probabilities of Project Participation for Non-Physics Seniors

	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 D.2.: Predicted Probabilities of Project Participation for Physics Seniors

3. Common Ownership and Entry in the Ontarian Cancer Drug Market

joint with Laura Grigolon

3.1. Introduction

One of the most debated issues in antitrust is the presence of large institutional investors whose investment strategy involves owning large stakes in competing rivals, especially in concentrated industries such as airlines, banking and pharmaceuticals. Antitrust scholars worry that those large owners may have an incentive to intervene in the competitive setting and induce the rivals to soften competition: according to the *common ownership hypothesis*, an investor holding a controlling stake in several competing firms might have nonzero profit weights among these firms. As a result, firms' entry, pricing, and investments decisions might be affected (Backus, Conlon, and Sinkinson, 2020; Gilje, Gormley, and Levit, 2019).[1]

This paper documents in a rigorous manner the features of a highly innovative industry characterized by a concentrated ownership structure, the cancer drug industry. The analysis has the objective of studying the effect of common ownership on the decision of a generic producer to enter the market. We focus on Ontario, the most populated province of Canada. In 2019, around 220,000 new cases of malignant cancer have been estimated in Canada. During the previous year, Ontario experienced circa 90,500 cases, with an age-standardized incidence rate of 571.1 cases per 100,000.[2]

[1] According to McCahery, Sautner, and Starks (2016), Fichtner, Heemskerk, and Garcia-Bernardo (2017), and Antón, Ederer, Giné, and Schmalz (2018), institutional investors may engage in active discussions with companies' board and management or vote against other investors with the purpose of influencing the companies' strategies.

[2] Ontario Cancer Statistics, 2018.

Investor in 1993		Investor in 2020	
TEVA Pharmaceutical			
American Cent Invt Mgmt, Inc.	7,7%	Capital Research & Mgmt Co.	15%
Fidelity Mgmt & Research Co.	6,6%	Berkshire Hathaway, Inc.	3,9%
Capital Research & Mgmt Co.	5,4%	BlackRock Fund Advisors	2,7%
Peregrine Capital Mgmt	4,6%	Abrams Capital Mgmt LP	2,2%
Columbia Wanger Asset Mgmt LLC	4,3%	Migdal Insurance Co. Ltd.	2%
Amgen			
Alliance Capital Mgmt	2,8%	The Vanguard Group, Inc.	8%
Sarofim Fayez	2,8%	Capital Research & Mgmt Co.	6%
Axa Financial, Inc.	2,6%	BlackRock Fund Advisors	5,3%
Tiger Mgmt Co.	2,3%	SSgA Funds Mgmt, Inc.	4,4%
Bankers Trust NY Co.	2,2%	PRIMECAP Mgmt Co.	3,1%
Pfizer			
Sarofim Fayez	2,8%	The Vanguard Group, Inc.	7,9%
Alliance Capital Mgmt	2,4%	SSgA Funds Mgmt, Inc.	5,3%
Axa Financial, Inc.	1,9%	BlackRock Fund Advisors	4,9%
State Str Co.	1,9%	Capital Research & Mgmt Co.	4%
Wellington Mgmt Co. LLP	1,8%	Wellington Mgmt Co. LLP	3,9%

Table 3.1.: Corporate Ownership Top Cancer Drugs

Common ownership linkages are a well-established feature of many industries, including the cancer drug industry. Table 3.1 shows the top five shareholders for the biggest cancer drugs publicly listed companies in Canada for the years 1993 and 2020. One can see that common ownership was present already in 1993, but the percentage of shares held is more concentrated nowadays. BlackRock and Capital Research, among the world's largest institutional investors together with the Vanguard Group, possess a significant amount of stocks in all the three companies.[3] Moreover, as of April 2020, the Vanguard Group is the top shareholder of Amgen Inc, Mylan Pharmaceuticals, Bristol Myers Squibb, Pfizer Inc., Johnson & Johnson and Merck & Co., that are the main pharma companies operating in the Canadian drug cancer industry.[4]

The cancer drug industry is an attractive setting for several reasons. First, markets are well defined in terms of the timing of entry. Second, cancer drugs are extremely expensive and generic medicines are crucial to lowering their prices. Third, cancer

[3]Sources: Thomson Reuters for 1993 and money.cnn.com for 2020.

[4]Sources: Bloomberg website, money.cnn.com, simplywall.st, marketscreener.com.

drugs budgeting is an important percentage of the public health expenditure and it is constantly growing over time: in line with the national trend, hospital and public drug program spending on cancer drugs in Ontario reached $550.7\$$ millions in 2018, accounting for 8.5% of the total public drug program spending, with a 25% increase relative to the previous year.[5] These factors, together with the high level of concentration, make it an appealing setting to understand the consequences of common ownership.

Market entry decisions are usually intended as substitutes due to competition. However, the presence of common ownership can reverse this idea. In our paper, we aim at understanding what are the main components of this phenomenon and how it links with firms' strategic decisions of entering a market. The analysis has important welfare implications and it can be helpful for policy-makers to understand the target of an eventual intervention.[6]

We use unique data on the timing of cancer drug entry in the market (branded and generics) and collect information on patents, drug approvals, and drug indications for the years 1993-2019. We complement our dataset by gathering ownership data mainly from 13F filings. In our sample, around half of the drug markets with generics experience common ownership between the brand and one or more generics during the years prior to the entry of the generic.

Following the recent literature (Backus et al., 2020; Newham, Seldeslachts, and Banal-Estanol, 2018), we calculate a theoretically based company-pair-specific measure of common ownership to empirically analyze this phenomenon in our setting. In particular, we look at pairs of companies and pairs of brand-generics for each drug market. Ownership concentration is a long-lasting feature of the cancer drug industry. In particular, we find that overlapping ownership between brands and generics two years before the generic entry is growing over time. Moreover, we show that the variation in relative investors' concentration explains a large part of the variation in common ownership in our sample. We also quantify the importance of the co-movement between overlapping ownership and investors' concentration when looking at variation

[5]Source: Canadian Institute for Health Information, page 27. https://www.cihi.ca/sites/default/files/document/pdex-report-2019-en-web.pdf

[6]Another relevant aspect of this industry is that cancer drugs are often used in bundles. Song, Nicholson, and Lucarelli (2017) for instance analyze the effects of a merger between two pharmaceutical firms selling complements for colorectal cancer treatment. They find that a merger may generate a price decrease. Following the same argument, common ownership might further affect the level of complementarity within products in the presence of bundling.

in common ownership for brands and generics prior to the generic entry. Our findings can be of interest for other innovative industries where common ownership represents an important threat to competition.[7]

Basing on the empirical evidence of this paper, we plan to develop a structural model of entry with strategic interaction, in the spirit of the entry models estimated in the empirical IO literature (Bresnahan and Reiss, 1991b). Using our data, we will test whether the presence of an incumbent, the branded drug, sharing a common owner with the entrant, has an impact on the entry probability of the generic and how the probability changes when allowing for strategic interaction between generics. With the structural model, we will be able to disentangle the economic effects of common ownership on the consumers' surplus, industry structure, and government expenditures. In a counterfactual experiment, we will consider how these components change if any engagement in corporate governance by institutional investors is limited or forbidden (Backus et al., 2020).

This work relates to several strands of literature. Many papers that analyze the determinants of generic entry decisions in drug markets find that generic entry is higher the larger the size of the branded drug's market before the patent expiration (Appelt, 2015; Morton, 1999, 2000; Saha, Grabowski, Birnbaum, Greenberg, and Bizan, 2006; Torres, Puig, Borrell-Arqué, et al., 2007). Regarding the Canadian pharmaceutical industry, McRae and Tapon (1985) analyzes the post-market barriers to entry, while Hollis (2002) finds that the first generic entrant has a lasting competitive advantage. None of these papers have looked at the interplay between common ownership and entry in Canada.

The theoretical link going from common ownership to competition, initially studied by Rubinstein and Yaari (1983) and Rotemberg (1984), has motivated new empirical research to find a significant effect (Azar et al., 2018; Backus, Conlon, and Sinkinson, 2018; Koch, Panayides, and Thomas, 2018; Schmalz, 2018). Newham et al. (2018) study how the presence of common ownership influences the decision of entering US pharmaceutical markets after the end of regulatory protection (off-patent drug markets).

[7]The pricing decision can be influenced by common ownership (Azar, Schmalz, and Tecu, 2018) and represents an additional threat to competition. In this particular market, prices are the result of a complex bargaining process which involves the government and the firms, therefore we abstract from prices assuming that they are set outside the setting.

They show that an increase in common ownership decreases the likelihood of the entry of generics.

We contribute to the literature in several aspects. First, we empirically assess the presence of common ownership in the Ontarian cancer drug industry. Second, we accurately document the differences in the ownership components if one considers linkages between companies or linkages between brands and generics. Third, we quantify which of these components mainly drive the link between common ownership and market entry.

The rest of the paper is organized as follows. Section 2 provides a description of the industry and the data. Section 3 describes the measure of common ownership used in the analysis. Section 4 presents the empirical evidence. Section 5 concludes.

3.2. Background and Data

In the pharmaceutical industry, brand companies engage in a process of research and development to discover new drugs, and if the process is successful they apply for drug approval. When the approval is granted, the brand is awarded *data exclusivity* for a period that goes from three to seven years, depending on the drug. This exclusivity protects the clinical data and runs concurrently with patent protection. *Market exclusivity* refers instead to the period between the end of data exclusivity and the expiration of the last patent.

Generic companies are able to enter a particular drug market once the regulatory protections have expired. These companies need to be marketed as brand-name products and afterward they produce replications of brand drugs at a much lower cost. During the market exclusivity period, generics can challenge the monopoly rights of the brand in court.

Health Canada is the department of the Canadian Government that is responsible for the country's federal health policy, under the administration of the Minister of Health. The department authorizes the sale and use of new drugs. Before the approval (or rejection) of a new medication, the process to review drug safety and efficacy information from clinical trials takes on average between two and four years. After the approval, each province must still decide whether or not to reimburse the cost of the new medica-

tion. Sometimes, patients can obtain an experimental therapy through Health Canada's Special Access Programme (SAP) before the process ends.[8][9]

Most cancer drugs in Ontario are publicly funded by the Ministry of Health. The New Drug Funding Program (NDFP), to which we have access, with information on sales and drug prices, directly covers the cost of many newer and often expensive injectable cancer drugs.[10]

We use data from several sources. General information on brands and generics, including the date of entry in the market, are taken mainly from Health Canada and drugbank.ca. Information regarding the ownership structure for the years of interest (1993-2019) comes from the Thomson Reuters dataset. In the following sections, we illustrate how we match the different datasets in more detail and provide a description of the data.

3.2.1. Pharmaceutical Data

We collect information on entry dates of brands and generics from drugbank.ca. The initial dataset contains 45 brands and 150 generics. We drop 1 brand and 22 generics as we do not have information on the entry date. After removing other inconsistencies, the final dataset consists of 44 drug markets: 24 markets do not have generics, while the remaining markets have in total 128 generics, with an average of 3 generics per drug. Table 3.2 lists the 44 drugs, the number of generic entrants and the date of the first generic entrant. For the drug markets with generics, the number of generics goes from 1 (Cabazitaxel and Rituximab) up to 17 (Zolendronic Acid).

We gather information on the end of data exclusivity from Health Canada and patent expiration from Health Canada and CIPO (Canadian Intellectual Property Office) websites. Notice that a brand can patent or renew the patent of part of a drug; following the literature, we refer to the last expiring date. Any product defined as a drug under the Canadian Food and Drugs Act must have an associated Drug Identification Number (or DIN). Once a drug has been approved, the Therapeutic Products Directorate issues

[8] https://laforcedmd.com/

[9] Once a drug is approved, there is the bargaining process on the price.

[10] Other programs are the Evidence Building Program and the Case-By-Case Review Program. For the non-funded drugs, patients may use private insurance or pay directly. Source: Cancer Care Ontario. https://www.cancercareontario.ca/en/cancer-treatments/ chemotherapy/funding-reimbursement/drug-funding-faqs

Drug Name	N. Generics	First Generic Entry
Aldesleukin	0	
Arsenic trioxide	2	20.12.2019
Azacitidine	2	25.10.2017
Bendamustine	0	
Bevacizumab	2	01.08.2019
Blinatumomab	0	
Brentuximab	0	
Bortezomib	8	17.11.2015
Cabazitaxel	1	18.12.2019
Cetuximab	0	
Clodronate	1	13.05.2004
Denosumab	0	
Docetaxel	8	01.03.2011
Epirubicin	0	
Eribulin	0	
Fludarabine	4	04.12.2006
Gemcitabine	11	20.11.2007
Interferon	0	
Ipilimumab	0	
Irinotecan	9	01.02.2006
Liposomal Doxorubicin	0	
Nab-Paclitaxel	0	
Nivolumab	0	
Obinutuzumab	0	
Oxaliplatin	11	16.12.2005
Paclitaxel	10	24.10.1997
Pamidronate	7	31.08.2001
Panitumumab	0	
Pembrolizumab	0	
Pemetrexed	8	03.06.2016
Pertuzumab	0	
Plerixafor	0	
Porfimer sodium	0	
Raltitrexed	0	
Ramucirumab	0	
Rituximab	1	11.12.2019
Romidepsin	0	
Siltuximab	0	
Temsirolimus	0	
Topotecan	10	02.10.2009
Trastuzumab emtansine	0	
Trastuzumab	3	06.06.2019
Vinorelbine	5	09.05.2007
Zoledronic Acid	17	25.07.2008

Table 3.2.: Drugs Information

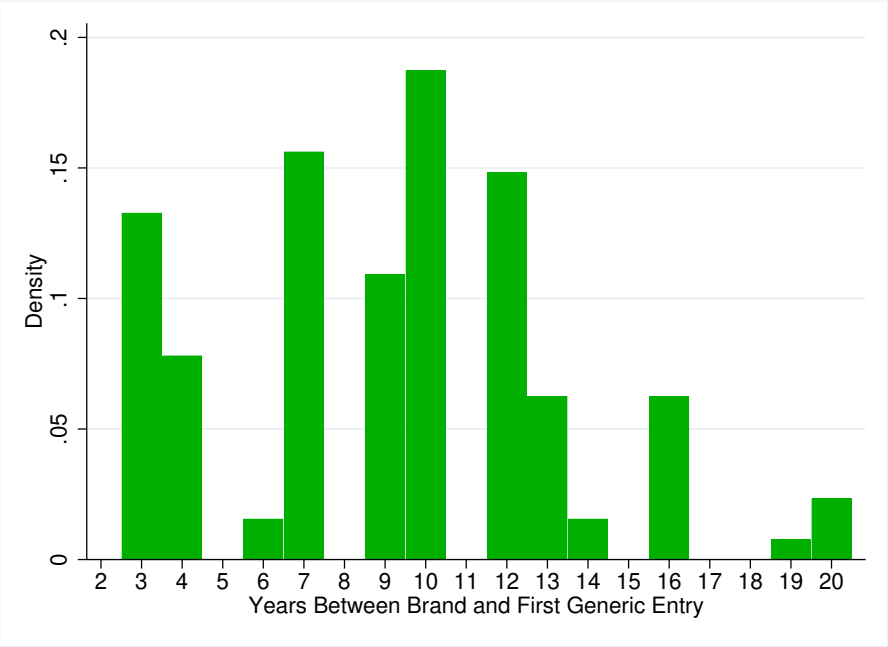


Figure 3.1.: Years Between Drug Launch and First Generic Entry in the Market

a DIN which permits the manufacturer to market the drug in Canada. The duration of the exclusivity period differs in the type of drug and is between three and seven years. In our sample, some drugs are newly issued, others have been granted new patents, thus many drugs' patents have not expired. Appendix A shows drugs DIN and patent expiration dates for a subset of products.

