Appliquée et collaborative ? : essais sur l'évolution de la recherche scientifique d'entreprise

Federico Bignone

THÈSE EN COTUTELLE PRÉSENTÉE
POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ET DE L'UNIVERSITÉ DE SWINBURNE

ÉCOLE DOCTORALE ENTREPRISE, ÉCONOMIE, SOCIÉTÉ

SCHOOL OF BUSINESS, LAW AND ENTREPRENEURSHIP

SPÉCIALITÉ: Science économiques

Federico BIGNONE

APPLIED AND COLLABORATIVE?
ESSAYS ON THE CHANGING NATURE OF CORPORATE SCIENTIFIC RESEARCH

Sous la direction de Russell THOMSON
et de Francesco LISSONI

Soutenue le 13/12/2023

Membres du jury :
M. Markus SIMETH, Associate Professor, Copenhagen Business School, Rapporteur
Mme. Katrin HUSSINGER, Full Professor, University of Luxembourg, Rapporteur
Mme. Pascale ROUX, Professeure, Université de Bordeaux, Rapporteur
Mme. Beth WEBSTER, Professor, Swinburne University, President

Membres invités :
M. Russell Thomson, Associate Professor, Swinburne University, Directeur de thèse
M. Francesco Lissoni, Professeur, Université de Bordeaux, Directeur de thèse

APPLIED AND COLLABORATIVE? ESSAYS ON THE CHANGING NATURE OF CORPORATE SCIENTIFIC RESEARCH

Federico Bignone

Co-supervisor: Russell Thomson
Co-supervisor: Francesco Lissoni

PhD in Economics

Dissertation submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Centre for transformative innovation, Swinburne University of Technology
Bordeaux School of Economics, Université de Bordeaux

2024

Abstract

Corporate science, broadly intended as scientific research conducted by business companies, is a key input to innovation worldwide. Some recent evidence has documented its quantitative decline since its heyday in the late 1960s / early 1970s, as measured by the shrinking share of scientific publications by business-affiliated authors, mostly due to the downsizing or closure of large corporate laboratories (Arora, Belenzon, and Patacconi 2018; Tijssen 2004). Despite this evidence, there are still many shadow zones on both the extent and features of this decline, as it is not clear whether firms are simply disengaging from scientific research or whether they are just changing their organisation by moving from vertically integrated R&D activities towards a more dynamic system of applied in-house research and collaborations with universities.

Based on an original dataset, I examine these changes and other qualitative features of corporate science in the United States from the 1980s until recent years. The dataset results from four different data sources. First, all scientific publications from Web of Science (WoS) with at least a US-based author from 1980 to 2014. Second company data and patents from the US edition of Orbis, a large business information database by Bureau van Dijk. Third, non-patent literature (NPL) citations from worldwide patents to scientific publications, as collected in the "reliance on science" database by Marx and Fuegi (2020). Last, USPTO, WIPO, and EPO patents from PATSTAT of the European Patent Office and Orbis.

I tackle three related questions. The first question addresses the increase in university-industry collaborative publications and the factors driving this trend. I show that the increase in collaborations is accompanied by a decrease in direct corporate involvement in scientific research, but this correlation is influenced by factors related to the nature of science, such as the increasing division of scientific labour and the "burden of knowledge", rather than commercial considerations of the companies. Specifically, the probability of firms collaborating with universities increases in fast-moving fields, and their relationship depends on firm size.

The second question concerns whether corporate science has become more applied, as suggested by the literature, and whether this has come at the expense of its ambition and scope, that is, whether it has also become less basic. I first revisit the distinction between basic

and applied research, based on Stokes' (1997) criticism of the counterposition between the two. Then, drawing again from Stokes (1997), I measure basicness and appliedness of both corporate and academic science at different points in time with two complementary indicators. I find that corporate science has progressively become more applied and less basic than academic science, after controlling for fields and journals and regardless of the firms' age and size.

The third question concerns whether corporate science's decline may be due, at least in part, to short-termism induced by shareholders' pressure. To this end, I investigate the impact of initial public offerings (IPOs) on firms' research activities. To do so, I use data on the population of firms filing for an IPO from 1996 to 2010 with at least one publication or patent. My identification strategy involves a treatment group of firms that completed an IPO and a control group of firms that filed for an IPO but afterwards decided to withdraw their filing. Identification is achieved with a stacked difference-in-difference specification and an instrumental variable. The results show a positive effect of IPOs on both corporate and collaborative publications. I find no effect on publications' forward citations, basicness or appliedness. I explain this result with firms' increased access to capital and the inflow of new scientists.

In conclusion, the analysis carried out in this thesis sheds light on the evolution of corporate science from 1980 to 2014, contributing to the recent literature on the decline of corporate science. The decline in corporate publications is accompanied by a notable increase in university-industry collaborations, signalling a change in companies' approach towards scientific research. Furthermore, companies have moved towards more applied and less basic science, supporting the hypothesis of increased short-termism. However, the evidence for shareholder pressure is less pronounced as companies that go public intensify their research efforts instead of constraining them.

Résumé

La science d'entreprise, largement entendue comme la recherche scientifique menée par des entreprises commerciales, constitue une contribution clé à l'innovation à l'échelle mondiale. Certaines preuves récentes ont documenté son déclin quantitatif depuis son apogée à la fin des années 1960 / début des années 1970, tel que mesuré par la réduction de la part des publications scientifiques d'auteurs liés à des entreprises, principalement en raison de la réduction de la taille ou de la fermeture de grands laboratoires d'entreprise (Arora, Belenzon, and Patacconi 2018; Tijssen 2004). Malgré ces éléments de preuve, il existe encore de nombreuses zones d'ombre concernant l'étendue et les caractéristiques de ce déclin, car il n'est pas clair si les entreprises se désengagent simplement de la recherche scientifique ou si elles en modifient simplement l'organisation, en passant d'activités de R&D intégrées verticalement à un système plus dynamique de recherche interne appliquée et de collaborations avec les universités.

À partir d'un jeu de données original, j'examine ces évolutions et d'autres caractéristiques qualitatives de la science d'entreprise aux États-Unis des années 1980 jusqu'à ces dernières années. Le jeu de données provient de quatre sources différentes. Tout d'abord, toutes les publications scientifiques provenant de Web of Science (WoS) avec au moins un auteur basé aux États-Unis, de 1980 à 2014. Deuxièmement, des données d'entreprise et des brevets provenant de l'édition américaine d'Orbis, une grande base de données d'informations commerciales de Bureau van Dijk. Troisièmement, des citations de littérature non brevetée (NPL) provenant de brevets du monde entier qui citent des publications scientifiques, telles que recueillies dans la base de données “reliance on science” par Marx and Fuegi (2020). Enfin, les brevets du USPTO, de l'OMPI et de l'OEB à partir de PATSTAT de l'Office européen des brevets et Orbis.

J'aborde trois questions connexes. La première concerne l'augmentation éventuelle du nombre de publications collaboratives entre universités et entreprises, ainsi que les facteurs qui les motivent. Je montre que l'augmentation des collaborations s'accompagne d'une diminution de l'implication directe des entreprises dans la recherche scientifique, mais que cette corrélation est influencée par des facteurs liés à la nature de la science, tels que la division croissante du travail scientifique et le “fardeau de la connaissance”, plutôt que par des considérations commerciales des entreprises. Plus précisément, la probabilité de

collaboration entre les entreprises et les universités augmente dans les domaines à évolution rapide, et leur relation dépend de la taille de l'entreprise.

La deuxième question s'intéresse au fait de savoir si la science d'entreprise est devenue plus appliquée, comme le suggère la littérature, et si cela s'est fait au détriment de son ambition et de sa portée, c'est-à-dire si elle est devenue moins fondamentale. Je revisite d'abord la distinction entre recherche fondamentale et recherche appliquée, en me basant sur la critique de Stokes (1997) de la dichotomie entre les deux. Ensuite, en m'appuyant de nouveau sur Stokes (1997), je mesure le caractère fondamental et appliqué de la science d'entreprise et académique à différents moments à l'aide de deux indicateurs complémentaires. Je constate que la science d'entreprise est progressivement devenue plus appliquée et moins fondamentale que la science académique, après avoir contrôlé par les domaines et les revues, indépendamment de l'âge et de la taille des entreprises.

La troisième question concerne la possibilité que le déclin de la science d'entreprise soit dû, au moins en partie, au court-termisme induit par la pression des actionnaires. À cette fin, j'étudie l'impact des introductions en bourse (IPO) sur les activités de recherche des entreprises. Pour ce faire, j'utilise des données sur la population des entreprises déposant une IPO entre 1996 et 2010 avec au moins une publication ou un brevet. Ma stratégie d'identification implique un groupe de traitement constitué d'entreprises ayant réalisé une IPO et un groupe témoin d'entreprises ayant déposé une IPO mais ayant ensuite décidé de retirer leur dépôt. L'identification est réalisée à l'aide d'un modèle staggered difference-in-difference et d'une variable instrumentale. Je trouve un effet positif des IPO tant sur les publications d'entreprise que sur les publications collaboratives. Je ne constate aucun effet sur les citations ultérieures des publications, le caractère fondamental ou appliqué. J'explique ce résultat par un accès accru des entreprises au capital et l'arrivée de nouveaux scientifiques.

En conclusion, l'analyse réalisée dans cette thèse éclaire l'évolution de la science d'entreprise de 1980 à 2014, contribuant ainsi à la littérature récente sur le déclin de la science d'entreprise. Le déclin des publications d'entreprise s'accompagne d'une augmentation notable des collaborations entre universités et entreprises, signalant un changement dans l'approche des entreprises vis-à-vis de la recherche scientifique. De plus, les entreprises se sont tournées vers une science plus appliquée et moins fondamentale, ce qui soutient l'hypothèse d'un court-termisme accru. Cependant, les preuves de la pression des actionnaires sont moins prononcées, car les entreprises qui se rendent publiques intensifient leurs efforts de recherche au lieu de les restreindre.

Acknowledgements

This thesis is the result of four years of hard work and many people deserve an acknowledgement.

I would like first to express my sincere gratitude to Professor Francesco Lissoni and Associate Professor Russell Thomson for their invaluable guidance throughout this journey. Despite the time difference between Australia and France and my stubbornness in going down the same rabbit holes for too long, they remained patient and consistently available. Thanks to them, I am a better researcher, and if the price to pay is being reminded that I support a bad football team and hearing stories about eels, so be it.

I would also like to extend my appreciation to the individuals at the Centre for Transformative Innovation (CTI) at Swinburne University, including Professor Beth Webster, Dr. Sarah Hegarty, Dr. Trevor Kollmann, Dr. Stephen Petrie, Associate Professor Alfons Palangkaraya, Professor Terry Healy, and Professor Tom Spurling.

Additionally, my gratitude goes out to my colleagues at BSE, namely Dr. Ernest Miguelez, Dr. Diego Useche, and Professor Valerio Sterzi, for their valuable support and collaboration.

I am deeply thankful to my fellow PhD students at Swinburne, including Babu, Aneeq, Danish, Achinthya, Jean Francois, Le, Tom, as well as those at Bordeaux, Deivyd, Valentina, Grazia, Joao, Jeremy, Jemal, Leo, Louis, Lucie, Coralie, Vito, Ruth, and Marcos.

A special thanks goes to Cecilia Maronero, David Paynter and Christian Chacua who were not only colleagues but also great friends and made this journey way easier.

I want to thank my family and friends who have supported me across three different countries, even amidst multiple relocations. Every time I return, it feels as if I never left.

I would finally like to thank Catherine Lis for standing by my side throughout this rollercoaster of a journey. It's not often that one receives the complete package of academic and emotional support — what a catch!

I also acknowledge the financial support received from Swinburne through the Swinburne University Postgraduate Research Award (SUPRA), Campus France for the Eiffel Scholarship, and the Project ISI and Pôle Doctorant of BSE, which funded some of my conference participation.

Lastly, I gratefully acknowledge the contribution of many people in developing the algorithm underpinning the database used in this thesis. The first iteration of the linking algorithm was developed on Swinburne University of Technology computing facility OzSTAR and was undertaken between 2016 and 2019 by a team led by Russell Thomson at Swinburne University of Technology. The bulk of coding at this stage was undertaken by Antonio Bibiano and Alex Codoreanu, with valuable input from T'Mir Julius, Steve Petrie and Patrick Doran. T'Mir Julius and Sarah Hegarty also made important contributions to assembling and cleaning both WoS and Orbis. Funding for the acquisition of key datasets was provided by Swinburne University and Commonwealth Scientific and Industrial Research Organisation (CSIRO).

Declaration

The candidate hereby certifies that:

- This thesis contains no material that has been accepted for the award to the candidate of any other degree or diploma, except where prior permission to do so has been received by the Associate Dean, Research and Development, and with due reference made about this in the text of the examinable outcome.
- To the best of the candidate's knowledge contains no material previously published or written by another person except where due reference is made in the text of the examinable outcome, and with permission received to republish the work in the thesis.
- Where the work is based on joint research or publications, this thesis discloses the relative contributions of the respective creators or authors.

Hawthorn VIC, Australia, January 22, 2024

Federico Bignone

Table of Contents

Abstract iv
Résumé vi
Acknowledgements viii
Declaration x
Table of Contents xi
List of Figures xiii
List of Tables xv
List of Abbreviations xviii
Introduction 1
Thesis' contribution 4
Chapter 1: The WoS-Orbis US corporate science database 7
1.1 Data 7
1.2 Matching Orbis and Web of Science (1) 9
1.3 Connecting patents to the Reliance on science database (2) 15
1.4 Linking all papers to the dual frontier (3) 15
1.5 Collecting patent information (4) 16
1.6 Descriptive statistics 16
1.7 Conclusions 25
Chapter 2: University-industry collaborations and the burden of knowledge: evidence from the US 26
2.1 Introduction 26
2.2 What drives university-industry collaboration? 30
2.3 Data and descriptives 33
2.4 Empirical strategy 49
2.5 Results and discussion 52
2.6 Alternative measures of the burden of knowledge 59
2.7 Conclusions 62
Chapter 3: In and out the Pasteur's quadrant: revisiting trends in corporate science 64
3.1 Introduction 64
3.2 Conceptual framework 66
3.3 Data and methods 68
3.4 Analysis 76
3.5 Conclusions 91
Chapter 4: Corporate science and IPOs 93
4.1 Introduction 93
4.2 The relationship between IPOs and innovation 95

4.3 Data and methods 100
4.4 Results 106
4.5 Mechanisms 110
4.6 IPOs and innovation 116
4.7 Conclusions 121
Conclusions 123
Policy implications 127
A. Appendix Chapter 1 130
B. Appendix Chapter 2 167
C. Appendix Chapter 3 172
D. Appendix Chapter 4 180
Bibliography 182

List of Figures

Figure 1.1: Schematic representation of the database linking procedure 9
Figure 1.2: Schematic representation of the matching algorithm 11
Figure 1.3: Distance metrics, visualisation 15
Figure 1.4: Patents, publications, collaborations, and connected publications, 1980-2014 17
Figure 1.5: Share of corporate publications as share of total WoS US publications, 1980-2014 18
Figure 1.6: Number of publications by broad scientific fields, 1980-2014 19
Figure 1.7: Number of publications by industry, 1980-2014 21
Figure 1.8: Share of corporate papers in WoS by field, 1980-2014 22
Figure 1.9: Rise and fall of corporate publishers, 1980-2014 24
Figure 2.1: Total publications, collaborations and solo publications, 1980-2013 37
Figure 2.2: Share of collaborations in firms' portfolio by broad scientific field, 1980-2013 ..... .39
Figure 2.3: Share of collaborations in firms' portfolio by industry, 1980-2013 41
Figure 2.4: Share of collaborations in firms' portfolio by firm size and age, 2000-2013 ... 46
Figure 2.5: Correlation between speed and Avg auth (left) and speed and IRA (right) 49
Figure 2.6: Speed marginal effects 55
Figure 2.7: Marginal effects of speed with scientific field fixed effects 56
Figure 2.8: Marginal effect of IRA firm and year FE (left), firm, year and scientific fields FE(right) 60
Figure 2.9: Marginal effect of Avg auth, firm and year FE (left), firm, year and scientific fields FE(right) 60
Figure 3.1: Dataset construction 70
Figure 3.2: Distance metrics, visualisation 72
Figure 3.3: Average appliedness and basicness by field, 1980-2014 75
Figure 3.4: Appliedness (left) and basicness (right), interaction term regression coefficients, 1980-2014 80
Figure 3.5: Appliedness interaction term regression coefficients by broad field, 1980-2014 81
Figure 3.6: Basicness interaction term regression coefficients by broad field, 1980-2014 . 82
Figure 3.7: Basicness and appliedness interaction coefficients in the Pasteur's quadrant .. 84
Figure 3.8: Basicness and appliedness interaction coefficients in the Pasteur's quadrant by field, corporate (left) and collaborations (right) 85
Figure 3.9: Appliedness (left) and basicness (right) by firm age, 1980-2014 89

Figure 3.10: Appliedness (left) and basicness (right) by firm size, 1980-2014 90
Figure 4.1: Overview of the data sources. IPOs information, Orbis data, patents, and publications 101
Figure 4.2: IPO impact on publications and collaborations, event study 107
Figure 4.3: IPO impact on patent number and citations, event study 118
Figure A 1: Matching algorithm flowchart 131
Figure A 2: Orbis cleaning workflow 133
Figure A 3: WoS cleaning flowchart 148
Figure A 4: Number of publications (left) and journals (right) in WoS by document type, 1980-2014 154
Figure A 5: Number of publications (left) and journals (right) by broad fields, 1980-2014 155
Figure A 6: Matching algorithm flowchart 156
Figure A 7: Simplified visualisation of the binning 158
Figure D 1: Appliedness (left) and paper citations (right) , event study 181

List of Tables

Table 1.1: Scoring intervals of the matches, description, and summary statistics..... 12
Table 1.2: Precision in a random sample of 100 affiliations per scoring interval 13
Table 1.3: Descriptive statistics of the final sample 16
Table 2.1: Corporate publications sample, 1980-2013 36
Table 2.2: Most common collaborations by broad scientific field, 1980-2013 43
Table 2.3: Speed of the five fastest and slowest scientific fields..... 48
Table 2.4: Summary statistics of the burden of knowledge variables 49
Table 2.5: LPM. Impact of speed on the probability of collaborating..... 52
Table 2.6: LPM. Impact of speed on the probability of collaborating with interaction 53
Table 2.7: Impact of speed on collaboration, LPM dividing SMEs and large firms 54
Table 2.8: Impact of speed on the probability of collaborating with quadratic interaction. 55
Table 2.9: LPM removing one industry with firm and year fixed effects 58
Table 2.10: LPM removing one industry with firm, year, and sc. fields fixed effects 58
Table 3.1: Summary statistics 75
Table 4.1: Summary statistics for the five years preceding an IPO filing and at the time of IPO, completed and withdrawn IPOs 103
Table 4.2: ATT of IPO on number of publications and collaborations 106
Table 4.3: ATT of IPO on patent impact, publication citations, appliedness, and basicness 108
Table 4.4: Instrumental variable regression on publication and collaboration number..... 109
Table 4.5: Instrumental variable regression on publication forward citations, appliedness and basicness 110
Table 4.6: Impact of IPO on publications and collaborations at the scientist level..... 113
Table 4.7: Impact of IPO on publication impact and appliedness at the scientist level 114
Table 4.8: Impact of IPO on the number of unique scientists and scientists-inventors..... 115
Table 4.9: Cash raised and number of publications, patents and R&D expenses..... 116
Table 4.10: ATT of IPO on number of patents and patent forward citations. 117
Table 4.11: Instrumental variable regression on patent number and forward citations..... 119
Table 4.12: IPO and innovation at the inventor level 120
Table 4.13: Impact of IPO on the number of unique inventors. 121
Table A 1: Orbis Historic variables 135
Table A 2: WoS abbreviations..... 136
Table A 3: Stopwords 137

Table A 4: Orbis table before abbreviating 138
Table A 5: Orbis table after abbreviating 138
Table A 6: Exceptions in abbreviations 138
Table A 7: Abbreviations dictionary 139
Table A 8: Inefficient Orbis tokens table 144
Table A 9: Orbis tokens table 144
Table A 10: Orbis financial data from all servers 145
Table A 11: Collapsed financial data 145
Table A 12: Final Orbis sample 146
Table A 13: Clean Orbis main table 147
Table A 14: WoS table before the cleaning operations 150
Table A 15: WoS table after the cleaning operations 150
Table A 16: Government and University keywords 151
Table A 17: Whole string substitution 152
Table A 18: Final WoS input table 152
Table A 19: Index file 157
Table A 20: Scores of the matching algorithm outcomes 160
Table A 21: Matching example 160
Table A 22: Government keywords 161
Table A 23: Specific government institutions 162
Table A 24: International organisations keywords 162
Table A 25: Medical centres keywords 163
Table A 26: Specific medical centres 163
Table A 27: Not-for-profit keywords 163
Table A 28: Specific not-for-profit 164
Table A 29: Educational institutions keywords 164
Table A 30: Specific educational institutions 165
Table A 31: Precision in a random sample of 100 affiliations per scoring interval 166
Table B 1: Distribution of subjects 167
Table B 2: Impact of Speed on the probability of collaborating, multi field and multi company. LPM 167
Table B 3: Impact of IRA and Avg Auth on the probability of collaborating, multi field and multi company. LPM 168
Table B 4: Impact of Speed on the probability of collaborating with interaction term, multi field and multi company. LPM 168

Table B 5: Impact of Speed on the probability of collaborating with interaction term, multi field and multi company. SMEs vs Large firms. LPM 169

Table B 6: Impact of Speed on the probability of collaborating with interaction term, multi field and multi company. LPM with quadratic term 169

Table B 7: Impact of IRA on the probability of collaborating with interaction term, multi field and multi company. SMEs vs Large firms. LPM 170

Table B 8: Impact of Avg auth on the probability of collaborating with interaction term, multi field and multi company. SMEs vs Large firms. LPM 170

Table B 9: Impact of IRA on the probability of collaborating with interaction term, multi field and multi company. LPM with quadratic term 171

Table B 10: Impact of Avg auth on the probability of collaborating with interaction term, multi field and multi company. LPM with quadratic term 171

Table C 1: Summary of the metrics 172

Table C 2: Corporate and collaborations appliedness and basicness, 1980, 2014 173

Table C 3: Corporate science appliedness by 10 broad fields, 1980-2014 174

Table C 4: Collaborations appliedness by 10 broad fields, 1980-2014 175

Table C 5: Corporate science basicness by 10 broad fields, 1980-2014 176

Table C 6: Collaborations basicness by 10 broad fields, 1980-2014 177

Table C 7: Appliedness and basicness by firm age, 1980-2014 178

Table C 8: Appliedness and basicness by firm size, 1980-2014 179

Table D 1: Number of papers and collaborations, IV calculated in the book building phase 180

Table D 2: Publications forward citations and appliedness, IV calculated in the book building phase 181

List of Abbreviations

2SLS	Two Stages Least Squares
A&J	Ahmadpoor & Jones
AAAS	American Association for the Advancement of Science
AI	Artificial Intelligence
ATT	Average Treatment effect on the Treated
AY	Absolute Year
CEO	Chief Executive Officer
DOI	Digital Object Identifier
DUO	Domestic Ultimate Owner
EDGAR	Electronic Data Gathering, Analysis, and Retrieval system
EPO	European Patent Office
FBI	Federal Bureau of Investigation
FE	Fixed effects
GE	General Electrics
GIA	Geological Institute of America
GSK	GlaxoSmithKline
GUO	Global Ultimate Owner
IBM	International Business Machines
ICT	Information and Communications Technology
IEEE	Institute of Electrical and Electronics Engineers
IMF	International Monetary Fund
INC	Incorporated
IP	Intellectual Property
IPC	International Patent Classification
IV	Instrumental Variable
IPO	Initial Public Offering
IRA	Inverse of References Age
LAY	Last Available Year
LPM	Linear Probability Model
LTD	Limited
MAG	Microsoft Academic Graph
MGH	Massachusetts General Hospital Medical
MIR	Machine Intelligence Research
MIT	Massachusetts Institute of Technology
NAICS	North American Industry Classification System

NASA	National Aeronautics and Space Administration
NASDQ	National Association of Securities Dealers Automatic Quotation System
NCAR	National Center for Atmospheric Research
NCGR	National Council for Geocosmic Research, National Center for Genome Resources
NIOSH	National Institute for Occupational Safety and Health
NMFS	National Marine Fisheries Service
NORC	Naval Ordnance Research Calculator
NOAA	National Oceanic and Atmospheric Administration
NPL	Non-Patent Literature
NRL	Naval Research Laboratory
NSF	National Science Fundation
OECD	Organization for Economic Cooperation and Development
OLS	Ordinary Least Squares
OTCBB	Over-The-Counter Bulletin Board
PATH	Program for Appropriate Technology in Health
PARC	Palo Alto Research Center
PI	Price Index
PPML	Poisson Pseudo Maximum Likelihood
PRBO	Point Reyes Bird Observatory
RAND	Research and Development
RIETI	Research Institute of Economy, Trade and Industry
RTOG	The Radiation Therapy Oncology Group
RY	Relative Year
SAMSI	Statistical and Mathematical Sciences Institute
SC	Science Category
SEO	Seasoned Equity Offering
SIC	Standard Industrial Classification
SISSA	International School for Advanced Studies
SMEs	Small and Medium Enterprises
SSL	Systems Science Lab
STEM	Science, Technology, Engineering, and Mathematics
TWFE	Two Way Fixed Effects
UK	United Kingdom
UNDP	United Nations Development Programme
US	United States
USD	United States Dollar

VA	Veteran Administration
VC	Venture Capital
WHO	World Health Organisation
WIPO	World Intellectual Property Organization
WOS	Web of Science
WWF	World Wide Fund for Nature
WWII	World War II
ZIP	Zone Improvement Plan

Introduction

From the advancements in solid state physics to the discoveries in materials science, there are few if any areas of the modern economy that do not owe a substantial debt to the research efforts of corporate America in the 20th century. From the beginning of the 20th century until the 1980s, large US corporations established and expanded internal R&D laboratories, increasing the time and resources dedicated to basic investigation and not just to applied research or development (Arora et al. 2021; Mowery 2009; The Economist 2007). This engagement was usually triggered by the need to solve technical problems that could not be easily solved through more trial-and-error engineering approaches (Rosenberg 1990) as well as by the need to establish and maintain the capacity to absorb the scientific advances produced by university and public laboratories (Cohen and Levinthal 1989).

In 1940, US company Du Pont spent $4.3 million in R&D to invent, develop, and commercialise nylon – the first synthetic substance that could be transformed into yarns, coatings, films, and plastic. Du Pont scientists completed the nylon project in 5 years and are nowadays credited to have started from zero a new field of research on synthetic fibres (Hounshell and Smith 1988). The radar development in the 1940s was a joint effort of the Naval Research Laboratory (NRL), General Electric, Radio Corporation of America, the Sperry Gyroscope Company, and later AT&T (Barton 2010). State and business collaborations in 1953 led to the invention of the Naval Ordnance Research Calculator (NORC), the world's most powerful computer for several years (IBM archives, 2020). Building upon decades of progress in solid-state physics and significantly contributing to it, Bell Labs' scientists William Shockley, John Bardeen and Walter Brattain invented the transistor, for which they received the 1956 Nobel Prize in Physics. Texas Instruments and Fairchild semiconductors built upon Bell Labs' research to create the first integrated circuit in 1959, a single piece of silicon (chip) containing transistors and resistors (Lojek 2006; Gertner 2012). The Systems Science Lab (SSL) of Xerox undertook projects on xerographic printing, computing, and optical memories that led to the first laser printer and the basic theory for the compact disc (Hiltzik 1999).

Historians and economists often refer to this productive period as the "golden age" of American corporate science. As the 20th century progressed, significant changes began to emerge. Large corporations started disengaging from large-scale, science-oriented research

programmes. The share of scientific publications authored by scientists with a business rather than an academic affiliation has declined ever since. Many companies reportedly focus more on short-term returns and development, neglecting basic science (Arora, Belenzon, and Patacconi 2018; Tijssen 2004).[1] The dominant cause of this decline is believed to be the disappointment with the results of the corporate investments in science of the 1970s and 1980s and the shareholders' pressure for immediate returns on investment (Hounshell and Smith 1988; Lazonick and O'Sullivan 2000; Pisano 2010; Arora, Belenzon, and Patacconi 2018). In this view, these factors contributed to forcing firms to reduce diversification, increase their focus on short-term results, and buy science-based inventions from universities and start-ups via the expanding markets for technologies rather than producing them internally (Arora, Belenzon, and Patacconi 2018).

Several authors have pointed at this decline of corporate science as one cause for the reduced rate of innovation and economic growth of the US and other established economies in the past few decades. The extent and qualitative features of the corporate science decline, however, have yet to be comprehensively measured. Studies such as Tijssen's (2004) refer to broad categories of authors' affiliations (business vs. academic), whereas firm-based studies such as Arora, Belenzon, and Patacconi (2018), Rafols et al. (2012), and Bhaskarabhatla and Hedge (2014) either refer only to large US companies, small samples, or single firm case studies. We know little about small and medium-sized firms that probably play a role in the outgoing process of vertical disintegration of corporate research.

Collaborations with universities are depicted as one possible strategy by which firms compensate for their disengagement from corporate science (Coombs and Georghiou 2002). The evidence, however, is not systematic, especially for small and medium enterprises. While the literature has spent significant effort describing how companies approach universities (Ankrah and AL-Tabbaa 2015), most of the focus was centred around the commercial environment of the company. This includes aspects like increased short-termism (Tijssen 2004), accessing technological opportunities (Zucker and Darby 1998), and collaborators proximity (D'Este, Guy, and Iammarino 2013). Conversely, relatively less attention has been

[1] In this thesis I focus on the United States because of their centrality in the corporate science debate, after the shift in global scientific leadership from Western Europe to the US after WW2 (Mowery 2009).

directed towards issues related to the scientific reasons why companies collaborate with universities.

Growth in collaboration with universities – whose traditional research focus is basic scientific understanding – is particularly puzzling if firms are losing interest in science. In Chapter 2, I argue that one reason for the shift from direct scientific discovery to collaborations is the increasing weight of the "burden of knowledge", which reflects the increasing complexity of science due to the accumulation of previous knowledge. This trend is anticipated to affect all firms, with a more pronounced impact in fast-moving fields. I argue that the extent to which the burden of knowledge influences a firm's propensity to collaborate depends on the internal resource constraints it faces. Firms with limited capacity find it challenging to address complex scientific problems within their R&D departments, compelling them to seek help from external institutions, such as universities.

The decline of corporate science is usually associated with an increased focus on short-term results and commercialisation (Lim 2004; Tijssen 2004; Arora, Belenzon, and Patacconi 2018). This shift is reflected by the type of research conducted by firms moving from longer-term basic-oriented research projects to more short-term and applied projects. Despite some studies confirming this shift (Arora, Belenzon, and Patacconi 2018; Lim 2004), there is ambiguity in the conceptualisation of basic versus applied research. Studies on the decline of corporate science built upon previous research on the economics of business R&D, which recognises that the two can be complementary. However, they implicitly assumed a linear view of the science-technology relationship, based on which basic and applied research would be mutually exclusive, with an increase in the latter always coming at the cost of a decline in the former. In Chapter 3, I argue that existing studies fail to adequately consider Stokes' (1997) critique of the linear model and its proposed taxonomy, according to which, depending on the discipline, their incentives, and the historical context, scientists can find themselves in the position of both pursuing fundamental research questions on the laws of nature and some immediate technological applications, as Luis Pasteur did at the dawn of microbiology. This implies that showing that corporate science has become increasingly applied does not suffice to prove its disengagement from basic science.

Shareholders' pressure and short-termism are contributing factors to the decline of corporate science and are prevalent attributes of publicly traded companies. Private companies may conduct scientific research with less pressure and legal requirements, thus achieving better results than public companies. Research points out that changing the organisational structure

from private to public negatively impacts innovation (Markovitch, Steckel, and Yeung 2005; Moorman et al. 2012; Wu 2012; Aggarwal and Hsu 2014; Bernstein 2015; Wies and Moorman 2015; Gao, Hsu, and Li 2018). However, there is no evidence of the impact on corporate science in more general terms and its relationship with technology.

THESIS' CONTRIBUTION

This PhD dissertation aims to bridge existing gaps in the corporate science literature discussed above by means of an original empirical work. Through data linkage, indicator development, and econometric analysis, this research sheds light on the changing nature of corporate scientific research.

Chapter 1 sets the scene by describing the data collection process, the methodology employed for data integration, and the resulting dataset, which I name "WoS-Orbis US corporate science database", with reference to its two main sources and geographical scope. I combined United States (US) publications from Web of Science (WoS), US company information from Orbis, patents citing scientific literature from the "reliance on science" database by Marx & Fuegi (2020), and patents from PATSTAT and Orbis.

The matching exercise was particularly challenging due to the extensive time period it has to cover, with company names and ownership structure changing over time. While a perfect ownership structure is hard to reconstruct, I draw upon publicly available data from Arora, Belenzon, and Sheer (2021b) and Orbis ownership data to minimise errors as much as possible. Additionally, the computational effort required to compute the matching algorithm is significant and requires the use of Swinburne's University supercomputer, OzStar.[2]

The resulting database comprises large US publicly listed companies and their subsidiaries as in Arora, Belenzon, and Patacconi (2018), Arora, Belenzon, and Sheer (2021a; 2021b), and it also includes small and medium firms as well as all firms with a US address, regardless of their headquarters location, sector, or patenting activity. The advantage of this database lies in its broader coverage, which enables me to access scientific publications from private SMEs and foreign subsidiaries. The final sample comprises 91,374 firms that published 979,171 publications and 5,095,749 patents from 1980 to 2014.

[2] See OzStar supercomputing, https://supercomputing.swin.edu.au/, Accessed 1 September 2023

My data reveal that publications with at least one author affiliated with a US-domiciled company have increased from 8% of all peer-reviewed publications to 10% in 2000 and then declined afterwards, returning to levels close to 8%. The industries with the most publications are business services, pharmaceuticals, electronic equipment, computers, and healthcare. The fields with the most publications are engineering, medical sciences, physics, chemistry, and biological sciences.

Corporate science has declined at least since 2000, but not for every scientific field. Medical, biological and agricultural sciences increased since 1980, moving from around 3% to 7%, 10% and 10%. This evidence is consistent with Arora et al. (2019). Physics, engineering, computer sciences and chemistry, instead, show downward patterns. This evidence supports the narrative on the decline of corporate science, as the declining fields align with those of companies experiencing declines, such as AT&T, Xerox, and Dupont. Last, geosciences and mathematical sciences show stable patterns.