Figure 3.1 shows the year range between the launch of the drug and the entry of the first generic in the market. Around 36% of generics enter between three 3 and 7 years after the drug launch because of the standard period of market exclusivity. Nevertheless, 45% of generics enter between 9 and 12 years, in line with the fact that, differently from the standard pharmaceutical industries, the cancer drug is an innovative industry and many drug patents are granted renewal. Figure 3.2 instead considers the entry for all the generics. One can notice that generics continue to enter the market even after many years from the brand entry, and the probability of entering after 20 years is 12%.

From drug formularies and monographs, we collect information on the main drugs indications (or therapeutic fields), which are presented in table 3.3. Six drugs are mainly

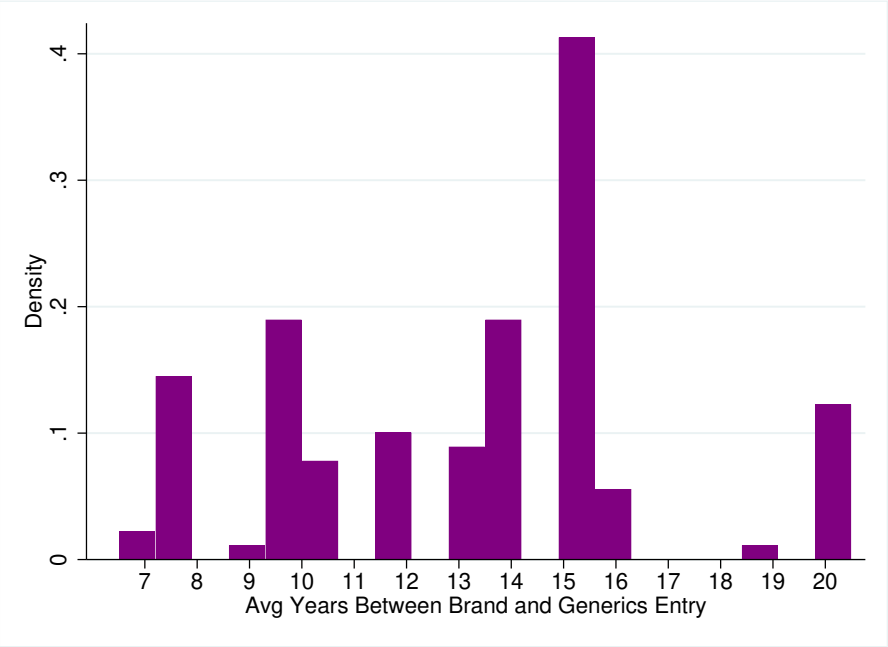


Figure 3.2.: Average Years Between Drug Launch and Any Generic Entry in the Market

used for breast cancer and five for colorectal or rectal cancer, reflecting that in Ontario the most commonly diagnosed cancers are breast and colorectal.[11]

3.2.2. Common Ownership Data

The Center for Research in Securities Prices (CRSP) dataset, available through the Wharton Research Data Services (WRDS), contains the identifier of each listed company and the total number of outstanding shares at the monthly level. For subsidiary companies, we go back to the listed parent or the listed holding the company belongs to, assuming that they are fully controlled by the parent (Newham et al., 2018). The CRSP dataset takes into account that companies may change their identity in the course of the sample period and that some of the companies may go public at some point in time and then become private again.

We use the identifier to match these data with ownership data from the Thomson Reuter's s34 database, also available in the WRDS (Gerakos and Xie, 2019; He and Huang, 2017; Schmalz, Azar, and Sahil, 2016). The database reports 13F filings, that

[11] As reported in 2018 from the Ontario Cancer Statistics.

Drug Name	Main Indication
Aldesleukin	Metastatic renal cell carcinoma
Arsenic trioxide	Acute promyelocytic leukemia
Azacitidine	Myelodysplastic syndrome and acute myeloid leukemia
Bendamustine	First line indolent non-Hodgkin's lymphoma and mantle cell lymphoma
Bevacizumab	First line metastatic colorectal cancer
Blinatumomab	Relapsed or refractory B-cell precursor acute lymphoblastic leukemia
Brentuximab	Hodgkin's lymphoma
Bortezomib	Relapsed or refractory myeloma
Cabazitaxel	Metastatic castration resistant prostate cancer
Cetuximab	Locally advanced squamous cell carcinoma of the head and neck
Clodronate	Metastatic breast cancer
Denosumab	Hormone-refractory prostate cancer
Docetaxel	Adjuvant breast cancer
Epirubicin	Adjuvant breast cancer
Eribulin	Metastatic or incurable locally advanced breast cancer
Fludarabine	Stage III-IV follicular lymphoma
Gemcitabine	Advanced pancreatic cancer
Interferon	Melanoma
Ipilimumab	Previously treated advanced unresectable melanoma
Irinotecan	First line metastatic colorectal cancer
Liposomal Doxorubicin	HIV positive Kaposi's sarcoma
Nab-Paclitaxel	Metastatic breast cancer
Nivolumab	Unresectable or Metastatic Melanoma
Obinutuzumab	Previously untreated chronic lymphocytic leukemia
Oxaliplatin	Adjuvant treatment of Stage III or high-risk Stage II colon and rectal cancer
Paclitaxel	Adjuvant breast cancer
Pamidronate	Plasma cell myeloma
Panitumumab	Metastatic colorectal cancer
Pembrolizumab	Unresectable or metastatic melanoma
Pemetrexed	First line or induction for the treatment of cell lung cancer
Pertuzumab	Positive unresectable locally recurrent or metastatic breast cancer
Plerixafor	Stem cell mobilization in non-Hodgkin's lymphoma or multiple myeloma
Porfimer sodium	Photodynamic therapy for NSCLC
Raltitrexed	Metastatic colorectal cancer (single agent)
Ramucirumab	Advanced gastric cancer or gastro-esophageal junction adenocarcinoma
Rituximab	Diffuse large B-cell lymphoma/aggressive
Romidepsin	Relapsed or refractory peripheral T-cell lymphoma
Siltuximab	Multicentric Castleman's disease (MCD)
Temsirolimus	Metastatic renal cell carcinoma
Topotecan	Advanced ovarian cancer
Trastuzumab emtansine	Unresectable locally advanced or metastatic breast cancer
Trastuzumab	Gastric cancer
Vinorelbine	Metastatic breast cancer
Zoledronic Acid	Hormone-refractory prostate cancer

Table 3.3.: Drugs Indications

are required by the SEC for all investment managers of US companies with over $\$100$ million in holdings.[12] In particular, we have quarterly data for the years 1993-2019 and compute the yearly shares averaging out across quarters. We focus on shareholders that own at least 1% of shares in the company.

There are 41 unique companies (including both brands and generics) in the dataset. Some companies are listed outside the US stock market, and the ownership information is not contained in the Thomson Reuters dataset.[13] For these companies, we gather information from other sources, such as company websites and financial websites, with the caveat that we are only able to find information on the current ownership. The remaining 14 are private companies not publicly listed on a stock exchange. We assume that they do not have common investors with any other. A public company may have no ownership information within a certain year in the Thomson Reuters dataset: in this case, we remove it from the analysis for that sample period (Newham et al., 2018).

The dataset consists of 968 company/year/country observations.[14] For the North American countries, the dataset contains 510 observations.[15] Table 3.4 shows the time series of the top 5 shareholders in the industry for the full sample. The industry seems very concentrated, as only 18 unique investors have been regular occurrences as top industry shareholders in over 25 years.

Because the industry we consider is highly innovative and many drug patents are not expired, we calculate the measure of common ownership in the year of generic entry, one year prior and two years prior, as it takes some time for the generic to prepare the entry in the market. [16] During the market exclusivity period, generics can challenge the monopoly rights of the brand in court, for instance through Paragraph IV certification, while brands might engage in strategic decisions (when choosing the advertising level for instance) to deter market entry (Ellison and Ellison, 2011). Our

[12] It has been recently noted that this dataset might have some gaps in coverage (Backus et al., 2018). By comparing the information contained in the dataset with other external sources such as Bloomberg, Thomson Reuters is a good source for the companies in our sample.

[13] In particular, one is only listed in India, two in Japan, one in Switzerland, and one in South Korea.

[14] It contains the following countries: Australia, Bermuda, Canada, Denmark, France, Germany, Hong Kong, Ireland, Japan, the Netherlands, Norway, Singapore, South Africa, Sweden, Switzerland, Taiwan, UK and US

[15] We also drop duplicates and inconsistent observations.

[16] Newham et al. (2018) look at common ownership for the year of the end of exclusivity, one year prior and two years prior.

approach has the advantage to capture possible links between companies also in case generics apply for Paragraph IV.[17]

In table 3.5 we show for each product the number of shareholders in common between generics and brand; it is calculated for the year of generic entry (t), the year prior ($t-1$) and two years prior ($t-2$). Common ownership is a prevalent feature for 8 drug markets. As Topotecan and Zoleandic Acid have the highest number of generics (table 3.2), their generics also experience the highest concentration of shareholders in common with the respective brands. Moreover, one can see that common ownership is larger in the years prior to the generics' entry. This points to the fact that different shareholders might exert control on the decision of a generic by owning large stakes of the company.[18] In the next paragraphs, we explore the link between common ownership and entry emerging from these descriptive statistics.

[17] We plan to collect evidence on Paragraph IV filings for our drug sample.

[18] Newham et al. (2018) for instance find a larger significant effect of common ownership on the generic entry probability one and two years prior the end of *market exclusivity* (Table C2 of the paper).

Year	Top 1 Shareholder	Top 2 Shareholder	Top 3 Shareholder	Top 4 Shareholder	Top 5 Shareholder
1993	Bankers Trust Corp	Mellon Bank Corp	Wells Fargo Inst Tr	Capital Research & Mgmt Co	Fidelity Management Research
1994	Wells Fargo Inst Tr	Bankers Trust Corp	Fidelity Management Research	Mellon Bank Corp	State Street Group
1995	Mellon Bank Corp	Wells Fargo Inst Tr	Bankers Trust Corp	Capital Research & Mgmt Co	Fidelity Management Research
1996	Barclays Bank Plc	Wells Fargo Inst Tr	Fidelity Management Research	Bankers Trust Corp	Mellon Bank Corp
1997	Barclays Bank Plc	College Retire Equities	Fidelity Management Research	Mellon Bank Corp	Bankers Trust Corp
1998	Fidelity Management Research	Barclays Bank Plc	Bankers Trust Corp	Mellon Bank Corp	College Retire Equities
1999	Fidelity Management Research	Barclays Bank Plc	Bankers Trust Corp	Mellon Bank Corp	Morgan Stanley Group Inc
2000	Fidelity Management Research	Barclays Bank Plc	State Street Group	The Vanguard Group	Bankers Trust Corp
2001	Fidelity Management Research	Barclays Bank Plc	Mellon Bank Corp	Citigroup Inc	Wellington Management Co
2002	Fidelity Management Research	Bankers Trust Corp	Wellington Management Co	Barclays Bank Plc	Mellon Bank Corp
2003	Fidelity Management Research	Bankers Trust Corp	Barclays Bank Plc	Citigroup Inc	State Street Group
2004	Fidelity Management Research	Bankers Trust Corp	Barclays Bank Plc	Mellon Bank Corp	State Street Group
2005	Wellington Management Co	Fidelity Management Research	Mellon Bank Corp	Barclays Bank Plc	Bankers Trust Corp
2006	Fidelity Management Research	Mellon Bank Corp	Barclays Bank Plc	Morgan Stanley Group Inc	Northern Trust Co
2007	Fidelity Management Research	Mellon Bank Corp	State Street Group	Barclays Bank Plc	Legg Mason Fund Advisors
2008	Fidelity Management Research	Barclays Bank Plc	State Street Group	The Vanguard Group	Axa Financial Inc
2009	Fidelity Management Research	Mellon Bank Corp	Axa Financial Inc	State Street Group	The Vanguard Group
2010	Northern Trust Co	Fidelity Management Research	Wellington Management Co	Blackrock Inc	Morgan Stanley Group Inc
2011	Blackrock Inc	Mellon Bank Corp	Northern Trust Co	The Vanguard Group	Wellington Management Co
2012	Blackrock Inc	Amverscap Plc London	Northern Trust Co	The Vanguard Group	Wellington Management Co
2013	Blackrock Inc	Fidelity Management Research	Northern Trust Co	The Vanguard Group	Wellington Management Co
2014	Blackrock Inc	Fidelity Management Research	Northern Trust Co	State Street Group	The Vanguard Group
2015	Blackrock Inc	Northern Trust Co	State Street Group	The Vanguard Group	Amverscap Plc London
2016	Blackrock Inc	State Street Group	The Vanguard Group	Northern Trust Co	Fidelity Management Research
2017	Blackrock Inc	Northern Trust Co	State Street Group	The Vanguard Group	Fidelity Management Research
2018	State Street Group	Northern Trust Co	The Vanguard Group	Fidelity Management Research	Wellington Management Co
2019	Blackrock Inc	State Street Group	Northern Trust Co	The Vanguard Group	Geode Capital Management

Table 3.4.: Top 5 Shareholders Over Time

3.3. Common Ownership Measure

Let b be the brand in the drug market, $g = \{1, ..., G\}$ the (potential and actual) generic and $i = \{1, ..., I\}$ the investors. Denote the shares held by investor i in brand b as β_{bi} and the shares held in generic g as β_{gi}. Investor i is a common owner if $\beta_{bi} > 0$ and $\beta_{gi} > 0$. The literature has proposed two measures of common ownership. The first one is based on the implicit assumption that investors actively engage with decision-making (production function approach). The second one is theoretically founded and poses that the generic's decision-makers take shareholders' portfolio interests explicitly into consideration.

Denote as s_i the profit of investor i, which is given by the sum of profits over the portfolio of investments weighted by cash-flow rights (Backus et al., 2020), so that:

$$s_i = \beta_{bi}\pi_b + \sum_{g=1}^{G} \beta_{gi}\pi_g. \tag{3.1}$$

In the presence of common ownership, companies maximize the profits of the shareholders (O'Brien and Salop, 2000; Rotemberg, 1984). In each market, the objective function V_g of generic g is proportional to the following term[19]:

$$\pi_g + \kappa_{gb}\pi_b + \sum_{f \neq g} \kappa_{gf}\pi_f, \tag{3.2}$$

where π_g is generic g's own profit, π_b is brand b's own profit and π_f is generic f's own profit, f being any generic company different from g. In particular,

$$\kappa_{gb} = \frac{\sum_i \beta_{gi}\beta_{bi}}{\sum_i \beta_{gi}^2} \tag{3.3}$$

is the profit weight. One can interpret it as the value of a dollar of profits accruing to brand b, relative to a dollar of profits for generic g, in g's maximization problem (Backus et al., 2020). An analogous interpretation applies to κ_{gf}.[20] We are interested in the effect of common ownership between brands and generics, therefore in the analysis,

[19] We assume that the rule *one share, one vote* applies (also defined as *proportional control*).

[20] In pages 10 and 11 of their paper, Backus et al. (2020) provide an example of ownership to show how the matrix of profit weights should look like.

		Common Owners		
Drug Name	**Generic Company**	*In t*	*In t-1*	*In t-2*
Bortezomib	Sandoz Canada Ulc	5	5	6
Bortezomib	Mda Inc	4	5	4
Bortezomib	TEVA Canada Ltd, Actavis Pharma		66	60
Cabazitaxel	Sandoz Canada Ulc		2	2
Fludarabine	TEVA Canada Ltd		1	4
Fludarabine	Hospira Healthcare Ulc		1	2
Gemcitabine	TEVA Canada Ltd	11	1	
Paclitaxel	Mylan Pharma	11	11	10
Paclitaxel	Sandoz Canada Ulc	12	13	15
Paclitaxel	TEVA Canada Ltd		5	6
Pamidronate	Fresenius Kabi Canada Ltd	9	13	8
Pamidronate	Sandoz Canada Ulc	11	10	7
Topotecan	Sandoz Canada Ulc	8	10	11
Topotecan	Mylan Pharma		14	16
Topotecan	TEVA Canada Ltd		14	16
Topotecan	Sandoz Canada Ulc		8	8
Topotecan	Actavis Pharma		5	4
Zoledronic Acid	Taro Pharmaceuticals Inc	2		
Zoledronic Acid	TEVA Canada Ltd		8	10
Zoledronic Acid	Mda Inc	8	9	9
Zoledronic Acid	Mylan Pharma	7	1	11

Notes: These numbers include shareholders outside the North American stock markets. For Bortezomib, we cannot distinguish the shareholders of TEVA and Actavis as they belong to the same listed company (TEVA).

Table 3.5.: Number of Shareholders in Common

we focus on κ_{gb}. Notice that when $\kappa_{gb} = 0$, the generic company is maximizing only its own profit, while mergers result in $\kappa_{gb} = 1$. Any level of $\kappa_{gb} > 0$ can arise in a common ownership setting.

Let $IHHI_g$ be the Herfindahl-Hirschman Index (HHI) for the investors in company g (or analogously in company b). Define $\cos(\beta_g, \beta_b)$ as the cosine similarity between vectors β_g and β_b. It represents the cosine of the angle between the positions that investors hold in b and those that investors hold in g. Backus et al. (2020) decompose the profit weights into two terms:

$$\kappa_{gb}(\beta) = \cos(\beta_g, \beta_b) \cdot \sqrt{\frac{IHHI_b}{IHHI_g}}. \tag{3.4}$$

The first part is defined as *overlapping ownership* and is the standard measure of common ownership analyzed by the literature: the closer the investor positions, the smaller the angle between the portfolios with the cosine similarity approaching one. The second is defined as *relative investor concentration* and is interpreted as the ability of common owners to exert control. Intuitively, *ceteris paribus*, generics with concentrated investors will place more weight on their own profits and less weight on brand profits, and vice versa.

3.3.1. Variance Decomposition

By decomposing κ, we are also able to understand what is the main source of empirical variation in common ownership profit weights. Again, following Backus et al. (2020), we take the log of κ and decompose the variance:

$$Var(\log \kappa_{gb}) = Var(\log \cos(\beta_g, \beta_b)) + Var\left(\log \sqrt{\frac{IHHI_b}{IHHI_g}}\right) + \tag{3.5}$$

$$+ 2Cov\left(\log \cos(\beta_g, \beta_b), \log \sqrt{\frac{IHHI_b}{IHHI_g}}\right). \tag{3.6}$$

The first term is the contribution of the overlapping ownership to the total variance of κ; the second term accounts for the variance in relative investor concentration; the

third term represents the co-movement between overlapping ownership and relative investor concentration.

3.4. Empirical evidence

In this paragraph, we report the results from the analysis on industry concentration and its potential link with market entry. First, we present empirical evidence for all pairs of companies in the sample. Then, we look at brand-generic pairs for each drug market. Last, we compute the variance decomposition of our common ownership measures. In all the specifications, we concentrate on the North American stock markets.

3.4.1. Common Ownership Paths

In Figure 3.3 we report the median together with the $75th$ and $95th$ quantiles of investor concentration.[21] Notice that investor concentration increases over time. For the most concentrated companies ($95th$ quantile), the index raises above 1000 in two different years. This value is not too far from 1500, which is considered the threshold for moderately concentrated markets according to the DOJ and FTC standards. However, one cannot conclude that the growing investor concentration automatically generates a rise over time in κ, as the common ownership index is proportional to the ratios of $IHHI$ for the pairs of companies considered and the cosine distance between vectors of shares.

In Figure 3.4 we illustrate the relationship between κ and $\cos(\beta_g, \beta_b)$ over time across pairs of companies. The trends are very similar, meaning that the cosine similarity tracks very well the average profit weight. Moreover, differently from Backus et al. (2020), for these measures we do not observe an upward sloping trend and the values move around relative large numbers (with a range similar to Backus et al. (2020) – between 0.2 and 0.7). This shows that common ownership has always been a feature of this industry compared to the set of S&P 500 firms analyzed by the authors.[22]

In our sample, the first generic enters the market in 1997, but most generics entered after 2001, as documented in Table (3.2). It is interesting to notice that after 2002

[21] The other percentiles do not show significant time variation.

[22] In both figures, we exclude 96 company pairs with an $IHHI$ above 0.5. They count for less than the 0.005% of the sample. The $IHHI$ is multiplied by 10,000.

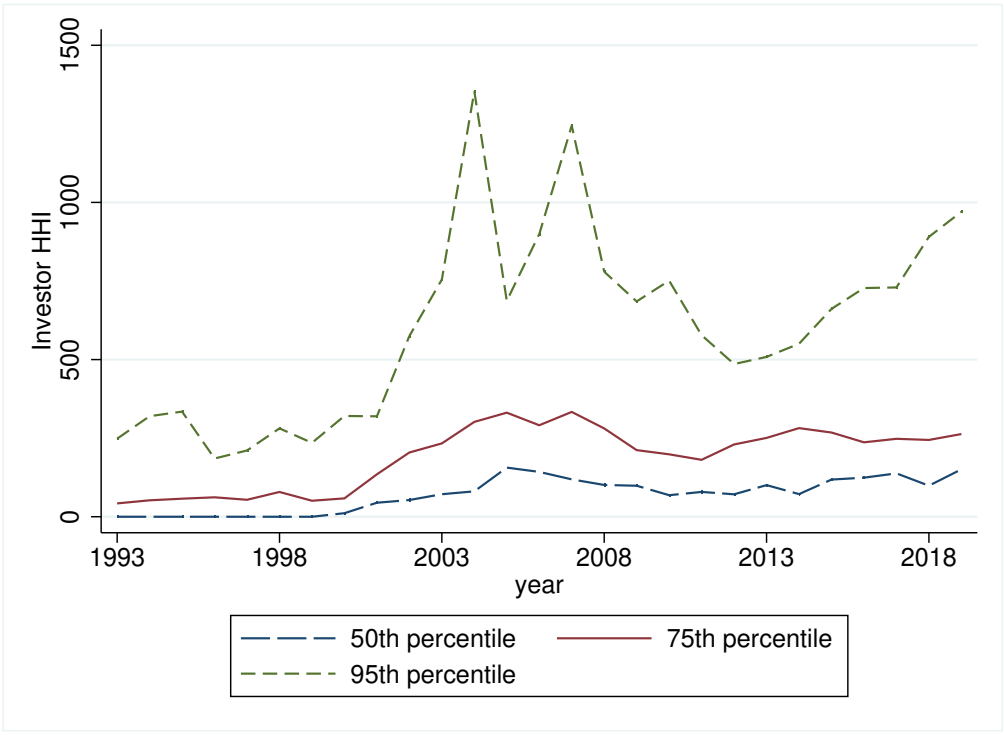


Figure 3.3.: Investor Concentration All Company-Pairs

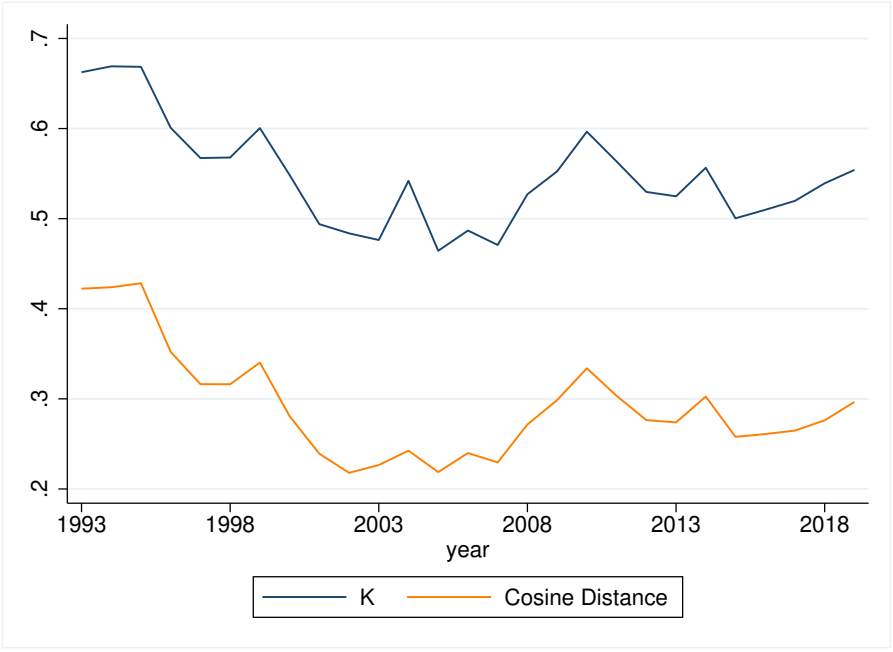


Figure 3.4.: κ and Similarity All Company-Pairs

both κ and the cosine similarity index start increasing over time. One might conclude that common ownership has been prevalent since the generics' entry in the market. However, this is not straightforward, as we need to look at common ownership between pairs of brands and generics. For this purpose, we construct profit weights κ_{bg}. In particular, in Figure 3.5 we show the trend over time of these weights for the year of the generics' entry, the year prior and two years prior. For an accurate comparison between these measures, we use balanced data from 2005 to 2016 because more than 90% of generic companies publicly listed are concentrated in this time-frame.[23] One can immediately see that common ownership is increasing over time when we look at the average profit weights two years before entry of the generic (Figure (c)). The upward-slope trend is not obvious instead for the profit weights computed the year of entry and one year prior. As it requires some years for a generic to prepare the entry in the market, this seems to suggest that shareholders buy shares of market rivals some time before the entry in order to intervene in their strategic decisions.

3.4.2. Results Variance Decomposition

In table 3.6, we show results from the variance decomposition for all company pairs and brand-generic pairs.[24] More than half of the variation in common ownership profit weights comes from investor concentration. This holds across all the specifications and points to the fact that corporate governance plays an important role in these weights, as a consequence of the high concentration of the industry. In line with the findings of Backus et al. (2020), for all the company pairs investor concentration has the largest impact in shaping the profit weights variation in the cross section while overlapping ownership explains the strongest percentage of common ownership change in the time series. This happens because, although the $IHHI$ grows over time, the index is such that for some κ the numerator becomes larger while for others it shrinks.

It is interesting to analyze the covariance term for brand-generic pairs. While for all company pairs the covariance has to be mathematically very close to zero since we observe for each log relative investor concentration also its inverse, this does not hold when looking at the profit weights κ_{bg}, which are unidirectional. Notice that

[23] Notice that these companies represent around 70% of the total population of generics.

[24] The results include only the actual generics, we plan to analyze these empirical regularities also for sets of potential entrants.

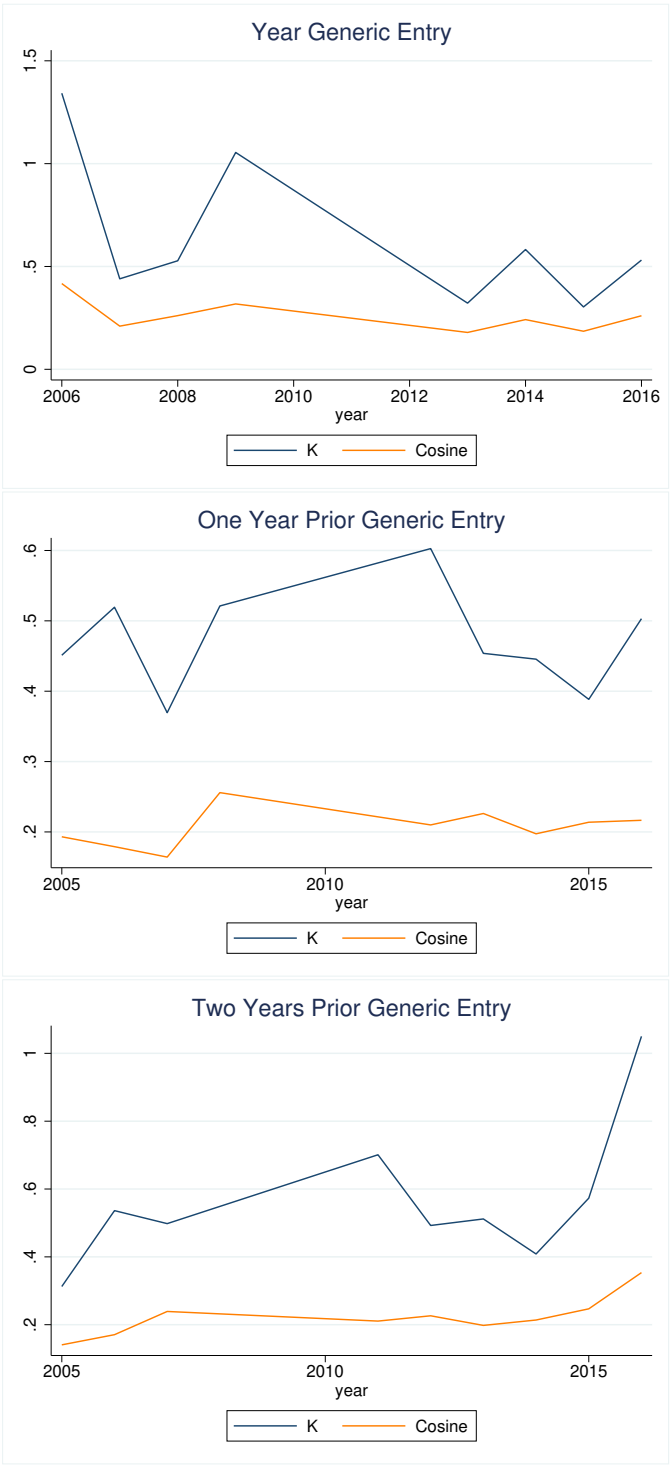


Figure 3.5.: (a) Year Generic Entry (b) One Year Prior (c) Two Years Prior

	Cosine	Relative IHHI	Covariance
		All Pairs	
Raw	35,86%	64,13%	0,01%
Cross-section	35,52%	64,4%	0,08%
Time-Series	42,1%	57,9%	0%
		Brand-Generic Pairs	
Raw	29%	51,97%	19,03%
Cross-section	29,2%	51,4%	19,4%
Time-Series	34,87%	47,09%	18%

Note: The cross-section variation is residualized on year fixed effects, the time-series is residualized on company-pair fixed effects. For brand-generic pairs, the decomposition is calculated for κ two years prior the generic entry.