Chapter 2 focusses on university-industry collaborations and their evolution in the general context of corporate science's decline. Using the WoS-Orbis US corporate science database from 1980 to 2013, I report an increase in university-industry collaborations from less than 2% of all peer-reviewed publications in 1980 to almost 6% in 2013. Conversely, publications with business affiliations have been shrinking from 6% to less than 2%. Exploiting variation within firms' portfolios of publications, I show that the propensity to collaborate with universities increases if the speed of scientific progress is faster. This evidence shows that the increase in collaborations is influenced by factors related to the nature of science, such as the burden of knowledge, and not only by commercial considerations of the companies. Furthermore, the relationship between speed and the likelihood of collaboration is not linear and is mediated by firm size. The effect of the speed of scientific progress on the likelihood of collaboration increases for higher values of firm size, reaching its peak around 2670 employees. Beyond that threshold, the likelihood of collaboration decreases.

Chapter 3 explores the relationship between basic and applied corporate science. I measure the basicness and appliedness of corporate science relative to academic science and its changes over time. Following Stokes (1997), I measure basicness and appliedness as two complementary indicators. Using the WoS-Orbis US corporate science database from 1980 to 2014, I find that corporate science became more applied and less basic than university science, regardless of firms' size or age. The scientific fields in which the increase in appliedness is more evident are agricultural sciences, biological sciences, computer science, geosciences,

medical sciences, and physics. While biological sciences, computer sciences, engineering, and physics declined in basicness. In most fields the magnitude of the coefficients increases in absolute value over time. University-industry collaborations also reflect this pattern, although the magnitude of the coefficients is relatively lower. I interpret these findings as indicative of firms moving away from Pasteur's quadrant, where fundamental understanding and consideration of use coexist, to Edison's quadrant, where consideration of use is the main rationale.

Chapter 4 addresses the hypothesis that shareholders' pressure may be among the causes of corporate science's decline. I study whether initial public offerings (IPOs) positively or negatively impact corporate science. I test the causal impact of going public on firms' scientific output, using data on the population of US IPOs from 1996 to 2010. I consider only firms that published or patented in the five years before going public. My empirical strategy involves a treatment group of firms that completed an IPO and a control group of firms that filed for an IPO but afterwards decided to withdraw their filing. Identification is achieved with a stacked difference-in-difference specification and an instrumental variable. The results show a positive effect of IPOs on scientific output, measured as scientific publications and university-industry collaborations. Firms' increased access to capital and the inflow of new scientists likely drive the effect. I find no effect on papers' number of forward citations, basicness, and appliedness.

In these four chapters, I contribute to the literature on the decline of corporate scientific research and open to future research avenues. Leveraging one of the largest available databases, I provide a systematic picture of US corporate science and university-industry collaboration. While my findings align with previous research, they reveal a more nuanced perspective on the decline, highlighting shifts from basic research towards more applied science, as well as a surge of university-industry collaboration. Additionally, for the first time, I connect the debate on the decline of corporate science with the literature on corporate finance and innovation, particularly in relation to firms' innovation strategies following their initial public offerings.

Chapter 1: The WoS-Orbis US corporate science database

This chapter introduces the database used throughout this thesis, to which I will refer as the WoS-Orbis US corporate science database, with reference to its two main sources and geographical scope. I provide a comprehensive description of the methodologies and algorithms I employed to build it and some descriptive statistics, which help delineate the general trends in corporate science one can infer from it. For the sake of synthesis, I only delve here into some of the technical details of the data collection process. Still, more details can be found in Appendix A.

This dataset spans from 1980 to 2014 and contains 91,374 firms, 979,171 publications, and 5,095,749 patents. The most comparable dataset in size and scope is the one built by Arora, Belenzon, and Sheer (2021b). Key differences include a more extensive set of firms, including all firms with US addresses, including foreign subsidiaries, regardless of their R&D spending, patenting activity or sector. The WoS-Orbis US corporate science database also includes all direct and indirect citation links between corporate publications and patents.

1.1 DATA

The database combines the following data sources.

Orbis: Orbis from Bureau Van Dijk is one of the largest datasets currently available on public and private firms. I used two different versions of Orbis: Orbis Historic (sold until 2012) and the current edition of Orbis (2017) in order to maximize the number of firms and historical series of financial information. I select more than 80 million firms across US public and private companies, plus foreign subsidiaries. I include all firms with US addresses from all sectors of activity at any point in time. Orbis contains balance sheet information and other information such as industry, size, incorporation date and industry. Additionally, Orbis contains ownership structure information and patents from 2005 to 2017. Further information on how to build a representative sample in Orbis and its coverage is available in Bajgar et al. (2020), and Kalemli-Ozcan et al. (2015, 2019).

Web of science: Web of Science (WoS) from Clarivate Analytics is a widely used source in academic studies concerning scientific publications. WoS contains 10,027,418 articles,

proceedings, and notes with at least one affiliation with address located in the US from 1980 to 2014. I exclude books, reviews, and editorial material, and content from social sciences and humanities. I exclude conference proceedings from the analyses in Chapter 2 and 3 because of the limited coverage. WoS contains information on publications date, journal, scientific fields, author names, and their respective affiliations and addresses. Further details on WoS coverage and comparison with other similar databases are available in Mongeon and Paul-Hus (2016), Martín-Martín et al. (2021), Singh et al. (2021).[3]

The "reliance on science" database: The "reliance on science" database by Marx and Fuegi (2020)[4] contains approximately 22 million non-patent literature (NPL) citations from worldwide patents to scientific publications from 1800 to 2018, both from the front page and in-text. The source of information on publications' authors, titles, journals, and other bibliographic data comes from MAG, the Microsoft Academic Graph.[5]

PATSTAT: PATSTAT from the European Patent Office (EPO) is one of the most used databases on patent data. I use the 2021 edition and I select USPTO, WIPO and EPO patent until 2014. I use PATSTAT information to complement Orbis and the reliance on science database that lack information such IPC code, inventors' names and addresses, and application date.[6]

Figure 1.1 shows graphically the database construction. The numbers in the figure indicate the sequence of data operations I performed and correspond to the following numbered bullet points.

1. First, I connected firms in Orbis to authors' affiliation in WoS by means of a original purpose built decision tree algorithm.
2. Second, I connected the WoS-Orbis matched companies to the "reliance on science" database, by identifying all patents in the latter, that directly cite one or more

[3] Web of Science. Clarivate analytics. Accessed 14 June 2022. https://clarivate.libguides.com/librarianresources/coverage

[4] Reliance on science in patenting (2022). Zenodo. Accessed 14 June 2022. https://zenodo.org/record/6629738. Currently there is new version available, however, it was not available at the time of the database construction.

[5] Microsoft Academic Graph (2022). Microsoft Corp. Accessed 14 June 2022. https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/. Processed though PostgreSQL using Chacua (2020)

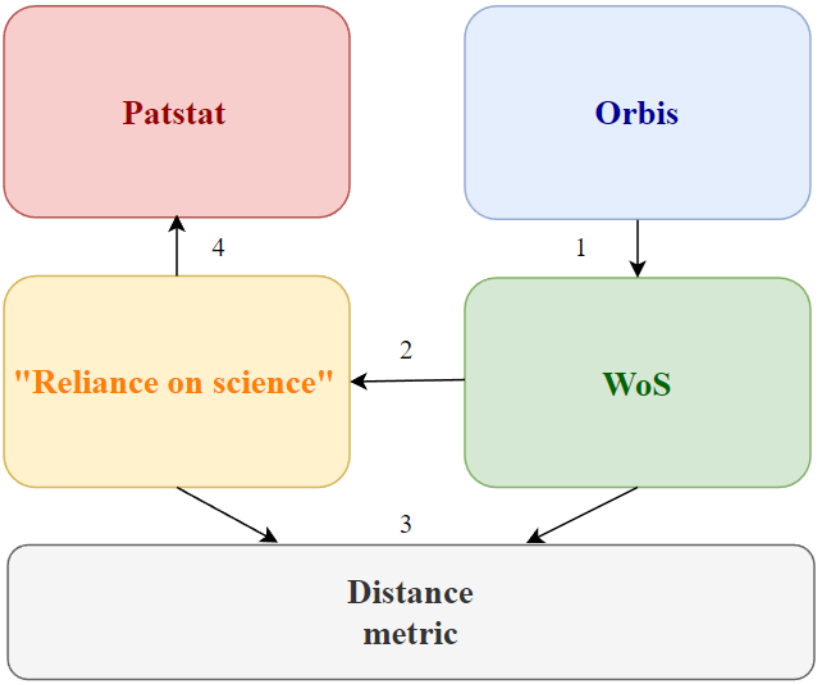
[6] PATSTAT (2021). European Patent Office (EPO), Accessed May 2023 https://www.epo.org/searching-for-patents/business/PATSTAT.html.

scientific publications in WoS, whether of the matched companies or not. Following Ahmadpoor and Jones (2017) terminology, I will refer to these patent-publication matches as the dual frontier or the science-technology frontier.

3. Third, based on Ahmadpoor and Jones' (2017) methodology, I connected all WoS publications to the publications at the science-technology frontier via their forward citations, either to the publications on the frontier or to the publications that cite them directly or indirectly via other publications.

4. Fourth, I enriched information on patents on the science technology frontier by matching them to patents in PATSTAT.

Figure 1.1: Schematic representation of the database linking procedure



Note: Schematic representations of the data sources. Orbis (blue), Web of Science (green) the "reliance on science" database (yellow), and PATSTAT (red). The numbers indicate the order in which I connected the data sources.

1.2 MATCHING ORBIS AND WEB OF SCIENCE (1)

Figure 1.2 shows schematically the matching technique. I employed an original decision tree algorithm to match affiliations reported in scientific publications in WoS to companies' names

in Orbis. The algorithm incorporates string similarity scores (Levenshtein distance[7]), shared non-dictionary words, and address information (same city or zip code).[8]

A perfect match ideally presents identical affiliation names, the same city and zip code, and at least one shared non-dictionary word (e.g., Siemens Med Solutions, Charlestown, 02129 in WoS and Siemens Med Solutions, Charlestown, 02129 in Orbis). A less than perfect match would present similar affiliation names, the same geography, or at least one shared non-dictionary word (e.g., Focus bio inova, Herndon, nan in Wos and Focus bioinova, Herdon, 20171 in Orbis).

One of the greatest challenges faced in matching WoS and Orbis was optimising the computational efficiency. Matching every affiliation in WoS with every company in Orbis would require computing 600k by 80 million interactions ($48*10^{12}$ possible matches). To reduce the number of possible permutations I reduced the number of potential matches in the following ways.

To begin with, I used recognisable keywords to exclude from WoS and Orbis universities (university, college, institute of technology, school, faculty etc.), government agencies (us department, us army, us navy, NSF, NASA etc.), not-for-profits (aquarium, botanical garden, zoo etc.) and medical centres (clinic, hospital etc.).[9]

Furthermore, I used web address information from Orbis to infer the legal status of the institutions. I assumed that web addresses ending with .edu belong to a university, .gov to the government and .org to a not-for-profit.

Last, I grouped the data in smaller groups (or "bins") performing the matching only on affiliations that have at least one word in common. For example, consider Merck Research Laboratories shown in Figure 1.2. Initially, I abbreviated the affiliation name into "merck", "res" and "labs". Then, I assessed the number of Orbis rows containing each of these words.

---

[7] The smallest number of insertions, deletions, and substitutions of a single character, needed to change one string into another (Levenshtein 1966; Yujian and Bo 2007).

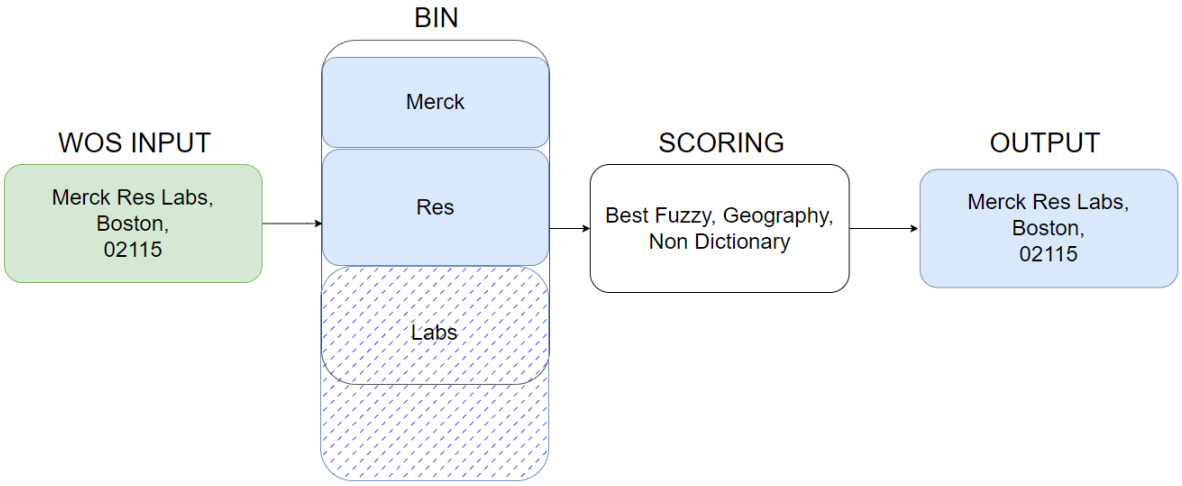
[8] I inherited the algorithm structure (the slicing and parallel job structure of section, the sequence of the actions performed by the algorithm, the way in which the results are saved as in Section A.3) from the work previously done by the Centre for Transformative innovation at Swinburne University, by Prof. Russell Thomson and Dr. Alex Codoreanu. My original contribution has been modifying and improving all the core sections of the algorithm: the abbreviations, the binning and the scoring. Furthermore, I worked with different inputs, merging together different versions of Orbis and a bigger set of Wos Articles.

[9] See the appendix in Section A.4.2 for a more comprehensive list of keywords.

Subsequently, I ordered those groups from the smallest to the biggest: “merck” with 100,000 rows, “res” with 200,000 rows, “labs” with 300,000 rows.[10] Next, I set the maximum bin size equal to 400,000 rows,[11] and I added the groups into the bin, starting from the smallest group and continuing adding groups until the 400k threshold is reached. In this case the rows with the word labs are excluded because they would exceed the maximum size of the bin. Last, I performed the matching on the remaining rows.

The last step involves the selection of the best results. The best results are those matches where there is the highest string similarity score, the highest similarity score combined with non-dictionary words in common, or the highest similarity score paired with shared geography.

Figure 1.2: Schematic representation of the matching algorithm



Note: This figure shows schematically the matching procedure and the “binning” (grouping data in smaller groups). On the left there is the WoS affiliation (green). At its right there is the bin, that is the group of firms in Orbis that have at least a word in common. In this case merck, res, and labs, sorted in order of bin size. The following step is scoring the match results and select the best Orbis match (blue).

1.2.1 Diagnostics

Following the match, I have tested the performance of the matching algorithm by calculating its precision grouping the potential matches in 27 scoring intervals.[12] The best matches present

[10] The figures are used as a hypothetical scenario to describe intuitively the binning procedure and they do not correspond to the real number of rows in Orbis.

[11] The choice of the maximum bin size is discretionary. After several trials, settling on 400,000 rows proved to be a favourable compromise between computational efficiency and precision.

[12] I cannot compute the more classical measure of recall (the ratio between true positives and the sum of true positives and false negatives) because I do not possess a “golden set”, i.e., a set of matches that are unambiguously

100% string similarity score, non-dictionary words and geography scores (1). Afterwards, I considered 100% string similarity score, non-dictionary words and shared zip code (2). Then 100% string similarity score, non-dictionary words and shared city (3), 100% string similarity score and shared geography (4), 100% string similarity score and shared zip code (5), 100% string similarity score and shared city (6), 100% string similarity score and non-dictionary words (7). The same mechanism applies to the scoring intervals with string similarity score from 99% to 90% (8-13). Intervals 14 and 15 require string similarity from 90% to 99% and shared non-dictionary words, or just 100% string similarity. From interval 16 to 27, instead, it is required the presence of a non-dictionary word and at least the city or zip code in common. The scoring proceeds as follows: 89% to 80% (16-18), 79% to 70% (19-21), 69% to 60% (22-24), and 59% to 50% (25-27). All the remaining cases are considered as unmatched.

The presence of non-dictionary words and geography scores serves as effective indicators for identifying a successful match. It allows keeping matches with lower string similarity with a little trade-off for precision.

Table 1.1: Scoring intervals of the matches, description, and summary statistics

Score	Description	N affiliations	N publications
1-4	Fuzzy 100, non-dictionary and geography scores	116,346	638,996
5-7	Fuzzy 100 and geography, Fuzzy 100 and non-dictionary	102,494	323,643
8-11	Fuzzy 90≤x<100 geography and non-dictionary	17,628	45,367
12-13	Fuzzy 90≤x<100 and geography	14,065	29,489
14-15	Fuzzy 100, Fuzzy 90≤x<100 and non-dictionary	74,532	179,889
16-21	Fuzzy 70≤x<90 geography and non-dictionary	16,772	43,446
22-27	Fuzzy 50≤x<70 geography and non-dictionary	12,746	33,914
Unmatched		200,130	526,553

Note: This table presents the scoring intervals, the scoring criteria and the number of affiliations and publications matched for each scoring interval.

true to compare my algorithms results with. A possible strategy would be, as in Marx and Fuegi (2020) to hire some research assistants and build manually a golden set and then compare it with the algorithm results.

1.2.1.1 Precision

Precision is calculated as the ratio of true positives and the sum of true positives and false positives. I selected a stratified random sample of 100 matched affiliations per scoring interval and I checked manually for true positives and false positives. When affiliation names are identical, and geographical information coincide, the assessment of the truthfulness of the match is self-explanatory. However, in instances where the assessment was less evident, I performed an internet search to verify whether the two affiliations were indeed the same.

Table 1.2: Precision in a random sample of 100 affiliations per scoring interval

Score	N affiliations	N publications	% True positives	Precision (cumulative)
1-5	122,846	656,533	100%	100%
6-11	102,494	323,643	94%	97.27%
12-13	14,065	29,489	87%	96.66%
16-21	16,772	43,446	90%	96.23%
22-27	12,746	33,914	67%	94.84%
14-15	74,532	179,889	?	?

Note: This table shows the precision for a random sample of 100 matches. Each row represents a different scoring interval. The fourth column shows the % of true positives within each scoring interval, while the last column displays the cumulative precision from score one up to the given scoring interval.

In Table 1.2 I show the accuracy of matches within every scoring interval, and the cumulative precision. Matches in the scoring interval 1-5 are almost error free. Precision lowers to 97.27% for the 6-11 interval, to 96.66% in the 12-13 interval, to 96.23% in the 16-21 interval and finally to 94.84 in the 22-27 interval. It is important to remark that the scoring interval 1-11 captures 77.82% of publications and 64.40 % of the affiliations with a precision of 97.27%. Assessing the precision of the scoring intervals 14-15 presents a more intricate challenge. Given the absence of shared geographical data, the only method to assess the quality of the match involves a manual internet search for both company names. However, despite implementing this method, occasionally the correctness of the match cannot be assessed with certainty, especially for older or lesser-known firms. While these matches are included into the sample, they necessitate further scrutiny and examination.

1.2.1.2 Unmatched companies

I found it impossible to match WoS to Orbis affiliations in two main scenarios. First, some companies may be absent in Orbis. This occurrence is more frequent for old firms when Orbis has lower coverage. Second, some potential matches have a low score. The algorithm is not

precise when handling unusual abbreviations, long company names with many uninformative tokens, and short single-word companies.

1.2.2 Ownership

Changes of ownership are frequent and are relevant to building this dataset. A big publisher like A&T was broken up into many pieces after the 1974 lawsuit. First, Western Electric became a separate entity, and then Lucent was created in 1996 from the former Bell Labs and AT&T technologies. Large companies are active in sectors such as pharmaceutics and biotech, often acquiring many promising firms. For example, Pfizer acquired Warner-Lambert in 2000, Wyeth in 2009, King Pharmaceuticals in 2010, and Hospira in 2015. Missing these ownership relations might underestimate/overestimate the firms' scientific engagement.

I carefully considered the ownership structure of the matched firms, using Orbis ownership data from 2005-2017 and complementing them with Arora, Belenzon, and Sheer (2021b) open-source data[13] and Orbis M&A.[14] A certain level of inaccuracy, due to the complexity of ownership arrangements and limited availability of ownership data, however, is inevitable. I minimised the errors by augmenting the data in the following way. First, I integrated data from different servers in Orbis, obtaining a 10-year series. Second, I used future ownership data to infer previous ownership. If the company name of the future owner has a high string similarity score regarding the past subsidiary name, I assume that the owner has not changed, and I use the future ownership data to induct the past ones. Last, I grouped together companies with a similar name and a non-dictionary word in common (ex: Exxon chem, Exxon chem america, Exxon Res & Engn) and manually checked the ownership of the firms with the most publications. When it is not possible to retrieve ownership data, the difference in company names calls to independence.

In addition to the data complexity, ownership is not binary, and due to ongoing M&A activity the boundaries between parent and subsidiaries are difficult to measure unambiguously. I assumed that the major shareholder is responsible for the research conducted by the firm, however, minor shareholders – such as a university or an independent institution – may be responsible too.

---

[13] Duke Innovation & Scientific Enterprises Research Network (DISCERN, 2020). Duke University. Accessed 14 June 2022. https://zenodo.org/record/4320782

[14] The database was previously known as Zephyr and it changed its name into Orbis M&A.

1.3 CONNECTING PATENTS TO THE RELIANCE ON SCIENCE DATABASE (2)

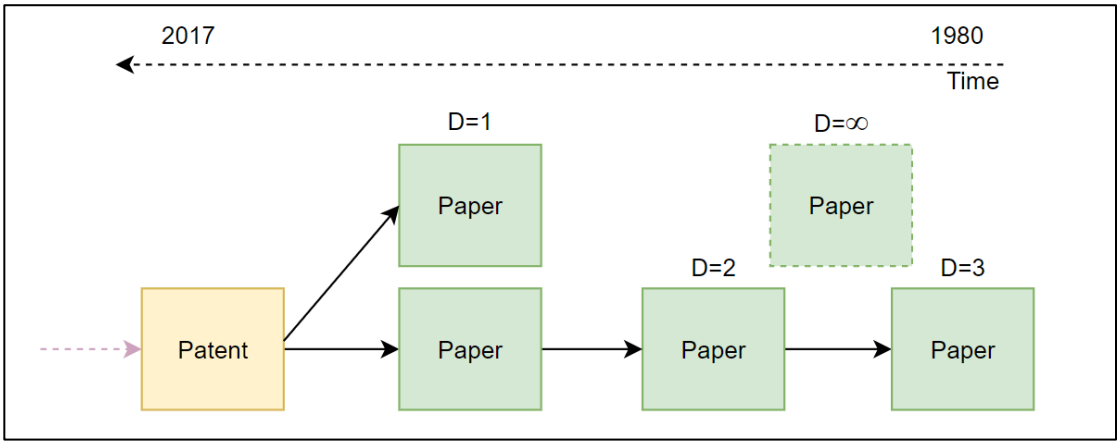
The next linking operation concerns the WoS publications (of which a subset is the WoS-Orbis matches) and the reliance on science database. Publications directly cited by patents constitute the "science-technology" or "dual" frontier, namely where science and technology connect.

The publications at the frontier come from MAG. To match WoS to MAG publications I establish matches based on criteria such as doi, title, volume, start and end page, authors, journal, and year. I was able to match 78.09% of the publications at the frontier to Web of Science for a total of 3,967,871 publications. Of those publications 266,177 are corporate publications.

1.4 LINKING ALL PAPERS TO THE DUAL FRONTIER (3)

Building on Ahmadpoor and Jones (2017) and using the network of backward citations generated from the science-technology frontier, I can calculate the distance between the publications in my database and the science-technology frontier. Any publication on a citation chain, ultimately leading to the science-technology frontier, stands at distance D=k from the frontier, where k is the number of publications along the chain. Publications that do not connect to the frontier by any path are considered *unlinked*. The distance metric is described in more detail in Chapter 3.

Figure 1.3: Distance metrics, visualisation



Note: Schematic representation of the distance metric. Patents are on the left in yellow while papers are in green. The dashed arrow indicates the time direction, while the solid arrow the direction of the backward citations. Papers directly cited by a patent are at distance D=1, papers cited by a paper at distance D=1 and not by a patent are at distance D=2. The same applies to D=3. The unlinked paper is at distance D=∞ and it is not included in the sample.

1.5 COLLECTING PATENT INFORMATION (4)

I obtained the firms' patent portfolios through Orbis. Orbis, however, does not report information such as IPC class, inventors, and applicants' names. In order to collect this information, I linked PATSTAT to Orbis through the patent publication number. I collected additional patents retrieving the patents' linked to PATSTAT's *person id*s corresponding to Orbis companies.

1.6 DESCRIPTIVE STATISTICS

The descriptive statistics of the resulting sample are presented in Table 1.3. I was able to identify 979,171 papers and conference proceedings, written by 91,374 companies. Of these firms 53,304 engaged in 397,951 publications that are co-authored with a university. Financial information is usually available from 2005 to 2014, with longer series for a small subset of large firms. Turnover and employees are the most available financial variables, while R&D data are available only for a subset of 3,997 firms.

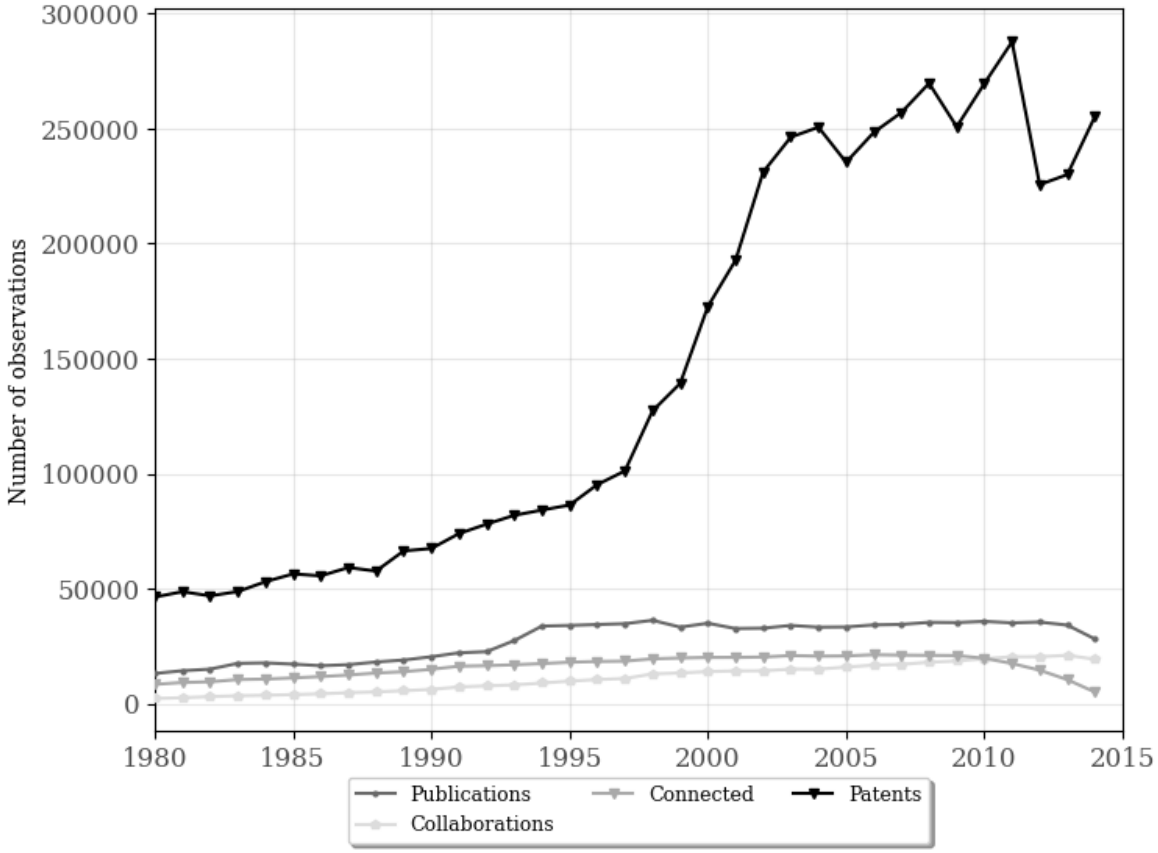
Table 1.3: Descriptive statistics of the final sample

Variables	# observations	# firms	Mean	S.D
Publications	979,171	91,374		
Collaborations	397,951	53,304		
Turnover	220,080	55,609	1.23E+09	1.04E+10
Employees	207,941	55,533	4.50E+03	2.96E+04
R&D	45,946	3,997	1.32E+08	6.58E+08
Patents	5,095,749	460,707		
Linked to science	567,585	53,307		

Note: This table presents the descriptive statistics for the thesis sample. Publications are papers and proceedings with at least one corporate affiliation. Collaborations are publications with at least one corporate and one university affiliation. The figures displayed for turnover, employees, R&D, and patent data show the values before consolidating the companies at the Ultimate Owner level.

Figure 1.4 shows the total number of publications, collaborations, papers connected to the frontier, and patents. As expected, innovative firms engage in patenting more than in scientific publications, and this is visible after the 1990s, when the growth in publications did not keep up with the surge in patenting. It is already noticeable that the number of university-industry collaborations has increased and represents a high share of the total corporate publications.

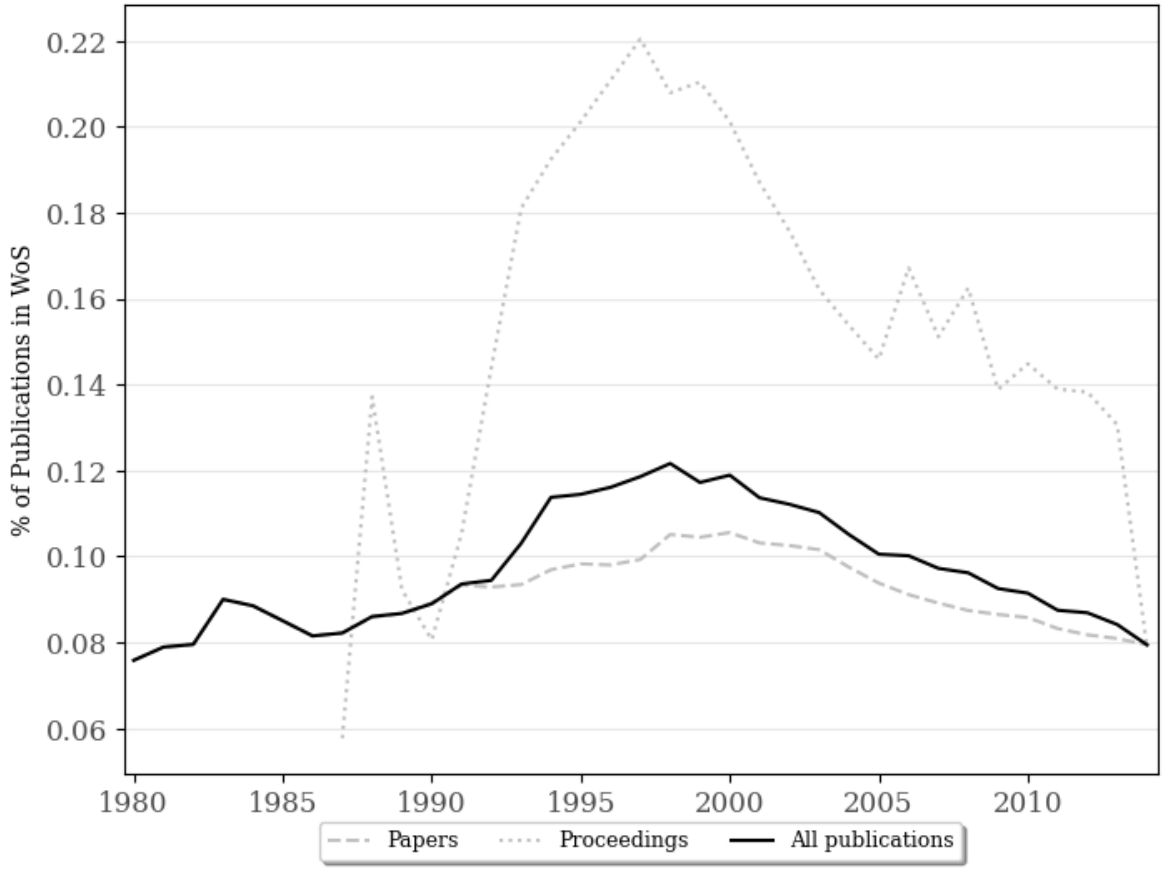
Figure 1.4: Patents, publications, collaborations, and connected publications, 1980-2014



Note: This graph shows the number of publications, collaborations, connected publications and patents in the sample. The sample spans from 1980 to 2014.

Figure 1.5 shows the share of corporate publications (with at least a corporate affiliation) in the whole WoS (including universities, government agencies, and not-for-profit). This approach allows us to see long-term differences between non-corporate and corporate science as it considers consistent journal coverage across years. The dashed line shows the share of papers, which exhibits a stable pattern ranging from 8% to 10%. The dotted line shows the conference proceeding shares, which present an inverted U-shaped trend, increasing from 1990 to 1997 and declining until 2014. Finally, the share of all corporate publications (solid line) is stable, ranging from 8% to 12% of total peer-reviewed publications. It represents the combined trend of papers and conference proceedings. The share of conference proceedings is higher than the papers as it is relatively more common for companies to publish in the former. However, due to their smaller absolute numbers, they have a relatively low influence on the overall publication trends.