Table 3.6.: Decomposition Variance Log κ

in this case the percentage of variation in $\log \kappa$ explained by the covariance term is large, exceeding 19% for the raw and cross-section variance. This means that the co-movement between overlapping ownership and relative investor concentration plays an important role in the change in common ownership between brands and generics. A possible interpretation is that the same shareholders participate in common companies and increase the percentage of shares in these companies.

3.5. Conclusions

Common ownership has become a central issue in recent debates on antitrust policies because the degree of common ownership grew in recent years, and some empirical studies show that it has a large effect on the strategic behavior of companies held by institutional shareholders. As pointed out by Boller and Morton (2019), common owners might have incentives to reduce competition so that industry outcomes such as prices, quantities, capacity, or new product introductions are closer to the monopoly level. In this paper, we analyze the features of the Ontarian cancer drug industry, a highly innovative industry characterized by a concentrated ownership structure. Using data on brands and generics characteristics and ownership information collected from different

sources, we empirically assess the presence of common ownership and quantify which components mainly drive the link between common ownership and market entry. In particular, we show that investors' concentration plays an important role in defining common ownership in the years prior to the entry of a generic in the market.

Common ownership may have anticompetitive effects and be harmful for welfare (Azar et al., 2018). However, the literature on welfare and the policy implications of common ownership is still underdeveloped (Sato and Matsumura, 2019). With the results of this paper, we make the first important step in identifying the target of eventual policy interventions to reduce common ownership, in this industry as well as in other innovative industries characterized by a high level of concentration. A full structural model that incorporates the analyzed features is ideal in order to quantify the drivers of market entry in the presence of common ownership. It will allow to disentangle the economic effects of this phenomenon on the consumers' surplus, industry structure, and government expenditures.

E. Patent Expiration Dates

Drugs DIN and patent expiration dates for a subset of products, as of June 2020. Sources: CIPO and Health Canada.

Drug Name	Brand	DIN	Patent Expiration
Aldesleukin	Proleukin	02130181	29.1.30
Arsenic trioxide	Trisenox	02407833	1.10.18
Bevacizumab	Avastin	02270994	22.2.33
Blinatumomab	Blincyto	02450283	26.11.24
Brentuximab	Adcetris	02401347	31.7.23
Denosumab	Xgeva	02368153	25.6.22
Ipilimumab	Yervoy	02379384	2.5.26
Nab-Paclitaxel	Abraxane	02281066	21.2.26
Nivolumab	Opdivo	02446626	2.5.26
Obinutuzumab	Gazyva	02434806	12.8.30
Pembrolizumab	Keytruda	02456869	13.6.28
Pertuzumab	Perjeta	02405016	28.1.29
Plerixafor	Mozobil	02377225	30.7.22
Ramucirumab	Cyramza	02443805	4.3.23
Siltuximab	Sylvant	02435128	26.10.22
Temsirolimus	Torisel	02304104	25.7.23
Trastuzumab emtansine	Kadcyla	02412365	23.6.20
Zoledronic Acid	Aclasta	02269198	18.6.21

Table E.1.: DIN and Patent Expiration Date for a Subset of Drugs

4. Bundling and Past Dependence of Sin Goods among Adolescents

joint with Liana Jacobi and Michelle Sovinsky

4.1. Introduction

The Drug Policy Alliance (DPA), one of the biggest US non-profit organizations in support of marijuana legalization, believes "marijuana should be removed from the criminal justice system and regulated like alcohol and tobacco",[1] thus eliminating once and for all the stigma of illegality for the consumption of the so-called *sin goods*.

Around two-thirds of Americans are in favor of marijuana legalization.[2] Several are the arguments in support of it. Some cannabis advocates stress the potential health benefits, as legalization would create standard requirements for all the marijuana products and promote consumer safety. Others point to the fact that making the drug more easily accessible would drastically reduce the contact with illegal drug dealers and thus decrease the probability of consumption of harder drugs. Another strong arguments is that legal weed would eliminate the black market. As of April 2020, it is legal to sell and buy marijuana for recreational use in Canada, Georgia, South Africa, Uruguay, the Australian Capital Territory in Australia, and eleven states, two territories, and the District of Columbia in the United States.[3]

[1] https://www.drugpolicy.org/issues/marijuana-legalization-and-regulation

[2] Pew Research Center, (2019). Published November 14, 2019 at https://www. pewresearch.org/fact-tank/2019/11/14/americans-support-marijuana-legalization. Accessed 10 March 2020.

[3] Another argument discussed in the policy debate is that homogeneous legalization across states would remove the instability related to the conflicting federal and states law, and this is at the basis of the MORE Act (Marijuana Opportunity Reinvestment and Expungement), which proposes a unification of the law at the federal level. The Judiciary Committee passed the Act in November 2019.

When analyzing the consequences of marijuana legalization, one cannot ignore two important factors. First, marijuana might be consumed with other products, such as alcohol and cigarettes. Between 2007 and 2017, alcohol and cigarettes use declined among US eighteen-years-old adolescents by ten percentage points while marijuana use increased by four percentage points.[4] Second, past use of one of the substances might have consequences for the consumption of that substance and other *sin goods*, especially if one considers complementarity in consumption. Addiction is a severe problem among adolescents. In 2017, for instance, around ten percent of young Americans aged 18 to 25 was diagnosed with an alcohol use disorder, defined as alcohol abuse or dependence, while nearly six percent had some sort of marijuana dependence.[5]

These pieces of evidence pose several important policy questions. Does the consumption of marijuana affect the consumption of the other *sin goods*? What happens when one considers the potentially addictive nature of the substances?

In this paper, we analyze the potential complementaritarities in use when individuals choose to consume bundles containing marijuana, alcohol, or cigarettes, taking into account persistence in behavior. In particular, we develop and estimate an economic model of multi-use of substances that incorporates complementarities in use and allows for habit formation. We then test our model empirically, by combining different data sources: pre-legalization time series data on multi-substance use from the Panel Study of Income Dynamics (PSID) survey for adolescents living in the United States, together with pricing data for marijuana, alcohol, and cigarettes, collected from different sources.

Our parameter estimates show that it is important to account for correlation across unobservables and persistence in behavior when analyzing the decision of using the *sin goods* in combination. Moreover, we find that the past use of a substance influences not only its current use but also the decision of using the substance together with other substances. These results provide insightful information on the long-run effect of legalizing marijuana for the consumption of potentially complementary addictive products, such as alcohol and cigarettes.

The paper refers to several strands of literature. In a related project, Allocca, Jacobi, and Sovinsky (2020) study the impact on youth of multi-substance use accounting for

[4]https://www.hhs.gov/ash/oah/adolescent-development/substance-use/marijuana/index.html

[5]National Survey on Drug Use and Health. https://www.samhsa.gov/data/sites/default/files/ cbhsq-reports/NSDUHNationalFindingsReport2018/NSDUHNationalFindingsReport2018.pdf

persistence within a dynamic structural model. Zhao and Harris (2004) estimate a joint participation model for cannabis, alcohol, and tobacco use in a static reduced-form framework, while the majority of the epidemiological literature that looks at the health implications of substance use relies on cross-sectional data and estimates reduced-form single equations models. Other studies have exploited policy changes such as change in minimum drinking age (Crost and Guerrero, 2012; DiNardo and Lemieux, 2001), introduction of marijuana for medical purposes (Wen, Hockenberry, and Cummings, 2014) or tax reforms (Pacula, 1998) to identify substitution or complementary relationships between substance uses.

Papers in the literature have looked at the relationship between substances and have found mixed evidence regarding the complementarities. For example, Crost and Guerrero (2012), Cameron and Williams (2001), and DiNardo and Lemieux (2001) find evidence that marijuana and alcohol are substitutes, while Zhao and Harris (2004), Wen et al. (2014), and (Pacula, 1998) find evidence that they are complements. Regarding marijuana and tobacco, there are fewer studies, with Zhao and Harris (2004) and Cameron and Williams (2001) finding a complementary relationship.[6]

This paper also refers to the literature that estimates complementarities within a structural framework, starting from Gentzkow (2007), who studies complementarity among newspapers. More recently, Thomassen, Smith, Seiler, and Schiraldi (2017) develop a multi-category multi-seller demand model to study the transportation costs in grocery shopping; Ershov, Laliberté, and Orr (2018) analyze the complementarity between soft carbonated drinks and potato chips, while Fosgerau, Monardo, and De Palma (2019) look at the complementarity between brands of RTE cereals. Moreover, Iaria and Wang (2020) provide instrument-free identification and estimation methods that solve the challenge of dimensionality related to bundling.

Several empirical studies have analyzed the effect of an early experimentation with legal substances such as alcohol or tobacco on the later consumption of more addictive illicit drugs (Kandel and Faust, 1975). Kenkel, Mathios, and Pacula (2001) underline the importance for policymakers of understanding whether reducing the demand for one drug has effects on the current and future use of other drugs. Pacula (1998) uses a multi-commodity habit formation model to show for instance that prior use of alcohol

[6]For a recent epidemiological survey of the literature about the outcomes associated with the use of cannabis and smoking as well as cannabis and alcohol, see Schlienz and Lee (2018).

and cigarettes increases the likelihood of current use of marijuana, while Pierani and Tiezzi (2009) do not find evidence that past consumption of alcohol reinforces current consumption of tobacco (and vice versa).

We contribute to the previous research in several ways. First, we present an economic model to analyze the potential complementarities in use of substances. Second, within this multi-use framework, we allow for persistence in behavior due to the potentially addictive natures of the products. Third, we combine different data-sources that are relevant for the empirical analysis. From a policy perspective, this paper represents the first step to study the impact of the legalization of an illegal substance such as marijuana on the contemporaneous and future consumption of the *sin goods*.

The paper is structured as follows. Section 2 describes the background and the datasets used. We present the model and the estimation methodology in Sections 3 and 4, and the identification strategy in Section 5. The results are reported in Section 6. Section 7 concludes.

4.2. Background and Data

The consumption of *sin goods* among adolescents is of primary importance for the policy debate. Several surveys have shown that in 2013 alcohol and tobacco were the drugs most commonly used by adolescents, followed by marijuana.[7]

For the empirical analysis, we use data from two primary sources. The first are individual-level panel data from the Transition into Adulthood Supplement (TAS) of the Panel Study of Income Dynamics (PSID) survey. This is a longitudinal household survey of US families conducted at the University of Michigan. The TAS began in 2005 and it is run biennially to collect information on young adulthood transitions in schooling, work, and family formation, including consumption behavior. The sample consists of respondents between 18 and 27 years old from all the states in five different waves: 2005, 2007, 2009, 2011, 2013. The panel data is required for the identification of the degree of complementarity/substitutability among substances, as discussed momentarily. The second source are pricing data for marijuana, alcohol and cigarettes for the same time frame. These data are collected from state-level administrative tax data and month-

[7]Monitoring the Future survey, National Survey on Drug Use and Health.

city-level transaction data. In the following subsections, we discuss the data sources in more detail and provide a description of marijuana regulation across the different states.

4.2.1. PSID Survey

The TAS of the PSID survey includes around 7,000 observations between 2005 and 2013. After dropping missing information, the sample consists of 6,440 observations over the years for about 2,900 different individuals. The data include individual characteristics of the respondents, as well as information on consumption decisions for the three products.

Table 4.1 gives summary statistics of demographic characteristics. Individuals age over time, as the average age goes from 21 years old to 23. The sample is almost evenly distributed between gender and consists of about 47 percent non-hispanic whites. Most of the individuals live in a metropolitan area (78%) and they report on average being in very good health. A large proportion of respondents have a high school degree and the percentage of adolescents going to University increases over time. Very few respondents are married in 2005, though this percentage increases substantially between 2005 and 2013, as expected. Moreover, we control for years of parental education. About 58% of the households involve parents with only a high-school degree, while 15% have a least one parent who has a university degree or higher.

	All years	**2005**	**2007**	**2009**	**2011**	**2013**
Age	21	19	20	21	22	23
Male	47%	46%	46.7%	46.3%	47.6%	47%
White and non-hispanic	47%	48%	47.5%	47%	45.8%	46%
Living in a Metropolitan Area	77.5%	75.7%	76.3%	77.5%	78.8%	77.8%
Health Quality (1 to 5)	3.8	3.79	3.81	3.81	3.79	3.75
Max education: High School	74.5%	85%	81.2%	75.5%	72%	66%
Max education: University	16.5%	1.7%	7.5%	14.9%	20.4%	27.3%
Married	9%	3%	5%	8%	10.7%	13.8%
Max Parents Education: High School	58%	60%	58.5%	58.6%	58%	57.3%
Parents Education: University or more	15%	14%	15%	15.6%	15%	15.6%
Observations	*6,440*	*695*	*1,078*	*1,493*	*1,796*	*1,378*

Table 4.1.: Summary Statistics Demographics

The purpose of the analysis is to study the consequences of the consumption of marijuana, alcohol, and tobacco among adolescents. The survey asks several questions regarding the consumption of these substances (the questions are reported in Appendix F). The binary variables used in this paper are constructed from the questions on consumption in the year of the interview. As table 4.2 shows, the average proportion that used marijuana in the past year remains steady at around 30 percent, with a similar trend of tobacco use. Alcohol use increases over time, growing from 62 to 82 percent. More interestingly, white non-hispanic respondents use of marijuana decreases between 2005 and 2013, and a similar path can be observed for adolescents above 20, despite the increase in age over time.[8] Whites consume more alcohol and tobacco relative to the average consumption, but the consumption of cigarettes decreases over time. This evidence suggests that demographics play a role in the consumption of these products.

	2005	**2007**	**2009**	**2011**	**2013**
% Used Marijuana in the Last Year					
On average	31	26	27	29	30
Male	37	31	33	36	38
White Non-Hispanic	38	33	32	31	31
Above 20	31	26	27	28	28
% Used Alcohol in the Last Year					
On average	62	66	70	74	82
Male	64	65	70	68	74
White Non-Hispanic	75	73	75	75	83
Above 20	77	73	76	73	74
% Used Tobacco in the Last Year					
On average	24	25	23	21	20
Male	26	27	27	24	23
White Non-Hispanic	29	28	25	22	20
Above 20	23	30	25	24	21
Observations	*695*	*1,078*	*1,493*	*1,796*	*1,378*

Table 4.2.: Descriptive Statistics by Use (in %).

Another key aspect of these substances is their addictive nature, which may create persistence in consumption. Table 4.3 presents for each year the percentage of respondents that consumed a substance also in the previous wave of interview. One can see

[8]The minimum legal drinking age in the US is 21.

that around half of those that use any of the products between 2007 and 2013 consumed them also in the previous wave, showing a persistent behavior that raises overtime for all the products. This evidence suggests that habit formation is an important feature that needs to be taken into account when estimating the consumption behavior for the substances.

	Year			
Used in t-1 as % Use in t	2007	2009	2011	2013
Marijuana	46	42	49	63
Alcohol	53	55	63	73
Tobacco	45	50	64	69

Notes: The total numbers of individuals reporting the consumption of the substances in 2007, 2009, 2011 and 2013 are respectively: 275, 403, 517, 411 for marijuana; 659, 957, 1162, 979 for alcohol; 267, 339, 374, 268 for tobacco.

Table 4.3.: Persistent Use (Used in the Last Wave as % of Users in the Previous Year).

Table 4.4 points to the presence of multi-use of substances. Comparing table 4.2 and table 4.4, less than half of the individuals who drink alcohol only use this substance. The majority who use alcohol either use it with cannabis (14 percent), with cigarettes (7 percent), or both (10 percent). Multi-substance use is even more pronounced for cannabis use. For example in 2005, 31 percent report using cannabis, but only 2 percent use it in isolation. Almost all cannabis users use it in combination with either alcohol or alcohol and cigarettes. The trends over time for the bundles containing marijuana are depicted in Figure 4.1. Notice that the use of marijuana consumed alone or with another substance slightly increases or remains substantially stable between 2005 and 2013, while the joint use of the three products (dot line) has a downwards slope. These simple descriptive statistics indicate the importance of considering consumption decisions for marijuana in combination with other substances.

Let us look at the correlation coefficients between dummies for consuming pairs of substances, displayed in table 4.5. The raw correlations between the substances are significantly positive at the 0.5 percent level. This suggests that an adolescent who smokes is on average more likely to consume alcohol and marijuana. However one cannot conclude from this that the substances are complement. It can be that those consumers that like smoking cigarettes like also drinking and using marijuana. Simple

% **Use in the Last Year**	All years	Year 2005	2007	2009	2011	2013
No Products	27.5	31	32	28	27	22
Marijuana Alone	2.4	2	1.8	2.1	3	2.7
Alcohol Alone	34	27.8	30.1	34	34.2	39.7
Tobacco Alone	3.2	3.4	3.8	3.3	3.4	2.3
Marijuana and Alcohol	14.2	15.7	11.7	12.9	14.8	16
Alcohol and Tobacco	7	7	9	7.4	6.4	6.1
Marijuana and Tobacco	1.8	1.4	1.7	2.1	1.8	1.9
All Products	9.8	11.8	10.3	9.8	9.2	9.2
Observations	*6,440*	*695*	*1,078*	*1,493*	*1,796*	*1,378*

Table 4.4.: Multi-Products Use (in %).

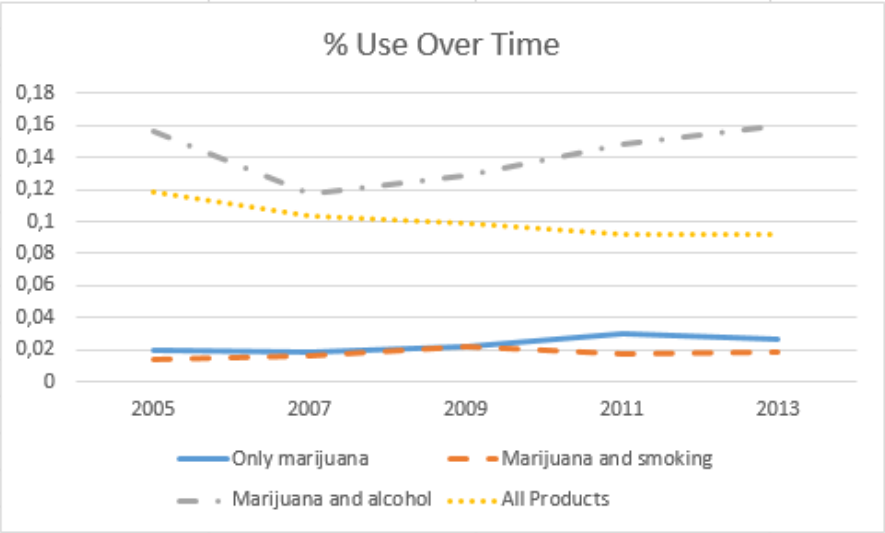


Figure 4.1.: Percentage Use Bundles Marijuana Over Time

evidence is given by the fact that for alcohol and tobacco the correlation disappears for individuals with certain demographic characteristics (white, male and over 20) and for marijuana and alcohol decreases drastically for individuals in the same group, while for alcohol and tobacco the correlation is reduced by half for individuals with a university degree. Once again, this shows that controlling for demographic characteristics is important when evaluating the complementarity/substituability between products. It

remains to see whether the rest of the correlation is due to complementarity or to unobserved tastes.

Correlations	Marijuana/Alcohol	Alcohol/Tobacco	Marijuana/Tobacco
Raw	0.2621*	0.1334*	0.2931*
White, Male, Over 20	0.0949*	-0.0025	0.267*
Max Education: University	0.2016*	0.0692*	0.2654*
Living in a Metropolitan Area	0.2702*	0.1355*	0.291*

Table 4.5.: Pearson Correlation Coefficients (significant 0.5 %)

4.2.2. Marijuana Regulation

There are important distinctions between marijuana decriminalization and legalization. Decriminalization consists of removing criminal penalties imposed for personal marijuana use. However, the production and sale of marijuana remain illegal, therefore it is still necessary to seek out suppliers in order to purchase the drug.[9] With legalization instead, laws banning the possession and personal use of marijuana are abolished or lifted. More importantly, the government can regulate and tax marijuana use and sales. The purpose of the analysis is to study the consequences of marijuana legalization. The sample period ends in 2013 and it does not cover the years of legalization. The data collected contain also state-level information on decriminalization.

In the US, the legalization of marijuana for recreational use started in 2012 with Colorado and Washington.[10] Alaska, Oregon and District of Columbia, implemented the same law in fall 2014, followed by Ohio in 2015. On November 8, 2016, adult-use recreational marijuana was approved in four other states (California, Maine, Massachusetts and Nevada). In 2018, Michigan voters approved to legalize, regulate, and tax marijuana in the state; in the same year, Vermont legalized marijuana for adult use, and the law

[9] As of April 2020, the non-medical use of cannabis is decriminalized in 16 states, whereas 9 states have decriminalized and later legalized it. Oregon was the first state to decriminalize marijuana in 1973, followed by Alaska, Maine, Colorado, California and Ohio in 1975. Decriminalization laws passed in Minnesota (1976), Mississippi (1977), New York (1977), North Carolina (1977), and Nebraska (1978), so that by the end of the 70s, 11 states in total decriminalized marijuana. A second wave of decriminalization began in 2001 with Nevada, followed by Massachusetts (2008), Connecticut (2011), Rhode Island (2012), Vermont (2013). Between 2014 and 2019, the District of Columbia, Maryland, Missouri, Delaware, Illinois, New Hampshire, New Mexico, North Dakota and Hawaii passed the law. In 2020, Virginia has adopted the same law.

[10] For those States, the actual sales started in 2014.

became effective in July. Very recently, on January 1, 2020, marijuana sales became legal in Illinois.

Moreover, twenty-one States considered bills for marijuana legalization in 2018.[11] The MORE (Marijuana Opportunity Reinvestment and Expungement) Act, which the Judiciary Committee passed in November 2019, proposes unified federal legislation to legalize cannabis. This suggests that in the near future more states will move in the direction of legalizing marijuana.

4.2.3. Prices

The pricing data for marijuana, alcohol, and cigarettes come from multiple sources. For alcohol and cigarette prices, we use administrative tax data. In particular, we use state exercise taxes placed on beer in the state of the individual interviewed, as beer is one of the most used alcoholic beverages among adolescents. All taxes are in dollars per gallon and are obtained from the Tax Foundation.[12] Cigarette prices are computed as a percentage of retail prices. We use the weighted (by market share) average prices per package.[13]

Marijuana prices are more difficult to find as marijuana is an illicit drug. We combine information from two different data sources, the High Times and PriceOfWeed.com. The High Times is a monthly magazine that publishes the prices of marijuana with the corresponding strain for multiple cities and states across the US.[14] These data are from 2004 to 2011. PriceOfWeed.com is a website that "crowd-sources" the street value of marijuana from the consumers. Site visitors anonymously input information about their most recent marijuana transactions: amount purchased, price and quality, choosing from low, medium, or high, together with data on the city, state, and country where the transaction took place.[15] These data are from September 2010 to September 2013.

[11]National Conference of State Legislators. Published October 17, 2019 at https://www.ncsl.org/research/civil-and-criminal-justice/marijuana-overview.aspx.

[12]https://taxfoundation.org/state-sales-gasoline-cigarette-and-alcohol-tax-rates

[13]Generic brands are included in the average calculation. Column 2 of Tables 13B, *The Tax Burden on Tobacco*, 2014. Each value is calculated as of November 1 of the corresponding year.

[14]This is in the section Trans High Market Quotations.

[15]Data source used also in Davis, Geisler, and Nichols (2016). In particular, we use a cleaned version of the data available on https://github.com/rlucioni/viz/tree/master/marijuana/data.

Given that the price of marijuana can vary a lot by quality, we include an index to capture the quality of the marijuana products in each state and year of observation. This index is constructed in the following way. For the data collected on the High Times, we use the information on marijuana strains to classify the purchase into high quality and low quality. To do so, we collect information on the level of delta-9-tetrahydrocannabinol (or THC) for each strain from multiple websites. THC is the major psychoactive chemical compound in marijuana. The amount of THC absorbed by marijuana users differs according to the part of the plant that is used (e.g., leaf, head), the way the plant is cultivated (e.g., hydro), and the method used to imbibe the drug.[16]

We classify marijuana strains with a level of THC above 15 percent as high-quality. When the level of THC is below 15 percent, we classified them as low-quality. For instance, skunk, haze, kind and crack contain a high level of THC while schwag and mids have a low level of THC.[17] Out of 1,454 unique strains, 1,153 are classified in terms of quality, and the remaining 301 are dropped from the sample as the amount of THC contained cannot be determined.[18]

To be consistent with the above-mentioned classification, for the data collected from PriceOfWeed.com, the transactions of medium quality are redefined to be low-quality transactions, so that we have purchases of low and high marijuana quality for the whole sample.[19] Based on this information, we construct an index of high-quality cannabis given by the percentage of high-quality marijuana purchases for each state and year of observation.

The combined dataset consists of around 138,000 observations for marijuana prices and quality, from 2005 to 2013. In our estimation, we use the price per gram for each state/year combination, by averaging over the months and cities.[20] Unfortunately, we

[16]These include the following websites (where we include only the portion before ".com") weedsmokersguide, marijuana-strains, organicann, dutch-passion, marijuana-seeds-weed, wikileaf, seedfinder, allbud, thebcsc, budderweeds, urbandictionary, leafly, cannasos, buddyboybrands, cannafo, cannabisreports, wikipedia, coloradocannabistours, libertycannabis, marijuanabreak, naturalcannabis, and medicalmarijuanastrains.

[17]Low-THC marijuana strains include also reggie, commercial, ditch weed, dirt, brick bud, shake, wack, bunk.

[18]These observations represent only the 0.3% of the entire sample.

[19]Because of some inconsistencies in prices and quality classification, we changed around 120 observations from low/medium into high quality.

[20]Observations include amounts in eighths (3.5 grams), quarters (7 grams), half ounces (14 grams), and ounces (28 grams). As pointed out by Davis et al. (2016), the limited range of reported quantities minimizes the price differences related to quantity discounts such as those found by Clements (2006).

do not observe prices in all years and states. To deal with missings, we use linear interpolation when we observe the prices in other years for the same state.

In table 4.6 we show descriptive statistics for cigarette prices, alcohol tax rates, marijuana prices, and the ratio of high-quality marijuana, across all states and years of observations. The average marijuana price is 12 dollars per gram and on average around 70% of marijuana transactions involved are reported as high quality.

Variable	**Mean**	**St. Dev.**	**Median**	**Min**	**Max**	*Observations*
Cigarettes Price ($/package)	5.12	1.38	4.94	3.04	10.03	*250*
Beer Tax ($/gallon)	0.29	0.25	0.2	0.02	1.17	*250*
Marijuana Price ($/gram)	12.06	2.42	11.68	1.76	21.16	*250*
Ratio High Quality Marijuana	0.72	0.22	0.7	0	1	*250*

Notes: Each observation is a combination of state and year. The District of Columbia is excluded.

Table 4.6.: Descriptive statistics prices and marijuana quality

The trend in the average marijuana price is shown in figure 4.2. The average price ranges from 11.2 to 13.5 price per gram between 2005 and 2013. One can see that the price decreases after 2009. This most likely reflects the quality trend, depicted in figure 4.3. Indeed, starting in 2009, there is a drop in the average ratio of high-quality marijuana.

Figure 4.2.: Marijuana Price

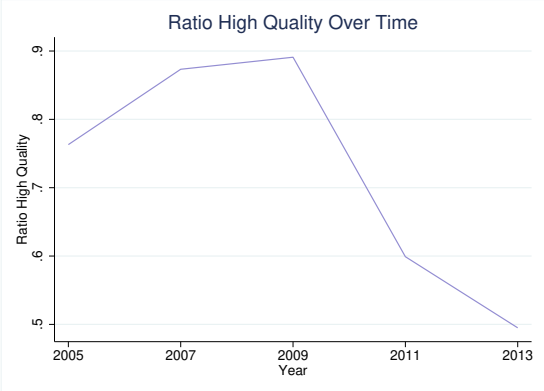


Figure 4.3.: High Quality Marijuana

Distinguishing between prices for high and low-quality marijuana (figure 4.4), one can see that the average price for high-quality marijuana (yellow line) exhibits a down-

ward trend over time, whereas the contrary is true for the low-quality marijuana average price (dark line).[21]

Figure 4.4.: Low and High Price Marijuana

4.3. Model

We model an individual's choice to consume marijuana, alcohol or cigarettes (and possible combinations of these products) over time, in the spirit of Gentzkow (2007). Differently from Gentzkow (2007), we incorporate persistence in consumption behavior because of the potentially addictive nature of the substances.

Individual i, for $i = \{1, ..., N\}$, chooses whether to consume good $j \in \{1, ..., J\}$ and whether to consume this product together with other products in market m, for $m = \{1, ..., M\}$. Denote the set of consumption bundles of product j at time t by $r_t \in \{0, ..., 2^J - 1\}$ where the bundles are ordered such that $r_t = 0$ refers to no consumption (i.e., an empty bundle) and $r_t \in [1, J]$ refers to a bundle that contains only good $j = r$. The indirect utility of i consuming j in market m at time t is given by

$$\overline{u}_{ijmt} = -\alpha_j p_{jmt} + 1(j \in r_{t-1})\delta_j + D_{it}\pi_j + X_{mt}\lambda_j + v_{ij}. \tag{4.1}$$

[21] Allocca et al. (2020) exploit the quality information contained in this reach dataset to construct an empirical price distribution (in the spirit of Jacobi and Sovinsky (2016)) for low and high-quality marijuana price that an individual faces. This method adds variation in terms of individual probability of use, based on the quality available in the market.

In this expression, p_{jmt} is the price of good j. Notice that the market prices for cigarettes and alcohol are observables; given that marijuana is an illicit product, we use the average marijuana prices by state from the data gathered.

An individual's decision to consume *sin good* j depends also on whether she consumed it last period (captured by $1(j \in r_{t-1})$). Therefore, the vector δ_j can be considered as a proxy of addiction that may be different for different substances and it is constant across time. We assume that the consumption of product j can be directly influenced by the past consumption of the same product but not by the past consumption of other products. However, the consumption of bundles containing two or three products can be influenced by the past consumption of each of the products in the bundles.

The vector D_{it} includes observed demographic characteristics which can influence use, such as gender, age, education variables, and health status. The parameter vector π_j allows for demographics to influence use differently depending on the substance. X_{mt} includes all the market-specific variables, the year in which the substance was purchased, the quality of marijuana available in the market, and whether marijuana is decriminalized in that market (this is important as there are different levels of enforcement for substance-related crimes and these may change across years). The parameter vector that captures the impact of the market variables, λ_j, varies by substance.