Figure 1.5: Share of corporate publications as share of total WoS US publications, 1980-2014



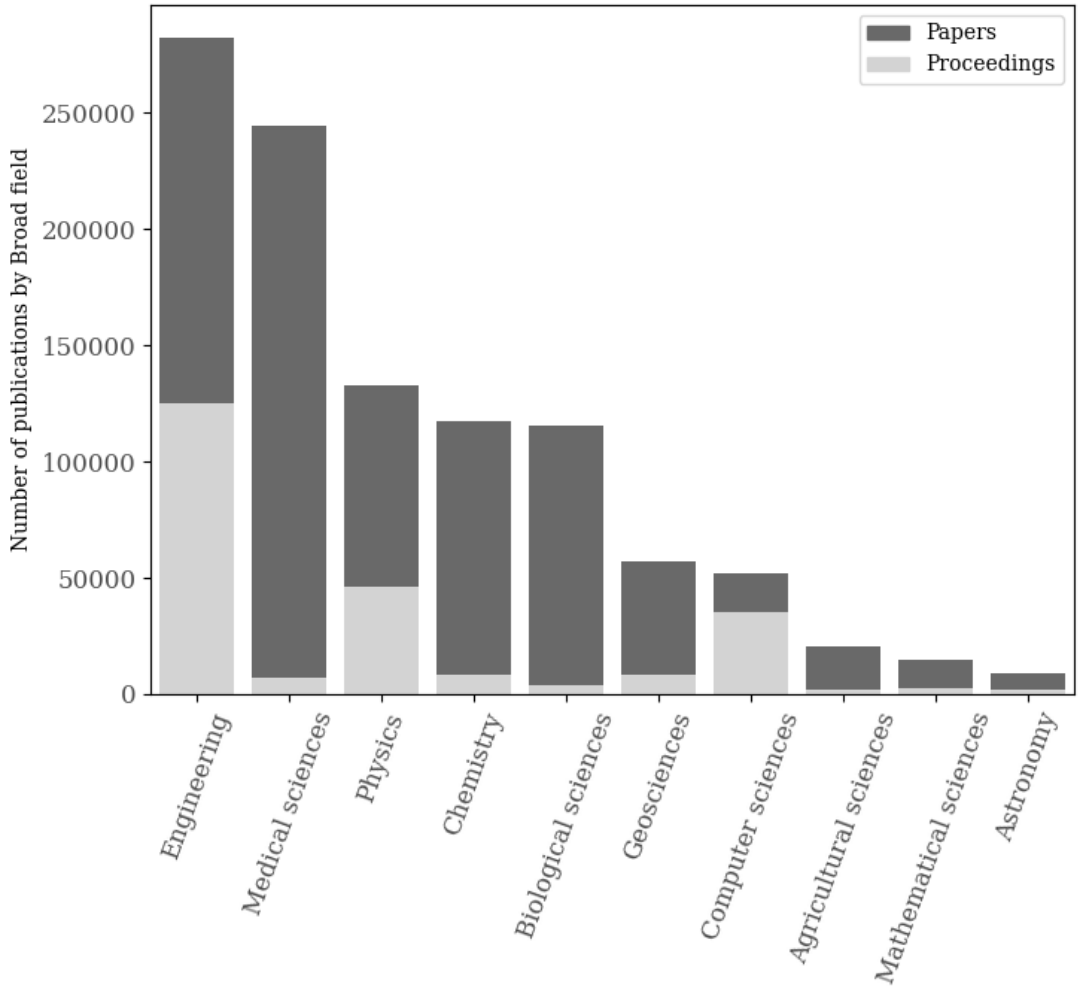
Note: This figure shows the share of publications (solid), papers (dashed), proceedings (dotted) as share of total WoS publications, papers, and proceedings. Only publications with a United States address are included.

Proceedings are published literature of conferences, symposia, seminars, colloquia, workshops, and conventions in a wide range of disciplines. IEEE outlets, the most important outlets in engineering subjects, are listed in WoS as papers and thus are present in our sample since the 1980s. The remaining proceedings are excluded from the analysis of the following chapters because of data quality issues. First, conference proceedings data are available only from the 1990s and not from the 1980s. Second, the coverage increases until 2000 and then declines sharply. This phenomenon creates imbalances in the shares of publications in WoS, given that companies have a higher propensity to publish in conference proceedings. I also excluded the year 2014 from the analysis due to right truncation.[15] More information about WoS coverage can be found in Section A.2.

[15] The year 2014 is excluded for the analyses in Chapter 2 and Chapter 4. It is retained for the analysis in Chapter 3 because the identification strategy involves comparing a group of corporate publications and collaborations with a control group of university publications. Since the right truncation affects both groups equally, it should not introduce bias in the estimates.

Figure 1.6 shows the number of papers and proceedings by broad scientific field. The WoS science categories are grouped in 10 broad fields using Milojević (2020).[16] Engineering is the field with most papers (281,872) and proceedings (124,500), followed by medical sciences with 240,087 papers and 7,134 proceedings. Next, physics has 132,280 papers and 45,979 proceedings (7,134), chemistry 116,192 papers and 8,255 proceedings, and biological sciences 113,117 papers and 3,785 proceedings. These 5 fields account for 85% of the total publications. Geosciences produces 56,149 publications and 8,151 proceedings, computer sciences 51,615 papers and 35,122 proceedings, agricultural sciences 20,408 papers and 2,053 proceedings, mathematical sciences 14,801 papers and 2,218 proceedings, and astronomy 8,603 papers and 1,521 proceedings.

Figure 1.6: Number of publications by broad scientific fields, 1980-2014



Note: This figure shows the number of papers (dark gray) and conference proceedings (light gray) by broad scientific field from 1980 to 2014. The scientific fields are grouped using Milojević (2020).

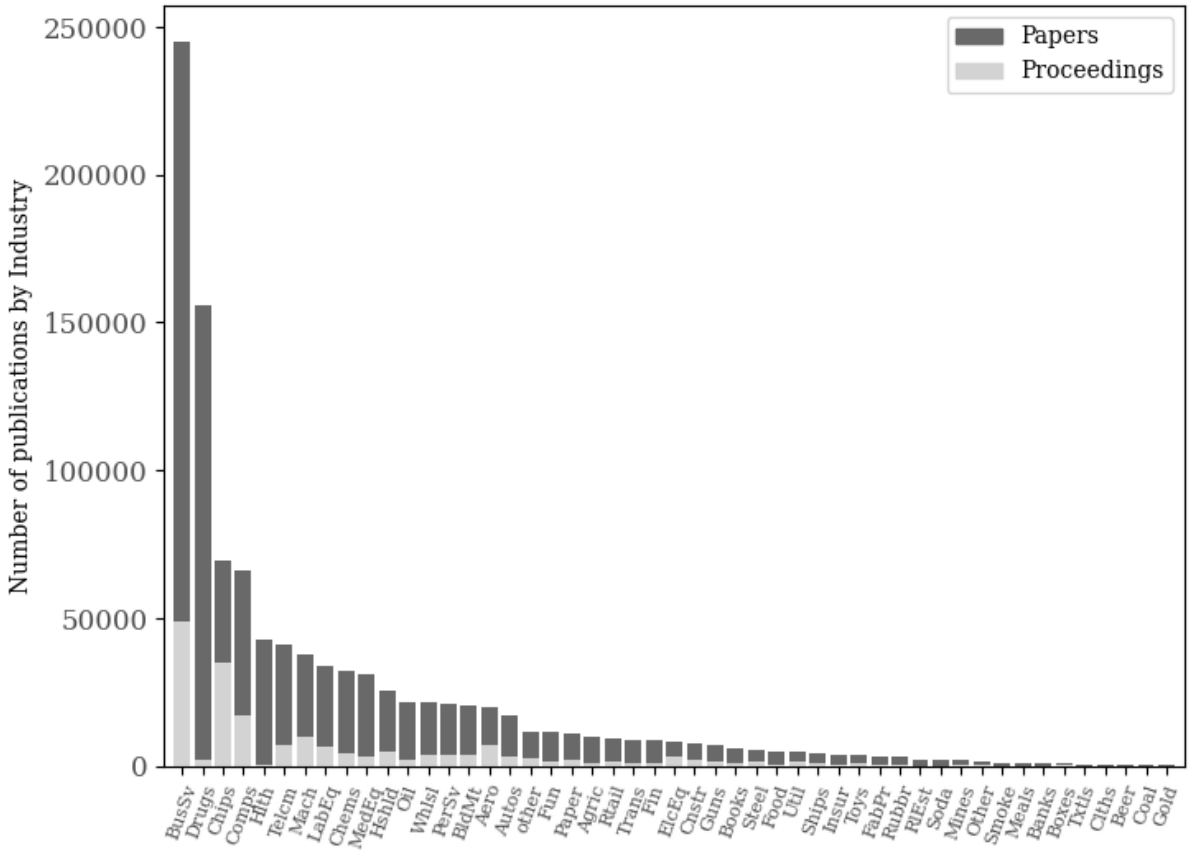
[16] The ten scientific fields are agricultural sciences, astronomy, biological sciences, chemistry, computer sciences, engineering, geosciences, mathematical sciences, medical sciences, and physics.

Figure 1.7 shows the number of publications by industry. Industries are grouped into 48 Fama-French industries[17] from the companies' NAICS industry codes. A company can belong to more than one Fama French industry, e.g., IBM's NAICS principal code is 5415 (Computer Systems Design and Related Services), that converted into the Fama French classification corresponds to "business services" and "computers". The two industrial sectors with the highest number of publications are business services (BusSv) and pharmaceutical products (Drugs) with 243,830 and 155,948 papers. Next follow electronic equipment (Chips, 69,579), computers (Comps, 66,033), healthcare (Hlth, 42,610), communication (Telcm, 41,265), machinery (Mach, 37,858), measuring and control equipment (LabEq, 34,086) , chemicals (Chems, 32,460), and medical equipment (MedEq, 31,189). The largest publishers of conference proceedings are in the ICT sectors, namely business services (49,151) electronic equipment (34,963), computers (17,054), machinery (10,143), and communication (7,309).[18]

[17] The conversion tables are publicly available at K. French (2023). Accessed 23 August 2023. https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_48_ind_port.html

[18] The remaining industries in Figure 1.7 are Consumer Goods (Hshld), Wholesale (Whlsl), Aircraft (Aero), Automobiles and Trucks (Autos), Medical Equipment (MedEq), Construction Materials (BldMt), Personal Services (PerSv), Retail (Rtail), Steel Works Etc (Steel), Trading (Fin), Construction (Cnstr), Electrical Equipment (ElcEq), Defense (Guns), Almost Nothing (Other), Food Products (Food), Utilities (Util), Agriculture (Agric), Shipbuilding, Railroad Equipment (Ships), Transportation (Trans), Insurance (Insur), Fabricated Products (FabPr), Entertainment (Fun), Business Supplies (Paper), Printing and Publishing (Books), Recreation (Toys), Real Estate (RlEst), Rubber and Plastic Products (Rubbr), Non-Metallic and Industrial Metal Mining (Mines), Tobacco Products (Smoke), Candy & Soda (Soda), Restaurants, Hotels, Motels (Meals), Apparel (Clths), Shipping Containers (Boxes), Textiles (Txtls), Beer & Liquor (Beer), Coal (Coal), Precious Metals (Gold), Banking (Banks), Petroleum and Natural Gas (Oil).

Figure 1.7: Number of publications by industry, 1980-2014

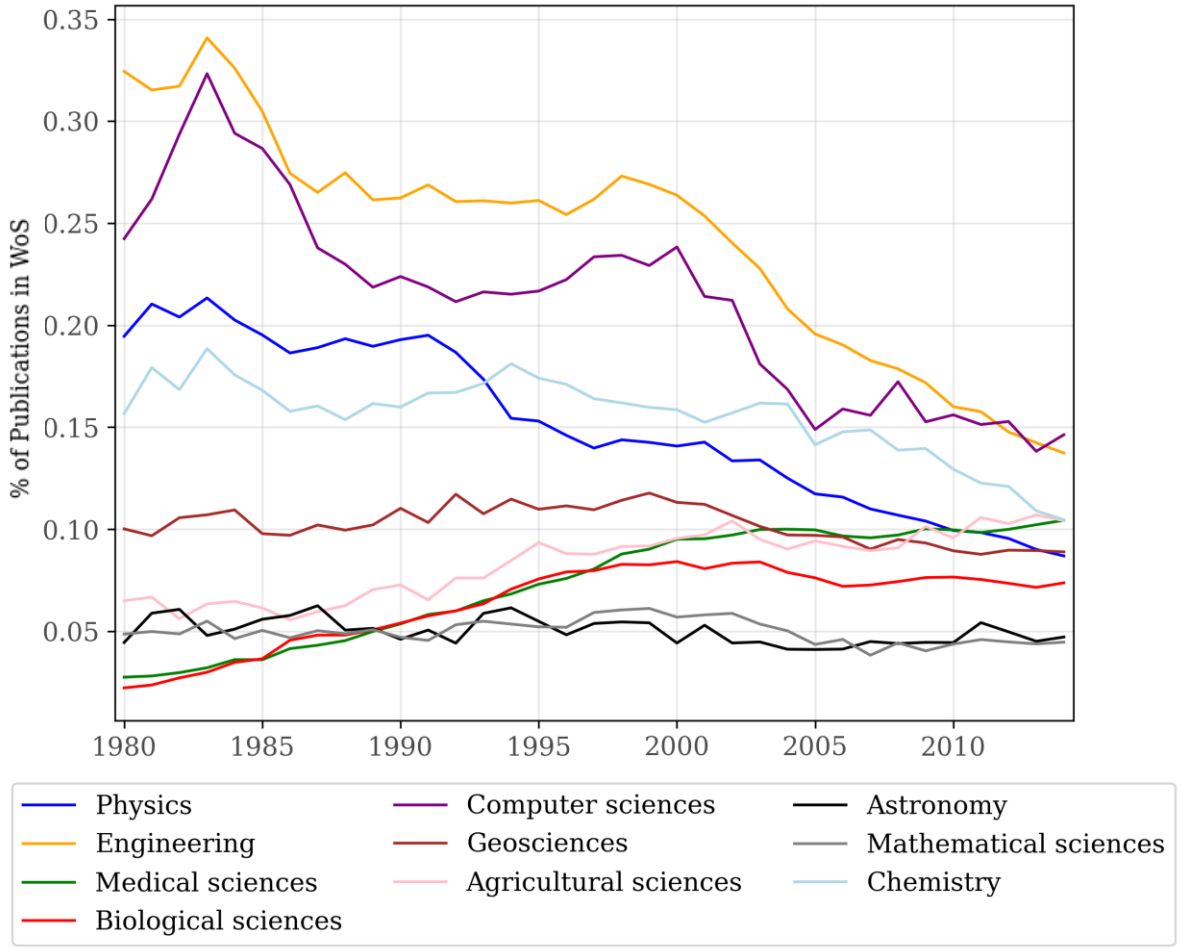


Note: This figure shows the number of papers (dark grey) and conference proceedings (light gray) by industrial sector from 1980 to 2014. Industries are classified using 48 industries Fama-French classification.

Figure 1.8 shows the share of corporate publications in WoS by broad scientific field from 1980 to 2014. The trends of corporate publications differ much by field. Physics, engineering, computer sciences, and chemistry show downward patterns, moving from around 32%, 25%, 20%, and 15% to 15%, 15%, 10%, and 9%. These fields correspond to those of the firms usually mentioned as examples of the decline of corporate science, such as AT&T, Xerox, and Dupont. Other fields, however, are not declining. Medical, biological and agricultural sciences moved from around 3% in 1980 to 7%, 10% and 10% in 2014. Given that the overall trends in corporate science are stable, ranging from 8% to 10%, this evidence suggests that the composition of the companies doing research is changing. The declining fields correspond to companies in mature industries such as semiconductors or chemicals, whereas the rising fields correspond to those in new and emerging sectors like biotechnologies or pharmaceuticals. Companies seem to persist in those sectors as Arora et al. (2019) reported. Last, geosciences, mathematical sciences, and astronomy show stable patterns, around 10%, 5%, and 5%.

The top publishers consist of old incumbents in a variety of different industries. With a publication count of 42,248, IBM holds the primate of the largest publisher and has been the world's largest patentee for 29 years. Similarly, AT&T, with 40,203 publications, is a historical company that contributed to the narrative of the golden era of corporate science. Pharmaceutical companies like Pfizer, Merck, Roche, Eli Lilly, Genentech, Novartis, and GlaxoSmithKline are heavily involved in R&D, and their publication count ranges from 31,011 to 7,403. Aerospace and aircraft industries also contribute to science, with Lockheed Martin and Boeing as main actors, with 9,400 and 7,209 publications. Alcatel-Lucent, Intel, HP, and Motorola lead the electronic equipment industry. Alcatel-Lucent, which spun off from A&T as Lucent in 1996 and then turned into Alcatel-Lucent in 2006, keeps high the commitment to science as its predecessors. General Electrics and General Motors are two companies in decline but consistently published over the years, with a total of 10,076 and 9,136 publications. Exxon and Dow Chemical petroleum and chemicals contribute with 9,693 and 7,320 publications.

Figure 1.8: Share of corporate papers in WoS by field, 1980-2014

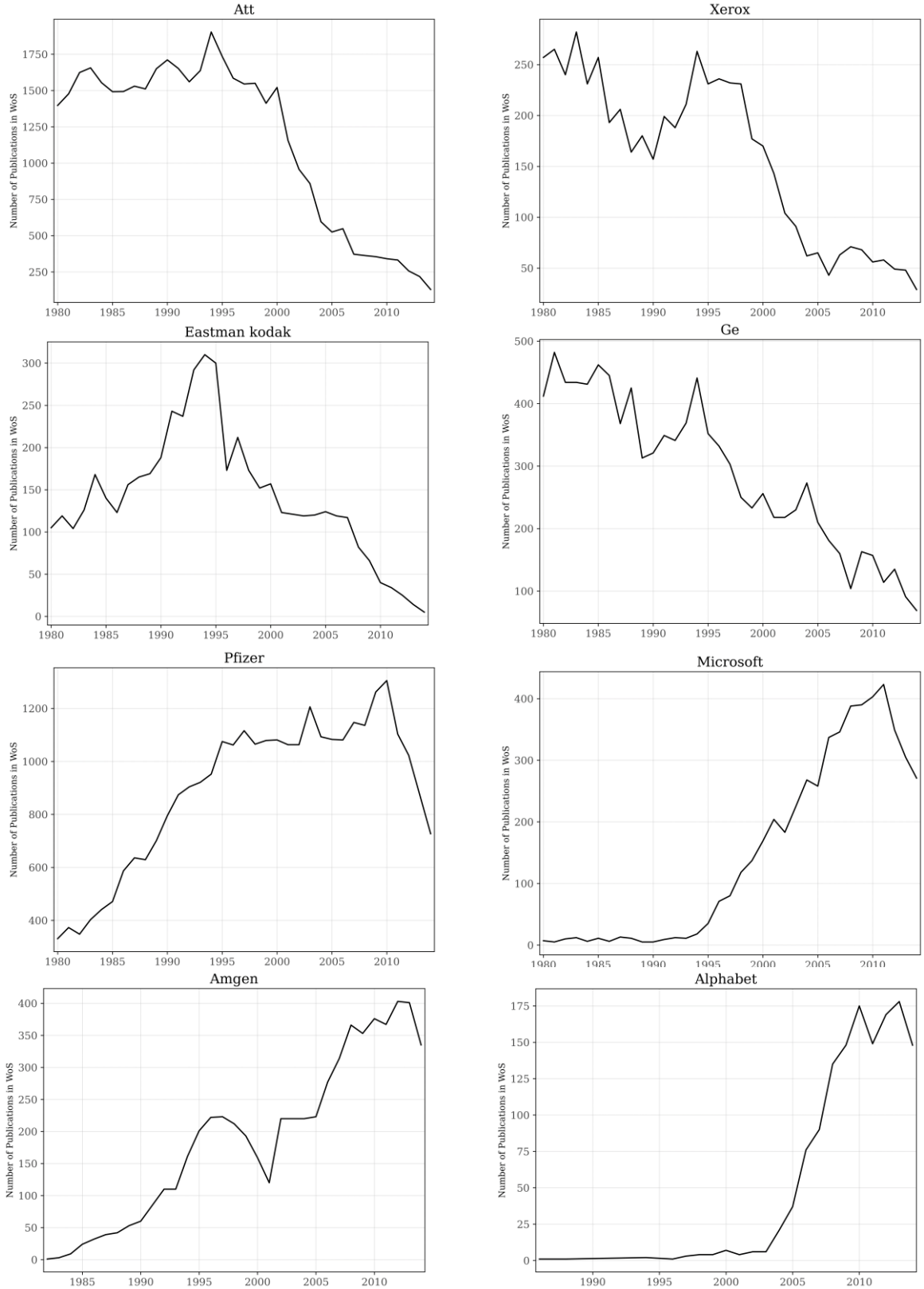


Note: This figure shows the share of corporate papers in WoS by broad scientific field from 1980 to 2014. The scientific fields are grouped using Milojević (2020).

Figure 1.9 provides a glimpse into the trajectories of various prominent firms, shedding light on their ups and downs over the years. One notable case is AT&T, which has remained a central figure in the discourse surrounding the decline of corporate science. This decline can be traced back to pivotal historical events, including the breakup of the Bell System in 1982 and the Federal Telecommunications Act of 1996, which dismantled competitive barriers in the telecommunications industry. Before these turning points, AT&T had been renowned for its groundbreaking research conducted at Bell Labs, earning Nobel Prize-level recognition, particularly in solid-state physics during the post-war era.

Another example is Xerox, which took a significant step by spinning off Xerox PARC as a wholly owned subsidiary in 2002. Likewise, Eastman Kodak, once a powerhouse with the highest number of inventors in Rochester, experienced a sharp decline in its stock price and employment following its peak in 1996. Interestingly, the publication trends align closely with these downturns, as highlighted by Moretti (2021). In contrast, some companies like General Electric chose to reorient their research strategies and did not decline because of a particular pivotal historical event. On the other hand, success stories abound with firms like Pfizer, Microsoft, Alphabet (Google), and Amgen, which increased their publication efforts in fields such as oncology, immunology, virology, and computer science.

Figure 1.9: Rise and fall of corporate publishers, 1980-2014



Note: This figure shows the publication trends of some firms actively engaged in corporate science. Starting from the top left the companies are AT&T, Xerox, Eastman Kodak, General electrics, Pfizer, Microsoft, Alphabet (Google), and Amgen. The first four companies are examples of firms shrinking their number of publications, while the remaining four increased their publication output.

1.7 CONCLUSIONS

In this chapter, I detailed how I built the WoS-Orbis US corporate science database and joined together firms' financial information, scientific publications, patents, and publications linked to science. The richness of this dataset consists in linking a large amount of information to a broad set of firms, including small and medium enterprises and foreign subsidiaries. The technical details of the database construction are listed extensively in Appendix A.

At first glance, no decline in corporate science is detectable until 2000, while a downward pattern can be seen afterwards. These trends vary by scientific field, as not every field experienced a decline. Publications in biological sciences and medical sciences increased from 1980 to 2014. This evidence is consistent with Arora et al. (2019) as companies persist in biomedical research. We find, however, a decline in Physics, engineering, computer sciences, and chemistry, the fields usually linked to the industries often mentioned in the narrative of the decline of corporate science (Arora, Belenzon, and Patacconi 2018). Companies like AT&T, Xerox, and Kodak have sensibly decreased the number of their scientific publications. In contrast, pharmaceuticals, biotechnologies, and IT companies like Pfizer, Amgen, Microsoft, and Alphabet took the opposite trajectory, increasing their scientific publications.

In Chapter 2, I unpack these trends to see more in detail the changing nature of corporate science. First, I will investigate the role of university-industry collaborations in relation to the corporate science literature. In particular, I will test if the growing burden of knowledge increases the likelihood of university-industry collaborations. In Chapter 3, I will test if corporate science is becoming more applied and less basic, moving from the so-called Pasteur's quadrant to Edison's one. In Chapter 4, I will focus on firms that go public and test the impact of IPOs on firms' scientific research.

Chapter 2: University-industry collaborations and the burden of knowledge: evidence from the US[19]

Research collaborations between university and industry are an important channel of innovation. In this chapter, we provide a systematic overview of US university-industry collaborations, drawing upon all US-based scientific publications from 1980 to 2013 in Web of Science, matched to the relevant business characteristics in BvD Orbis. We find that university-industry collaborations increased from 2% to 6% of all indexed publications. A decrease in direct corporate involvement in scientific research accompanies this surge. Furthermore, we show that this increase is influenced by factors related to the nature of science such as the burden of knowledge, rather than companies' commercial considerations. Specifically, the likelihood of collaboration increases as the speed of scientific progress increases, with differences by firm size.

2.1 INTRODUCTION

Research collaborations between industry and academia have been the source of many radical, high-impact technological innovations. In the 1940s, MIT and Bell Labs collaborated to develop the cavity magnetron, which allowed for more compact powerful radar units for use in aircraft and ships, heralded as "the technology that won the war" (Fisk, Hagstrum, and Hartman 1946; Gertner 2012). More recently, AstraZeneca and the University of Oxford achieved breakthrough success in developing the world's first vaccine for SARS-COV-2. Using novel data derived by linking bibliometric with enterprise data, we present systematic trends in university-industry research collaboration for US firms from 1980-2013 and measure the extent to which the changing nature of science itself, relating to the "burden of knowledge", is shaping incentives to collaborate.

Despite the commonly held view that university-industry collaboration has increased over recent decades, existing evidence is more qualified. Reported trends in collaboration derived from bibliometric data show conflicting results across fields, countries, and time. Seminal work

[19] Publication version co-authored with Prof. Russell Thomson (https://orcid.org/0000-0002-1359-7542)

by Tijssen (2004) reports that the number of university-industry co-authored publications (globally) showed no overall growth between 1996 and 2001, rising in some fields and falling in others. Similarly, the NSF (2018, 120) reports that the number of US university-industry co-authored scientific journal articles remained relatively flat from 2006 to 2016, with co-authored publications declining as a share of all published scientific research. In contrast, Calvert and Patel (2003) found that university-industry collaboration in the UK (as measured by co-authored publications) increased between 1981 and 2000 – both in absolute number and as a share of all scientific publications. The authors argued that these collaborations are driven by foreign firms tapping into cutting-edge science with UK universities. Motohashi (2005) reports survey data indicating that the share of Japanese firms engaged in collaboration with universities increased between 1997 and 2002.[20]

To provide a systematic picture of patterns of university-industry collaboration in the US over the long run we link all articles published between 1980 and 2013 and indexed in the Web-of-Science (WoS) to enterprise financial data from Bureau van Dijk Orbis. We identify more than 85 thousand US-based companies. Authors who nominated these firms as their affiliated organizations produced over 800 thousand articles. A benefit of our data is that we identify articles by small, unlisted, and sometimes short-lived companies as well as large research-intensive corporate actors which have been more extensively studied. These data reveal that university-industry collaborative articles increased in number over the period. More strikingly, we find that collaborative articles increase more rapidly than those produced by university-affiliated authors alone. University-industry collaboration contributes more than six percent of all published work indexed in WoS, up from only two percent in 1980. We find the increase in every scientific field and industrial sector, with medical sciences and pharmaceuticals accounting for the highest number of collaborations. Regarding the share of collaborations in companies' portfolios, engineering and machinery experienced the highest collaboration surge. Our data also confirm the sharp decline in sole industry publications observed for large listed corporations; such papers comprise only two percent of all indexed work in 2013, down from six percent in 1980.

We then turn to consider the underlying drivers of university-industry research collaboration. The expansion of university-industry collaboration is often understood within an

[20] Park and Leydesdorff (2010) report that university-industry-government collaboration increased in South Korea between 1975 to 1995.

evolving "new industrial ecology" (Coombs and Georghiou 2002) and a new division of innovative labour (Arora, Belenzon, and Sheer 2021a) whereby companies are withdrawing from vertically integrated scientific research – the decline of top-flight corporate labs such as Bell Labs, Du Pont, and IBM and instead looking to universities for "ideas and new products" (Arora, Belenzon, and Patacconi 2018, 6).[21] Companies are transitioning from the vertically integrated R&D model to a more dynamic system of long-term alliances and technology-based joint ventures with universities. The reasons behind this trend are not yet fully understood. However, most arguments to date focus primarily on commercial and competitive considerations in which firms operate, such as impatient capital, the need to generate commercial results in a time frame faster than basic research (i.e., short-termism), and the exploitation of external technological opportunities (Rosenberg and Nelson 1994; Etzkowitz and Leydesdorff 1997; D'Este, Guy, and Iammarino 2013; West et al. 2014). It has also been argued that weakened antitrust consideration, strengthening IP rights, and the Bayh-Dole Act, underpinning deepening in markets for technology, have effectively reduced the benefits of internal science relative to acquisition (Arora, Belenzon, and Patacconi 2018; Arora, Belenzon, and Sheer 2021a).

In this chapter, we examine how the changing nature of science itself – rather than commercial environment – may be contributing to incentives to collaborate. Specifically, we consider the extent to which the burden of knowledge plays a role in firms' decision to collaborate as scientific discoveries are getting increasingly difficult to achieve. The role of burden of knowledge in scientific collaboration between university scholars has been widely examined/speculated (Jones 2009; Uzzi et al. 2013; Agrawal, Goldfarb, and Teodoridis 2016; Zhu, Liu, and Yang 2021). Scientific fields are becoming increasingly intertwined, and scientists from different backgrounds must work together to overcome potential bottlenecks and stand-stills affecting many sectors (Arora, Belenzon, and Patacconi 2018; Jones 2009). A single person, department, or laboratory cannot possess all the necessary skills to tackle a problem; thus, the necessity to collaborate emerges. Correspondingly, the average number of authors and affiliations per paper is increasing, and individual scientists' career paths are lengthening and becoming more specialised. Despite the evidence that the burden of knowledge

[21] See also Arora, Belenzon, and Patacconi (2018), Hartmann and Henkel (2020), Krieger et al. (2021)

is driving academic collaboration, its role in corporate-university collaboration has not been examined.

We argue that the influence of the burden of knowledge on collaboration propensity will depend on firms' internal resource constraints. Our hypothesis suggests that this relationship may not be linear. We suggest that the propensity to collaborate in fast-paced fields increases as firm size increases, but only up to a certain threshold. Beyond that threshold, the propensity to collaborate decreases. Smaller firms require a minimum level of capacity to collaborate effectively with universities. For this reason, they may be less sensitive to the burden of knowledge because they lack the capacity to initiate collaborations in the first place. In contrast, very large companies with self-sufficient R&D labs can collaborate in a broad range of fields and are less affected by the burden of knowledge. This can be attributed to their greater capacity, including larger teams, a higher number of well-educated scientists, as well as more efficient knowledge hierarchies (Astebro, Braguinsky, and Ding 2020).

To study the effect of the burden of knowledge on collaboration, we model the influence of the speed of scientific progress on companies' propensity to collaborate in research. By focusing on variation in collaboration within each company's research portfolio, we hold constant company-level commercial, capital-market, and managerial considerations. We use a two-way fixed effect identification strategy based on two sources of cross-sectional variation: variation in the speed of scientific progress across fields and variation in firm capacity.[22] This estimation strategy allows us to control for company-level fixed effects as well as field of research fixed effects. Next, we introduce an interaction term with speed and firms' size, trying to capture how firms of different sizes are sensitive to changes in speed. As we hypothesized a non-linear relationship, firm size is squared to take this into account.

We measured the speed of scientific progress using the Price Index (Milojević 2012; Price 1976). The measure of speed is computed exclusively using university publications (excluding corporate publications) in order to capture the state-of-the-art of scientific research. Including corporate publications would create a spurious measure of speed, as it would mix broader trends in scientific research with companies' commercial considerations. As part of our

[22] In the spirit of Rajan and Zingales (1998) who use two sources of cross-sectional variation instead of combining time variation with cross-sectional variation (in their case they use variation across countries combined with variation across industrial sectors).

robustness checks we consider the average team size and the inverse of references age in purely academic collaborations as an alternative proxy for complexity.

Our results show that the burden of knowledge is positively correlated with the likelihood of collaboration. Our analysis also reveals distinct patterns among SMEs and large firms; we observe an inverted U relationship between speed and collaborations, mediated by firm size. For the same level of burden of knowledge, the likelihood of collaboration increases until it reaches a particular threshold. Beyond this point, which we observe being around 2670 employees, the likelihood of collaboration declines. In essence, our results suggest that the speed of scientific progress affects all kinds of firms, but very small and very large companies are less sensitive to speed.

2.2 WHAT DRIVES UNIVERSITY-INDUSTRY COLLABORATION?

Why do firms collaborate with universities? Much of the existing literature has focused on factors that can be broadly classified as relating to commercial considerations and the desire to tap into university technology for commercial development. Less well understood is how the changing nature of scientific research is influencing the process of collaboration. In this chapter, we argue that increasing complexity in fast-moving fields is associated with the "burden of knowledge" and may be driving the need for firms to collaborate. Furthermore, this effect is anticipated to be dependent on firm size. This section provides a concise overview of factors known to play a role in university-industry collaboration and develops our argument that the burden of knowledge should play a role.

The existing literature has stressed attention to university-industry collaborations in relation to the commercial considerations of the firm, principally related to the issue of "impatient capital" and the need to deliver shareholder value within a time frame inconsistent with basic research. Under pressure from shareholders, firms face the need to see market results quickly. This "short-termism" is believed to lead large firms to withdraw from internal research and replace the research once conducted in-house with external research acquired through university-industry collaborations, VC investments, or acquisitions (Tijssen 2004; Arora, Belenzon, and Patacconi 2018). Coombs and Georghiou (2002) describe this ecosystem as a new industrial ecology of corporate R&D. Similar conclusions are taken by Varma (2000), who saw university labs as a potential "virtual lab" for industry, where firms let the university scientists undertake research to try then to absorb it afterwards.

It has also been argued that a declining value of science has contributed to the preference for external knowledge acquisition. Scientific discoveries might be less useful for innovation, and companies can innovate by developing or recombining previous technologies rather than engaging in basic research in the first place (Arora, Belenzon, and Patacconi 2018; Arora, Belenzon, and Sheer 2021a). This argument finds support in cases where companies innovate without publishing a single research paper despite substantial R&D expenditures (Lim 2004).

Increased competition is another motivation inducing firms to reduce their in-house research and focus more on external knowledge acquisition. Companies have to balance between the benefits of undertaking in-house scientific research, which may eventually find application in their own innovations, and the danger of disclosing valuable knowledge to rivals, who can exploit it without baring the costs of research (Arora, Belenzon, and Patacconi 2018; Arora, Belenzon, and Sheer 2021a).

Another well-articulated driver of university-industry collaboration relates to firms' desire to tap into technological opportunities arising from research being undertaken in universities, benefiting from knowledge spillovers (e.g., Rosenberg and Nelson 1994; Etzkowitz and Leydesdorff 1997; D'Este, Guy, and Iammarino 2013). Universities usually possess knowledge that is predominantly tacit and naturally excludable. Joint research is a way to overcome the complexity related to tacit knowledge transfer and allow effective knowledge exchange (L. G. Zucker, Darby, and Armstrong 2002).

Less well studied is how the changing nature of science itself is affecting incentives to collaborate. Here, we focus on the impact of increasing complexity in fast moving scientific fields associated with "the burden of knowledge" (Jones 2009). There is increasing awareness that scientific research is growing in complexity, and that this "burden of knowledge" is impacting research productivity and driving greater specialization and collaboration between academic scientists. Bloom et al. (2020) documents evidence that research productivity is falling across a range of scientific and technical domains, including microprocessor design (and end of Moore Law), agricultural crop yields, the number of new molecular entities, and the changes in life expectancy. Jones (2009) argues that the burden of knowledge is driving an increase in team size, the age of first innovation, and the time lag between patents filed by the same inventor. Additionally, it is becoming more difficult for scientists to switch across fields when innovating (Jones 2009). We argue that this is exacerbated in fast-moving research areas.

The effect of the burden of knowledge on collaboration between scientists is well established (Jones 2009; Uzzi et al. 2013; Agrawal, Goldfarb, and Teodoridis 2016; Zhu, Liu, and Yang 2021), by extension, we anticipate an analogous effect on rates of university-industry collaboration. Indeed, several scholars argue that in fast moving, complex fields, companies will find it increasingly challenging to possess all the necessary skills to remain engaged with this rapidly expanding frontier spread both geographically and across disciplines (Grindley and Teece 1997; Howells 2000; Santoro and Chakrabarti 2002; Almeida, Hohberger, and Parada 2011; Arora, Belenzon, and Sheer 2021a). Collaboration with science-intensive institutions such as universities can complement and extend firms' scientific capability (Arora, Belenzon, and Suh 2021; Jones 2009; Narula 2004).

While there is a paucity of systematic econometric evidence regarding the role of the burden of knowledge in driving university-industry collaboration, industry research leaders have clearly articulated a belief that collaboration is increasingly necessary to overcome the burden of knowledge. For example, Nancy Krunic, Global Head, Diagnostic Sciences and Partnerships at Novartis, interviewed by Nature (Savage 2018), reports that collaborations are vital because a single person, department or laboratory cannot possess by themselves all the necessary skills to tackle a problem. The GSK Immunology network, started in 2015, goes in the same direction. GSK incentivises researchers, including university researchers on sabbatical, to temporarily join their company and work on cutting-edge science. Science (2021), interviewing John Lepore, head of pharmaceutical research, reports that the GSK immunology network is aimed at "access[ing] to information to advance immunology research, faster and better". Further empirical evidence comes from Agrawal, Goldfarb, and Teodoridis (2016), who looked at publications of mathematicians before and after the collapse of the Soviet Union. After the Iron Curtain's collapse, previously unavailable Russian expertise became suddenly accessible. Collaborations increased in response to this exogenous increase in the burden of knowledge, especially in those subfields of mathematics that were Soviet-rich.

Collaboration is not a panacea to scientific complexity; instead, firms require adequate capacity to identify the relevant frontier, and to engage in productive collaboration. Firm capacity plays a crucial role in determining the available benefits from collaboration. On the one hand small firms need collaborate more to compensate the lack of internal capabilities. On the other hand, they need some threshold of capability to engage effectively in university-industry collaboration.

If – as their R&D managers suggest – large firms like Novartis and GSK struggle to keep up with the rapid pace of scientific advancement, and engage in collaborations, we argue that the benefits of collaboration to overcome the burden of knowledge will even be greater for smaller, more resource constrained firms. Small firms are less likely to possess internally all the knowledge necessary to complete a research project, especially in a fast-moving field. As a result, they are more likely to collaborate with universities (Audretsch and Belitski 2021; Durst and Runar Edvardsson 2012) and to rely on universities' laboratories to access the necessary machinery or equipment (Onida and Malerba 1989; López-Martínez et al. 1994).

The resource constraints, however, may also limit firms' ability to collaborate successfully. Small firms may lack absorptive capacity (Cohen and Levinthal 1990), scope, resources, and expertise needed to absorb external knowledge and learn from collaborations. Evidence shows that SMEs with higher R&D and absorptive capacity increase the probability of successful collaborations (Bougrain and Haudeville 2002; Muscio 2007), and that external learning increases with startup size (Almeida, Hohberger, and Parada 2011). Last, it is argued that large organisations can exploit better the layers of their organisation to face the effects of the burden of knowledge (Astebro, Braguinsky, and Ding 2020).

In summary, we argue that firms will have a greater need to collaborate in fast moving complex fields. The role of complexity in academic collaboration has been well documented, we consider here whether the same factors are driving university-industry collaborations. Moreover, we expect that complexity's role will be conditioned on firm capacity. Specifically, it will follow an inverted U shape – complexity will be a less important consideration for very large firms due to their scale and capacity, enabling them to adapt more effectively to rapid scientific progress. The speed of scientific progress will also be less important for smaller firms that lack the capacity to engage successfully in collaborations with universities.

2.3 DATA AND DESCRIPTIVES

A research collaboration involves researchers working together and sharing a common goal to produce new scientific knowledge (Katz and Martin 1997; Sonnenwald 2007). More complicated is instead its measurement. A common practice is to define collaborations through co-authorship. The primary assumption is that each co-author brings something into the collaboration and participates actively in the research project. We measure *collaboration* as

any corporate publication that includes at least one author affiliated to a university.[23] Conversely, we define a *solo publication* as a corporate publication with solely a company affiliation, indicating firms that did not collaborate and undertook the research within their company boundaries.

To provide a systematic view of university-industry research collaboration, we construct a novel dataset by linking company names and institutional affiliations of authors of papers indexed in WoS with company data from Bureau Van Dijk Orbis for all US based firms (including foreign subsidiaries) from 1980 to 2013. We measure university-industry collaboration using multiple affiliations of authors of peer-reviewed scientific papers.

WoS covers about 8 million research articles in hard sciences, reporting at least one US affiliation corresponding to 48,168 publication outlets. We consider 164 WoS subject categories (SC) in STEM sciences, with each publication being associated with one or multiple science categories. We exclude social sciences and humanities. We focus on STEM sciences for two reasons. First, we are interested in firms performing R&D activities, which is more likely in hard sciences. Second, as Clarivate Analytics claims, there is stronger coverage of natural sciences, health sciences, engineering, computer science, and materials sciences. In Section 2.3.2, we group the WoS SC into ten broad scientific fields using Milojević's (2020) classification[24] for visualisation purposes. The comprehensive discussion of WoS coverage can be found in Appendix A.

WoS data include indication of "document type" as either conference proceedings or journal articles. We exclude publications coded in WoS as proceedings from our selection as their coverage is unsatisfactory.[25] However, we note that IEEE conference papers[26] – arguably the most important for engineering and computer science companies – are coded as journal articles and are therefore included in our analysis.

Our firm level company information comes from BvD Orbis. Orbis contains around 80 million unique firms' identifiers coming from 13 different releases from 2005 to 2017.

[23] We keep authors with multiple affiliations. If a paper is written by one author with a corporate and academic affiliation, I still consider it as a collaboration. The issue is addressed directly in Section 2.3.6.

[24] Refer to footnote 16 for the list of fields.

[25] The coverage is satisfactory only from 1995 to 2006, to then drop afterwards. See Section A.2.3 for further information.

[26] Ex: IEEE transactions on magnetics, nuclear science, electron devices, etc.

Collecting different versions of Orbis allows us to have better coverage and to match a higher number of companies, especially small and medium-sized enterprises. Orbis includes balance sheet information as well as sector of activity, size, ownership, and other economic variables. We carefully considered the ownership structure of the matched firms, using Orbis data from 2005-2013, one of the most comprehensive sources currently available, particularly for medium and small enterprises, Arora, Belenzon, and Sheer (2021b) open-source data,[27] and Orbis M&A[28] as described in Section 1.2.2.

To link WoS and Orbis, we developed a novel decision tree algorithm incorporating string similarity scores (Levenshtein distance), shared non-dictionary words, and address information (same city or zip code). The biggest challenge of the matching process consisted in harmonising the affiliations and company names in WoS and Orbis. The same company may be recorded with many name variations (Abbott, Abbott Laboratories, Abbott lab, Abbott labs etc.), incorporation type (inc, corp, ltd, llc etc.), and abbreviations (General Electrics, GE, Gen Elect, General Elect etc.). For more extensive information about the matching algorithm refer to Section 1.2 and Appendix A.

We then divided the affiliations according to their type (firms; universities, not-for-profit organisations, medical centres and clinics, governmental agencies) using two main approaches. First, we used recognisable keywords for universities (university, college, institute of technology, school, faculty, etc.), government (us department, us army, us navy, NSF, NASA, etc.), not-for-profits (aquarium, botanical garden, zoo, etc.) and medical centres (clinic, hospital etc.).[29] Second, we used web address information from Orbis to guess the legal status of the institutions. We assumed that a web address ending with .edu belongs to a university, .gov to the government and .org to a not-for-profit. Last we checked manually that large firms had the right label.

Table 2.1 summarises the availability of the financial variables. Financial information is usually available from 2005 to 2013, with longer series for a small subset of large firms.

[27] Duke Innovation & Scientific Enterprises Research Network (DISCERN, 2020). Duke University. Accessed 14 June 2022. https://zenodo.org/record/4320782

[28] The database was previously known as Zephyr and it changed its name into Orbis M&A.

[29] See the appendix in Section A.4.2 for a more comprehensive list of keywords.

Table 2.1: Corporate publications sample, 1980-2013

Variables	# observations	# firms	Mean	S.D
Publications	798,421	80,723		
Collaborations	376,832	50,978		
Employees	743,334	13,901	99,250.21	117,768.08
Industry	1,098,773	57,370		

Note: This tables preset the descriptive statistics of the sample. Each observation is a publication, that can appear in the sample more than once if it belongs to more than one scientific field or industry. Publications is the number of publications with at least one corporate affiliation, while collaboration with at least one corporate and one university affiliation. Industry is available with the NAICS classification.

2.3.1 Trends in research collaboration

Figure 2.1 shows that the share of collaborations between universities and industry is increasing. Publications are shown as shares of the total publications in WoS to minimise the bias caused by new journals being included in the sample, given that WoS coverage has been expanding.[30] The share of corporate publications that are produced by a collaboration rose steadily from 2% in 1980 to 6% in 2003. This indicates that during this period, collaborations between universities and industry outpaced publications produced solely by university-affiliated authors. The share remained stable at around 6% until 2013, suggesting that university-industry collaborations grew as fast as university publications. In absolute terms, collaborations increased from 2,317 in 1980 to 21,266 in 2013, while university publications went from 160,462 publications in 1980 to 351,846 in 2013. Conversely, solo publications saw a decrease from 6% in 1980 to 2% in 2013. Corporate collaborations exceeded sole corporate publications after 1997. In absolute terms, solo publications decreased from 10,426 in 1980 to 9,116 in 2013. These results highlight the increasing industry dependency on academia and the decline in in-house corporate R&D.

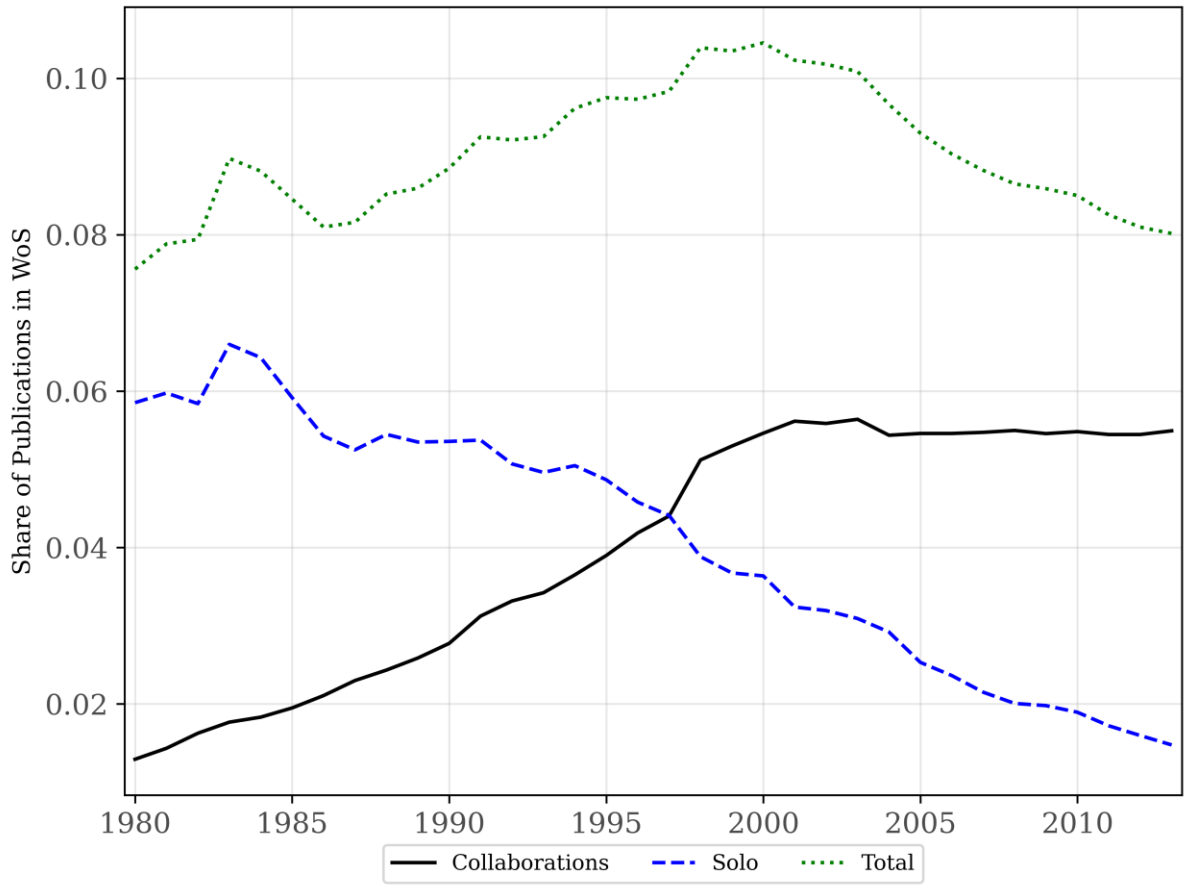
The total share of corporate publications has not experienced a consistent decline; however, a downward trend is noticeable starting from 2000. Until that point, the total share increased, but it subsequently started declining, falling to around 8% in 2013. Regarding publication counts, the number of publications increased from 13,471 to 34,381.

The data reveal that most firms followed a mixed research strategy, publishing a mix of both solo publications and collaborations. Virtually all firms (99.995%) with more than ten

[30] See appendix Section A.2.3.

publications followed a mixed research strategy. Most companies that only collaborate or publish alone have less than ten publications. Among firms with less than ten publications over the period, we find that 35.65% published solo, while 36.59% only collaborated. The remaining share followed a mixed approach. Consequently, it is less common for companies to adopt a mixed approach when they have not published many publications yet. A priori, this pattern suggests that to understand drivers of collaboration, it is necessary to look beyond firm specific factors such as capital market demands and consider why they choose to collaborate on some projects and not on others.

Figure 2.1: Total publications, collaborations and solo publications, 1980-2013



Note: This figure shows the share of corporate publication in WoS (dotted), the share of collaborations (dashed, blue) and of solo publications (solid)

The scientific fields with the most collaborations are the same ones with the most publications (Figure 1.6), and are medical sciences (148,460), engineering (104,623), biological sciences (62,465), physics (57,120), and chemistry (38,980). The same applies to industries, where the most collaborative industries coincide with those with most publications

(Figure 1.7). Those industries are pharmaceuticals (84,767), business services (54,098), electronic equipment (28,952), computers (23,850), and healthcare (21,791).

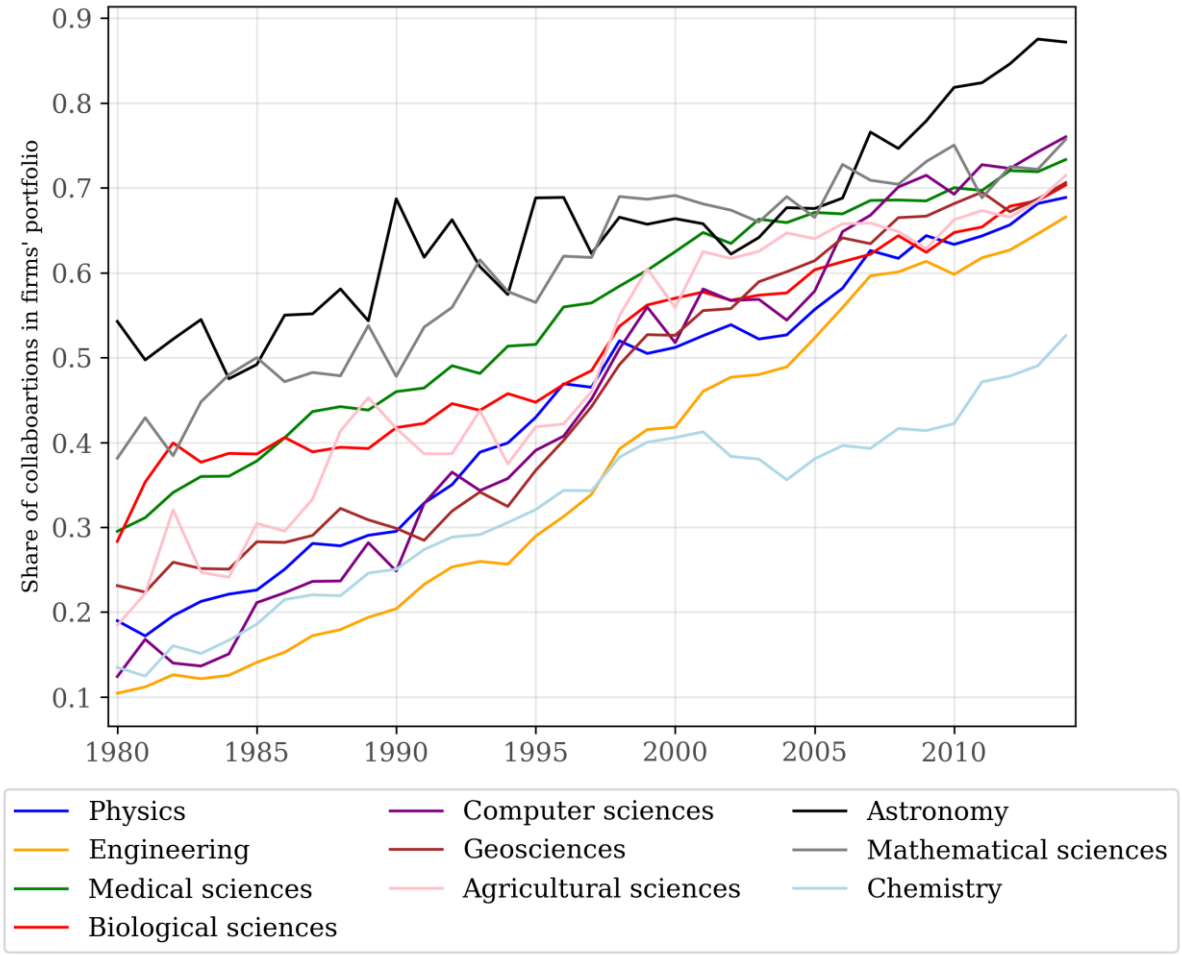
2.3.2 Trends by scientific field

Besides the absolute number of collaborations, it is interesting to look the different propensity to collaborate within each field. To explore this aspect, we show in Figure 2.2 the share of collaborations in firms' portfolio by broad scientific field from 1980. All scientific fields present upward trends throughout the whole period, but with significant differences across fields. In descending order, the fields with the most collaborations are astronomy (54%), mathematical sciences (38%), medical sciences (30%), biological sciences (28%), geosciences (23%), physics (19%), agricultural sciences (18%), chemistry (13%), computer sciences (12%), and engineering (10%).

In 2013, the most collaborative field was always astronomy (88%, +34 percentage points), but the relative order of the other fields changed substantially. The second most collaborative field is computer sciences, which moved from 12% to 74%, with an impressive difference of 62 percentage points. Another field that considerably increased is engineering, which moved from 10% to 65% (+55pp). The other fields in descending order are mathematical sciences (72%, +34pp), medical sciences (72%, +42pp), agricultural sciences (68%, +50pp), geosciences (69%, +46pp), biological sciences (69%, +41pp), physics (68%, +49pp) and chemistry (49%, +36 pp).

The gap between fields reduced over the period. In 1980, the gap between astronomy and engineering was 40 percentage points; in 2013, it reduced to approximately 23 percentage points. Excluding chemistry, which presents lower levels of collaboration, all fields are comprised in the interval 65%-88% in 2013, while in 1980, the range was 10%-54%.

Figure 2.2: Share of collaborations in firms' portfolio by broad scientific field, 1980-2013



Note: This figure shows the share of collaborations in firms' portfolio by broad scientific field from 1980 to 2013. The scientific fields are grouped using Milojević (2020).

2.3.3 Trends by industry

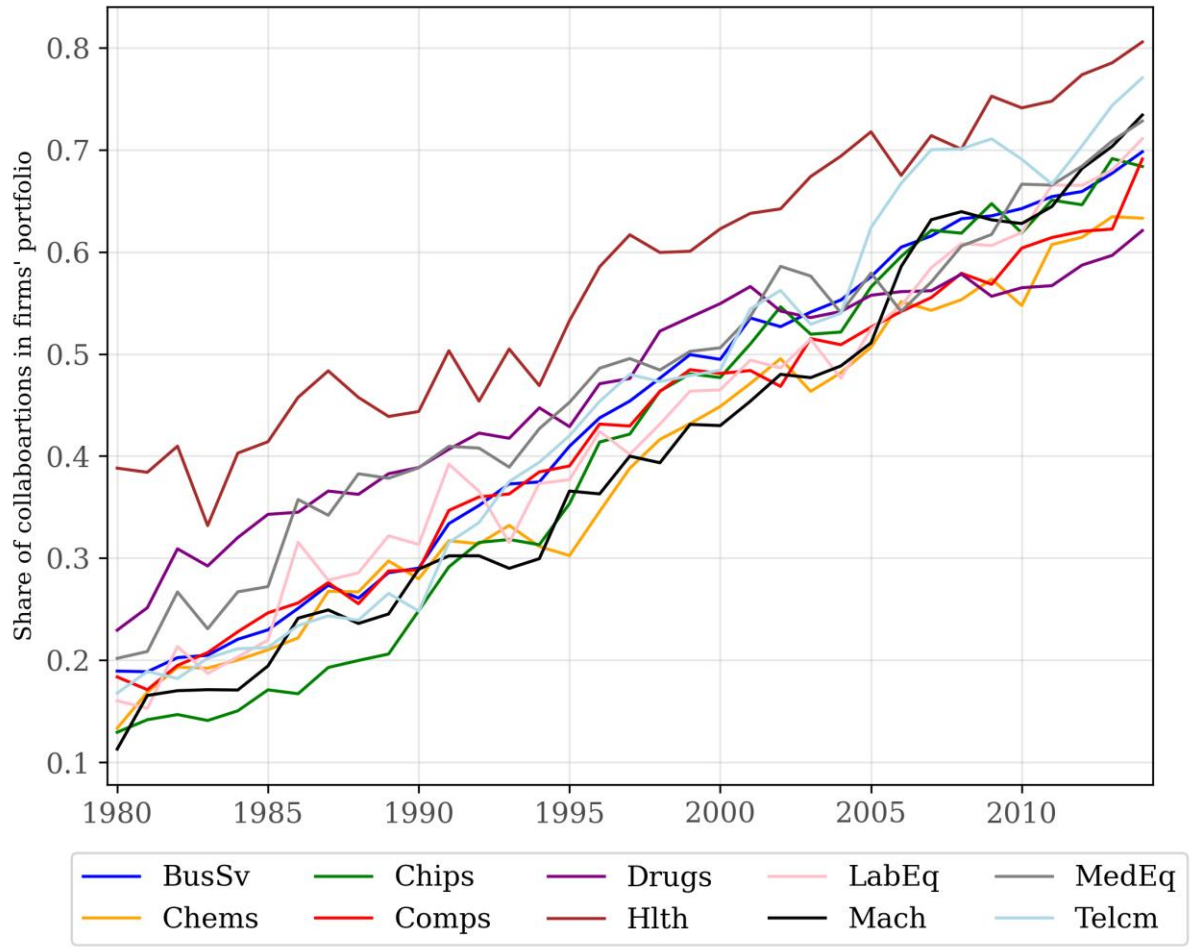
Figure 2.3 shows the share of collaborations in firms' portfolio by 48 Fama-French industries from 1980-2013. Firms are assigned to their primary industry code. When it is not possible all industry codes are kept. Only the ten industries with the most publications are displayed graphically. Analogous to trends observed in scientific fields, all industries show upward patterns and a significant increase.

The most collaborative industry is healthcare with 39% collaborations in 1980 and 79% in 2013 (+40 percentage points). The other industries in 1980 have the following share of collaborations: pharmaceuticals (23%), followed by medical equipment (20%), business services (19%), computers (18%), communication (17%), measuring and control equipment (16%), chemicals (13%), electronic equipment (13%), and machinery (11%). In 2013, all

industries experienced significant growth in collaboration. In descending order, we can find communication (74%, +57percentage points), medical equipment (71%, +51pp), machinery (70%, +69pp), Electronic equipment (62%, +49pp), measuring and control equipment (68%, +52pp), business services (68%, +49pp), chemicals (63%, +50pp), computers (62%, +44pp), and pharmaceuticals (60%, +37pp)

Differently from scientific fields trends, the gap between industries is almost identical, as the difference between the most and least collaborative industry is 28% in 1980 and 29% in 2013. It must be cautioned that many firms in the sample are large multinational and multi-industry corporations; thus, assigning each firm to a single industry is not trivial. General Electrics, for example, operates in six different 3-digit NAICS sectors (336, 334, 333, 522, 335, 532). This broad scope is reflected by the scientific fields in which the firms publish. Merck published in 143 out of 164 WoS SC, IBM 140, General Electrics in 131, and Hewlett Packard 108.

Figure 2.3: Share of collaborations in firms' portfolio by industry, 1980-2013



Note: This figure presents the share of collaborations in firms' portfolio by industry from 1980 to 2013. Industries are classified using 48 industries Fama-French classification. To see the list of all industries, refer to footnote 18.

2.3.4 Most collaborative universities and companies

The companies with the most collaborations once again coincide with those with the most publications. Among the top twenty collaborators eight are pharmaceutical companies: Pfizer (13,652), Merck (8,581), Roche (6,869), Ely Lilly (5,003), GlaxoSmithKline (4,985), Johnson & Johnson (3,587), Abbott Laboratories (3,391), Novartis (3,320), and Bristol Myers Squibb (3,124). Five firms belong to the ICT sectors, namely IBM (13,603), AT&T (11,807), Intel (3,440), General Electrics (2,897), and Microsoft (2,391). The remaining companies are Genentech and Amgen (3854 and 3255, biotechnologies), Lockheed Martin (2,870, aircraft), Dow Chemical (2,443, chemicals), General Motors (3,516, automotive), and Exxon (3,735, petroleum and gas).

The most collaborative universities are the following. Harvard University (18,344), Stanford University (11,430), University of North Carolina (10,107), University of California Los

Angeles (9,913), University of Washington (9,421), University of Washington Tacoma (9,311), University of Washington Seattle (9,276), Johns Hopkins University (8,721), University of Michigan (8,704), and University of California San Diego (8,119).

Table 2.2 shows the pairs of company-university that collaborated the most from 1980 to 2013. The results are divided by broad scientific field. In the field of physics, the most common collaboration involves Princeton University and General Atomics with 504 collaborations. In computer sciences, Stanford University and IBM (172); in medical sciences, Harvard University and Pfizer (668); in chemistry, Stanford University and IBM (187); in mathematical sciences Princeton University and AT&T (88); in biological sciences University of California San Francisco, and Roche (268); in geosciences University of Maryland Baltimore and Science Systems and Applications (169); in engineering University of Michigan and Ford Motors (282); in agricultural sciences University of Illinois Urbana-Champaign and Monsanto (47) and in astronomy Universities of Perugia, Padua and Paris Sorbonne with Nycb Realtime Computing (116).

Companies often seek collaboration with academic institutions situated in close proximity to their R&D centres. Some examples are AT&T and Princeton (New Jersey, 795 collaborations), Merck and Harvard university (Boston, 770), Pfizer and Harvard University (Boston, 838), Roche and University of California San Francisco (Pleasanton, Santa Clara, San Jose, 762), and IBM and Stanford University (Almaden, 884).

Table 2.2: Most common collaborations by broad scientific field, 1980-2013

Broad Area	University Name	Company Name	Collab.	Broad Area	University Name	Company Name	Collab.
Physics	Princeton Univ	General Atomics	504	Computer Sciences	Stanford Univ	IBM	172
Physics	Stanford Univ	IBM	448	Computer Sciences	Carnegie Mellon Univ	IBM	154
Physics	Princeton Univ	AT&T	417	Computer Sciences	Univ Of Illinois Urbanachampaign	IBM	152
Medical Sciences	Harvard Univ	Pfizer	668	Chemistry	Stanford Univ	IBM	187
Medical Sciences	Harvard Univ	Merck	575	Chemistry	Univ Of Rochester	Eastman Kodak	134
Medical Sciences	Univ Of Michigan	Pfizer	475	Chemistry	Univ Of Minnesota Twin Cities	3M	114
Mathematical Sciences	Princeton Univ	AT&T	88	Biological Sciences	Univ Of Calif San Francisco	Roche	268
Mathematical Sciences	Tel Aviv Univ	IBM	88	Biological Sciences	Tufts Univ	Jmi Labs	251
Mathematical Sciences	Univ Of Calif Berkeley	Microsoft	71	Biological Sciences	Univ Of Michigan	Pfizer	248
Geosciences	Univ Of Maryland Baltimore	Sci Syst Applications	169	Astronomy	Univ Of Perugia	Nycb Realtime Comp	116
Geosciences	Univ Of Maryland Baltimore Cty	Sci Syst Applications	155	Astronomy	Univ Of Padua	Nycb Realtime Comp	116
Geosciences	Univ Of Maryland Coll Park	Sci Syst Applications	117	Astronomy	Pres Univ Sorbonne Paris Cite	Nycb Realtime Comp	116
Engineering	Univ Of Michigan	Ford Motor	282	Agricultural Sciences	Univ Of Illinois Urbanachampaign	Monsanto	47
Engineering	Stanford Univ	IBM	263	Agricultural Sciences	Univ Of Illinois Urbanachampaign	Pfizer	44
Engineering	Univ Of Michigan	GM	254	Agricultural Sciences	Rutgers State Univ	Pure Seed Testing	43

Note: This table shows the most common pairs of firm-university collaborations. The scientific fields are grouped using Milojević (2020).

2.3.5 Trends by size and age

Figure 2.4 shows that there are some differences by firm size. Firm size is calculated as the average number of employees in the period 2000-2018. For this reason, we show trends for firm size only starting from 2000. Four categorical variables describe firm size. Very large companies (orange) have more than 1000 employees. Large companies (blue) are not very large and have more than 250 employees. Medium-sized companies (red) have more than 15 employees and are not very large companies or large companies. Residually, small companies (green) have less than 15 employees.[31]

Very large firms have the lowest share of collaboration in their publication portfolio ranging from 50% in 2000 to 63% in 2013. Small firms are the more collaborative, ranging from 52% in 2000 and 73% in 2013. In the middle stand large and medium enterprises, with 53% and 56% in 2000, and 70% in 2013.

We identify firm age as the difference between the incorporation date and the article's publication date. Similarly, we calculate the firm age starting from 2000. Young firms (solid) are companies younger than ten years. Old firms (dashed) are companies older than ten years. Figure 2.4 shows that the two trends are distinct, as both old and young firms proceed parallel from 2000 to 2013. Young firms collaborate more than old firms, with the share of collaborations within their portfolios that ranges from 56% in 2000 to 74% in 2013. Old firms, instead, range from 50% in 1980 to 66% in 2013.

2.3.6 Multi-affiliated authors

A particular form of collaboration is when a single scientist is affiliated with two different affiliations, a business and an academic one. Scientists with double affiliation may represent cases of university spinoffs, but also of scientists who kept their university affiliation after having left their faculty position for a matter of prestige and recognition. We are able to gather this information only from 2008 since WoS started linking affiliation addresses to their

[31] We use 250 as threshold that separates SMEs to large companies using the EUROSTAT definition of SMEs, EUROSTAT (2023), accessed 11 September 2023, https://ec.europa.eu/eurostat/web/structural-business-statistics/information-on-data/small-and-medium-sized-enterprises#:~:text=micro%20enterprises%3A%20less%20than%2010,250%20or%20more%20persons%20employed

respective authors only starting from this date. We find that 15.69% of collaborations have an author with a double affiliation.[32]

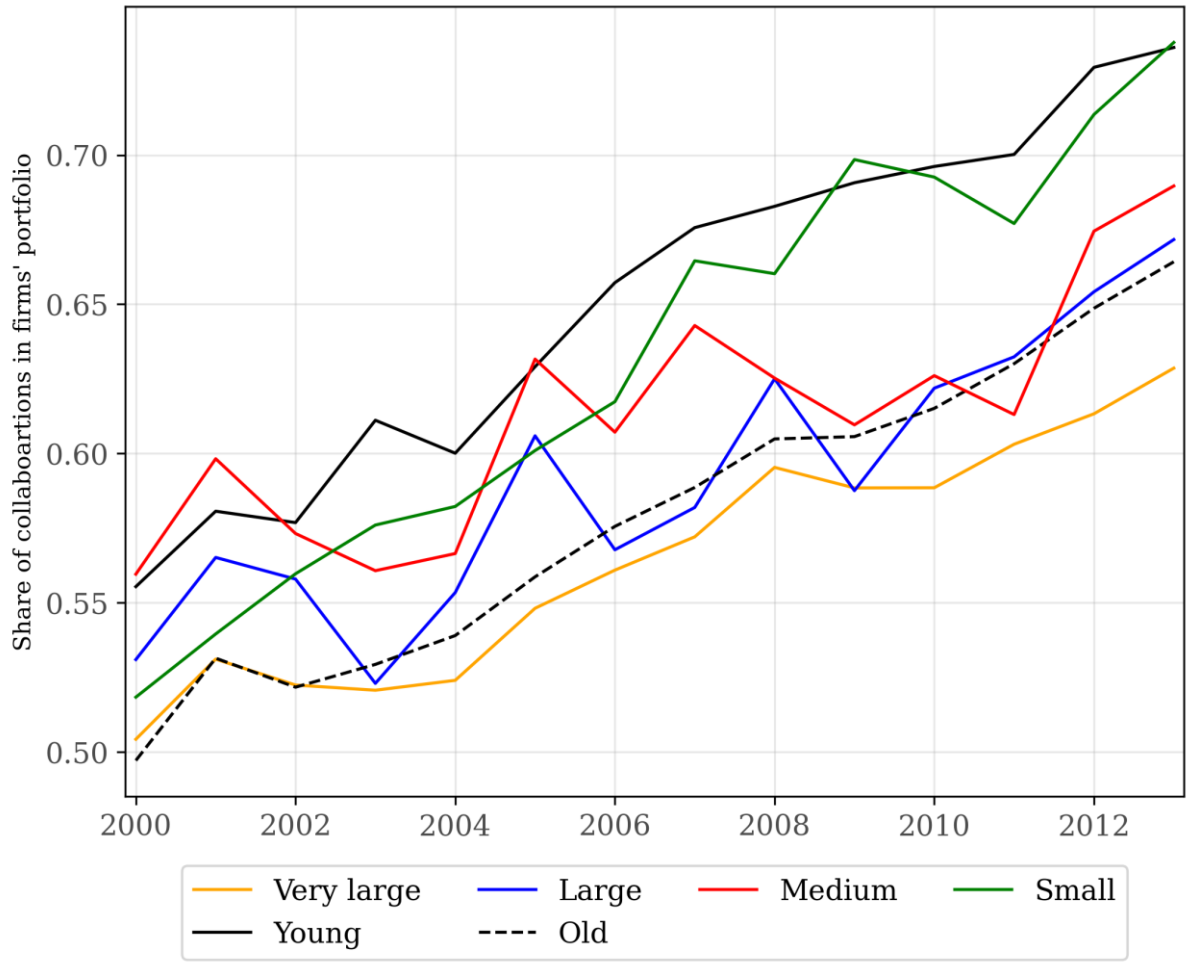
It is worth mentioning that most of these publications with double affiliation are likely to be a collaboration, as in 45.80% of those publications there is at least another author with a university affiliation distinct from the authors with university-corporate affiliations. If we exclude these cases, we are left with 8.50% of papers authored by individuals affiliated with both a company and a university, with no other authors having university affiliations. Given the limited availability of data, we cannot estimate our models removing this special set of collaborations from the sample. However, we are confident that this subset of publications does not affect the results. First, we believe that an individual with dual affiliations still maintains some ties with both institutions, making it a collaborative effort. Second, 8.50% represents a relatively small percentage of the total collaborations.