The idiosyncratic term v_{ij} represents the unobserved heterogeneity that may influence a person's choice of smoking cigarettes, drinking, or using marijuana.[22] Because unobserved heterogeneity is person-specific and likely to be correlated across the three products, we assume that the random effects are distributed tri-variate normal with symmetric covariance matrix.

Notice that our model allows for persistence in consumption so, as in all models of dynamics, we need to consider the initial period of consumption. One way to address this is to allow the mean of the unobserved effects to be a function of some data that are relevant in an initial period (for example, the age at which the individual first tried marijuana). We discuss both of these issues in detail in the following section.

[22] Gentzkow (2007) introduces also a time-varying idiosyncratic term to rationalize the remaining unexplained variation. In our context, this term could potentially capture time-specific shocks which affect the consumption of all the substances in the same way, like for instance a yearly shock in the health condition of the respondent. Notice that we include in our regression a variable related to health, which can partially explain this variation. The potentially unexplained time variation still present will be captured by a time-individual-bundle specific error term.

Individuals obtain an indirect utility from consuming the goods in combination. Again, following Gentzkow (2007), the utility an individual obtains from consuming bundle r is given by

$$u_{irmt} = \sum_{\mathcal{S}_j \in r_t} \overline{u}_{ijmt} + \gamma_r + \epsilon_{irmt} \tag{4.2}$$

where the $\mathcal{S}_j \in r_t$ denotes the set of goods containing j that belong to bundle r. We assume that the idiosyncratic terms ϵ_{irmt} follow a normal distribution. Notice that the consumption of bundles containing two or three products can be influenced directly by the past consumption of each of the products in the bundles. There may be heterogeneity that we do not observe in the data that influences choices and has a persistent nature. This is captured by the idiosyncratic terms. The utility from the outside option of consuming none of the products is $u_{i0mt} = \epsilon_{i0mt}$ where all non-stochastic terms are normalized to zero because we cannot identify relative utility levels.

We are interested in measuring the substitutability among alcohol (A), cigarettes (C), and marijuana (M). Notice that in this framework, we assume that each individual has free access to marijuana. Denote by γ_r the interaction term that relates to the substitutability among the products in bundle r.[23]

The interaction terms between pairs are γ_{AM}, γ_{AC}, γ_{MC} and the interaction among all three goods is given by γ_{ACM}. Gentzkow (2006) shows (under certain regularity conditions) that the substitutability between two goods (say A and C) can be expressed as the sum of a direct component γ_{AC} and an indirect component which has the same sign as the product $\gamma_{AM}\ \gamma_{MC}$.

When γ_{ACM} is negative (positive) the two goods are substitutes (complements); when it is zero the two goods are independent. Substitutability between the pairs $\{A, M\}$ and $\{M, C\}$ is measured by an analogous sum of the direct and indirect terms.

4.4. Econometric Methodology

At each time period a subject can consume any combination of marijuana, alcohol and tobacco, or none. Denote by $y_{itm} = r$ a categorical outcome variable that reflects consumption of bundle r of products from $j \in J$, where $J = \{1(m), 2(a), 3(t)\}$ is

[23] When goods are not consumed in combination $\gamma_r = 0$.

the set containing the products marijuana, alcohol and tobacco. To simplify the discussion of the econometric model, define the set of consumption bundles $r \in R$ as $R = \{1, 2, 3, 4, 5, 6, 7\}$, where

$$R = \begin{cases} 1 & \text{marijuana} \\ 2 & \text{alcohol} \\ 3 & \text{cigarettes} \\ 4 & \text{marijuana, alcohol} \\ 5 & \text{marijuana, cigarettes} \\ 6 & \text{alcohol,cigarettes} \\ 7 & \text{marijuana, alcohol, cigarettes} \end{cases} \tag{4.3}$$

and $r = 0$ is the base category of no substance consumption. The set R^M define the bundles containing marijuana and the restricted set R^R of bundles without marijuana use.

The observed choice $y_{itm} = r$ is modelled in terms of the bundle utilities u_{irmt} as

$$y_{itm} = \begin{cases} r & \text{if } u_{irmt} > max\{u^{-r}_{irmt}, 0\} \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

where u_{irmt} is defined as in equation (1) in terms of the product specific utilities $\bar{u}_{ijmt}$ of the products contained in bundle r and bundle effects γ_r for bundles with more than one product. For simpler notation, let us rewrite the latent utility for bundle r as

$$u_{irmt} = \sum_{j \in r} (\mathbf{x}'_{imt}\boldsymbol{\beta}_j + p_{ijmt}\alpha + \nu_{ij}) + I[r > J]\gamma_r + \epsilon_{irmt} = \mu_{irmt} + \varepsilon_{irmt} \tag{4.5}$$

where $\mathbf{x}_{imt} = (I(j \in r_{t-1}), D_{it}, X_{mt})$ and $\boldsymbol{\beta}_j = (\delta_j, \pi_j, \lambda_j)$. The γ_r parameters are mean shifter and adjust the intercept either up or down reflecting that goods are complements or substitutes respectively. Note that X_{mt} contains all but first year controls so that we can include a constant as first element of the demographics vector D_{it} and the intercept as first element of π_j. The unobservables terms ν_{ij} capture time constant consumer characteristics. Following Gentzkow (2007), we assume that the $J \times 1$ vec-

tor of individual effects follows a multivariate Normal distribution $\nu_i \sim N(0, \Omega)$. For identification purposes, we restrict the covariance matrix to be

$$\Omega = \begin{pmatrix} 1 & \sigma_1 & \sigma_2 \\ \sigma_1 & 1 & \sigma_3 \\ \sigma_2 & \sigma_3 & 1 \end{pmatrix}$$

The inclusion of random effect gives rise to initial conditions problem as one cannot assume that the decision to consume a substance at $t = 0$ is uncorrelated with the unobserved individual effects $\nu_i = \{\nu_{i1}, \nu_{i2}, \nu_{i3}\}$. We follow Wooldridge (2002) that builds upon Chamberlain (1984) and model the mean of the random effects as a function of initial value y_{i0} and time-invariant covariates.[24]

We define a joint distribution of the unobserved random effects for good ν_i as

$$\nu_i \sim N(\mu_i^{\nu}, \Omega) \tag{4.6}$$

where the means are defined as $\mu_{ij}^{\nu} = \gamma_{j1} y_{i0} + \bar{x}' \gamma_{j2}$ and depend on the starting age for the use of substance j, y_{i0}, and the average of time-varying demographics/controls over the sample $\bar{x}$. For marijuana consumption, y_{i0} includes the age when i started using marijuana. However, the PSID does not contain such information for alcohol and smoking, so we include the behavior observed in the first period of the data for these two substances.

Under the multinomial probit model, the $R \times 1$ vector of latent bundle utilities $\mathbf{u}_{imt} = (u_{i1mt}, u_{i2mt}, ..., u_{iRmt})$ for subject i is then given by

$$\mathbf{u}_{imt} = \boldsymbol{\mu}_{imt} + \boldsymbol{\epsilon}_{imt} \ , \quad \boldsymbol{\epsilon}_{imt} \sim N(0, \Sigma) \tag{4.7}$$

with the mean vector defined as $\boldsymbol{\mu}_{imt} = (\mu_{1rmt}, ..., \mu_{7rmt})$. The error vector $\boldsymbol{\epsilon}_{imt}$ follows a multivariate Normal distribution with covariance matrix Σ. Since we capture the correlation across bundles via product-specific individual random effect, Σ is a diagonal

[24] For instance, Wooldridge (2005) discusses how the random component of the fixed effect then can be integrated out to yield the likelihood function of the random effects probit model.

matrix $\sigma * I_{R\times R}$. For identification purposed we set $\sigma_{11} = 1$ (see for example Geweke and Keane 1994). Under these assumptions, the likelihood in terms of latent utilities is

$$f(u_{mt}, y_{mt}|\theta, \nu) = \prod_i \prod_t \int_{A_r} N(u_{imt}|\boldsymbol{\mu}_{imt})\, d\,\boldsymbol{u}_{imt}$$

where the integration region is given by $A_r = \{u_{irmt} > max\{0, u_{i-rmt}\} : r \in R\}$.

4.5. Identification

Imagine we observe consumers using marijuana, cigarettes and alcohol together. One possibility is that these products are complements. Another could be that an individual has unobserved tastes for marijuana, alcohol and cigarettes that are correlated, such as a taste for *feeling high*, which is constant over time. Gentzkow (2006, 2007) provide an extensive discussion of how to separately identify the substitution/complement parameter (γ) from the unobserved covariance.

Exclusion restrictions, namely something that impacts the utility of one product but not the other product, is one source of identification. In our analysis, prices represent valid exclusion restrictions. For example, if the fact that many individuals consume marijuana together with alcohol is driven by complementarity, the prices will influence the consumption of these products.

Another source of identification comes from the panel data. We assume that correlation in the unobservable tastes do not vary over time,[25] therefore the panel data aspect of the PSID allows us to separate complementarity in purchases from correlation in unobservables. The idea is that the choices in previous time periods will be more correlated with the choices at time t the larger the variance of the random effects.

A common concern in discrete choice models of product choice is that there is some component not included in the utility function (such as quality) which enters the error term. To the extent that prices are correlated with quality, this presents a potential endogeneity problem.

[25] This assumption might sound non-obvious for bundles with marijuana, given that the unobservable taste may vary over time because of changes in the regulation. In our analysis, we control for these changes, including dummies for decriminalization or legalization.

We have three products with associated prices - alcohol, cigarettes, and marijuana. Price endogeneity is less of an issue in our framework. First, we are modeling the decision to consume a substance, but not which brand of the substance. Given that, the price is applicable to all brands of that substance, namely the tax rate, and it is plausibly not correlated with quality. Second, for marijuana we include a control for quality in the consumer's utility function, using the information we gather on the amount of high-quality marijuana observed in each year/state observation.

4.6. Results

In this section, we present results from several regressions where we show the importance of past consumption and we measure the degree of substitution across products.

4.6.1. Preliminary Regressions

We start by analyzing the results from preliminary regressions where we do not consider the use of the substances in combination. The dependent variables are dummies for the consumption of alcohol, tobacco and marijuana. We include product-level characteristics, demographic characteristics, and fixed effects.

Table 4.7 presents results from standard probit regressions. Notice that we do not take into consideration the panel nature of the data, and we treat every observation as independent. Nevertheless, the coefficient estimates provide information on what may impact the use of the three substances.

In particular, for each product, specification (1) includes product and individual characteristics and year fixed effects. Not surprisingly, alcohol and cigarette prices affect negatively the probability of drinking and smoking respectively. The price of marijuana comes out not significant. This is probably related to the fact that we do not observe individual prices.[26] Male and white non-hispanic are more likely to consume a drug, while the contrary holds for married people. All demographics influence in the same direction the use of marijuana, alcohol and tobacco, except for the education variables:

[26] Allocca et al. (2020) deal with this issue by simulating individual prices using empirical price distributions for high and low-quality marijuana.

for marijuana and tobacco, having a high school or a university degree impacts in a significantly negative way the probability of using the substances, whereas for alcohol this impact is significantly positive.

In specification (2) we exclude the variables related to individual education and include parental education. One can see that having a parent with a university degree increases the probability of drinking but decreases the probability of smoking. Specification (3) includes all the regressors used in the previous specifications and further controls for region fixed effects. There are no significant changes relative to the previous specifications.

	Marijuana Use			**Alcohol Use**			**Tobacco Use**		
Variable	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Product Charachteristics									
Marijuana Price	-0.001	-0.003	-0.009						
	(0.01)	(0.01)	(0.01)						
Alcohol Price (Beer Tax)				-0.295***	-0.284***	0.067			
				(0.0727)	(0.073)	(0.09)			
Cigarette Price (Package)							-0.076***	-0.07***	-0.047*
							(0.0197)	(0.02)	(0.026)
Marijuana Decriminalization	0.09***	0.075***	0.081**						
	(0.036)	(0.036)	(0.039)						
Ratio High Quality Marijuana	0.237	0.26	0.295						
	(0.16)	(0.16)	(0.19)						
Demographics									
Age	0.02***	-0.01	0.02***	0.124***	0.141***	0.13***	0.104***	0.063***	0.1***
	(0.008)	(0.008)	(0.008)	(0.0086)	(0.0082)	(0.0087)	(0.009)	(0.0085)	(0.009)
Male	0.39***	0.405***	0.39***	0.278***	0.244**	0.268***	0.21***	0.266***	0.23***
	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)	(0.037)	(0.037)	(0.037)
White Non-Hispanic	0.347***	0.294***	0.321***	0.5953***	0.613***	0.527***	0.412***	0.322***	0.45***
	(0.035)	(0.035)	(0.037)	(0.0362)	(0.0361)	(0.037)	(0.0394)	(0.039)	(0.042)
Health Quality (1 to 5)	-0.134***	-0.147***	-0.136***	-0.096***	-0.086***	-0.103***	-0.247***	-0.27***	-0.24***
	(0.018)	(0.019)	(0.018)	(0.019)	(0.019)	(0.019)	(0.02)	(0.02)	(0.02)
Married	-0.739***	-0.73***	-0.73***	-0.447***	-0.426***	-0.422***	-0.173***	-0.195***	-0.22***
	(0.074)	(0.074)	(0.074)	(0.0625)	(0.062)	(0.063)	(0.07)	(0.067)	(0.069)
Max education: High School	-0.296***		-0.305***	0.284***		0.243***	-0.872***		-0.82***
	(0.058)		(0.058)	(0.0587)		(0.06)	(0.058)		(0.059)
Max education: University	-0.511***		-0.528***	0.593***		0.527***	-1.71***		-1.59***
	(0.075)		(0.075)	(0.076)		(0.077)	(0.085)		(0.086)
Max Parents Education: High School		0.024	-0.006		-0.097**	-0.055		0.477***	0.39***
		(0.04)	(0.04)		(0.04)	(0.041)		(0.046)	(0.047)
Parents Education: University or more		0.04	0.05		0.25***	0.238***		-0.124*	-0.093
		(0.054)	(0.054)		(0.0557)	(0.0564)		(0.066)	(0.067)
Region fixed effects	No	No	Yes	No	No	Yes	No	No	Yes

Number of obs: 6,440. Robust standard errors in parenthesis. All specifications include year fixed effects and a constant. ***p<0.01, **p<0.5, *p<0.1.

Table 4.7.: Standard Probit Regressions

Table 4.8 presents results from dynamic probit models for marijuana, tobacco and alcohol consumption. These regressions use the information on the behavior for all the periods, thus the sample size is reduced (1,945 individuals for which we observe consumption choices for at least two periods of time). For all the substances, past behavior is a positive significant indicator of current behavior, and these effects hold

after controlling for initial conditions à la Wooldridge (2005) (specifications (2) for each substance). This suggests that persistence in behavior is an important factor to consider when analyzing the use of legal and illegal substances. Control variables related to age, gender, race, health condition, and marital status have the same effect found in Table 4.7 on the probability of use. However, in contrast to the results of the previous table, when we include the initial conditions, having a high school degree or a university degree decreases not only the probability of smoking or consuming marijuana but also the probability of drinking.

4.6.2. Multi-Substance Use Regressions

As we are interested in the possible complementarity/substitutability among substances, we then estimate the model of multi-substance use where individuals make choices among all bundles, allowing thus for correlation among errors in choices.