2.3.7 University spinoffs

We use the double affiliation in combination to firms' age to identify potential university spinoffs. We select all companies younger than two years that engaged in university-industry collaborations and have a scientist with a double affiliation. We were able to identify 416 firms that published 728 papers from 2008 to 2013. Since the share of potential university spinoffs is relatively small, we are confident that this subpopulation does not drive the trends in collaborations.

[32] Alternative sources such as Scopus, Open Alex or MAG could be used to calculate the share of multi-affiliated authors before 2008.

Figure 2.4: Share of collaborations in firms' portfolio by firm size and age, 2000-2013



Note: This figure shows the share of collaborations in firms' portfolio by firm size and age from 2000-2013. Firm size is described by four categorical variables. Very large companies (orange) have more than 1000 employees. Large companies (blue) have more than 250 employees and are not very large companies. Medium-sized companies (red) have more than 15 employees and are not very large companies or large companies. Residually, small companies (green) have less than 15 employees. Firm age is calculated as the difference between incorporation date and the article publication date. Young firms (solid) are companies younger than 10 years. Old firm (dashed) are companies older than 10 years.

2.3.8 Measuring the burden of knowledge

There is no universally accepted definition of the "burden of knowledge". We focus here on measures relating to specific research fields. We argue that the topic of research is pre-determined at the time of team formation; in contrast, any measure of complexity based on each specific research output is endogenous to the team itself – i.e., whether or not the team included collaboration between university and corporate or whether the corporate was "going it alone". Of course, this raises the complexity of separating the role of field specific burden of knowledge form other field specific factors (such as co-authoring norms) which we address

using a two-way fixed effects approach in the spirit of R. Rajan and Zingales (1996) and discussed below in this section. Our primary approach captures differences in the speed of the scientific frontier between scientific disciplines. As part of our robustness checks we replicate results implementing two alternative indicators based on research team size and reference age in academic science.

Our preferred measure of the burden of knowledge is the speed of scientific progress (*speed*), measured through the Price Index (*PI*) from Milojević (2012). The Price Index (1970) measures the fraction of *n* years old references as showed by equation 2.1. Citing the most recent literature reflects a fast-moving scientific field.

$$Speed_{ajt} = \frac{\#\ references\ that\ are\ d\ years\ old_{jt}}{\#\ references_{jt}} \tag{2.1}$$

We set *d* to be equal to five, as it is the most used time bracket. *Speed* is measured at the scientific field level from the whole WoS using the years 2000 and 2001. The scientific fields used are the ones presented at the beginning of Section 2.3. When a paper belongs to multiple fields, we compute the arithmetic mean of those fields' speed. *Speed* is aimed at capturing the overall differences between fields academic science. To measure *speed*, we include in the measure only articles written by universities, excluding any papers authored, or co-authored by companies, to avoid outcomes of the research under investigation directly influence the measure of speed. Similarly, we measure *speed* based on publications in 2000-2001. Since we aim to capture the state of play at the time research team formation is begin decided. Using speed at the start of the period also prevents any outputs by created over the period influencing the measure. Finally, this approach avoids the complex problem of measuring intertemporal changes in speed. As reported by Egghe (2010) and Larivière et al. (2008), the Price Index is decreasing over time due to the ageing of the body of the scientific literature available with few exceptions in the case of paradigm shifts (Milojević 2012). For this reason, we chose a measure speed that is time-invariant, being able to capture differences across fields without the noise of the intertemporal variation.

Table 2.3 shows descriptively the 164 scientific fields speed. The fastest fields are environmental studies (0.56), cell biology (0.55), computer science information systems (0.54), physics particles and fields (0.53), and infectious diseases (0.53). A price index of 0.56 indicates that 56% of the references in that field are younger than 5 years. The slowest fields

are mineralogy (0.28), ornithology (0.28), geology (0.28), mathematics (0.27) and palaeontology (0.24). In the sample the mean speed value is 0.44, and the standard deviation is 0.06.

Table 2.3: Speed of the five fastest and slowest scientific fields

Scientific Field	Speed
Environmental Studies	0.557739
Cell Biology	0.548372
Computer Science, Information Systems	0.537893
Physics, Particles & Fields	0.527219
Infectious Diseases	0.527121
[…]	[…]
Mineralogy	0.284343
Ornithology	0.27876
Geology	0.275962
Mathematics	0.274093
Paleontology	0.245459

Note: This table shows the 5 slowest and fastest scientific fields. Speed is measured with the Price Index (PI) from Milojević (2012).

We replicate the previous results implementing two alternative indicators. The first indicator is the average number of authors (*Avg Auth*, eq 2.2) measured as the average number of authors in a scientific field. The average number of authors is also one of the measures used in Jones (2009) to introduce the concept of burden of knowledge. The second indicator used as part of our robustness checking is the Inverse of References Age (*IRA*, eq 2.3), calculated as the inverse of the reference age in a scientific field. Both measures are calculated in the same way of *speed*, so considering only articles written by universities, excluding companies, and spanning the years 2000 to 2001.

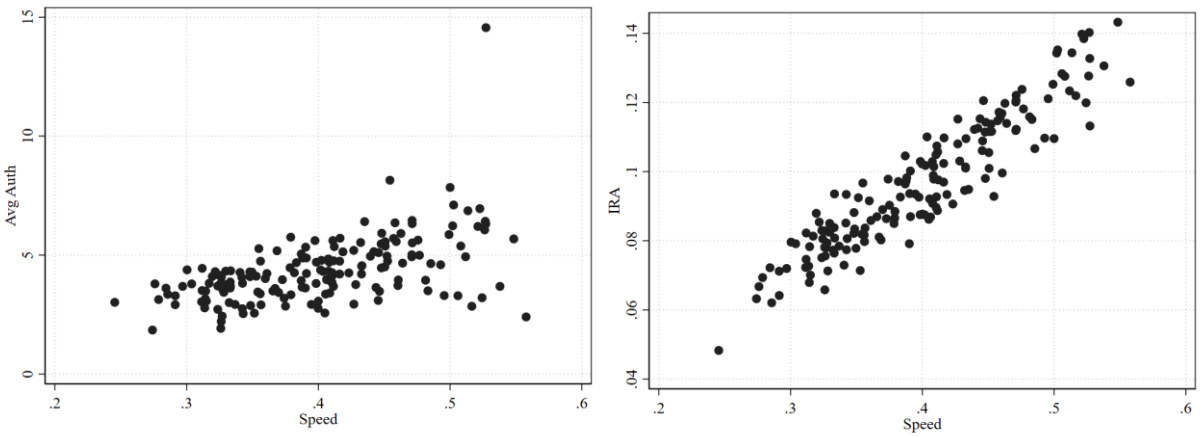
$$Avg\ Auth = \frac{1}{n}\sum authors\ per\ paper_{jt} \tag{2.2}$$

$$IRA = \frac{1}{\frac{1}{n}\sum references\ age_{jt}} \tag{2.3}$$

Figure 2.5 shows the correlation between *speed* and *IRA*, and between *speed* and *Avg Auth*. In both cases speed is positively correlated with the two alternative measure of the burden of

knowledge, suggesting that at least descriptively the measure capture similar aspects. Table 2.4 presents some descriptive statistics about all measures of the burden of knowledge, namely *Speed*, *Avg Auth* and *IRA*

Figure 2.5: Correlation between speed and Avg auth (left) and speed and IRA (right)



Note: This figure shows the scatterplots of the correlations between *speed* and *Avg auth* (left) and *speed* and IRA (right)

Table 2.4: Summary statistics of the burden of knowledge variables

Variables	# observations	Mean	S.D	Min	Max
Speed	454,310	0.433717	0.059021	0.245459	0.557739
Avg Auth	454,310	4.666678	1.130705	1.849877	14.55935
IRA	454,310	0.104858	0.016848	0.04828	0.143225

Note: This table shows the descriptive statistics for the burden of knowledge measures. The sample extends from 2000 to 2013 and comprises only papers (no proceedings) and firms with at least two employees.

2.4 EMPIRICAL STRATEGY

In Section 2.2, we hypothesised that changes in the knowledge burden might affect firms' collaboration propensity. In this section, we assess the relationship between the speed of the scientific frontier and companies' propensity to collaborate with universities. Since we are primarily interested in abstracting from factors affecting commercial or competitive environment of innovating firms, we approach the issue by modelling decision to collaborate on various research topics within a given firms portfolio of projects; that is, controlling for each firm's average propensity to collaborate. This goal can be achieved by introducing firm, year

and scientific field fixed effects. We estimate a Linear Probability Model (LPM), a common approach with models with rich fixed effects (Bellemare et al., 2015; Greene, 2004), as follows.

$$Collaboration_{ifjt} = \beta_0 + \beta_1 speed_{\ j} + \lambda_f + \mu_t + u \quad (2.4)$$

For paper *i* published by firm *f*, in field *j*, and in year *t*. *Collaboration* is a dummy variable equal to 0 if a publication is solo, and equal to one if it is a collaboration. Our definition of collaboration is presented in Section 2.3. We focus on papers that belong to a single scientific field and a single corporate author. Results are consistent when we use an expanded sample, including papers belonging to multiple firms and subjects, which are included in Appendix B.

The coefficient β_1 can be interpreted as the probability increase in collaborating with a university if *speed* increases.

We choose a LPM instead of a logit model for the following reasons. Fixed effects are crucial to our empirical strategy for addressing endogeneity concerns, and nonlinear models like probit and logit are not good choices for dealing with large numbers of fixed effects due to the incidental parameter problem (Bellemare, Novak, and Steinmetz 2015; Greene 2004). In addition, LPM with fixed effects allows to calculate the coefficient without dropping the groups with no variability in the dependent variable. Moreover, LPM does not impose a distribution on the error term and thus prevents identification via the specific functional form (Bellemare et al., 2015). Last, the results are directly interpretable as marginal effects without the need to transform the odds ratios (Cameron and Trivedi 2005; Gomila 20200924).

There are some concerns related to the use of LPM, but we take a number of steps to mitigate them. First, linear regressions with binary dependent variables are heteroskedastic. Empirically, the problem can be solved by estimating robust standard errors. Second, LPM can estimate probabilities that fall outside of the interval (0,1) and the bias and inconsistency increase as the number of predictions falling outside the interval increases (Horrace and Oaxaca 2006). Wooldridge (2002), Angrist and Pischke (2008), and Hellevik (2009), however, suggest that the issue might not be relevant if interested in causality and not in predictions.

We include firm fixed effects λ_f and year fixed effects μ_t, to avoid potential endogeneity in the difference publication strategies across firms and across years. We calculate robust standard errors to correct for heteroskedasticity, inevitable in linear probability models.

In Section 2.2 we anticipated that the benefits of collaboration would be conditioned on firm size, which we use to proxy firm capacity. To estimate the coefficients of interest, we estimate the following equation.

$$Collaboration_{ifjt} = \beta_0 + \beta_3 speed_j * firm\ size_f + \lambda_f + \mu_t + \gamma_j + u \quad (2.5)$$

For paper *i* published by firm *f*, in field *j*, and in year *t*. *Collaboration* is a dummy variable equal to 0 if a publication is solo and equal to one if it is a collaboration. We measure *firm size* as the natural logarithm of the number of employees. As remarked in Section 2.3 firm size does not change in time and is calculated as the average number of employees from 2000-2018. β_3 is the coefficient of the interaction of *speed* and *firm size* and it can be interpreted as the probability increase in collaborating if size and speed increase. We exploit two levels of cross sectional variation as in R. Rajan and Zingales (1996), exploiting variation across firms and scientific fields, rather than time variation with cross-sectional variation. Introducing firm (λ_f) and scientific fields (γ_j) fixed effects we can separate the effect of speed from the individual companies' propensity to collaborate and field-specific characteristics. β_1 and β_2 are not reported in the equation as they are absorbed by the fixed effects.

In Section 2.2 we argued that we do not expect the relationship between *collaboration* and *firm size* may not be linear. More specifically we expect the effect of speed to be less pronounced for small and very large firms. In order to test nonlinearity, we estimate the

$$Collaboration_{ifjt} = \beta_0 + \beta_4 speed_j * firm\ size_f + \beta_5 speed_j * firm\ size_f^2 + \lambda_f + \mu_t + \gamma_j + u \quad (2.6)$$

following equation.

For paper *i* published by firm *f*, in field *j*, and in year *t*. Differently from equation (2.5) we introduce a quadratic term *firm size*^{2} to account for a nonlinear relationship between firms' size and the probability of collaborating. Interpreting the regression coefficients is not as straightforward as in the previous model. To determine the impact of *speed* on the probability of collaboration, we need to compute the partial derivative of *Collab* with respect to *firm size*. In this case, the partial derivative will be equal to:

$$\frac{\partial collab}{\partial speed} = \beta_4 firm\ size + \beta_5 firm\ size^2 \quad (2.7)$$

Modelling the equation in this manner, the marginal effect will depend on. β_4 and β_5 and will vary by firm size. β_1, β_2 and β_3 are not reported in the equation as they are absorbed by the fixed effects.

2.5 RESULTS AND DISCUSSION

Table 2.5 shows the regression results of equation (2.1). Column 1 introduces year fixed effects, while column 2 introduces year and firm fixed effects. All the coefficients are positive and statistically significant, suggesting a positive association between *speed* and likelihood of collaboration. In particular, column 2 shows that after controlling firm fixed effects, the positive correlation between *speed* and the probability of collaborating holds. Interpreting the coefficient of column 2, an increase of *speed* of 0.1 would increase the probability of collaborating by 0.06. This means, within a firm's portfolio of research projects, they are more likely to collaborate with universities on work that is on topics where the literature is evolving relatively quickly.

It may be useful to quantify what a 0.1 increase in *speed* means by providing an example. Let us consider a scientific field with 100 references per paper on average. If 15% of the references are younger than five years speed equals to 0.15. A 0.1 increase in *speed* would mean that speed moves from 0.15 to 0.25, so that 25% of the references are younger than five years. In this case the likelihood of collaborating would increase by 6 percentage points.

Table 2.5: LPM. Impact of speed on the probability of collaborating.

	(1) Collaboration	(2) Collaboration
Speed	0.511***	0.600***
	(0.0182)	(0.0274)
Observations	137,874	124,555
R-squared	0.023	0.262
Year FE	Yes	Yes
Firm FE	No	Yes
Sc. Field FE	No	No

Note: This table presents the regression results of equation $Collaboration_{ifjt} = \beta_0 + \beta_1 speed_{\ j} + \lambda_f + \mu_t + + u$ introducing different level of fixed effects: year (column 1), year and firm (column 2).

In Table 2.6, we introduce the interaction term *speed*firm size* to test if the impact of *speed* on the probability of collaborating varies by firm size. The model is identical to the one in Table 2.5 for what regards the fixed effects. *Speed* is still positively correlated with the

probability of collaborating in all columns. The interaction coefficients present conflicting results. The interaction *speed*firm Size* is negative and statistically significant in column (2) with only year fixed effects, positive and statistically significant when introducing firm fixed effects, and non-significant when including scientific fields fixed effects.

Table 2.6: LPM. Impact of speed on the probability of collaborating with interaction

	(1) Collaboration	(2) Collaboration	(3) Collaboration	(4) Collaboration
Speed	0.578***	1.001***	0.474***	
	(0.0207)	(0.0451)	(0.0734)	
Firm size	-0.0107***	0.0133***		
	(0.000348)	(0.00243)		
Speed*Firm size		-0.0558***	0.0164**	0.00773
		(0.00549)	(0.00810)	(0.00807)
Observations	113,719	113,719	105,760	105,758
R-squared	0.029	0.030	0.230	0.330
Year FE	Yes	Yes	Yes	Yes
Firm FE	No	No	Yes	Yes
Sc. Field FE	No	No	No	Yes

Note: This table presents the regression results of equation $Collaborations_{ifjt} = \beta_0 + \beta_3 speed * \text{firm size}_i + \lambda_f + \mu_t + \gamma_j + u$ introducing different level of fixed effects: year (column 1 and 2), year and firm (3) and year firm and field (4).

Recall that we anticipated the conditioning role of firm size on the relationship to be highly non-linear, with the role of speed increasing as firm size increase above some minimal threshold to benefit from collaboration to a maximum and then decreasing as firm size increase to the extent that capacity constraints become less binding. An underlying non-linear relationship would make the value of the coefficients of the interaction *speed*firm size* undetermined, and indeed we see that it is not significant at conventional levels when both firm and field specific fixed effects are included (column 4). To explore this possibility further, divide the sample in two, separating large firms (with 250 or more employees) and SMEs (with less than 250 employees). Columns 1 to 3 of Table 2.7, show the regression results for SMEs, while columns 4 to 6 show those for large firms. First, the coefficients of *speed* are always positive and statistically significant except for column 2. It can be noticed that this baseline result is robust, and we can be confident about a positive correlation between speed and probability of collaborating. Concerning the interaction term, the coefficient is negative and statistically significant for large firms while statistically non-significant for small firms.

Table 2.7: Impact of speed on collaboration, LPM dividing SMEs and large firms

	<250 employees			≥250 employees		
	(1) Collaboration	(2) Collaboration	(3) Collaboration	(4) Collaboration	(5) Collaboration	(6) Collaboration
Speed	1.044***	-0.0242		0.849***	1.234***	
	(0.0849)	(0.149)		(0.137)	(0.189)	
Firm size	0.00838			0.0126**		
	(0.0110)			(0.00607)		
Speed*	-0.0548**	0.128***	0.0136	-0.0413***	-0.0532***	-0.0405**
Firm size	(0.0248)	(0.0434)	(0.0425)	(0.0135)	(0.0180)	(0.0175)
Observations	36,909	30,018	30,014	76,768	75,705	75,704
R-squared	0.040	0.499	0.526	0.017	0.136	0.266
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	Yes	No	Yes	Yes
Sc. Field FE	No	No	Yes	No	No	Yes

Note: This table presents the regression results of equation $Collaboration_{ifjt} = \beta_0 + \beta_3 speed *$ firm size $_i + \lambda_f + \mu_t + \gamma_j + u$ discerning by firms size. Columns 1-2-3 show the results for SMEs while 4-5-6 for large firms. Speed is time invariant and calculated as the average speed from 2000 to 2001.

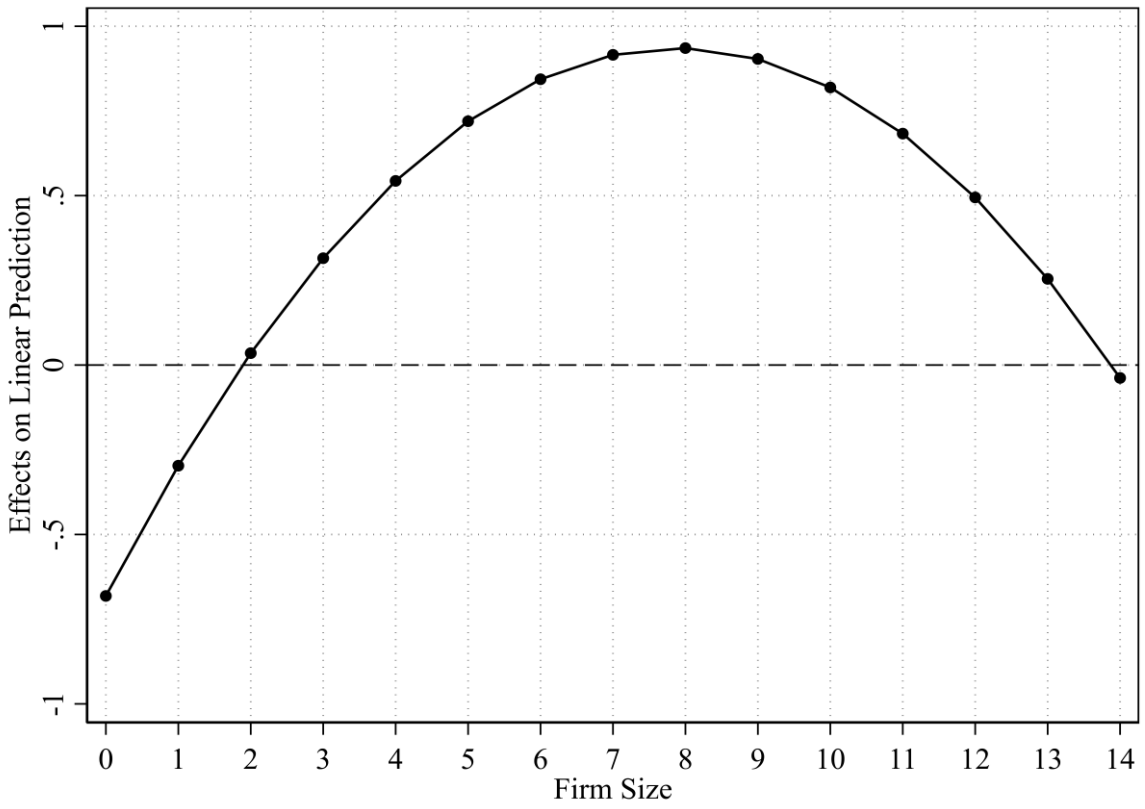
In Table 2.8, we test the nonlinearity in the relationship between firm size and collaboration. Both in columns 1 and 2 β_4 is greater than zero, while β_5 is lower than zero, suggesting an inverted U shape relationship. Figure 2.6 shows graphically the marginal effects at different values of *firm size*. The effect is positive for values between 2.3 and 13.68 and negative for values between zero and 2.3 and between 13.68 and 14. The effect is increasing until 8.03 and then declining until 14. This evidence confirms the suggestions of Table 2.8 as the effect increases as firms get bigger, but only until a certain threshold; after that, it starts declining. It has to be noted that 8.03, the maximum, corresponds to 3,072 employees.

Table 2.8: Impact of speed on the probability of collaborating with quadratic interaction

	(1) Collaboration	(2) Collaboration
Speed	-0.681***	
	(0.139)	
Speed*Firm size	0.410***	0.134***
	(0.0464)	(0.0445)
Speed*Firm size^2	-0.0260***	-0.00839***
	(0.00317)	(0.00305)
Observations	105,723	105,721
R-squared	0.230	0.330
Year FE	Yes	Yes
Firm FE	Yes	Yes
Sc. Field FE	No	Yes

Note: This table shows the results of equation $Collaboration_{ijft} = \beta_0 + \beta_4 speed_{ij} * firm\ Size_f + \beta_5 speed_j * firms\ Size_f^2 + \lambda_f + \mu_t + \gamma_j + u$. This model presents a double interaction, one with *Firm size* and one with *Firm size*^{2}. Papers with multiple companies and scientific fields are included.

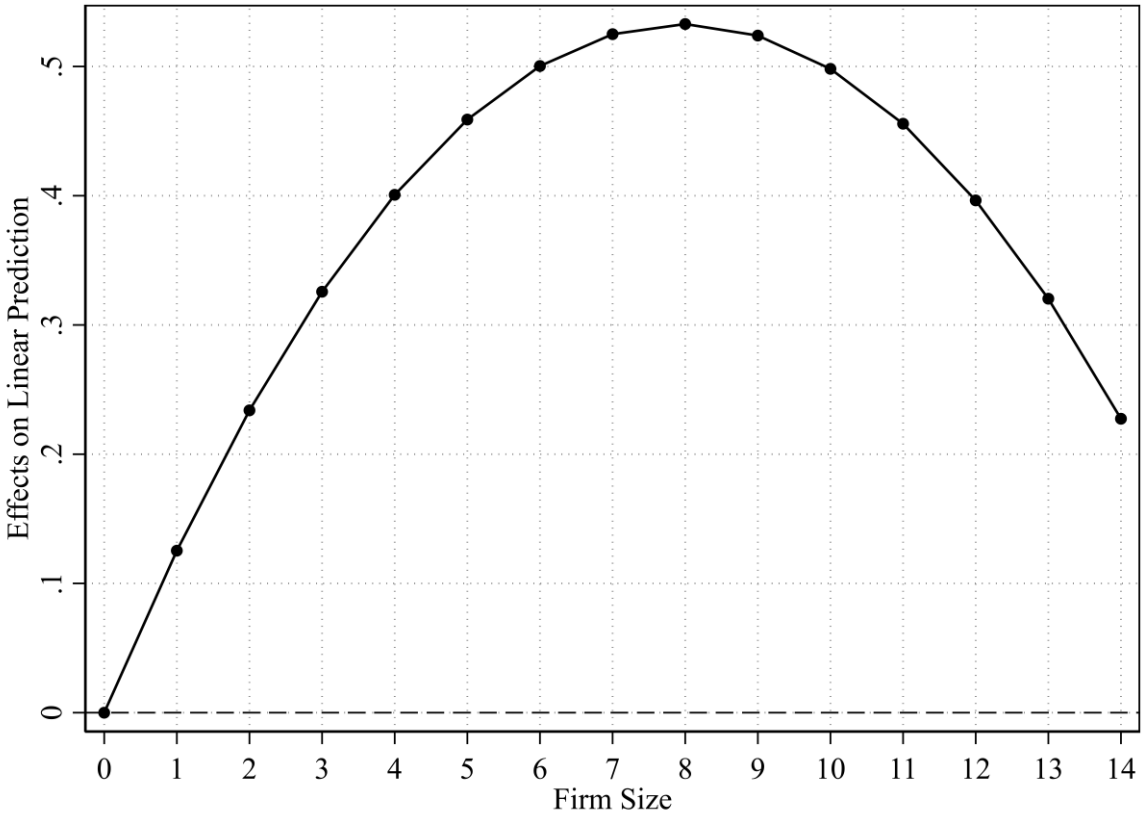
Figure 2.6: Speed marginal effects



Note: This figure shows the average marginal effect of speed estimated from the equation $Collaboration_{ijft} = \beta_0 + \beta_4 speed_{ij} * firm\ Size_f + \beta_5 speed_j * firm\ Size_f^2 + \lambda_f + \mu_t + u$. The marginal effects are calculated for values of bigger that range from 0 to 14.

Figure 2.7 shows the average marginal effects speed with all fixed effects, including subject fixed effects. The effect is positive for all values greater than 0. The effect increases until 7.89 and then declines until 14. The relationship between collaborations and speed is similar to the one described in Figure 2.6, with the marginal effects increasing until 7.89 (the maximum, equivalent to 2670 employees) and declining afterwards.

Figure 2.7: Marginal effects of speed with scientific field fixed effects



Note: This figure shows the Average marginal effect of speed estimated from the equation $Collaboration_{ijft} = \beta_0 + \beta_4 speed_{ij} * Firm\ Size_f + \beta_5 speed_j * Firm\ Size_f^2 + \lambda_f + \mu_t + \gamma_j + u$. The marginal effects are calculated for values of bigger that range from 0 to 14.

In the previous specification, we did not control for industry variables. To eliminate potential concerns over the results being driven by a particular industrial sector, we estimate the model of equation (2.6) twelve times, each one excluding a different Fama French industrial sector from the estimation. Table 2.9 shows the estimations for the model with firm and year fixed effects. Every estimated coefficient aligns with the baseline specification.

Table 2.10 shows model estimation with firm, year, and scientific field fixed effects. In this case as well, the coefficients follow the same pattern, supporting the presence of an inverted U

shape relationship between speed and the likelihood of collaboration. We repeat the same robustness check using 48 Fama French industries and the results are unchanged.

The regression results for papers with more than one firm and more than one scientific field are available in Appendix B and support the baseline results.

Table 2.9: LPM removing one industry with firm and year fixed effects

VARIABLES	(1) BusEq	(2) Chems	(3) Durbl	(4) Enrgy	(5) Hlth	(6) Manuf	(7) Money	(8) NoDur	(9) Shops	(10) Telcm	(11) Utils	(12) other
Speed	-0.770***	-0.911***	-0.909***	-0.906***	-0.531***	-1.026***	-0.878***	-0.968***	-0.889***	-0.876***	-0.891***	-1.049***
	(0.136)	(0.130)	(0.129)	(0.129)	(0.143)	(0.131)	(0.131)	(0.132)	(0.132)	(0.129)	(0.129)	(0.175)
Speed*Firm Size	0.464***	0.510***	0.508***	0.506***	0.383***	0.562***	0.500***	0.530***	0.504***	0.496***	0.503***	0.505***
	(0.0464)	(0.0430)	(0.0425)	(0.0426)	(0.0469)	(0.0437)	(0.0430)	(0.0434)	(0.0431)	(0.0425)	(0.0425)	(0.0561)
Speed*Firm Size^2	-0.0286***	-0.0339***	-0.0337***	-0.0335***	-0.0298***	-0.0376***	-0.0334***	-0.0352***	-0.0336***	-0.0329***	-0.0335***	-0.0316***
	(0.00320)	(0.00291)	(0.00288)	(0.00288)	(0.00316)	(0.00298)	(0.00290)	(0.00293)	(0.00291)	(0.00288)	(0.00288)	(0.00376)
Observations	98,685	116,735	120,466	118,736	81,414	111,004	120,514	117,590	119,064	121,125	121,212	90,309
R-squared	0.263	0.261	0.254	0.261	0.310	0.257	0.257	0.251	0.251	0.257	0.257	0.215
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sc. Field FE	No	No	No	No	No	No	No	No	No	No	No	No

Table 2.10: LPM removing one industry with firm, year, and sc. fields fixed effects

VARIABLES	(1) BusEq	(2) Chems	(3) Durbl	(4) Enrgy	(5) Hlth	(6) Manuf	(7) Money	(8) NoDur	(9) Shops	(10) Telcm	(11) Utils	(12) other
Speed*Firm Size	0.173***	0.141***	0.162***	0.152***	0.211***	0.173***	0.161***	0.171***	0.151***	0.163***	0.160***	0.198***
	(0.0447)	(0.0416)	(0.0411)	(0.0412)	(0.0459)	(0.0422)	(0.0416)	(0.0420)	(0.0418)	(0.0411)	(0.0411)	(0.0540)
Speed*Firm Size^2	-0.0112***	-0.00942***	-0.0104***	-0.00990***	-0.0139***	-0.0111***	-0.0104***	-0.0109***	-0.00983***	-0.0106***	-0.0103***	-0.0125***
	(0.00308)	(0.00283)	(0.00280)	(0.00280)	(0.00315)	(0.00289)	(0.00283)	(0.00285)	(0.00283)	(0.00280)	(0.00280)	(0.00363)
Observations	98,683	116,733	120,464	118,734	81,412	111,002	120,512	117,588	119,062	121,123	121,210	90,308
R-squared	0.365	0.355	0.347	0.355	0.363	0.352	0.347	0.344	0.344	0.349	0.349	0.323
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sc. Field FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: Table 2.10 shows the results of equation $Collaboration_{ijft} = \beta_0 + \beta_4 speed_j * firm\ Size_f + \beta_5 speed_j * firm\ Size^2 + \lambda_f + \mu_t + u.$ Table 2.10 shows the results of equation $Collaboration_{ijft} = \beta_0 + \beta_4 speed_j * firm\ Size + \beta_5 speed_j * firm\ Size_f^2 + \lambda_f + \mu_t + \gamma_j + u.$ Each column presents a separate regression without the industrial sector in the label.

2.6 ALTERNATIVE MEASURES OF THE BURDEN OF KNOWLEDGE

In Section 2.3.8 we introduced that the burden of knowledge can be measured in different ways, as there is no agreement on its definition. In this section we will use two alternative measures of burden of knowledge, the Inverse Reference Age (*IRA*) and the Average number of Authors (*Avg Auth*) per field.

Figure 2.8 shows the marginal effects for the model $Collaboration = \beta_0 + \beta_4 IRA_{ij} * firm\ Size_f + \beta_5 IRA_j * firm\ Size_f^2 + \lambda_f + \mu_t + \gamma_j + u$. The figure on the left includes firm and year fixed effect, while the figure on the right includes firm, year, and scientific field fixed effect. In the left figure the coefficients are positive and increasing between 1.86 and 8.35, and positive and decreasing for values lower than 8.35. The coefficient is negative for values lower than 1.86. In the right figure the coefficients are always positive, increasing until 9 and decreasing afterwards.

Figure 2.9 shows the marginal effects for the model $Collaboration = \beta_0 + \beta_4 Avg\ Auth_{ij} * firm\ size_f + \beta_5 Avg\ Auth_j * firm\ size_f^2 + \lambda_f + \mu_t + \gamma_j + u$. The figure on the left includes firm and year fixed effect, while the figure on the right includes firm, year, and scientific field fixed effect. In the left figure the coefficients are positive and increasing between 1.95 and 8.90, and positive and decreasing for values smaller than 8.90. The coefficient is negative for values lower than 1.95. In the right figure the coefficients are always positive, increasing until 8.94 and decreasing afterwards. In both Figure 2.8 and Figure 2.9 is clearly visible the inverted U relationship found in the baseline results. The regression tables are available in Appendix B.