We first run a multinomial probit regression of multi-substance use to estimate the choice probabilities implied by equation 4.2 without considering habit formation. The results of table 4.9 show that substance use is less likely the higher are the prices. Concerning marijuana usage, individuals who live in a decriminalized state (note this is before legalization) are more likely to use marijuana. We also see that demographic characteristics matter but not necessarily in the same direction for all products. For example, having a university degree increases the probability of using alcohol alone or together with marijuana but decreases the probability of choosing the other bundles. Moreover, individuals who have higher educated parents are more likely to drink alcohol or use marijuana but less likely to smoke cigarettes. The constant terms for the bundles of pairs represent the direct substitution effects. For any pair of substances, these effects are negative. However, we cannot conclude that goods are substitutes, as we have to look also at the indirect effect. As discussed earlier, the sign of the indirect effect is given by the sign of the product of the other two interaction terms. In all the cases, the indirect effects are positive. Once we compute the bundle interactions following Gentzkow (2007), we find that all the products in pairs are complements, as the γs are positive.[27]

[27]For each bundle of pairs, these effects are computed by subtracting the estimated constants for the bundles of products consumed alone to the estimated constant of the pair bundles.

	Marijuana Use		Alcohol Use		Tobacco Use	
Variable	(1)	(2)	(1)	(2)	(1)	(2)
Lagged Use						
Marijuana Use Last Period	1.5***	1.09***				
	(0.0526)	(0.086)				
Alcohol Use Last Period			1.14***	0.66***		
			(0.062)	(0.095)		
Smoker Last Period					1.93***	0.573***
					(0.075)	(0.148)
Product Characteristics						
Marijuana Price	-0.018	-0.006				
	(0.018)	(0.022)				
Alcohol Price (Beer Tax)			-0.2	0.216		
			(0.122)	(0.15)		
Cigarette Price (Package)					-0.018	0.004
					(0.035)	(0.06)
Marijuana Decriminalization	0.079	0.068				
	(0.052)	(0.069)				
Ratio High Quality Marijuana	0.507*	0.413				
	(0.271)	(0.332)				
Demographics						
Age	-0.009	-0.0028	0.02	0.0225	0.049***	0.092*
	(0.012)	(0.034)	(0.014)	(0.0336)	(0.014)	(0.05)
Male	0.3***	0.39***	0.224***	0.287***	0.177**	0.33***
	(0.048)	(0.064)	(0.051)	(0.066)	(0.066)	(0.104)
White Non-Hispanic	0.15***	0.238***	0.447***	0.478***	0.17***	0.24**
	(0.052)	(0.071)	(0.0583)	(0.072)	(0.066)	(0.113)
Health Quality (1 to 5)	-0.103***	-0.026	-0.108***	-0.128***	-0.173***	-0.188**
	(0.025)	(0.047)	(0.027)	(0.045)	(0.035)	(0.059)
Married	-0.508***	-0.555***	-0.381***	-0.407***	-0.155**	-0.221
	(0.095)	(0.124)	(0.0752)	(0.091)	(0.08)	(0.138)
Max education: High School	-0.155*	-0.256	0.273***	-0.708***	-0.377***	-0.616*
	(0.081)	(0.224)	(0.084)	(0.224)	(0.117)	(0.32)
Max education: University	-0.37***	-0.591***	0.50***	-0.522***	-0.89***	-0.76**
	(0.099)	(0.257)	(0.101)	(0.251)	(0.177)	(0.37)
Max Parents Education: High School	-0.025	0.015	-0.109*	-0.127*	0.19***	0.227*
	(0.055)	(0.073)	(0.059)	(0.075)	(0.073)	(0.125)
Parents Education: University or more	0.047	0.117	0.241***	0.267**	-0.128	-0.27
	(0.071)	(0.096)	(0.004)	(0.104)	(0.073)	(0.176)
Initial conditions included	No	Yes	No	Yes	No	Yes

Number of obs: 4,486. Number of individuals: 1,954. Robust standard errors in parenthesis.
All specifications include region and year fixed effects, a constant and individual panel-level variance.
***p<0.01, **p<0.5, *p<0.1.
The initial conditions specifications include the mean over time of all time varying regressors.

Table 4.8.: Dynamic Probit Regressions

Variable	*Only Marij*	*Only Tobacco*	*Only Alcohol*	*Tobacco & Marij*	*Alcohol & Marij*	*Alcohol & Tobacco*	*Use All*
Product Charachteristics							
Marijuana Price	0.00286			0.00286	0.00286		0.00286
	(0.0112)			(0.0112)	(0.0112)		(0.0112)
Alcohol Price (Beer Tax)			-0.291***		-0.291***	-0.291***	-0.291***
			(0.0804)		(0.0804)	(0.0804)	(0.0804)
Cigarette Price (Package)		-0.0774***		-0.0774***		-0.0774***	-0.0774***
		(0.0474)		(0.0474)		(0.0474)	(0.0474)
Marijuana Decriminalization	0.103***			0.103***	0.103***		0.103***
	(0.0393)			(0.0393)	(0.0393)		(0.0393)
Ratio High Quality Marijuana	0.26			0.26	0.26		0.26
	(0.173)			(0.173)	(0.173)		(0.173)
Demographics							
Age	-0.0148***	0.0906***	0.124***	0.0757***	0.109***	0.214***	0.2***
	(0.00695)	(0.00947)	(0.009)	(0.012)	(0.0114)	(0.012)	(0.0141)
Male	0.352***	0.114***	0.201***	0.466***	0.553***	0.315***	0.667***
	(0.0384)	(0.0408)	(0.0387)	(0.0512)	(0.0491)	(0.0541)	(0.0577)
White Non-Hispanic	0.161***	0.393***	0.532***	0.554***	0.693***	0.925***	1.086***
	(0.0406)	(0.0435)	(0.0411)	(0.0545)	(0.0522)	(0.0579)	(0.0618)
Health Quality (1 to 5)	-0.0688***	-0.229***	-0.0756***	-0.298***	-0.144***	-0.304***	-0.373***
	(0.0198)	(0.0215)	(0.0205)	(0.0272)	(0.0263)	(0.0283)	(0.0309)
Married	-0.471***	-0.0581	-0.245***	-0.286	-1.005***	-0.243**	-1.237***
	(0.201)	(0.140)	(0.0879)	(0.185)	(0.123)	(0.11)	(0.146)
Max education: High School	-0.173***	-0.843***	0.395***	-1.016***	0.223***	-0.447***	-0.62***
	(0.0621)	(0.0617)	(0.0606)	(0.0792)	(0.0801)	(0.0835)	(0.0913)
Max education: University	-0.406***	-1.708***	0.912***	-2.115***	0.506***	-0.796***	-1.203***
	(0.0871)	(0.0948)	(0.0881)	(0.119)	(0.111)	(0.122)	(0.131)
Max Parents Education: High School	0.287***	0.318***	-0.0784	0.606***	-0.202***	0.474***	0.239***
	(0.111)	(0.103)	(0.0591)	(0.146)	(0.0677)	(0.0861)	(0.0774)
Parents Education: University or more	0.342**	-0.608**	0.329***	0.141	0.257***	0.158	0.205*
	(0.153)	(0.253)	(0.0815)	(0.24)	(0.0902)	(0.128)	(0.109)
Constant Terms	-1.45***	-1.562***	-2.591***	-1.935***	-2.865***	-4.001***	-3.395***
	(0.119)	(0.228)	(0.205)	(0.3)	(0.263)	(0.275)	(0.344)
Bundle Interactions (γ)				1.077***	1.176***	0.152***	

Number of obs: 6,440. Standard errors in parenthesis. Included are year fixed effects. ***p<0.01, **p<0.5, *p<0.1.

Table 4.9.: Multinomial Probit Regressions

Finally, in table 4.10 we present results from a multinomial probit model where we include lagged values of the dependent variables. For these regressions, we use panel-adjusted standard errors clustered at the individual level. We continue to find a significant impact of past behavior on the consumption of single products, which is not surprising, but past behavior of a substance now impacts also the consumption of bundles containing the substance. The estimates on demographics are consistent with those of table 4.9. However, notice that the coefficient on marijuana price is now negative, though not significant. Once again, the direct substitution effects are negative, whereas the bundle interactions have positive signs, meaning that products in pairs are complements in use.

Appendix G presents additional estimates from multivariate probit regressions. In conclusion, our results show that it is important to control for correlation in unobservables for the choices of consuming the sin goods in combination. Controlling for past use also influences the degree of complementarity/substitutability of the products.

4.7. Conclusion

As illicit substances move into legal product space, substitution patterns with legal products become more salient. Two-thirds of Americans are in favor of marijuana legalization. In this paper, we make the first step to assess the impact of marijuana legalization on the consumption of potentially addictive *sin goods*, as marijuana might be consumed together with other substances.

Specifically, we study an individual's choice to consume marijuana, alcohol or cigarettes, and possible combinations of these products, allowing for persistence in behavior. We combine longitudinal data from the PSID survey (2005-2013) with pricing data for the three substances to estimate the parameters associated with the consumption decisions in the context of a dynamic model of multi-substance use. We show that it is important to control for unobserved correlation across behaviors and persistence in use to evaluate the potential complementarities among substances. We find that the past consumption has an impact on the current use of the drug consumed alone and in combination with other *sin goods*.

With this research we contribute to the literature on bundling and participate to the current debate on the effects of marijuana legalization. Some important caveats

Variable	*Only Marij*	*Only Tobacco*	*Only Alcohol*	*Tobacco & Marij*	*Alcohol & Marij*	*Alcohol & Tobacco*	*Use All*
Lagged Use							
Marijuana Use Last Period	1.375***			1.375***	1.375***		1.375***
	(0.0631)			(0.0631)	(0.0631)		(0.0631)
Alcohol Use Last Period			1.004***		1.004***	1.004***	1.004***
			(0.0597)		(0.0597)	(0.0597)	(0.0597)
Smoker Last Period		1.841***		1.841***		1.841***	1.841***
		(0.0756)		(0.0756)		(0.0756)	(0.0756)
Product Charachteristics							
Marijuana Price	-0.00005			-0.00005	-0.00005		-0.00005
	(0.0155)			(0.0155)	(0.0155)		(0.0155)
Alcohol Price (Beer Tax)			-0.128		-0.128	-0.128	-0.128
			(0.111)		(0.111)	(0.111)	(0.111)
Cigarette Price (Package)		-0.0615***		-0.0615***		-0.0615***	-0.0615***
		(0.0285)		(0.0285)		(0.0285)	(0.0285)
Marijuana Decriminalization	0.0697			0.0697	0.0697		0.0697
	(0.0535)			(0.0535)	(0.0535)		(0.0535)
Ratio High Quality Marijuana	0.267			0.267	0.267		0.267
	(0.235)			(0.235)	(0.235)		(0.235)
Demographics							
Age	-0.0148	0.0632***	0.126***	0.0484***	0.111***	0.189***	0.174***
	(0.0098)	(0.014)	(0.0133)	(0.0169)	(0.0169)	(0.0183)	(0.0216)
Male	0.243***	0.0707	0.179***	0.314***	0.422***	0.250***	0.493***
	(0.0519)	(0.0558)	(0.0518)	(0.0704)	(0.0672)	(0.0725)	(0.0789)
White Non-Hispanic	0.102**	0.208***	0.422***	0.309***	0.523***	0.629***	0.731***
	(0.055)	(0.061)	(0.0552)	(0.0764)	(0.0695)	(0.0795)	(0.0854)
Health Quality (1 to 5)	-0.053*	-0.158***	-0.0572**	-0.211***	-0.11***	-0.216***	-0.269***
	(0.029)	(0.0308)	(0.0286)	(0.0387)	(0.0383)	(0.0419)	(0.046)
Married	-0.196	-0.0644	-0.134	-0.406	-0.870***	-0.137	-1.101***
	(0.238)	(0.189)	(0.115)	(0.297)	(0.168)	(0.148)	(0.2)
Max education: High School	-0.161*	-0.508***	0.348***	-0.669***	0.187	-0.160	-0.321***
	(0.0867)	(0.0919)	(0.0896)	(0.113)	(0.119)	(0.128)	(0.139)
Max education: University	-0.359***	-1.11***	0.789***	-1.468***	0.439***	-0.312*	-0.671***
	(0.121)	(0.138)	(0.129)	(0.17)	(0.156)	(0.181)	(0.189)
Max Parents Education: High School	0.366**	0.332**	-0.0517	0.663***	-0.165*	0.328***	0.150
	(0.146)	(0.153)	(0.0769)	(0.235)	(0.088)	(0.112)	(0.108)
Parents Education: University or more	0.0364	-0.817*	0.308***	0.11	0.196*	0.207	0.176
	(0.225)	(0.422)	(0.105)	(0.415)	(0.117)	(0.164)	(0.154)
Constant Terms	-1.943***	-2.242***	-3.327***	-3.209***	-4.068***	-5.239***	-5.264***
	(0.176)	(0.329)	(0.298)	(0.451)	(0.39)	(0.421)	(0.519)
Bundle Interactions (γ)				0.976***	1.202***	0.33***	

Number of obs: 4,486. Robust standard errors in parenthesis (adjusted for 1,954 clusters). Included are year fixed effects.
***p<0.01, **p<0.5, *p<0.1.

Table 4.10.: Multinomial Probit Regressions with Lags

need to be mentioned: first, we do not observe individual prices for marijuana; second, we do not consider who has access to marijuana. To make a clear welfare assessment, one would need a full structural empirical model where these aspects are incorporated. The parameter estimates can inform the policy debate regarding the long-run impact of the legalization of marijuana by considering the concurrent and future effect on other potentially substitutable products.

F. Questions on consumption of substances PSID

Question	Values
Marijuana	
Have you ever taken marijuana?	Yes/No
On how many occasions have you used marijuana in your lifetime?	1-2 occasions 3-5 occasions 6-9 occasions 10-19 occasions 20-39 occasions 40 or more occasions
On how many occasions (if any) have you used marijuana in the past 12 months?	1-2 occasions 3-5 occasions 6-9 occasions 10-19 occasions 20-39 occasions 40 or more occasions
On how many occasions (if any) have you used marijuana in the past 30 days?	1-2 occasions 3-5 occasions 6-9 occasions 10-19 occasions 20-39 occasions 40 or more occasions
Smoking	
Did you ever smoke cigarettes?	Yes/No
Do you smoke cigarettes?	Yes/No
Drinking	
Do you ever drink any alcoholic beverages such as beer, wine, or liquor?	Yes/No
In the last year, on average, how often did you have any alcohol to drink?	Less than once a month About once a month Several times a month About once a week Several times a week Every day

Table F.1.: Original Questions from the TAS survey.

G. Other Regressions

In this Appendix, we show results from multivariate probit models. The first three columns of table G.1 present results from a model which considers that the choices are made together (and errors are correlated across products) but does not explicitly consider choices of product bundles. In the second three columns we allow for individuals to make choices among all bundles. The results indicate that demographic characteristics matter when choosing which substances to use but not necessarily in the same direction for all substances. For example, individuals with a higher level of education are more likely to drink alcohol but less likely to smoke cigarettes. Relative to marijuana, the quality of the product matters in a positive way. Finally, substance use is less likely the higher are the prices. When we control for joint use (last three columns), the effects of age and race on the consumption of marijuana change: if an individual is older or white and non-hispanic, the propensity to consume marijuana is lower. One can see that the joint use effects are all negative and significant.