Figure 2.8: Marginal effect of IRA firm and year FE (left), firm, year and scientific fields FE(right)

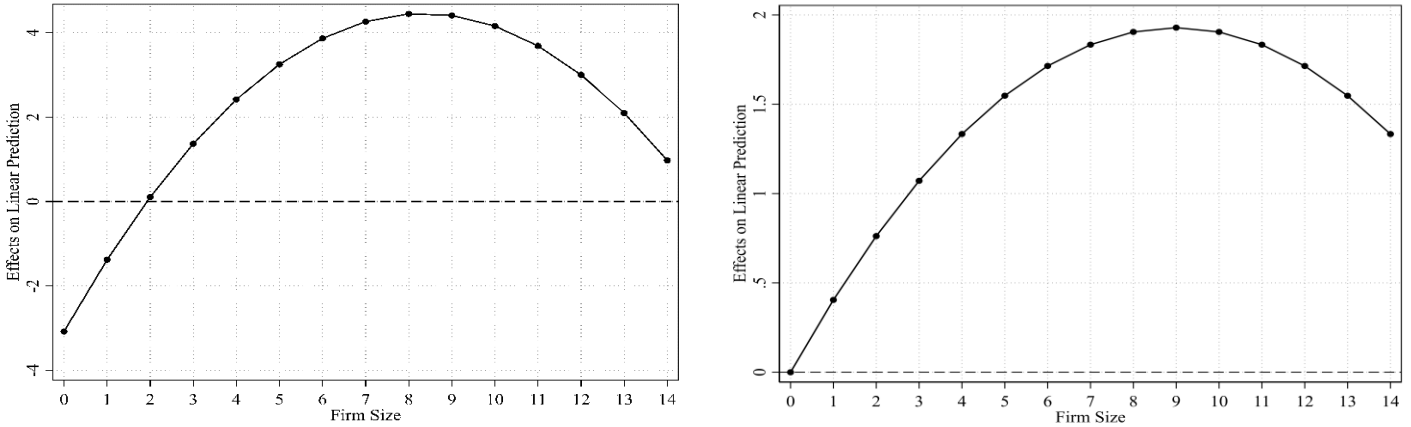
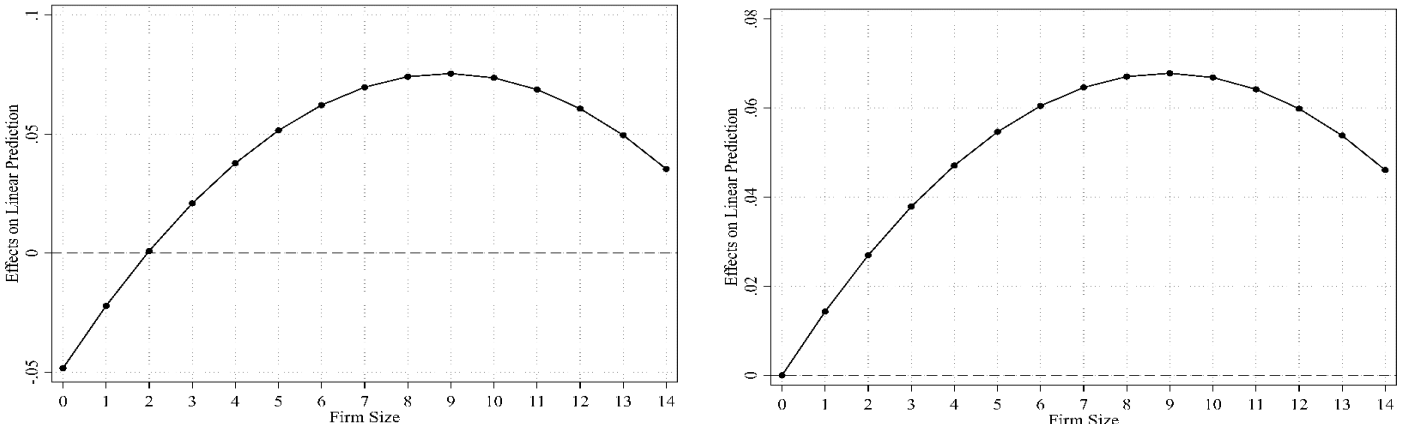


Figure 2.9: Marginal effect of Avg auth, firm and year FE (left), firm, year and scientific fields FE(right)



Note: Figure 2.8 shows the average marginal effect of time invariant speed estimated from the equation $Collaboration_{ijft} = \beta_0 + \beta_4 IRA_j * firm\ size_f + \beta_5 IRA_j * firm\ size_f^2 + \lambda_f + \mu_t + u$. The marginal effects are calculated for values of bigger that range from 0 to 14. The figure on the left shows the model with year and firm fixed effects, while the one on the right includes also sc. fields fixed effects. Figure 2.9 shows the average marginal effect of time invariant speed estimated from the equation $Collaboration_{ijft} = \beta_0 + \beta_4 Avg\ Auth_j * firm\ size_f + Avg\ Auth_j * firm\ size_f^2 + \lambda_f + \mu_t + u$. The marginal effects are calculated for values of bigger that range from 0 to 14. The figure on the left shows the model with year and firm fixed effects, while the one on the right includes also sc. fields fixed effects.

Before providing some concluding remarks there are a number of caveats and limitations that should be acknowledged, principally relating to limitations in the data construction and the empirical analysis.

While the data used in this analysis is arguably the most comprehensive available in the world at the time of writing, some problems remain perennial and will no doubt benefit from ongoing refinement. There is no single perfect method for identifying the nature of the institutions. Although universities and government agencies are easily recognisable due to recurrent keywords, not-for-profit firms, public medical centres, or clinics often have distinct names indistinguishable from for-profit firms if not checked manually. Second, tracking the changes in ownership is a universal challenge facing firm level data. Firms, especially in science-intensive sectors, might experience many changes of ownership in a few years, leading us to attribute the scientific publications to the wrong firms or to consider as independent a firm that is not independent anymore. Finally, if mislabelled, all the cases mentioned above might inflate the number of companies in our sample and, consequently, the number of collaborations.

Another broad challenge in measuring collaboration is that not all collaborations result in an observable output – whether through a publication or otherwise. Here we defined collaborations as co-authored publications, and as highlighted broadly by the literature (Melin and Persson 1996; Katz and Martin 1997; Calvert and Patel 2003; Ankrah and AL-Tabbaa 2015), this measure presents some limitations. For example, we cannot capture other formal (such as contract research, consultancy, and licencing) and informal (such as forums, lectures, and personal contact) collaborations.

Next, our variable measuring the speed of scientific progress, known as the Price Index, tends to decrease over time due to the aging of the existing body of scientific literature, as demonstrated by Egghe (2010). Consequently, we cannot assert whether scientific fields have become faster or slower but only compare fields to each other in a regression framework.

Fourth, we infer firms' size using the average number of employees. In future version of this chapter, we will address this issue performing the same empirical strategy on a subset of firms with more reliable financial information and using time varying firm size. Alternatively, we could employ alternative measures of firms' capacity, such as their patent portfolios, to capture firms' technological capacity regardless their size.

Last, we may employ a time varying measure of speed to test also if within-field intertemporal speed variation impacts the likelihood of collaboration.

2.7 CONCLUSIONS

This chapter focuses on the university-industry collaborations of US companies, given their rich history of science and innovation and their big achievements in the so-called golden era of corporate science from the post-war period to the end of the 1970s. This chapter approaches the study of university-industry collaborations systematically, matching companies from Bureau Van Dijk Orbis and publications from Clarivate Analytics WoS, obtaining one of the most extensive samples currently available in the literature containing large firms, SMEs, and foreign subsidiaries.

Our results highlight increased university-industry collaborations from 1980 to 2013 across most industrial sectors and scientific fields. The share of university-industry collaborations in relation to all peer-reviewed research articles increased from 2% to 6%. This surge in collaborations leaves no doubt — it is not possible to comprehend contemporary corporate scientific research without understanding the role of collaboration with universities.

In publication numbers, medical and biological sciences drove the increase in collaborations. Importantly, we find that the overwhelming majority of companies are pursuing a mixed research strategy – collaborating on some research projects while also producing scientific research authored entirely in-house. A priori, this pattern suggests that to understand drivers of collaboration, it is necessary to look beyond firm specific factors such as capital market demands or product market competition to try to understand which areas of research are subject to.

Our evidence suggests a positive relationship between the burden of knowledge and collaborations. We also find that this relationship is mediated by firm size, as there is an inverted U-shaped relationship between speed and collaboration. For the same level of burden of knowledge, the likelihood of collaboration increases as size increases, until around 2670 employees. After this threshold, the likelihood of collaboration decreases as size increases. Robustness checks with alternative measures of the burden of knowledge support this evidence.

How can we reconcile the increase in university-industry collaborations with the decline of corporate science driven by the lower value associated with research (as discussed for example in Arora, Belenzon, and Patacconi (2018))? Companies that assign less value to science should

substitute some of their in-house R&D with collaborations with universities. Our evidence, however, shows that collaborations are more than just replacing solo publications. Collaborations are becoming the dominant means of knowledge production, and companies are not losing interest in them. The burden of knowledge may explain both the decline in solo publications and the increase in collaborations, and it should be included in the ongoing debate on the decline of corporate science.

The results have clear policy implications. Significant involvement of US companies in scientific research is socially desirable because it directly fuels the economy's productivity growth (Rosenberg 1990). If science is really becoming open and most of the industrial publications come from collaborations, the government must ease the barriers between university and industry, keeping firms interested in long-term research projects and closer to the frontier of science.

It is still unclear, however, to what extent this increase in collaborations is related to the overall decline in corporate science and if, indirectly, it changed the firms' reliance on science in their innovations. As we looked exclusively at publications and not patents, we cannot say if the output of the collaborations is useful for firms' innovation, or if it is published just for strategic reasons, and possibly not strictly relevant to the companies' core business. Arora, Belenzon, and Sheer (2021), for example, find that firms produce more research when it is used internally, but less research when it can benefit rivals. Our evidence opens to further research to study the causal link between firms' collaborations and the internal use of scientific research.

Chapter 3: In and out the Pasteur's quadrant: revisiting trends in corporate science [33]

Using data on 900,000 corporate publications, 7 million university publications, and 2.5 million patents, we examine trends in corporate science from 1980 to 2014. Drawing from both Stokes (1997) and Ahmadpoor and Jones (2017), we measure appliedness and basicness with separate indicators under no assumption of a trade-off between the two. We find that publications signed by corporate scientists have progressively become more applied and less basic than academic ones, after controlling for fields and journals, regardless of the firms' age or size.

3.1 INTRODUCTION

Science is at the basis of many technological innovations, and business companies have systematically exploited it since at least the XIX century. The way they have done it and still do it, however, has changed considerably over time, with one particular form of organisation, the industrial R&D laboratory, having attracted considerable attention. A vast historical literature has documented the rise and expansion of these laboratories and their increasing engagement in scientific research (from which the term "corporate science"), up at least until the 1970s (Arora et al. 2021; Gertner 2012; Mowery 1983; Reich 2002; Smith 1990). More recently, the restructuring or demise of many such laboratories, along with a progressive outsourcing of scientific research to universities and start-ups, has attracted most of the scholarly attention (Arora, Belenzon, and Patacconi 2019; Block and Keller 2009; Mowery 2009). Besides documenting the rise and fall of famous R&D labs such as those of AT&T, Dupont and Xerox, several scholars have also suggested that corporate science, apart from shrinking, has also drifted away from fundamental research and towards more applied targets, including recombining established technologies rather than trying to create new ones. This would have long-term consequences on the innovation potential of both the business companies

[33] Publication version co-authored with Prof. Francesco Lissoni (https://orcid.org/0000-0002-2966-1414). Currently under review at Industry and Innovation.

and the overall economy (Arora, Belenzon, and Patacconi 2018; Arora, Belenzon, and Sheer 2021a; Bhaskarabhatla and Hegde 2014; Tijssen 2004).

While the general quantitative trends are unquestionable, the evidence on the basic-to-applied shift suffers from three major weaknesses. First, the literature mostly focuses on well-established incumbent firms and sectors, potentially overlooking the contributions and dynamics of new entrants and emerging industries. Second, the indicators used to measure basicness or appliedness of research mostly rely on outdated journal classifications (e.g. Hamilton, 2003). Last, and more fundamentally, the indicators chosen and the narrative around them betray a linear view of science-technology links, which implies a trade-off between basic and applied research. This is in contrast with Stokes' (1997) well-known argument on the importance of what he named Pasteur's Quadrant, namely the non-negligible set of circumstances under which the quest for a fundamental understanding of natural phenomena and for practical applications of research results can coexist, and even mutually reinforce each other.

In this chapter, we revisit the evidence on the basic-to-applied shift in corporate science by making use of what we deem being more appropriate indicators, which both operationalise Stokes' (1997) well-known taxonomy and exploit recent bibliometric advancements by Ahmadpoor and Jones (2017). Most notably, we measure research's applied and basic orientation with two independent and orthogonal indicators. Furthermore, based on a more inclusive dataset of business companies than those prevailing in the literature, we search for differences across sectors, firm size, and age.

Our dataset includes all the publications signed by authors with a US-based business firm affiliation from 1980 to 2014 in all sectors of activity. We distinguish between purely corporate publications, in which all authors have a business affiliation, and collaborative ones, which include authors affiliated with a university. As a result, our dataset is more extensive than that of Arora, Belenzon, and Sheer (2021b), as it covers not only the "core" corporate actors (large and persistent producers of corporate science whose R&D labs' decline has been widely documented) but also many smaller and shorter-lived firms that often go undetected. Besides, we consider every firm with a US address, thus also including the subsidiaries of foreign companies.

Our identification strategy relies on comparing the corporate and collaborative publications to a control sample of academic ones, in which all authors have university affiliations. To

ensure comparability, we select the control sample from journals in which both universities and companies publish. Using journal fixed effects and a time trend in our estimation, we can isolate and identify relative changes in appliedness and basicness while eliminating field-specific variations or heterogeneity. Notice that our methodological choice has the advantage of disentangling trends that may be specific to corporate science from general trends in science *tout court*. However, it comes at a price; we do not provide evidence on absolute values. Whenever we find a diverging trend between academic and corporate science, we cannot ascertain whether it is due to changes in the former, the latter, or both.

Our results show that corporate science is becoming more applied and less basic than academic science regardless of firms' size or age. Nevertheless, we do observe some variations across different fields. Specifically, agricultural sciences, biological sciences, computer sciences, geosciences, medical sciences, and physics are shifting towards applied science. Concurrently, biological sciences, computer sciences, engineering, and physics are becoming less basic. Collaborations also mirror this trend; however, the magnitude of the coefficients is relatively lower.

Taken together, our results show that corporate science is changing according to the trends identified by literature and moving away from Pasteur's quadrant, especially since the 2000s.

The remainder of this chapter proceeds as follows. Section 3.2 presents our conceptual framework. Section 3.3 introduces our data sources and methods. Section 3.4 tests if corporate science has become more applied and less basic. Section 3.5 concludes.

3.2 CONCEPTUAL FRAMEWORK

In innovation studies, terms such as basic research and science are often used interchangeably (Godin 2003; 2006). This ambiguity is a problem in the context of recent studies on the decline of corporate science, as this is usually associated with for-profit companies' engagement in scientific research, usually basic or discovery-driven (Hartmann and Henkel 2020; Zahra, Kaul, and Bolívar-Ramos 2018), implicitly assuming that a decline in corporate science reflects a decline in basic research. Many such studies tend to frame corporate science, more or less implicitly, within a linear model of innovation, in which the latter proceeds along a hierarchical process of subsequent steps from basic to applied research and then development. The model's origin is usually traced back to Vannevar Bush's Endless Frontier, and despite some criticism (Rosenberg 1994; Stokes 1997), it is still holding up and variously supported (Balconi, Brusoni, and Orsenigo 2010).

The most recent literature on the decline of corporate science argues that many large firms have been increasingly focusing on applied R&D for at least the last 20 years while neglecting basic research. Companies in the semiconductor sector, for example, are able to innovate and patent only by performing applied research and very little basic research, if not none (Lim 2004) or by recombining pre-existing technologies (Arora, Belenzon, and Patacconi 2018). This view subscribes implicitly to the linear view of basic and applied research as conceptually distinct activities and possibly mutually exclusive.

Furthermore, many bibliometric studies, despite providing an engaging historical narrative on the evolution of firms' scientific publications, are limited at attesting their decline in quantitative terms, with limited insights on the underlying changes in science and its relationships with technological progress (Arora, Belenzon, and Patacconi 2018; Arora, Belenzon, and Sheer 2021a; Lim 2004; Tijssen 2004).

Earlier research on corporate science, however, was largely based on the rejection of the clear-cut distinction between basic and applied research for a number of reasons. First, the direction of scientific progress is not unidirectional: while some basic research results need to be mulled over by downstream activities, it is also the case that discoveries in applied projects might be the starting point for more basic research efforts (Fabrizio 2009; Fleming and Sorenson 2004; Rosenberg 1990).

Second, more science, regardless its nature, helps build the company's absorptive capacity, namely the capacity to use and understand the knowledge produced elsewhere (Cohen and Levinthal 1990). Zahra (2002) remarks that the skills acquired in performing science allow the firms both to better deal with the tacit and hard-to-appropriate knowledge (Mowery and Oxley 1995) and to improve their problem-solving abilities (Kim 1998).

Stokes (1997) has provided one of the most influential critiques of the linear model, particularly in relation to the post-WW2 government policies that prioritised investment in basic research, viewed as the "pacemaker of technological progress" (Stokes 1997, 3). Stokes' key argument is that considerations of use, typical of applied research, do not necessarily contrast with the search for fundamental natural principles, which characterise basic research. The relationship between the two depends on both the intrinsic features of the research object and the socio-economic context in which the research is conducted. As a result, basic and applied science can be treated as two distinct concepts, each lending itself to separate measurements. This generates a taxonomy composed of three "quadrants," which Stokes names

after as many scientists, with each one epitomising the type of research in their respective quadrant. Most famously, research in Pasteur's quadrant is both inspired by consideration of use and a quest for a fundamental understanding of nature, as it was the case for microbiology in the XIX century and is still common in contemporary life science (Latour 1993; Murray and Stern 2006).[34]

Despite Stokes' taxonomy being often acknowledged, the metrics used to assess the extent of basicness and appliedness in recent corporate science literature do not yet incorporate its key implications, as it still treats the two dimensions as mutually exclusive. For example, several authors continue to rely on journal classifications based on experts' qualitative assessments, which rank journals linearly from the most basic to the most applied (e.g., Hamilton, 2003). More recently, Ahmadpoor and Jones (2017) have proposed a new metric, based on the citation distance between scientific publications and patents. While they still adhere to the linear model's perspective by defining appliedness and basicness as two opposing and distinct concepts, the proposed appliedness indicator can be used independently.

3.3 DATA AND METHODS

Our dataset combines three different data sources:

1. **Publications from Web of Science** (WoS, by Clarivate)**.** We use the 2015 edition of the WoS Core collection and select scientific articles from 1980 to 2014, with at least one author with a US address, for a raw total of more than 7 million papers.[35]
2. **Orbis company data.** We first disambiguate all company data for US firms from 13 different Orbis editions from 2005 to 2017 to obtain ~80 million unique identifiers over a longitudinal dimension. For most of them, we obtain from Orbis information on their sector of activity, size, ownership, and other economic and financial variables.[36]

[34] Stokes names the other quadrants in his taxonomy after Niels Bohr, representing exclusively basic research, in which the quest for fundamental principles trumps over any consideration for practical applications; and Thomas Edison, where the opposite holds, and practical applications take precedence over the pursuit of fundamental principles.
[35] Information about coverage and representativity is available in Web of Science, Clarivate analytics, accessed 14 June 2022, <https://clarivate.libguides.com/librarianresources/coverage>
[36] Information about coverage and representativity is available in Bajgar et al. (2020) and Kalemli-Ozcan et al. (2015).

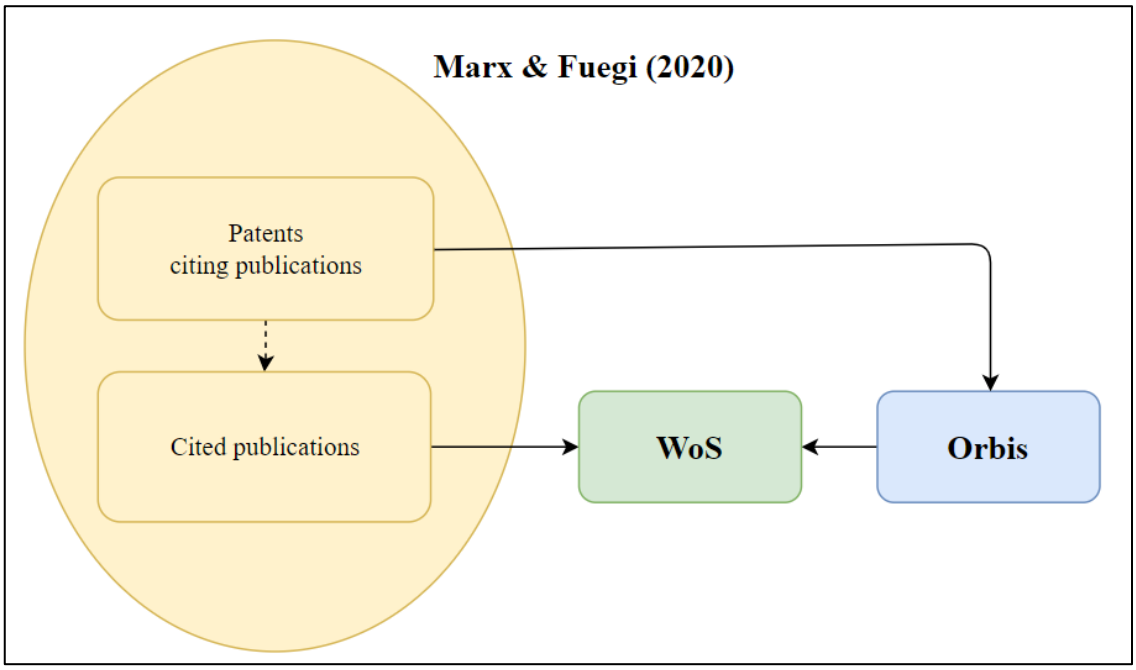
3. **The "reliance on science" database** by Marx and Fuegi (2020).[37] The database contains approximately 22 million non-patent literature (NPL) citations from worldwide patents to scientific publications from 1800 to 2018, both from the front page and in-text. The source of information on publications' authors, titles, journals, and other bibliographic data comes from MAG, the Microsoft Academic Graph.[38]

We first match all the publications from datasets 1 and 3 based on their doi, title, authors, journal and year. Second, we identify corporate publications by matching WoS authors' affiliations to Orbis company names, using a decision tree algorithm that incorporates string similarity scores, the presence of shared non-dictionary words, and address information (same city or zip code). The resulting matched sample contains 913,113 articles and proceedings and 93,365 firms (4,636 public, 88,435 private, 294 foreign with US subsidiaries). Further information about the matching techniques is available in Chapter 1 and Appendix A. [39]

Figure 3.1 provides a synthetic view of our data linkage strategy. It is important to note that our corporate sample is more extensive than that of Arora, Belenzon, and Patacconi (2018) and Arora, Belenzon, and Sheer (2021b), which include only US publicly listed firms, with positive R&D and at least one patent. We consider instead all US public and private firms including foreign subsidiaries. This choice allows us to examine a larger and more heterogeneous set of firms, which includes, in particular, a large number of small and medium enterprises with and without patents.

[37] Reliance on science in patenting (2022), Zenodo, accessed 14 June 2022, <https://zenodo.org/record/4235193#.YPqNn-j7RPa>.
[38] Microsoft Academic Graph (2022), Microsoft Corp, accessed 14 June 2022, <https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/>. Processed though PostgreSQL using Chacua (2020).
[39] The descriptive statistics differ slightly to WoS-Orbis US corporate science database presented in Chapter 1 as we used a previous version of the database when drafting this chapter.

Figure 3.1: Dataset construction



Note: Schematic representation of the dataset construction. The yellow oval shows the data from Marx and Fuegi (2020), while the green and blue rectangles publications form Web of Science and Orbis company information.

We also identify, more summarily, all the non-corporate publications as coming from universities and other non-profit organisations, as well as the publications coming from collaborations between firms and universities. This allows us to classify each publication into one of the following groups: *corporate*, *university* and *collaborations*. *Corporate* publications include all papers with at least one business affiliation but no university or non-profit affiliation. *University* conversely includes papers published by universities only without any co-author with a business affiliation. Residually, *collaborations* refer to those publications with at least a university and a corporate affiliation.

University publications serve as a benchmark, identifying trends in academic science *tout court* — undisturbed by the involvement of corporate scientists in the research process and resulting from universities' sole research effort. *Corporate* publications, and *collaborations* are compared to this benchmark of *university* publications to see if over time appliedness and basicness among the three groups diverge. One positive aspect of this methodology is that the structural biases of the appliedness and basicness indicators, both influenced by right truncation, affect in the same way all groups, allowing an easier comparison. However, when comparing groups to each other, we can only draw relative conclusions, unable to make assertions about the absolute values of the bibliometric indicators.

The corporate and collaborative publications in our dataset are published across 12,175 unique journals within STEM fields. For our bibliometric analysis, we specifically choose all university publications from the same set of journals. This methodological approach enables us to compare the three publication groups without introducing a composition effect, given that firms and universities might publish in different journals. Indeed, it is plausible that certain journals exclusively feature publications from either universities or corporate entities. By restricting our focus to journals where both universities and corporations contribute, we aim to investigate whether, while maintaining consistent journal characteristics, there are tangible differences between corporate and university publications. This leaves us with 6,967,425 publications.

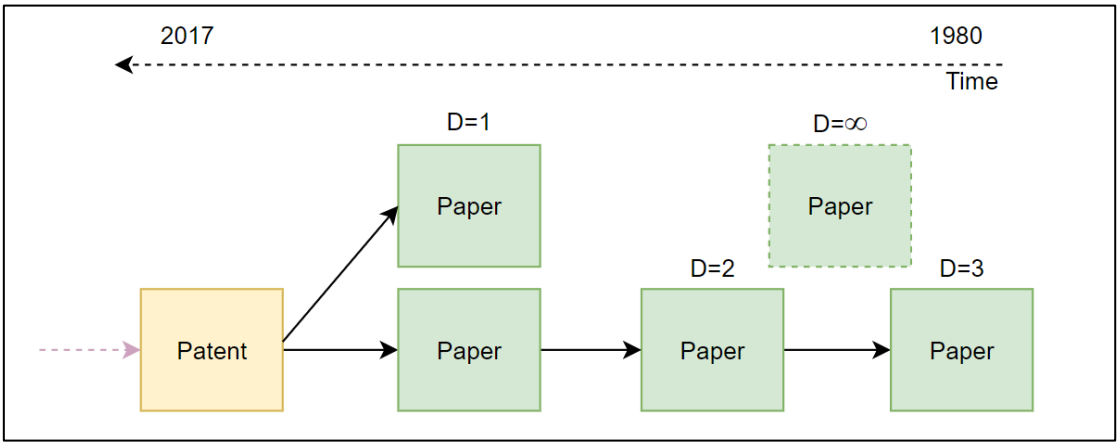
We then selected those publications that are also included in our third data source to assess how many of them are directly or indirectly cited by patents. This further reduces our dataset to 5,490,568 publications. For each one of the publications, we produce two indicators of, respectively, "appliedness" and "basicness".

3.3.1 Appliedness: inverse distance from the "dual frontier"

We build on Ahmadpoor and Jones (2017) (from now on A&J) and measure research appliedness as the inverse distance of the resulting publications from what we will call the "science-technology" or "dual" frontier, namely the set of patent and publication pairs linked by a direct citation running from the former to the latter. Starting from the publications on the frontier, we then follow the trail of backward citations of each publication in our dataset to compute the citation-based distance from the frontier of each publication in our dataset (see Figure 3.2). The value of appliedness remains constant over time and is determined by utilizing the network of backward citations in one specific point in time. In this case the network covers all publications up to 2014.

If a publication is cited directly by a patent (it belongs to the frontier), we consider the patent-publication distance equal to one (D=1). If a publication is cited by a publication on the frontier and by no patent, we assign distance two (D=2). To all publications neither at D=1 or D=2, but cited by publication at D=2, we assign distance three (D=3), and so forth. Any publication on a citation chain ultimately leading to the dual frontier, thus stands at distance D=k from the frontier, where k is the number of publications along the chain. As for publications that are not connected to the dual frontier by any chain, we classify them as *unlinked*, and we do not include them in the sample.

Figure 3.2: Distance metrics, visualisation



Note: Schematic representation of the distance metric. Patents on the left in yellow and papers in green. The dashed arrow indicates the time direction, while the solid arrow the direction of the backward citations. Papers directly cited by a patent are at distance D=1, papers cited by a paper at distance D=1 and not by a patent are at distance D=2. The same applies to D=3. The unconnected paper is at distance D=∞ and it is not included in the sample.

It is important to notice that each pair of publications in our dataset may be either linked by a direct citation (one publication appears in the reference list of the other) or several indirect ones (one publication also appears in the reference list of a further publication cited by the other). This means that each publication may be linked to the frontier by "citation paths" of different lengths. When this occurs, we retain only the shortest path.

By taking the inverse of distance, we obtain a measure of proximity to the dual frontier, which we rename *appliedness* and measure as follows.

$$\boldsymbol{Appliedness} = \frac{\ln(20) - \ln(distance)}{\ln(20)} \tag{3.1}$$

Where 20 is the maximum finite distance we find in our sample. This measure has the advantage of ranging from 0 to 1, where 0 represents the lowest appliedness and 1 the highest.

This indicator provides us with a quantitative, non-binary measure of how much science is applied in the sense of following considerations of use, such as those that are necessary to obtain a patent.[40] Publications connected to the dual frontier and close to it imply a higher applicability to technology, and thus are considered to be more applied than publications away

[40] It is worth recalling that, in most patent legislations worldwide, for an invention to be patentable it must satisfy three criteria, namely: novelty, inventive step and, which is of our interest, industrial application or usability. See for example paragraph 1483 of the USPTO consolidate patent rules (https://www.uspto.gov/web/offices/pac/mpep/consolidated_rules.pdf).

from it. Differently from A&J, however, we do not interpret appliedness as the opposite of basicness, that is we do not take for granted that more distant (but linked) publications are more basic. In line with Stokes' (1997) taxonomy, we treat the two dimensions as orthogonal, which requires inferring basicness from other indicators, as follows.

3.3.2 Basicness

We measure basicness with an indicator derived from Trajtenberg et al. (1997). This is based on the assumption that the more basic the research, the more its results will be abstract, general, and relevant for a larger set of disciplines. We measure this aspect by looking at the diversity of scientific disciplines of the publications citing the focal publication.

In particular, we apply a variant of the Rao Stirling index by Porter and Rafols (2009), which reflects three different aspects of diversity: variety, balance, and disparity (Wang, Thijs, and Glänzel 2015). The formula for the indicator is given by equation (3.2).

$$\boldsymbol{Basicness} = 1 - \sum_{ij} s_{ij}\, p_i p_j \tag{3.2}$$

where p_i is the proportion of papers belonging to Science Category (SC) i among all papers citing a focal publication and s_{ij} is the cosine measure of similarity between SCs i and j. The cosine s_{ij} ensures that SCs different from each other have a higher weight. Intuitively, basicness is higher if a paper is cited contextually by references from engineering and biology than engineering and acoustics.

We build a matrix for each year containing all the s_{ij} cosine values of all the possible combinations of science categories in Web of science. The SCxSC matrix is calculated as follows:

$$\frac{\sum xy}{\sqrt{\sum x \sum y}} \tag{3.3}$$

Where xy is a co-citation of science categories and $\sum x$ and $\sum y$ are the sum of the papers with at least a citation in that science category. If two science categories are always together in every paper in the sample, the resulting value of the matrix will be 1. On the contrary, if two science categories are never cited together, the resulting value of the matrix will be zero.

3.3.3 Descriptives

Table 3.1 provides some summary statistics. Figure 3.3 reports instead the average values of both appliedness and basicness, respectively on the horizontal and vertical axis, for ten broad disciplinary fields, each one consisting of an aggregation of multiple WoS science categories,

as per Milojević (2020).[41] Each paper can belong to multiple broad fields, and it is counted multiple times if this occurs. For each field, we have three points, one for each group of publications. The distribution of points on the graph suggests two observations. First, appliedness and basicness are largely complementary, in line with what was suggested by Stokes (1997) and contrary to the trade-off assumption typical of the linear model. Second, by roughly partitioning the graph into four quadrants, we can see that the most populated one is the top right one, namely Pasteur's, in which both appliedness and basicness tend to be very high. Not surprisingly, the points that populate it correspond to the disciplines close to Pasteur's ones (biology and medicine), plus computer sciences and collaborations in engineering.

Edison's quadrant, in which considerations of use prevail over the quest for fundamental understanding (see footnote 34), is mostly populated by corporate publications, most notably in engineering, physics, chemistry and agricultural sciences. Besides, the corporate publications are consistently more applied than university publications and collaborations.

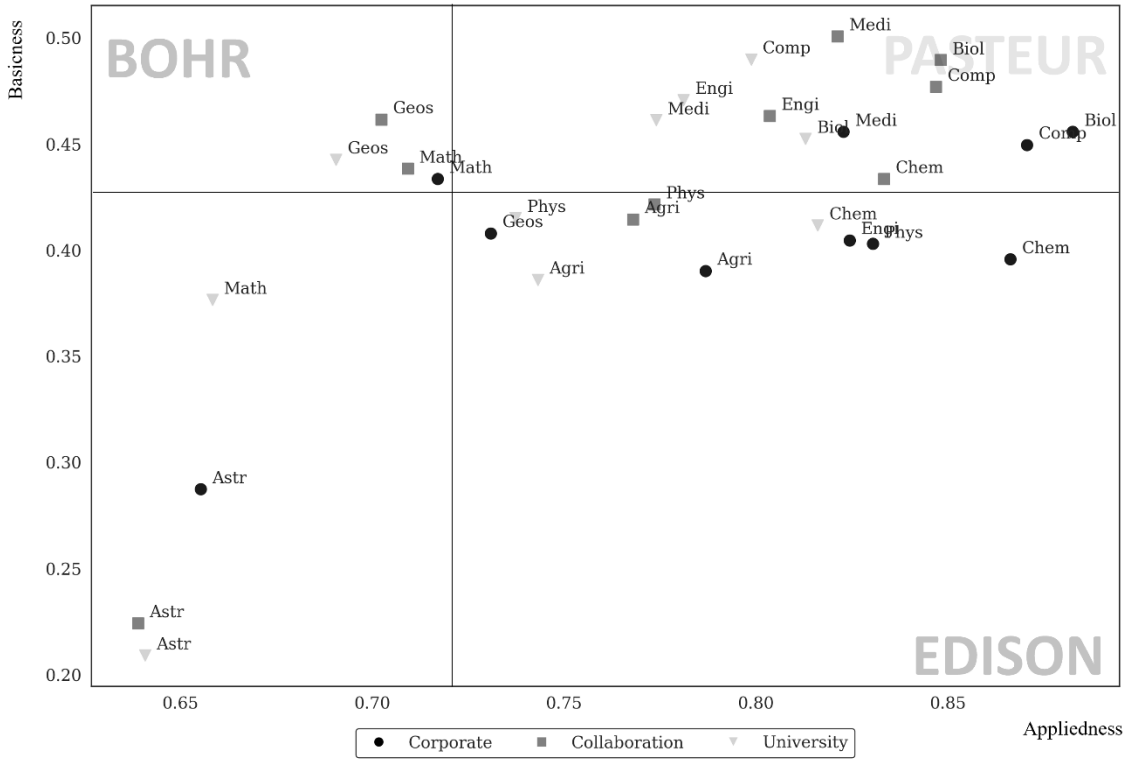
As for Bohr's quadrant, in which it is the fundamental understanding that counts, it is populated by mathematical sciences and geosciences. University publications in mathematics and astronomy do not belong to any of the three quadrants, which suggests some noise in our indicators with respect to Stokes' theory. Ideally, there should not be publications with low values for both appliedness and basicness. These abnormalities may happen for the following reasons. First, differences in the network of backward citations may lead to different metric outcomes. Second, our selection of journals in which both companies and universities publish potentially excludes some more basic journals where university disproportionately publish. Last, basic publications in mathematics and astronomy may receive most citations within the same field. It is important to note that any differences in the metric affect corporate and university publications equally. Since our primary goal is to estimate the differences across groups, none of the previous issues affects our estimates.

[41] Refer to footnote 16.

Table 3.1: Summary statistics

Variable		N	Mean	Std Dev	Min	Max
Appliedness	Corporate	282,648.00	0.84	0.15	0.14	1.00
	University	5,946,476.00	0.77	0.14	0.04	1.00
	Collaborations	428,183.00	0.81	0.15	0.17	1.00
Basicness	Corporate	245,120.00	0.41	0.19	0.00	0.90
	University	5,517,833.00	0.44	0.20	0.00	0.93
	Collaborations	401,167.00	0.46	0.18	0.00	0.92

Figure 3.3: Average appliedness and basicness by field, 1980-2014



Note: Average appliedness and basicness by field and group, pooling all publications together. Appliedness is on the x-axis while basicness on the y-axis. Both metrics range between 0 and 1. The circles represent the results for corporate science, the squares for collaborations, while the triangles for university. The top left is Bohr's quadrant, the top right Pasteur's quadrant, and the bottom right Edison's quadrant.

3.4 ANALYSIS

3.4.1 Methodology

We examine the trends in corporate science by means of an event study regression at the publication level, based on a sample including only university and corporate publications through the following equation:

$$Y_{\text{itj}} = \beta_o + \beta_1 Corp_i + \beta_2 \sum_{t=1}^{7} \text{I(year = t)} + \beta_3 Corp_i * \sum_{t=1}^{7} * \text{I(year = t)} + \mu_j + u_{itj} \quad (3.4a)$$

where Y_{itj} is our outcome variable, either *appliedness* or *basicness* of publication *i*, in time period *t* and journal *j*. Time periods are measured as groups of five years, so *t* can take values from 1 (1980-1984) to 7 (2010-2014). $I(year = t)$ is a dummy variable equal to one when the time period is *t* and zero otherwise, with *t= 1* as the omitted category. $Corp_i$ is a dummy variable equal to one if publication *i* belongs to the corporate group, zero if it belongs to universities. University publications serve as the reference category and act as a benchmark to observe differences with corporate publications. Journal fixed effects μ_j control for potential differences across publication layouts. Coefficients corresponding to β_2 capture the general time trends of the outcome variable (which we assume to coincide with trends of university publications), while those corresponding to β_3 capture the interaction between the time trend and the $corp_i$ dummy: if significant, they signal a divergence between trends in corporate science and general trends.

To eliminate potential field-specific effects, we implement two approaches. First, we run the model by pooling all fields together. Second, we run separate regressions, one for each scientific field.

We further investigate the possibility that trends differ across firms of different size and age, by classifying all firms in our sample by the age and size at time period *t*, as follows: young (less than 5 years) versus old (more than 5 years); and small (less than 150 employees) versus large (more than 150 employees).

To test the differences between collaborations and university science we retain in our sample only the collaborative and university publications and estimate the following equation:

$$Y_{\text{itj}} = \beta_o + \beta_1 Collab_i + \beta_2 \sum_{t=1}^{7} \text{I(year = t)} + \beta_3 Collab_i * \sum_{t=1}^{7} * \text{I(year = t)} + \mu_j + u_{itj} \quad \text{(3.4b)}$$

This is identical to equation (3.4a), except for the substitution of $corp_i$ with $collab_i$, a dummy variable equal to one if a publication belongs to the collaboration group, and zero if it belongs to universities.

3.4.2 Baseline results

Figure 3.4 shows graphically the estimated β_3 coefficients for both equations (3.4a,b), for the fields-pooled regression. The circles represent the results for corporate science, while the squares for collaborations.

Figure 3.4a shows the estimated coefficients when the dependent variable is appliedness. The full regression results are available in Appendix Table C 2. For the first four time periods, from 1980 to 1999, the coefficients for corporate science are stable and always close to zero, which is a parallel trend with university science, albeit for higher level of appliedness (the estimate for β_1, unreported in the figure, being in fact equal to 0.0300 and significant). Starting in 2000, however, the coefficients become positive and increasing, thus signalling a diverging trend, with corporate science becoming more and more applied. For example, the estimated β_3 coefficient for the period 2005-2009 is 0.0075, meaning that in total the corporate-university difference reached 0.0375.

The same applies, but to a lesser extent, to university-industry collaborations. The estimated β_1 is equal to 0.0164 while the estimated β_3 for the period 2005-2009 is 0.0032. This means once again a positive collaboration-university baseline difference, but equal to only half of the corporate-university difference and a much milder increase over time.

Figure 3.4b shows the equivalent results for basicness. The full regression table is available in Appendix Table C 2 . The estimated β_1 coefficient (unreported in the figure) is 0.0075, meaning that in 1980-1984 corporate science was on average more basic than university science. But the β_3 coefficients clearly exhibit a downward pattern from the very start, with a value of -0.0144 in the period 2005-2009. This implies, for the same period, a corporate-university basicness gap of -0.0069. In other words, after the period 2005-2009 corporate science is less basic than university science. Once again, the trend for collaborations is similar, but less pronounced.

One limitation of the evidence produced so far is that it may suffer of a composition effect, with the average results being driven by a few disciplinary fields, in which either corporate science has become more applied and less basic, or vice versa, university science has done the opposite. In Figure 3.5 and Figure 3.6 we break down the evidence by field, based on separate field regressions. The regression tables can be found in Appendix Table C 3, Table C 4, Table C 5, and Table C 6.

Figure 3.5 shows the results for appliedness. The fields exhibiting positive trends in the β_3 for corporate science are agricultural sciences, biological sciences, computer science, geosciences, medical sciences, and physics. More specifically, agricultural sciences and geosciences show a notable increase with β_3 in 2010-2014 equal to 0.0409 and 0.0497. biological sciences, computer sciences, and physics follow with β_3 equal to 0.0211, 0.0280, and 0.0290 respectively. Medical sciences experience an increase but to a lesser extent. The β_1 are positive and significant for all fields but agricultural sciences and geosciences. For example, consider biological sciences. The estimated β_1 is 0.0337, meaning that in average corporate biological sciences are more applied than university science. The β_3 show an upward pattern, with a value equal to 0.0211 in the period 2010-2014. As a consequence, corporate science is 0.0548 more applied than university science in that period.

No trend is evident for collaborations. Only geosciences have an upward trend after 2005. astronomy, biological sciences, chemistry and engineering have positive β_3 only in the last period. Medical sciences have a positive but stable trend. The other fields have positive but non-significant coefficients. In conclusion, even within field trends in appliedness are more evident for corporate publications than for collaborations.

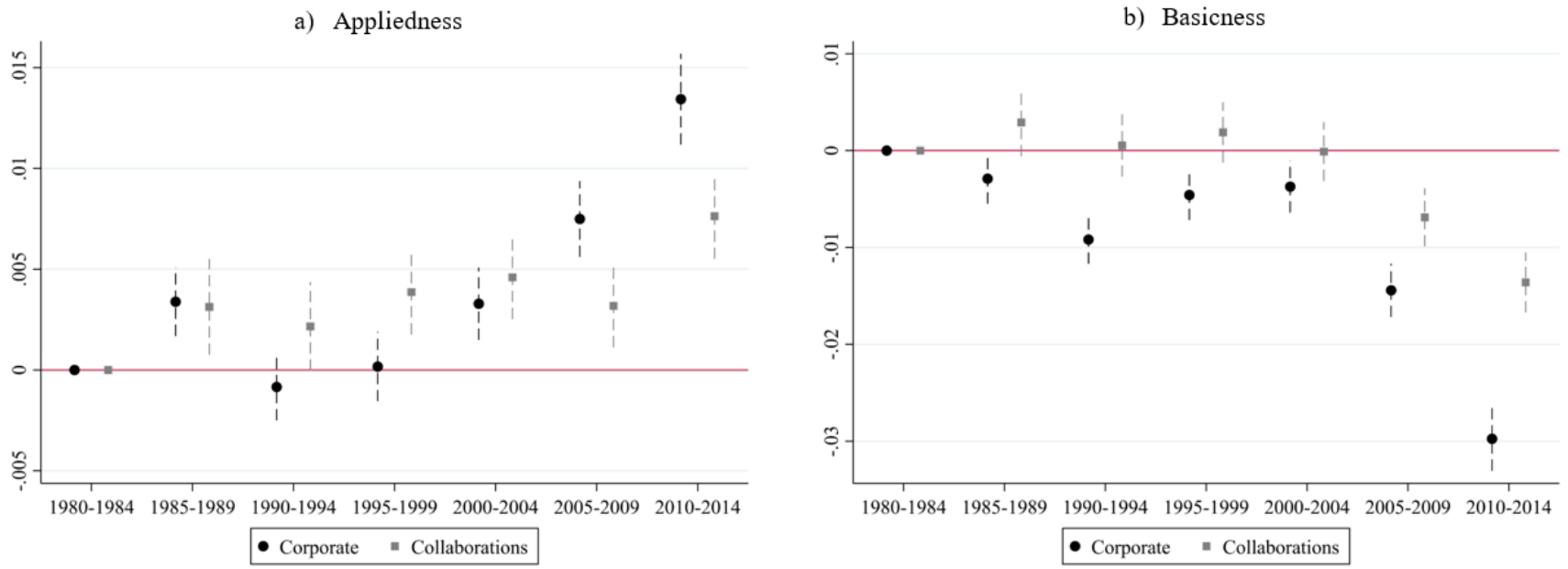
Figure 3.6 shows the results for basicness. For what concerns corporate publications, biological sciences, computer sciences, engineering, and physics exhibit negative trends, while astronomy is the only field that shows a positive trend. The β_1 are negative and statistically significant for astronomy, chemistry, and geosciences, meaning that in average corporate science is less basic. Conversely, biological sciences, computer sciences, engineering, mathematical sciences, and physics have positive and statistically significant β_3. Let's consider biological sciences again as an example. The estimated β_1 is 0.0155, meaning that in average corporate biological sciences are more basic than university science. The β_3 show an upward pattern, with a value equal to -0.0198 in the period 2005-2009. Therefore, after 2005 biological sciences are less basic than universities. Moreover, all fields with positive β_1 become less basic

than universities after 2005. Chemistry and geosciences were already less basic in 1980-1985 and have since continued to widen the gap over time.

Collaborations exhibit clear negative trends only in chemistry, computer sciences, medical sciences and physics. Biological sciences, chemistry, medical sciences, and physics have positive and statistically significant β_1, engineering a negative β_1, while the other fields show non-significant results. Medical sciences and physics experience the largest decrease. In 2010-2014, despite a downward trend, only physics is less basic than universities. On the other hand, chemistry and medical sciences, besides narrowing the gap, remain slightly more basic than universities.[42]

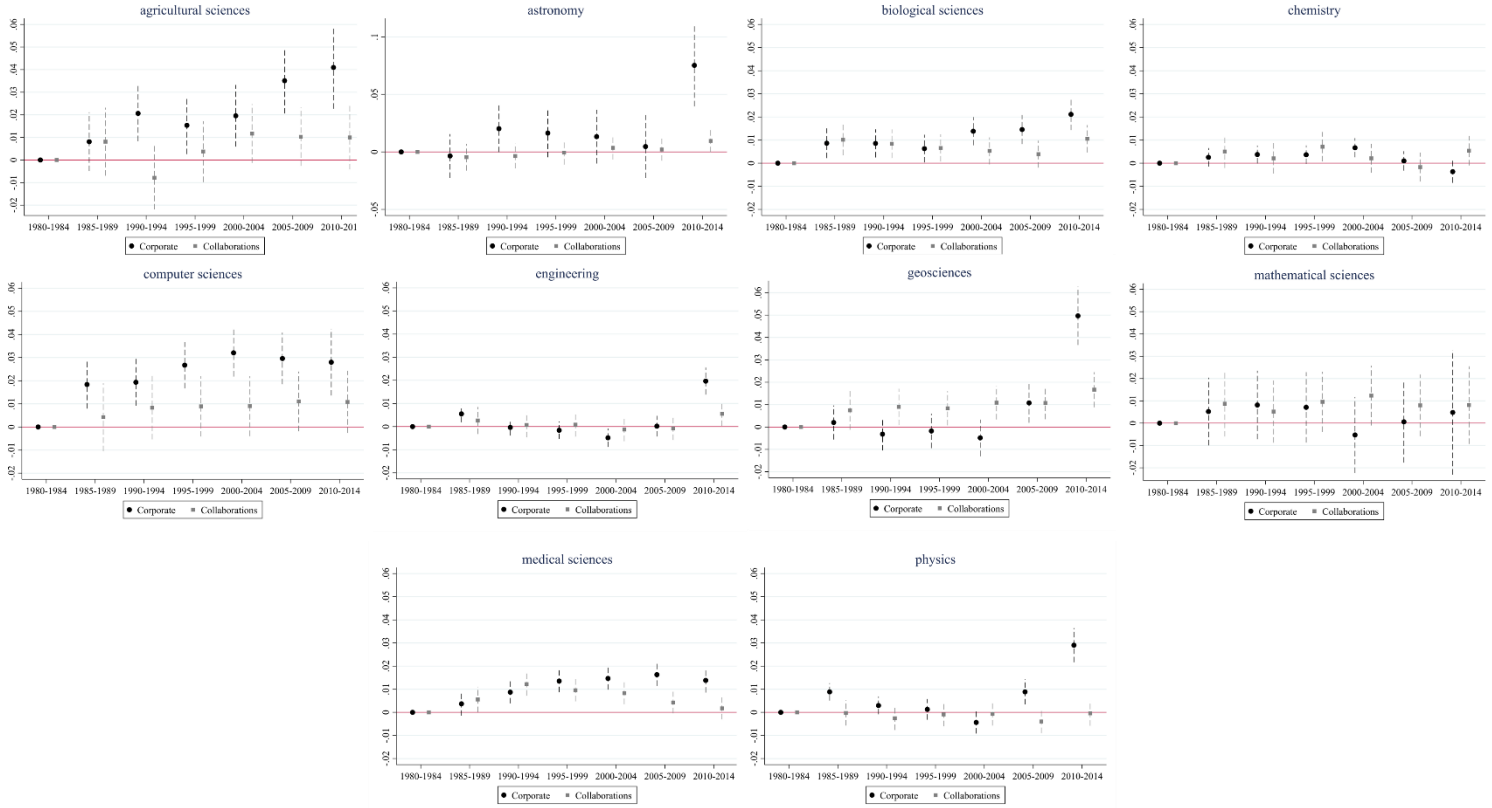
[42] Both measures of appliedness and basicness, are affected by right truncation, given that they are both based on backward citations. To address potential concerns about the right truncation of the database influencing our estimates, statistically significant where the right truncation happens, it may appropriate to correct the metric. Appliedness and basicness can be measured allowing each publication in our sample only five years to connect to the science-technology frontier. In this way we would give each publication the same chances to connect and mitigate the potential bias given by intertemporal differences in the backward citation network size.

Figure 3.4: Appliedness (left) and basicness (right), interaction term regression coefficients, 1980-2014



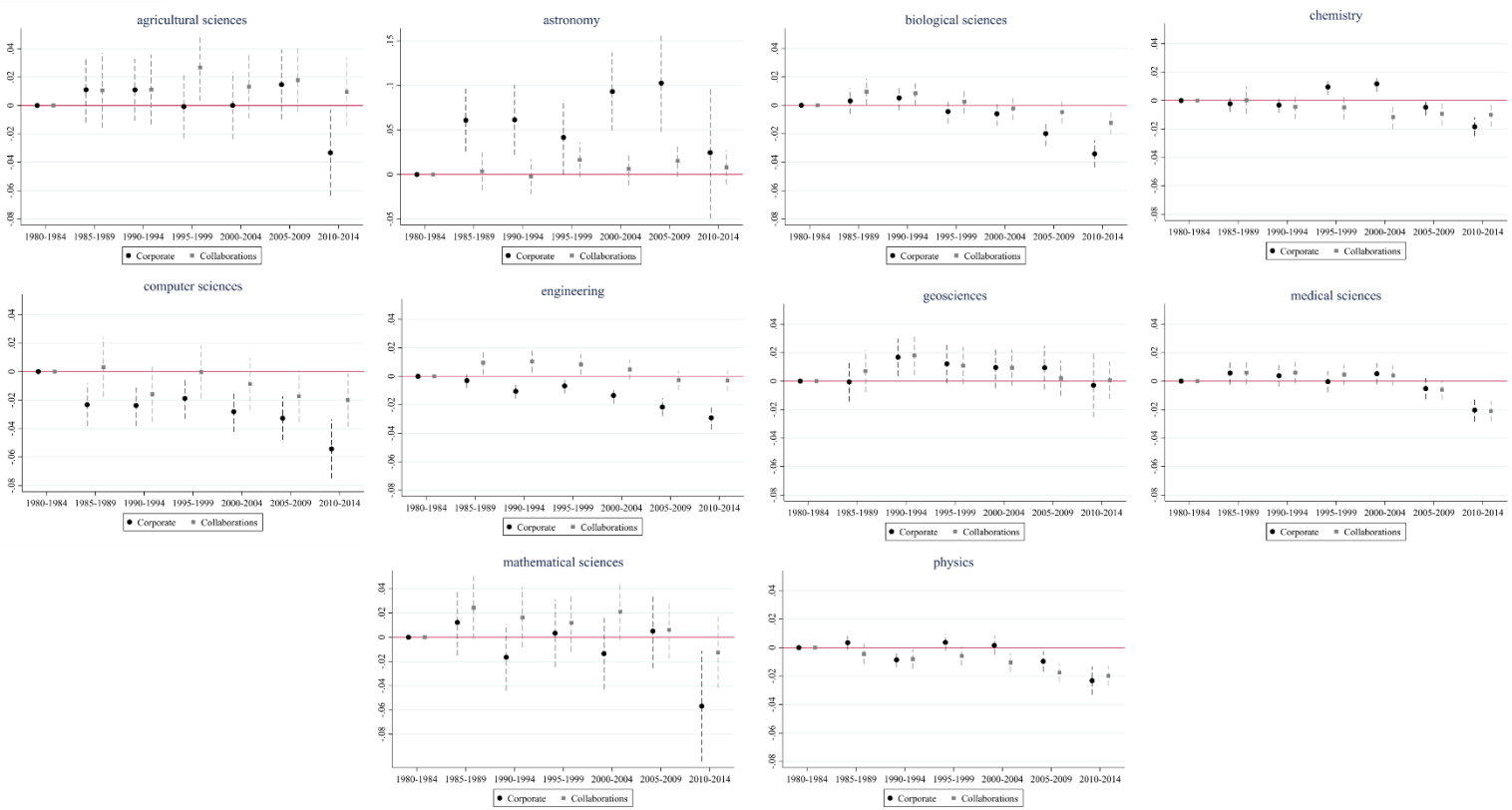
Note: This figure shows the coefficients for the regression $Y_{\text{itj}} = \beta_o + \beta_1 corp_i + \beta_2 \sum_{t=1}^{7} \text{I(year} = \text{t)} + \beta_3 corp * \sum_{t=1}^{7} * \text{I(year} = \text{t)} + \mu_j + u_{itj}$ and $Y_{\text{itj}} = \beta_o + \beta_1 collab_i + \beta_2 \sum_{t=1}^{7} \text{I(year} = \text{t)} + \beta_3 collab * \sum_{t=1}^{7} * \text{I(year} = \text{t)} + \mu_j + u_{itj}$. The dependent variable is appliedness in figure a) while basicness in figure b). The time trend is identified by 7 dummy variables for groups of 5 years. Confidence intervals are p<0.95. Circles represent the coefficients for corporate publications, while squares for collaborations. The β_1 not reported in the figure are available in Appendix Table C 2 and are the following for corporate and collaborations: appliedness 0.0300***and 0.0164***, basicness 0.0075*** and 0.0165***.

Figure 3.5: Appliedness interaction term regression coefficients by broad field, 1980-2014



Note Each figure shows a separate regression for 10 broad fields. The coefficients displayed come from the regression $Appliedness_{\mathrm{itj}} = \beta_o + \beta_1 corp_i + \beta_2 \sum_{t=1}^{7} \mathrm{I(year = t)} + \beta_3 corp * \sum_{t=1}^{7} * \mathrm{I(year = t)} + \mu_j + u_{itj}$ and $Appliedness_{\mathrm{itj}} = \beta_o + \beta_1 collab_i + \beta_2 \sum_{t=1}^{7} \mathrm{I(year = t)} + \beta_3 collab * \sum_{=1}^{7} * \mathrm{I(year = t)} + \mu_j + u_{itj}$. The time trend is identified by 7 dummy variables for groups of 5 years. Confidence intervals are p<0.95. Circles represent the coefficients for corporate publications, while squares for collaborations. The β_1 not reported in the figure are available in Appendix Table C 3 and Table C 4 are the following for corporate and collaborations: agricultural sciences 0.0009 and 0.0091, astronomy -0.0114* and 0.0049, biological sciences 0.0337*** and 0.0185***, chemistry 0.0274*** and 0.0049*, computer sciences 0.0085** and 0.0081, engineering 0.0274*** and 0.0075***, geosciences 0.0036 and -0.0031, mathematical sciences 0.0242*** and 0.0153***, medical sciences 0.0219*** and 0.0304***, and physics 0.0384*** and 0.0128***.

Figure 3.6: Basicness interaction term regression coefficients by broad field, 1980-2014



Note: Each figure shows a separate regression for 10 broad fields. The coefficients displayed come from the regression $Basicness_{\mathrm{itj}} = \beta_o + \beta_1 corp_i + \beta_2 \sum_{t=1}^{7} \mathrm{I(year = t)} + \beta_3 corp * \sum_{t=1}^{7} * \mathrm{I(year = t)} + \mu_j + u_{itj}$ and $Basicness_{\mathrm{itj}} = \beta_o + \beta_1 collab_i + \beta_2 \sum_{t=1}^{7} \mathrm{I(year = t)} + \beta_3 collab * \sum_{t=1}^{7} * \mathrm{I(year = t)} + \mu_j + u_{itj}$ Time trend is a monotonic transformation of time, grouping years by 5 years. Confidence intervals are p<0.95. Circles represent the coefficients for corporate publications, while squares for collaborations. The β_1 not reported in the figure are available in Appendix Table C 5 and Table C 6 are the following for corporate and collaborations: agricultural sciences -0.0133 and -0.0146, astronomy -0.0495*** and 0.0056, biological sciences 0.0155*** and 0.0202***, chemistry -0.0082*** and 0.0151***, computer sciences 0.0183***and 0.0104, engineering 0.0052*** and -0.0063*, geosciences -0.0282*** and 0.0006, mathematical sciences 0.0214** and 0.0063, medical sciences -0.0003 and 0.0284***, and physics 0.0119*** and 0.0190***.

Our results so far confirm what suggested by literature on the decline of corporate science we reviewed in Section 3.2, but in a more nuanced way. By making use of separate indicators for appliedness and basicness, we can portray it as migrating from Pasteur's to Edison's quadrant.

In order to do so, we consider jointly the estimated β_3 coefficients for equation (3.4a) (on appliedness) and the same values for equation (3.4b) (on basicness). In Figure 3.7 we report them on a quadrant representation comparable to that in Figure 3.3 with the horizontal axis representing values of the estimated coefficients (β_3) for appliedness and on the vertical axis those for basicness. Each circle and square in the plot represent a combination of coefficients for appliedness and basicness, ranging from period 2 (1985-1989) to period 7 (2010-2014), respectively for corporate science (black circles) and collaborations (grey squares). The reference category is 1980-1984. The arrows indicate the time directions. They clearly indicate that both corporate publications and collaborations are moving from Pasteur's towards Edison's quadrant. However, the direction of the movement is clear only for the last two periods, and collaborations are moving less.[43]

Figure 3.8 reports similar information, for the fields that exhibited statistically significant trends in appliedness or basicness in Figure 3.5 and Figure 3.6. On the left-hand side, we have the estimated pairs of coefficients for corporate publications (β_3), and on the right-hand side, the same for collaborations. For visualisation purposes we show only the coefficients of the last three periods (after 2000). It is evident that corporate publications are moving towards Edison's quadrant, while the behaviour for collaborations is less clear. Nonetheless, one consistent finding is that collaborations are progressing at a slower pace compared to corporate publications.[44]

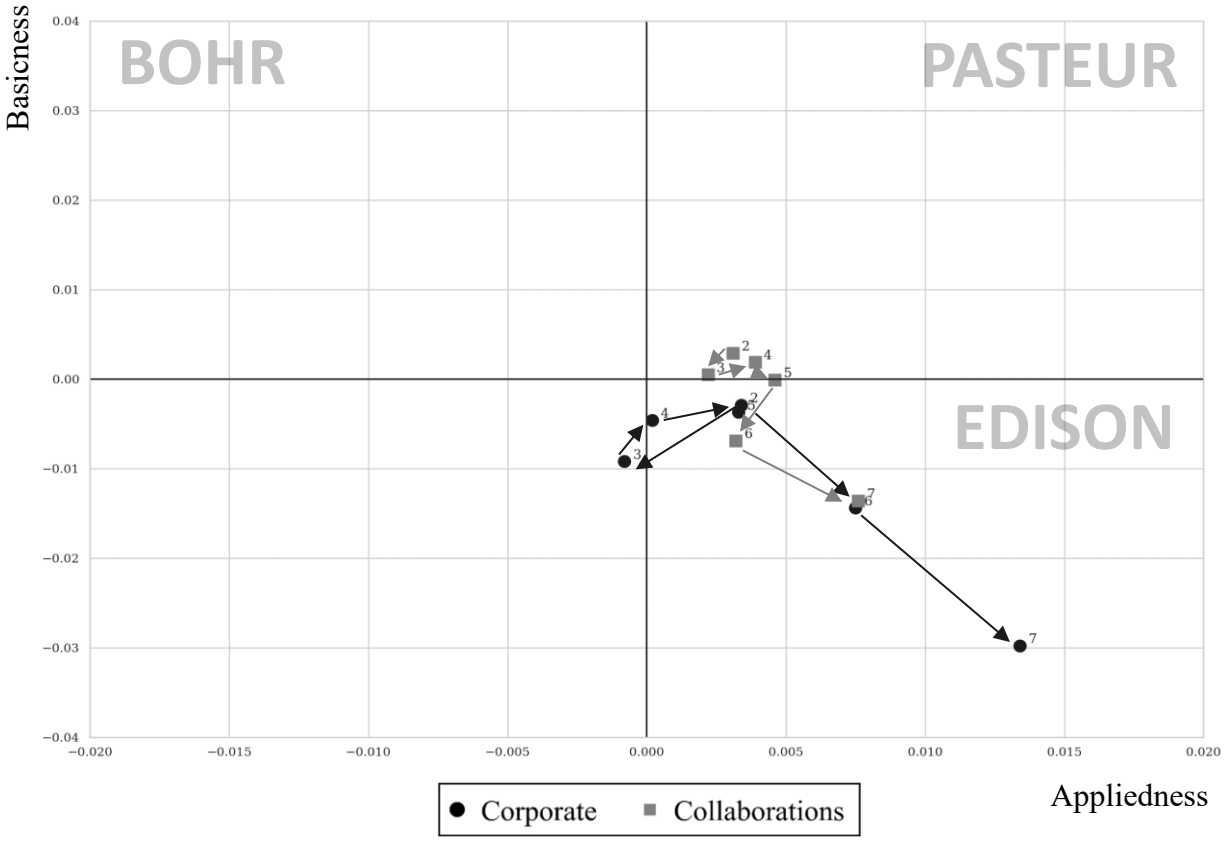
Our results are in line with Arora, Belenzon, and Patacconi (2018). Companies are indeed becoming more applied. Additionally, they are also becoming less basic. However, it is possible that the trend might be driven by a particular set of firms. In the next section we explore if there are any differences in term of appliedness and basicness by firm size and age. The narrative surrounding the decline of corporate science often revolves around established

[43] No confidence interval in displayed in the plot. To have further information about the standard errors and the confidence interval refer to Appendix Table C 2.

[44] No confidence interval in displayed in the plot. To have further information about the standard errors and the confidence interval refer to Appendix Table C 3,Table C 4, Table C 5, and Table C 6.

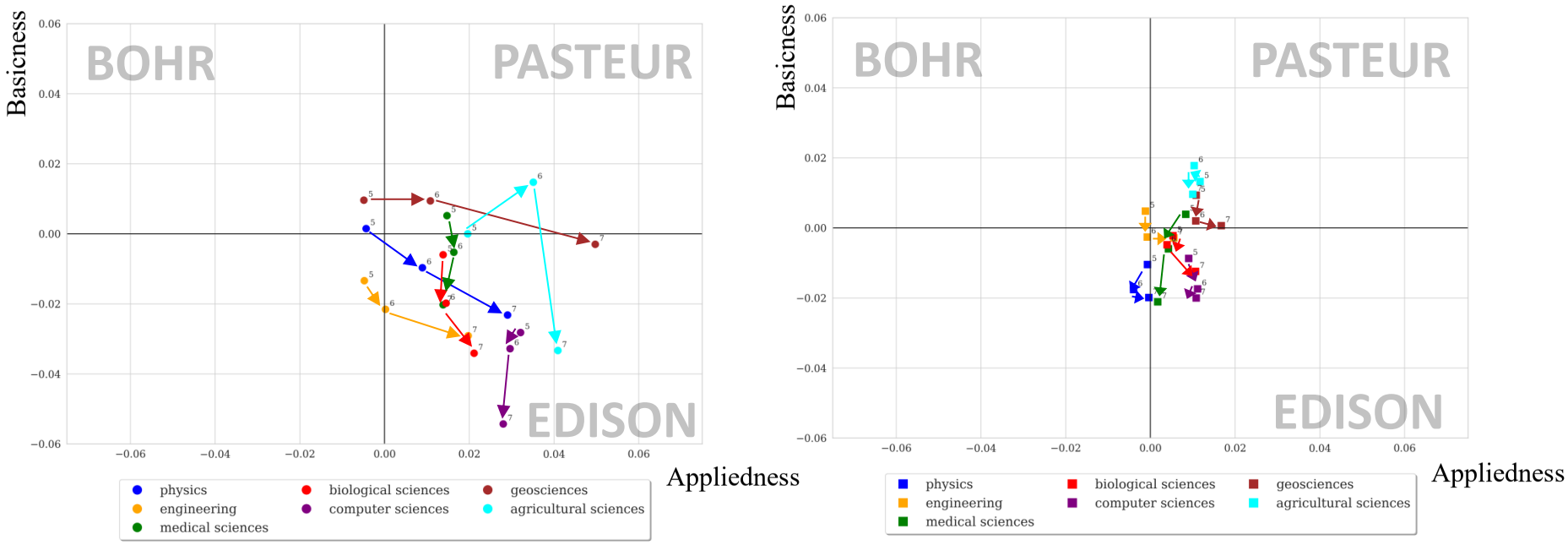
incumbents like AT&T, Xerox, Dupont, etc. These companies have reportedly been disengaging from science, however, young and science-intensive companies like those in the biotech sector may follow different R&D strategies, which may contribute to different trends in appliedness and basicness.

Figure 3.7: Basicness and appliedness interaction coefficients in the Pasteur's quadrant



Note: x-axis appliedness, y-basicness. The scatterplot represents the coefficients of the interaction terms of the models $Y_{\text{itj}} = \beta_o + \beta_1 corp_i + \beta_2 \sum_{t=1}^{7} \text{I(year} = \text{t)} + \beta_3 corp * \sum_{t=1}^{7} * \text{I(year} = \text{t)} + \mu_j + u_{itj}$ and $Y_{\text{itj}} = \beta_o + \beta_1 collab_i + \beta_2 \sum_{t=1}^{7} \text{I(year} = \text{t)} + \beta_3 collab * \sum_{t=1}^{7} * \text{I(year} = \text{t)} + \mu_j + u_{itj}$. The confidence intervals are not displayed. The direction of the arrow indicates the direction in which the coefficient of the interaction term is moving.

Figure 3.8: Basicness and appliedness interaction coefficients in the Pasteur's quadrant by field, corporate (left) and collaborations (right)



Note: Note: x-axis appliedness, y-basicness. The scatterplot represents the coefficients of the interaction terms of the models $Y_{\text{itj}} = \beta_o + \beta_1 corp_i + \beta_2 \sum_{t=1}^{7} \text{I(year} = \text{t)} + \beta_3 corp * \sum_{t=1}^{7} * \text{I(year} = \text{t)} + \mu_j + u_{itj}$ and $Y_{\text{itj}} = \beta_o + \beta_1 collab_i + \beta_2 \sum_{t=1}^{7} \text{I(year} = \text{t)} + \beta_3 collab * \sum_{t=1}^{7} * \text{I(year} = \text{t)} + \mu_j + u_{itj}$. Every colour represents a different broad scientific field. The confidence intervals are not displayed. The direction of the arrow indicates the direction in which the coefficient of the interaction term is moving. We report only the fields that exhibit statistically significant trends in appliedness or in basicness.

3.4.3 Firm age and size

Figure 3.9 explores differences in appliedness and basicness by firm age. We identify as young firms those with less than five years of life in each period, and as old firms all others. We estimate changes in appliedness and basicness with the following model:

$$Y_{\text{itj}} = \beta_o + \beta_1 firmage_i + \beta_2 \sum_{t=1}^{7} \text{I(year} = \text{t)} + \beta_3 firmage_i * \sum_{t=1}^{7} * \text{I(year} = \text{t)} + \mu_j + u_{itj} \quad (3.5)$$

where Y_{itj} is our outcome variable, either *appliedness* or *basicness* of publication *i*, in time period *t* and journal *j*. Time periods are measured as groups of five years, so *t* can take values from 1 (1980-1984) to 7 (2010-2014). $I(year = t)$ is a dummy variable equal to one when the time period is *t* and zero otherwise, with *t= 1* as the omitted category. $Firmage_i$ is a categorical variable equal to one if publication *i* belongs to a young firm, two if it belongs to an old firm, and zero if it belongs to universities. Journal fixed effects μ_j control for potential differences across publication layouts. As in equation (3.4) coefficients corresponding to β_2 capture the general time trends of the outcome variable, while those corresponding to β_3 capture the interaction between the time trend and the $firmage_i$ variable. If significant, they would signal us any divergence between young and old companies, compared to university publications.

Figure 3.9a shows the results with appliedness as dependent variable. The full regression table is available in Appendix Table C 7. The coefficients show similar patterns as the ones showed in Figure 3.4. For the first four time periods, from 1980 to 1999, the coefficients for young and old firms are stable and always close to zero, which is a parallel trend with university science, albeit for higher level of appliedness (the estimate for β_1 is equal to 0.0338 and significant for young firms, and equal to 0.0326 and significant for old firms). Starting from 2000, the coefficients become positive and increasing, both for young and old firms. The coefficients for young firms are higher but not statistically different from old firms. Figure 3.9b shows the results with basicness as dependent variable. All the coefficients of interest are not statistically significant for young firms. Old firms, instead exhibit a positive β_1(0.0112), and a downward β_3 trend from t=2. It is possible that this lack of statistical significance is due to the smaller sample size given that 6,435 old firms published 316,088 publications, while 2,633 young firms published only 6,697 publications.

To sum up, appliedness is increasing for both young and old firms, and there are not statistically significant differences between the two groups. For what concerns basicness, instead, only old firms experienced a decline in basicness, while young firms show non-significant coefficients.