Table G.2 presents the estimate for multivariate probit regressions which allow for correlation in the decision to consume the substances in combination and control for persistence in behavior.[1] The covariance results are consistent with those of the previous table; moreover, we find that it is important to take into account persistence, whose impact remain significant when controlling for the initial conditions (last three columns). In general, past use increases the current probability of consuming the substances, as expected.

[1] Contoyannis, Jones, and Rice (2004) also specify a random effects dynamic ordered probit model controlling for endogeneity á la Wooldridge. In their model, the outcome variable (and thus the lagged outcome variable) is a vector of J dummy variables for the $J + 1$ possible outcomes (dropping one of them).

	Multivariate Probits no Joint Use			Multivariate Probits		
Variable	*Alcohol*	*Marijuana*	*Tobacco*	*Alcohol*	*Marijuana*	*Tobacco*
Product Charachteristics						
Marijuana Price		-0.009			-0.013	
		(0.009)			(0.022)	
Alcohol Price (Beer Tax)	-0.214***			-0.131***		
	(0.07)			(0.068)		
Cigarette Price (Package)			-0.076***			-123***
			(0.198)			(0.044)
Marijuana Decriminalization		0.058*			0.112	
		(0.034)			(0.075)	
Ratio High Quality Marijuana		0.302**			0.683*	
		(0.151)			(0.367)	
Demographics						
Age	0.125***	0.023***	0.102***	0.064***	-0.064***	0.027*
	(0.008)	(0.008)	(0.009)	(0.008)	(0.017)	(0.015)
Male	0.263***	0.39***	0.216***	-0.028	0.078	-0.042
	(0.034)	(0.034)	(0.037)	(0.032)	(0.072)	(0.067)
White Non-Hispanic	0.576***	0.347***	0.491***	0.191***	-0.351***	0.091
	(0.036)	(0.036)	(0.04)	(0.034)	(0.08)	(0.071)
Health Quality (1 to 5)	-0.1***	-0.129***	-0.24***	0.071***	0.048	0.005
	(0.0184)	(0.018)	(0.02)	(0.018)	(0.038)	(0.034)
Married	-0.446***	-0.73***	-0.23***	-0.039	-0.138	0.17
	(0.06)	(0.072)	(0.069)	(0.055)	(0.168)	(0.113)
Max education: High School	0.255***	-0.305***	-0.82***	0.614***	-0.294***	-0.756***
	(0.057)	(0.058)	(0.059)	(0.07)	(0.102)	(0.08)
Max education: University	0.55***	-0.531***	-1.59***	1.026***	-0.652***	-1.44***
	(0.075)	(0.075)	(0.085)	(0.08)	(0.177)	(0.18)
Max Parents Education: High School	-0.082***	-0.004	0.39***	-0.123***	0.143	0.257***
	(0.04)	(0.04)	(0.046)	(0.038)	(0.091)	(0.084)
Parents Education: University or more	0.23***	0.056	-0.11*	0.187***	0.115	-0.651***
	(0.057)	(0.054)	(0.067)	(0.05)	(0.449)	(0.211)
Joint Use Effects						
Smoke Cigarettes and Use Marijuana				-2.09***		
				(0.037)		
Drink and Use Marijuana				-1.07***		
				(0.02)		
Drink and Smoke Cigarettes				-1.46***		
				(0.023)		

Number of obs: 6,440. Robust standard errors in parenthesis. All specifications include year fixed effects and a constant. ***p<0.01, **p<0.5, *p<0.1.

Table G.1.: Multivariate Probit Regressions

	MV Probits			MV Probits Initial Cond		
Variable	*Alcohol*	*Marijuana*	*Tobacco*	*Alcohol*	*Marijuana*	*Tobacco*
Lagged Use						
Marijuana Use Last Period		0.297***			0.169*	
		(0.099)			(0.101)	
Alcohol Use Last Period	0.0506***			0.595***		
	(0.041)			(0.053)		
Smoker Last Period			0.957***			0.773***
			(0.103)			(0.14)
Product Charachteristics						
Marijuana Price		-0.023			-0.017	
		(0.03)			(0.032)	
Alcohol Price (Beer Tax)	-0.037			-0.06		
	(0.078)			(0.078)		
Cigarette Price (Package)			-0.086			-0.089
			(0.056)			(0.057)
Marijuana Decriminalization		0.109			0.085	
		(0.097)			(0.102)	
Ratio High Quality Marijuana		0.88*			0.88	
		(0.51)			(0.59)	
Demographics						
Age	-0.0074	-0.073***	0.0355	-0.023	-0.01	0.031
	(0.01)	(0.023)	(0.024)	(0.022)	(0.06)	(0.048)
Male	-0.051	0.091	-0.108	-0.047	0.083	-0.113
	(0.037)	(0.092)	(0.09)	(0.038)	(0.098)	(0.09)
White Non-Hispanic	0.115***	-0.325***	-0.153	0.1**	-0.31***	-0.142
	(0.04)	(0.104)	(0.099)	(0.04)	(0.109)	(0.103
Health Quality (1 to 5)	0.07***	0.031	0.0364	-0.012	0.051	0.07
	(0.02)	(0.047)	(0.046)	(0.032)	(0.74)	(0.067)
Married	0.012	-0.106	0.199	0.008	-0.058	0.205
	(0.058)	(0.188)	(0.136)	(0.058)	(0.201)	(0.137)
Max education: High School	0.573***	-0.443***	-0.565***	-0.212	-0.504	0.05
	(0.085)	(0.123)	(0.108)	(0.2)	(0.38)	(0.26)
Max education: University	0.965***	-0.705***	-1.07***	-0.069	-0.53	0.094
	(0.092)	(0.187)	(0.205)	(0.218)	(0.475)	(0.408)
Max Parents Education: High School	-0.147***	-0.328**	0.057	-0.124***	0.355***	0.027
	(0.043)	(0.128)	(0.111)	(0.044)	(0.135)	(0.113)
Parents Education: University or more	0.163***	0.207	-1.02***	0.164***	0.23	-1.02***
	(0.057)	(0.17)	(0.37)	(0.057)	(0.18)	(0.37)
Joint Use Effects						
Smoke Cigarettes and Use Marijuana	-2.09***			-2.09***		
	(0.044)			(0.44)		
Drink and Use Marijuana	-1.05***			-1.56***		
	(0.023)			(0.023)		
Drink and Smoke Cigarettes	-1.39***			-1.39***		
	(0.027)			(0.027)		

Number of obs: 4,486 (individuals interviewed in one year only are dropped from the sample). Standard errors in parenthesis. All specifications include year fixed effects and a constant. ***p<0.01, **p<0.5, *p<0.1.
The initial conditions specification includes the mean over time of all the time varying regressors.

Table G.2.: Multivariate Probit Regressions with Lags

Bibliography

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

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

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

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

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

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

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

A. Allocca, L. Jacobi, and M. Sovinsky. Dynamics and complementarities among sin goods. 2020.

M. Antón, F. Ederer, M. Giné, and M. C. Schmalz. Common ownership, competition, and top management incentives. *Ross School of Business Paper*, (1328), 2018.

S. Appelt. Authorized generic entry prior to patent expiry: reassessing incentives for independent generic entry. *Review of Economics and Statistics*, 97(3):654–666, 2015.

J. Azar, M. C. Schmalz, and I. Tecu. Anticompetitive effects of common ownership. *The Journal of Finance*, 73(4):1513–1565, 2018.

M. Backus, C. Conlon, and M. Sinkinson. Common ownership and competition in the ready-to-eat cereal industry. *New York University Stern Working Paper*, 2018.

M. Backus, C. Conlon, and M. Sinkinson. Common ownership in america: 1980-2017. *American Economic Journal: Microeconomics*, 2020.

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

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

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

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

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

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

L. Boller and F. S. Morton. Testing the theory of common stock ownership, 2019.

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

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

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

L. Cameron and J. Williams. Cannabis, alcohol and cigarettes: substitutes or complements? *Economic Record*, 77(236):19–34, 2001.

G. Chamberlain. Panel data. *Handbook of econometrics*, 2:1247–1318, 1984.

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

K. W. Clements. Pricing and packaging: the case of marijuana. *The Journal of Business*, 79(4):2019–2044, 2006.

P. Contoyannis, A. M. Jones, and N. Rice. The dynamics of health in the british household panel survey. *Journal of Applied Econometrics*, 19(4):473–503, 2004.

B. Crost and S. Guerrero. The effect of alcohol availability on marijuana use: Evidence from the minimum legal drinking age. *Journal of health economics*, 31(1):112–121, 2012.

A. J. Davis, K. R. Geisler, and M. W. Nichols. The price elasticity of marijuana demand: Evidence from crowd-sourced transaction data. *Empirical Economics*, 50 (4):1171–1192, 2016.

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

J. DiNardo and T. Lemieux. Alcohol, marijuana, and american youth: the unintended consequences of government regulation. *Journal of health economics*, 20(6):991–1010, 2001.

G. Ellison and S. F. Ellison. Strategic entry deterrence and the behavior of pharmaceutical incumbents prior to patent expiration. *American Economic Journal: Microeconomics*, 3(1):1–36, 2011.

D. Ershov, J.-W. Laliberté, and S. Orr. Mergers in a model with complementarity. *Mergers in a Model with Complementarity (December 3, 2018)*, 2018.

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

J. Fichtner, E. M. Heemskerk, and J. Garcia-Bernardo. Hidden power of the big three? passive index funds, re-concentration of corporate ownership, and new financial risk. *Business and Politics*, 19(2):298–326, 2017.

M. Fosgerau, J. Monardo, and A. De Palma. The inverse product differentiation logit model. *Available at SSRN 3141041*, 2019.

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

M. Gentzkow. Valuing new goods in a model with complementarity: Online newspapers. *Available at SSRN 607401*, 2006.

M. Gentzkow. Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review*, 97(3):713–744, 2007.

J. Gerakos and J. Xie. Institutional horizontal shareholdings and generic entry in the pharmaceutical industry. *Tuck School of Business Working Paper*, (3285161), 2019.

E. P. Gilje, T. A. Gormley, and D. Levit. Who's paying attention? measuring common ownership and its impact on managerial incentives. *Journal of Financial Economics*, 2019.

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

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

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

B. H. Hamilton, J. A. Nickerson, and H. Owan. 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, 2003.

J. J. He and J. Huang. Product market competition in a world of cross-ownership: Evidence from institutional blockholdings. *The Review of Financial Studies*, 30(8): 2674–2718, 2017.

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

A. Hollis. The importance of being first: evidence from canadian generic pharmaceuticals. *Health economics*, 11(8):723–734, 2002.

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

A. Iaria and A. Wang. Identification and estimation of demand for bundles. 2020.

L. Jacobi and M. Sovinsky. Marijuana on main street? estimating demand in markets with limited access. *American Economic Review*, 106(8):2009–45, 2016.

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

D. Kandel and R. Faust. Sequence and stages in patterns of adolescent drug use. *Archives of general psychiatry*, 32(7):923–932, 1975.

D. Kenkel, A. D. Mathios, and R. L. Pacula. Economics of youth drug use, addiction and gateway effects. *Addiction*, 96(1):151–164, 2001.

A. Koch, M. A. Panayides, and S. Thomas. Common ownership and competition in product markets. In *29th Annual Conference on Financial Economics & Accounting*, 2018.

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

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

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

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

J. A. McCahery, Z. Sautner, and L. T. Starks. Behind the scenes: The corporate governance preferences of institutional investors. *The Journal of Finance*, 71(6): 2905–2932, 2016.

J. J. McRae and F. Tapon. Some empirical evidence on post-patent barriers to entry in the canadian pharmaceutical industry. *Journal of Health Economics*, 4(1):43–61, 1985.

F. M. S. Morton. Entry decisions in the generic pharmaceutical industry. *The Rand Journal of Economics*, 10:421–440, 1999.

F. M. S. Morton. Barriers to entry, brand advertising, and generic entry in the us pharmaceutical industry. *international Journal of industrial Organization*, 18(7): 1085–1104, 2000.

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

M. Newham, J. Seldeslachts, and A. Banal-Estanol. Common ownership and market entry: Evidence from pharmaceutical industry. 2018.

D. P. O'Brien and S. C. Salop. Competitive effects of partial ownership: Financial interest and corporate control. *Antitrust LJ*, 67:559, 2000.

R. L. Pacula. Does increasing the beer tax reduce marijuana consumption? *Journal of health economics*, 17(5):557–585, 1998.

P. Pierani and S. Tiezzi. Addiction and interaction between alcohol and tobacco consumption. *Empirical Economics*, 37(1):1–23, 2009.

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

J. Rotemberg. Financial transaction costs and industrial performance. 1984.

A. Rubinstein and M. E. Yaari. Repeated insurance contracts and moral hazard. *Journal of Economic theory*, 30(1):74–97, 1983.

A. Saha, H. Grabowski, H. Birnbaum, P. Greenberg, and O. Bizan. Generic competition in the us pharmaceutical industry. *International Journal of the Economics of Business*, 13(1):15–38, 2006.

S. Sato and T. Matsumura. Free entry under common ownership. 2019.

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

N. J. Schlienz and D. C. Lee. Co-use of cannabis, tobacco, and alcohol during adolescence: policy and regulatory implications. *International review of psychiatry*, 30 (3):226–237, 2018.

M. C. Schmalz. Common-ownership concentration and corporate conduct. *Annual Review of Financial Economics*, 10:413–448, 2018.

M. C. Schmalz, J. Azar, and R. Sahil. Ultimate ownership and bank competition. 2016.

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

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

M. Song, S. Nicholson, and C. Lucarelli. Mergers with interfirm bundling: a case of pharmaceutical cocktails. *The RAND Journal of Economics*, 48(3):810–834, 2017.

Ø. Thomassen, H. Smith, S. Seiler, and P. Schiraldi. Multi-category competition and market power: a model of supermarket pricing. *American Economic Review*, 107 (8):2308–51, 2017.

I. M. Torres, J. Puig, J.-R. Borrell-Arqué, et al. Generic entry into a regulated pharmaceutical market. Technical report, Department of Economics and Business, Universitat Pompeu Fabra, 2007.

H. Wen, J. Hockenberry, and J. R. Cummings. The effect of medical marijuana laws on marijuana, alcohol, and hard drug use. Technical report, National Bureau of Economic Research, 2014.

J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press: Cambridge, Mass., 2002.

J. M. Wooldridge. Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of applied econometrics*, 20(1):39–54, 2005.

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

X. Zhao and M. N. Harris. Demand for marijuana, alcohol and tobacco: Participation, levels of consumption and cross-equation correlations. *Economic Record*, 80(251): 394–410, 2004.

Curriculum Vitae

Alessandra Allocca

— 2014–2020	*University of Mannheim* Ph.D. Student in Economics
— 2013–2014	*CSEF - Università degli Studi di Napoli Federico II* Research Assistant
— 2011–2013	*Università degli Studi di Napoli Federico II* *Laurea Magistrale* in Economics and Finance
— 2011–2012	*Goethe University Frankfurt Am Main* LLP Erasmus Program in Economics
— 2008–2011	*Università degli Studi di Napoli Federico II* *Laurea* in Economics

Ehrenwörtliche Erklärung

Ich versichere hiermit, dass ich die Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ahnlicher Form noch keiner Prufungsbehorde vorgelegen.

Mannheim, den 13 August 2020 Alessandra Allocca