Figure 3.10 repeats the same econometric exercise but differentiates firms by size. We estimate changes in appliedness and basicness as follows.

$$Y_{\mathrm{itj}} = \beta_o + \beta_1 firmsize_i + \beta_2 \sum_{t=1}^{15} \mathrm{I(year = t)} + \beta_3 firmsize_i * \sum_{t=1}^{15} * \mathrm{I(year = t)} + \mu_j + u_{itj} \quad (3.6)$$

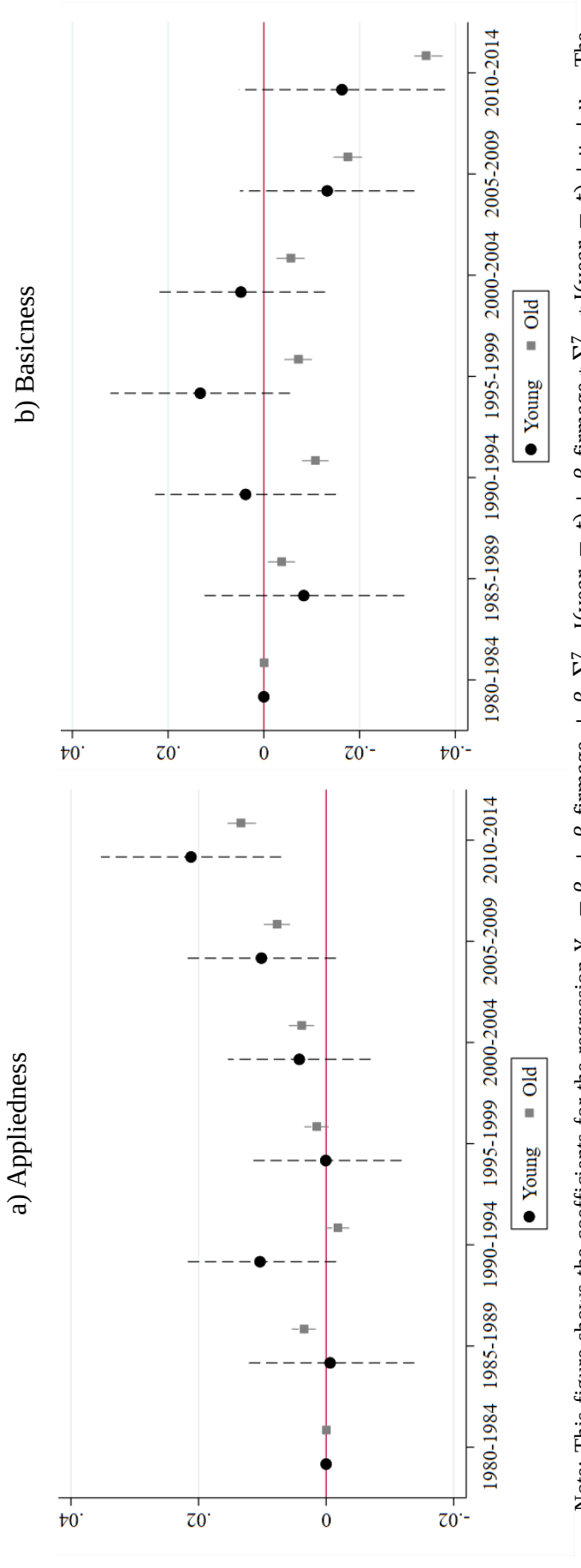
where all variables are identical to those in equation (3.5), except *firmsize*$_i$, which replaces *firmage*$_i$ and is a categorical variable equal to one if publication *i* belongs to a small firm, two if it belongs to a large firm, and zero if it belongs to universities. We classify firms as small if they have less than 150 employees and large if more than that. Given that it is not easy to find information on firm size from 1980 to 2014, we restrict the sample from 2000 to 2014 to limit missing values and improve the precision of our estimates. When exact data for time t are unavailable, we infer the size using the employees' number closest in time for t>0. Choosing financials for t<0 may lead to underestimating the size of firms that experienced growth. For example, if a firm published in 2005 and we have only financial information for 2000, we do not include that publication in the sample. Conversely, if financials are available in 2010, we do include it. We acknowledge that this is a stringent criterion, but for small firms, financial information is often available only for one year, and our aim is to maximize the number of observations for small firms.

Figure 3.10a shows the results with appliedness as dependent variable. The full regression table is available in Appendix Table C 8. Once again, the results are similar to the baseline one. When appliedness is the dependent variable, both small and large firms exhibit positive and statistically significant coefficients starting from 2006-2007. The β_1 are 0.0222 and 0.0310. However, the coefficients of the two groups are statistically different only in 2012 and 2014.

Figure 3.10b shows the results when basicness is the dependent variable. Small and large firms exhibit negative and statistically significant coefficients starting from 2005. The coefficients of the two groups are statistically different only in 2014.

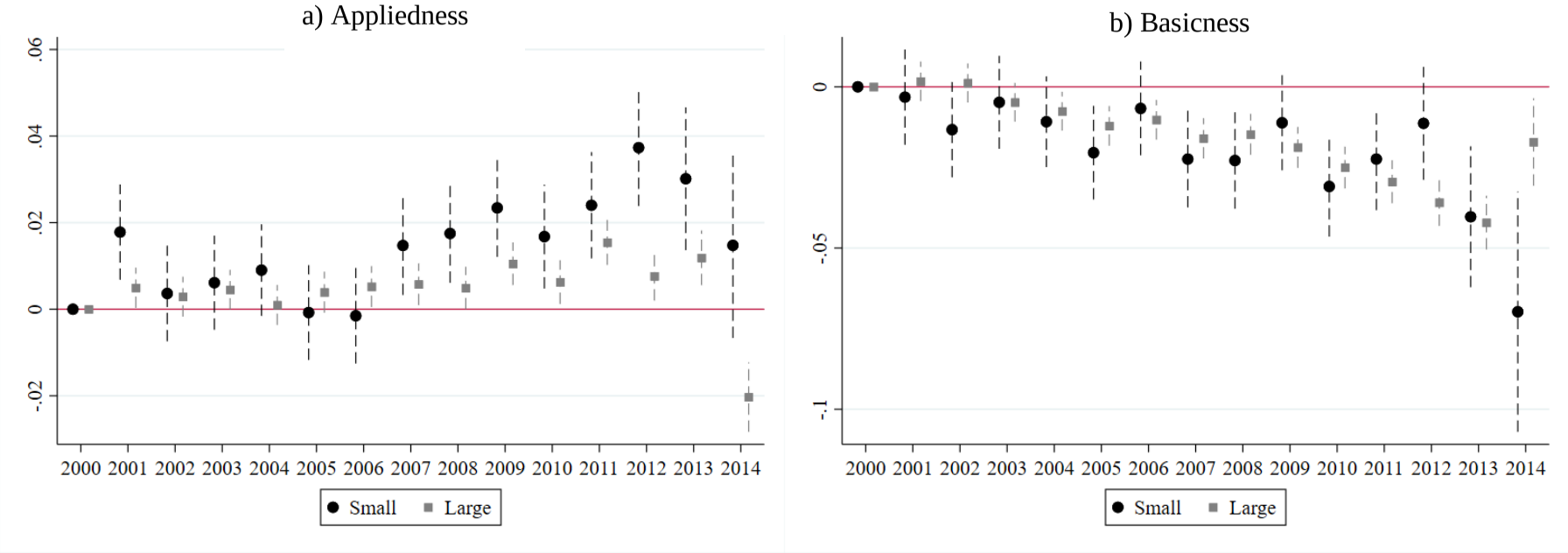
To conclude, the baseline evidence is confirmed as young and old, small and large firms show similar patterns.

Figure 3.9: Appliedness (left) and basicness (right) by firm age, 1980-2014



Note: This figure shows the coefficients for the regression $Y_{itj} = \beta_o + \beta_1 \text{firmage} + \beta_2 \sum_{t=1}^{7} I(\text{year} = t) + \beta_3 \text{firmage} * \sum_{t=1}^{7} I(\text{year} = t) + \mu_j + u_{itj}$. The dependent variable is appliedness in a) while basicness in b). The time trend is identified by 7 dummy variables for groups of 5 years. Confidence intervals are p<0.95. Circles represent the coefficients for young firms, while squares for old firms. The β_1 not reported in the figure are available in Appendix Table C 7 and are the following for young and old firms: appliedness 0.0338***and 0.0326***, basicness -0.0040 and 0.0112***.

Figure 3.10: Appliedness (left) and basicness (right) by firm size, 1980-2014



Note: This figure shows the coefficients for the regression $Y_{itj} = \beta_o + \beta_1 \text{firmsize} + \beta_2 \sum_{t=1}^{15} I(\text{year} = t) + \beta_3 \text{firmsize} * \sum_{t=1}^{15} * I(\text{year} = t) + \mu_j + u_{itj}$. The dependent variable is appliedness in figure a) while basicness in figure b). The time trend is identified by 15 yearly dummy variables. Confidence intervals are p<0.95. Circles represent the coefficients for small firms, while squares for large firms. The β_1 not reported in the figure are available in Appendix Table C 8 and are the following for small and large firms: appliedness 0.0222*** and 0.0310***, basicness 0.0036 and 0.0086***.

3.5 CONCLUSIONS

We have revisited the evidence on the decline of corporate science, in particular for what concerns its supposed shift away from basic research, with an original metric inspired by Stokes' (1997) taxonomy. In particular, we consider the basicness and appliedness of research as distinct and independent concepts, which require being measured by independent indicators. We also extend our analysis beyond the firms that have attracted most of the literature's attention, namely the large corporations with large industrial R&D labs, often occupying long-dating incumbent positions in their sectors of activity. In particular, we examine also entrant firms and firms of small size.

We provide some evidence of the difference in *appliedness* and *basicness* of these publications and especially of their changes over time. Our results are consistent with the narrative of the decline in corporate science (Arora, Belenzon, and Patacconi 2018) but more nuanced. First, we find that most corporate research can be characterised as belonging to Pasteur's quadrant, in which considerations of use coexist with the quest for a fundamental understanding of natural principles. This is especially true for fields related to the life sciences and computer engineering. The remainder of it is almost exclusively located in the Edison's quadrant, where considerations of use are dominant. We then document how, especially since 2000, corporate research has moved from Pasteur's quadrant in the direction of Edison's one due to a mix of increase in appliedness and decline in basicness. The scientific fields in which the increase in appliedness is more evident are agricultural sciences, biological sciences, computer science, geosciences, medical sciences, and physics. While biological sciences, computer sciences, engineering, and physics declined in basicness. In most fields the magnitude of the coefficients increases in absolute value over time. Similar patterns apply to collaborations; however, the coefficients are lower in magnitude. We find none when searching for significant differences between firms of different sizes or ages.

This chapter makes two main contributions to the literature. First and foremost, we apply to corporate science a metric that operationalises a critique of the linear model of innovation that had heavily influenced the early research on corporate science but has been rather neglected by the more recent one. Second, and more derivatively, we validate the use of corporate publications as a proxy indicator of research effort. This use has received criticism because scientific publications might be used for strategic purposes (Hicks, 1995) or simply consist of disclosure of information that firms may deem not relevant for their patent portfolio. Instead, we find that the majority of corporate publications both contribute to scientific advancement

(as measured by citation indicators) and are close to technology (cited by patents, either directly or indirectly).

This chapter comes with some limitations. First, our empirical strategy allows us to identify changes in corporate science and collaborations relative to university publications. However, both metrics rely on a network of backward citations and suffer from right truncation. As a result, we cannot draw any conclusions about the absolute value of appliedness and basicness. Second, measuring basic science is challenging. While our measure of basicness captures some of its aspects, it does not cover them all. Future research may address the issue by measuring basicness with other indicators. Last, in this chapter, we calculated the appliedness and basicness of scientific publications. However, the network of backward citations leading to the dual frontier can be equally exploited to calculate patents' appliedness and basicness.

Finally, validating our measures of appliedness and basicness is challenging. Although it may be tempting to directly compare our results to studies utilizing alternative metrics for distinguishing basic and applied science, such as those by Krieger et al. (2021), and Lim (2004), this approach faces conceptual limitations. We operate under the assumption that there is no inherent trade-off between basic and applied science. In contrast, many existing metrics view basic and applied science as substitutes, suggesting a linear model of innovation, which is incompatible with our conceptual framework.

Chapter 4: Corporate science and IPOs

Initial public offerings (IPOs) help companies raise capital, but the requirement to disclosure and pressure from the shareholders can affect the company's scientific strategies. I test the causal impact of going public on firms' scientific output, using data on the population of firms that undertook an IPO from 1996 to 2010 and had at least one publication or patent. My empirical strategy involves a treatment group of firms that successfully completed an IPO and a control group of firms that filed for an IPO but then withdrew their filing. Identification is achieved with a stacked difference-in-difference, instrumenting the decision to go public by the post IPO filing market returns. The results show a positive effect of IPOs on scientific output, measured as scientific publications and collaborations. Going public impacts not only innovation outcomes but also the underlying scientific research through the influx of new scientists and inventors joining the company and the capital raised at the IPO.

4.1 INTRODUCTION

IPOs are an important milestone in the life of science-intensive companies that seek growth. However, IPOs are often found to have a negative impact on innovation, both in terms of quantity and quality. Several mechanisms have been identified as contributing to this negative relationship. These include agency problems, short-termism, and disclosure requirements. Conversely, a smaller portion of the literature finds a positive effect, usually linked to the greater availability of funds after an IPO or the ability to attract individuals to the company.

Surprisingly, even if these reasons have been identified as contributing to the decline of companies' engagement in science and technology, to date, the only metric used to measure firm post-IPO R&D performance is patent count or patent quality (Aggarwal and Hsu 2014; Bernstein 2015; Gao, Hsu, and Li 2018; Markovitch, Steckel, and Yeung 2005; Moorman et al. 2012; Wies and Moorman 2015; Wu 2012). Patenting, however, is only a portion of companies' R&D, that measures technology and applied research, but neglects more long-term oriented research. To capture this aspect of firms' scientific involvement, I will measure firms' scientific engagement with scientific publications. Scientific publications, as patents, reflect the firms' R&D capabilities and investment in science and are equally affected by the decision to go public.

In this chapter, I focus on firms' IPO decisions and test the causal impact of going public on firms' propensity to invest in scientific research, measured as scientific publications. A more holistic view of companies' R&D has to consider publication output; otherwise, it would be measured only a partial effect of IPOs on firms' R&D activities. First, I test if an IPO affects the number of scientific publications (including collaborations with universities). Second, I test if an effect is observable also in terms of forward citations. Third, measuring applied and basic science as two separate indicators, I test if IPOs drive the direction of firms' scientific research.

My identification strategy relies on a stacked difference-in-differences (Baker, Larcker, and Wang 2022; Cengiz et al. 2019). This empirical specification suits well the IPO context, given that the firms in my sample are not treated at the same time, but in different periods from 1996 to 2010. I compare a control group of 183 firms that filed for an IPO but at last withdrew from it to a treatment group of 636 firms that successfully completed an IPO. I consider only firms that published at least one publication or patent in the 5 years before an IPO and exclude foreign and financial firms. Firms that undertake an IPO are extremely self-selected and likely to be successful business cases. Firms that showed willingness to undertake an IPO but then withdrew constitute a more comparable group of firms and more likely to have similar characteristics to the firms that completed an IPO (similar tech cycle, age, etc.).

To mitigate endogeneity concerns I instrument the decision to go public by the average market returns in the 2 months after the IPO filing following the empirical strategy of Bernstein (2015) and Larrain et al. (2021). Bad market returns in the post filing period are strongly correlated with the likelihood of withdrawing from an IPO but are not correlated with the companies' long-term science and innovation strategies.

My main empirical findings are two. First, I find a positive effect of IPOs on the firms' number of publications, including those in collaboration with universities. Second, I do not find any effect of IPOs on either publications' forward citations, appliedness or basicness.

Following the same empirical strategy, I support my empirical evidence looking at two potential mechanisms. First, shifts in corporate science might be driven by the change in workforce after the IPO. I test changes in research output of scientists who stayed in the company after the event, and I find negative results for the number of publications. I find no effect for collaborations, forward citations, appliedness or basicness. I find not significant results for the subset of scientists-inventors. Nevertheless, the number of unique scientists and scientists-inventors increases more in the firms that go public that those that remain private. Second, a positive effect might be caused by the increased capital after the IPO. However, I

find a negative, but not economically significant, relationship between the amount of capital raised at the IPO and the post-IPO outcomes.

To sum up my results indicate a positive effect of IPO on corporate science, suggesting that companies use the capital collected during an IPO to fuel post IPO scientific research. This effect is likely to be driven by the influx of new scientists after the IPO and not by an increase in productivity of incumbent ones.

The remainder of this chapter proceeds as follows. Section 4.2 presents the background literature on IPO and corporate science. Section 4.3 introduces the data, the variables of interest, and the empirical strategy. Section 4.4 presents the results, while Section 4.5 shows the potential mechanisms. Section 4.6 investigates whether the mechanisms are different for patents. Section 4.7 concludes.

4.2 THE RELATIONSHIP BETWEEN IPOS AND INNOVATION

In this section, I initially explore the mechanisms that impact innovation following an IPO. Subsequently, I explain the rationale behind incorporating scientific publications, in addition to patents, when assessing the outcomes related to science and innovation.

4.2.1 IPOs and innovation

The positive relationship between IPOs and innovation, is often attributed to the access to public funding and the capacity to attract new workforce.

The first mechanism revolves around the use of capital raised at the IPO by entrepreneurial firms, which is often directed to further pursuing internal R&D projects. Hall and Lerner (2010) argue that public equity is a better source of funding for innovative projects than debt because innovative firms' intangible assets are not a good collateral. Allen and Gale (1999), indicate that public equity markets, which allow investors with diversified opinions to participate, enable the financing of innovative projects with uncertain probabilities of success. Rajan (2012) finds that the ability to secure capital alters the innovative nature of firms. Easier access to equity capital makes firms more likely to conduct capital-intensive fundamental innovation. Acharya and Xu (2017), based on a set of 2,214 public and private firms, find that IPOs have a positive effect on innovation in industries depending on external finance (like equity), while they have no effect on those that rely more on internal funding (own profits and assets). This implies that IPOs are effective only for those companies that need external financing to fuel their innovation. Vismara (2014), shows that firms with different level of innovative activity decide to go public for different reasons. Those with higher R&D investments tend to invest

more after IPO, while firms with large patent portfolios invest less. This correlation suggests that larger patent portfolios are linked to higher technological maturity and more risk aversion. Atanassov et al. (2007), studying a panel of US companies from 1974-2000, find that doing an IPO or a Seasoned Equity Offering (SEO) increases the innovative activity for the subsequent two years.

The second mechanism is the capacity to attract new workforce. Bernstein (2015) reports that acquiring external innovation and attracting new inventors are mitigating factors of the negative impact of IPO on innovation. Similarly, Borisov et al. (2021) find that employment increases after an IPO, especially in industries with high skilled labour and high dependence on external finance. The positive effect is due to the relaxation of financial constraints, improved access to dept, and the ability to acquire external firms.

Besides the positive effects, most of the literature on IPOs focussed on how agency problems, short termism and disclosure requirements negatively affect innovation outcomes.

Agency problems have for long being investigated as a possible force acting against innovation, due to conflict and differences in views of managers and shareholders (Berle and Means 1932; Jensen and Meckling 1976). One first agency problem often goes under the name of the "*lazy manager hypothesis*", by which managers would resist putting the effort required for undertaking the innovative projects expected by the shareholders (Aghion, Van Reenen, and Zingales 2013; Bernstein 2015; Bertrand and Mullainathan 2003). A second problem has to do with the managers' *career concerns*. CEOs – especially if risk-averse – might decide, to follow more short-term strategies because they are concerned of losing their job. Even for pure stochastic reasons, bad outcomes might convince the shareholders that the CEO is not behaving well. This behaviour generates a natural aversion to innovation (Aghion, Van Reenen, and Zingales 2013; Bernstein 2015). Private ownership can alleviate agency problems. With tighter control and improved monitoring, it becomes possible to effectively manage the CEO's performance and encourage them to pursue more innovative strategies without concerns about their career (Aghion et al., 2013; Bernstein, 2015). Bernstein (2015), studying a sample 1,599 withdrawn US IPO between 1985 and 2003, finds that IPO negatively affects innovation quality due to agency problems. More specifically, career concerns cause an exodus and a productivity decline of inventors. However, this negative impact can be offset by attracting new inventors and acquiring external innovation. Wu (2012) supports a similar claim by looking at changes in productivity and citation behaviour of pre and post-IPO new hires. After

an IPO, firms explore less and exploit more. Nevertheless, explorative innovation search based on scientific knowledge increased, indicating that companies may augment their explorative efforts in certain key areas.

Short-termism refers to all managerial behaviours prioritising short-term performances instead of long-term strategies. Overall, the stock market reacts positively when firms move from value-creation activities, such as innovation and development, to more intensive product marketing or other value-appropriation activities (Mizik and Jacobson 2003). Asker, Farre-Mensa, and Ljungqvist (2015), studying a sample of 2,595 public firms and 1,476 private US firms over the period 2001–2011, find that short-termism makes managers of public firms behave less efficiently than those of private firms, both in terms of investment allocation and sensitivity to changes in investment opportunities. Bernstein (2015) highlights that when funding is not enough for simultaneous innovation and commercialisation, access to IPO money is mostly directed to commercialisation, especially in capital-intensive industries. This tendency is clear for biotech firms that having already spent heavily in R&D, are still years away from commercialisation and need funding for development (Deeds, Decarolis, and Coombs 1997). Wu (2012) analyses the characteristics of institutional investors in 205 US medical device companies that received venture capital and went public from 1980 to 2008. She found that one year after the IPO, only about a third of the institutional shareholders were long-term focused. After five years, the percentage of long-term focused shareholders dropped to one-fifth. The study also revealed that CEO ownership decreased, and only 28% of the founders remained as CEOs after the IPO. Wies and Moorman (2015), studying 40,000 US product introductions of consumer-packaged goods firms from 1980 to 2011, find a positive effect of going public on the number of innovations and innovation variety, while a negative effect on breakthrough innovations and new-to-the-firm[45] innovations. The results are attributed to the enhanced access to financial and strategic resources and the stock market incentives provided by going public. However, the negative impact can be mitigated for firms with a strong focus on appropriability. Markovitch et al. (2005) find that firms alter the risk profiles of innovation projects conditional on their prior period's industry-adjusted stock returns. Moorman et al. (2012) observe that stock market incentives drive firms to time their innovation strategies through a ratcheting strategy that sacrifices revenues in product markets

[45] Markets in which the firm does not have existing brands.

but reaps benefits in financial markets. Private ownership is also found to be more effective when firms are more aggressive in their innovation strategies, trying to explore new ideas and have a broader scope, while public ownership suits better more conservative strategies where companies exploit existing knowledge, in already explored technological classes. In both cases, the innovativeness of private firms derives from the different tolerance levels for failure, and different investment horizons (Ferreira, Manso, and Silva 2014; Gao, Hsu, and Li 2018).

Due to **regulatory requirements**, public firms must regularly disclose to shareholders the status of ongoing projects. Given the uncertain nature of innovative projects, managers would be more likely to choose less risky projects to report more tangible results (Ferreira, Manso, and Silva 2014). A similar effect might be generated by an increase in the analyst coverage that hinders long-term innovation projects, applying pressure on managers to meet short-term goals, at the cost of a decline in the quantity and quality of patents (He and Tian 2013). Aggarwal and Hsu (2014), for a sample of biotech firms, find evidence that the quality of innovation, measured by patent forward citations, is negatively impacted by analyst attention when the companies have many projects in the early stages of development. Bhattacharya and Ritter (1983), Maksimovic and Pichler (2001), Spiegel and Tookes (2008) show through theoretical models that disclosing information after going public hinders long-term strategies. Consequently, private ownership might be preferable when disclosing product or service innovation details might help competitors (Wies and Moorman 2015). This is particularly evident in sectors like biotechnology, where the disclosure of R&D details might seriously hinder the future of the company (Guo, Lev, and Zhou 2004; Wies and Moorman 2015).

4.2.2 IPOs and corporate science

Innovation is a complex process that, at least partially and for certain technologies, relies on scientific research. Such research can be conducted in-house by firms or outsourced to external entities such as universities or carried out through collaborations with these institutions. Moreover, research can vary in applicability, meaning it can have direct implications for innovation or be more fundamental, aiming to establish general natural laws rather than explaining specific phenomena. The existing literature, which primarily focuses on patents as a measure of innovation, has yet to extensively explore whether and to what extent IPOs influence firms' propensity to invest in scientific research or the nature and objectives of such research. To investigate these aspects, it is necessary to employ another indicator, such as firms' scientific publications.

Furthermore, scientific publications may provide insight into other various aspects of a firm's R&D strategies. First, publications can provide information about the commercialisation of a new product. Scientific publications can help companies achieve their commercialisation goals and improve their reputation. In the context of an IPO, companies may focus on commercialisation after raising capital. Scientific publications can suggest a superior quality of a product, generate publicity and interest, and accelerate the commercialisation process. This has been studied mainly in the pharmaceutical, biotech, and chemical industries, where publications in top journals can help convincing doctors and hospitals about the effectiveness of a drug. Scientific publications may also be part of the regulatory approval process (Hicks 1995; Polidoro and Theeke 2012; Simeth and Lhuillery 2015; Arora, Belenzon, and Sheer 2021a; Rotolo et al. 2022).

Second, scientific publications can also be used to support IP strategies as a form of defensive (or strategic) publishing. Publishing involves creating new prior art to extend patent races, reduce expected patent value, broaden patent scope, or prevent the privatization of inventions (Rotolo et al. 2022).

Third, publications may be part of a reward system to attract and retain new researchers. According to Merton (1973), scientists are naturally inclined towards scientific pursuits. When a company hires talented scientists, it may allow them some freedom to pursue their research interests since they value the freedom to pursue them and often require time to publish their work (Roach and Sauermann 2010; Stern 2004).

Last, scientific publications may also constitute a signal. Simeth and Cincera (2016) found that scientific publications are positively correlated with market value after controlling for R&D activity, patent stock, and patent quality. Some studies also find that a company's science base has a positive signalling effect on investors (Colombo, Meoli, and Vismara 2019; Darby and Zucker 2002; Fukugawa 2022).

If firms publish as a propaedeutical activity to patenting, one could expect similar dynamics, and that any changes in scientific publications would mirror changes in patenting. However, it is also possible for the relationship to go in the opposite direction. Many studies showed that patenting and publishing could follow different incentives and not always go hand-in-hand. Bhaskarabhatla and Hegde's (2014) study on IBM, for example, showed that a shift in intellectual property management in 1989 caused a threefold increase in patents and a decline in publications.

Moreover, the underlying mechanism may be similar for publications and patents. It is known that IPOs allow firms to attract new inventors who may produce higher quality patents than those already in the company. Since the R&D department is responsible for both patents and publications, and inventors themselves may also contribute to publications, changes in inventor mobility can impact both publication and patent outcomes.

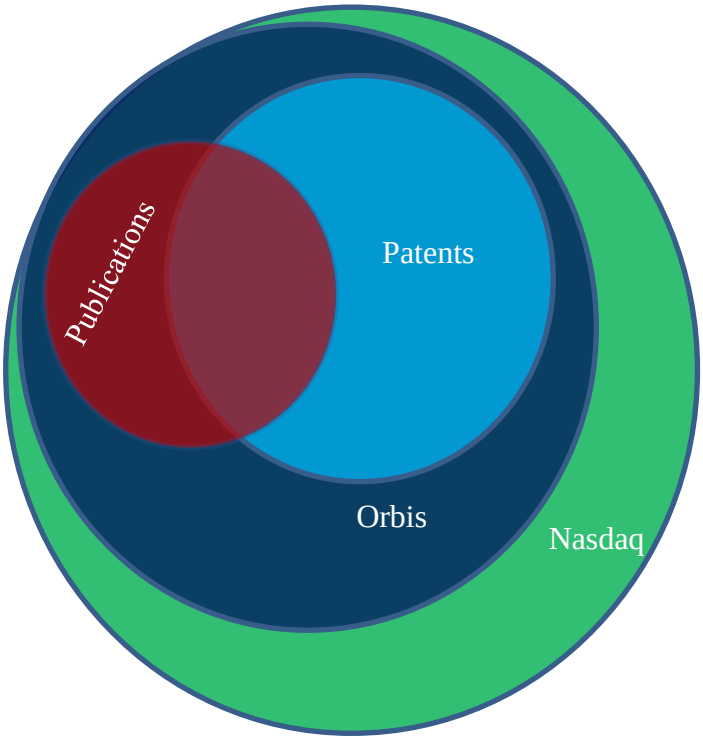
4.3 DATA AND METHODS

Data are collected from five main sources:

- *IPO deals from NASDAQ*. I collected data on 10,336 US deals from 1996 to 2014 among completed and withdrawn IPOs from NASDAQ, NYSE and OTCBB. Financial companies and foreign firms are excluded. NASDAQ reports the data in the public database US SEC EDGAR. Afterwards I select only the 819 companies with at least one publication or patent in the 5 years before the IPO.
- *Publications from Web of Science*. I identify corporate publications matching WoS authors' affiliations to Orbis company names, using a decision tree algorithm incorporating string similarity scores (Levenshtein distance), presence of shared non-dictionary words as well as address information (same city or zip code). Further details about the matching algorithm are presented in Appendix A and Chapter 1. I obtain a set of 71,970 publications and conference proceedings from 1996 to 2010 belonging to 1,372 firms.
- *Financial information from Orbis*. IPO firms are subsequently matched with Orbis company information using the same algorithm presented previously. I retrieve incorporation date, financial information at the time of the IPO filing, and industry.
- *Patents from PATSTAT and Orbis*. I combine patent data from PATSTAT and Orbis in order to obtain the patent portfolio of all the IPO companies. I complement this set of patents matching company names to applicant names on PATSTAT. I obtain a set of 371,278 patents from 1996 to 2010 belonging to 2,142 firms.
- *Stock market information*. I retrieve monthly NASDAQ returns from 1996 to 2010 from Yahoo finance.

The final panel dataset includes 819 IPOs (636 completed, 183 withdrawn) from 1996 to 2010 of firms that published at least one publication or patent before the IPO filing. The sample is available until 2010, allowing for a five-year period following the IPO. The sample contains 8,964 firm-year observations. Figure 4.1 presents a schematic representation of the data.

Figure 4.1: Overview of the data sources. IPOs information, Orbis data, patents, and publications



Note:10,366 IPOs are listed on the NASDAQ website and EDGAR. It is the largest set of firms, and it is likely the universe of US IPO firms from 1996 to 2014. Not every firm that files to go public is in Orbis. I could retrieve the Orbis identifier for 7,378 firms. Among those firms 819 have at least one patent or one publication before the IPO filing.

Despite my main interest laying with publications, I collect data on patents, in addition to scientific publications, for two reasons. First, patents allow me to calculate an indicator of appliedness of scientific publications , based on Ahmadpoor and Jones (2017) and analogous to the one used in Chapter 3. If a publication is directly cited by a patent, it constitutes a direct application of science in technology. Starting from this frontier I calculate the distance of each publications using the network of backward citations. If a publication is directly cited by a patent, it will be at distance D=1. If a publication is cited by a publication at the frontier, it will be at distance D=2. To generalise, a publication cited by other publications along the chain leading to the science-technology frontier will have a distance of D=k from the frontier, where k represents the number of publications in the chain (see Chapter 3.3 for further details). The

measure of appliedness is the inverse distance of a publication from the science-technology frontier. Second, using patents I can replicate the previous studies on IPO and innovation (Bernstein 2015; Larrain et al. 2021) in order to test the consistency of my findings with the existing literature.

The main variables used in the empirical analysis are the following:

IPO: A dummy variable equal to 1 for treated firms in the post IPO periods. Zero otherwise.

 # Publications: Number of articles or conference proceedings. A publication is assigned to a company if at least one of its authors is affiliated to it. It includes collaborations.

 # Collaborations: Number of publications in collaboration with universities. I measure collaboration as co-authored publications that list at least one corporate affiliation and university affiliation.

Pub citations: Citations received by publications in the three years following publication. I normalise the citation received diving the number of citations by the expected number of citations received by a document of the same type, in the same year, and in the same subject (WoS Science Category). When a publication is assigned to multiple subjects, I take an average of the ratios for each subject.

Appliedness: The inverse distance from the science-technology frontier measured as in Ahmadpoor and Jones (2017). See Section 3.3.1. for further details.

Basicness: I consider basicness as complementary to appliedness, implying no trade-off between the two measures. Basicness is an adaptation of Trajtenberg et al. (1997) metric, consisting in a Rao Sterling index (that captures diversity) of the publications citing the focal publication. Diversity of references is a characteristic usually associated with basic science. The metric is analogous to the one presented in Section 3.3.2.

Table 4.1 shows the summary statistics of the two groups for the whole sample, while presents the summary statistics of the two groups for the five years leading to the IPO. The first part of the table shows the independent variables summary statistics The final column indicates whether there are any statistically significant differences in the means of the two groups for each variable. The second part of the table shows the financials at the time of the IPO. I extract the financial information from the NASDAQ website and directly from the R&D prospectuses.

Table 4.1: Summary statistics for the five years preceding an IPO filing and at the time of IPO, completed and withdrawn IPOs

Corporate science measures	Completed			Withdrawn			
	count	mean	std	count	mean	std	diff
# Publications	2475	2.334	14.729	490	1.102	4.316	1.232*
# Collaborations	901	3.248	13.635	151	1.762	4.335	1.486
Pub Citations	901	1.842	6.862	151	1.668	2.381	0.174
Appliedness	693	-0.649	0.694	111	-0.609	0.680	-0.041
Basicness	618	0.667	0.203	98	0.682	0.212	-0.015
# Patents	3180.	3.756	24.260	915	3.121	7.366	0.634
Patent Citations	1124	1.026	1.162	338	1.119	1.303	-0.093

	Completed			Withdrawn			
Financials at the IPO	count	mean	std	count	mean	std	diff
Employees	417	2786.8	19661.7	102	474.4	1190.5	2312.4
Revenue ($mm)	389	306.8	1610.1	85	80.3	204.9	226.4
Stockholders Equity ($mm)	414	191.1	954.7	100	72.7	248.2	118.4
Total Assets ($mm)	414	529.0	2761.9	100	181.9	570.2	347.1
Total Liabilities ($mm)	414	337.3	1552.9	97	124.8	338.2	212.6
Net Income ($mm)	414	21.1	77.9	100	11.4	20.1	9.7

Note: This table shows the descriptive statistics for the five years before the IPO from 1990 to 2010 as well as the financials at the time of the IPO. The first part of the table displays figures from a panel of firm-year observations. The second part displays the financials in the year of the IPO. The last column shows a t-test to test if the averages of the two groups are not statistically different. Confidence intervals are *** p<0.01, ** p<0.05, * p<0.1.

To mitigate selection bias I follow the empirical strategy proposed by Bernstein (2015) and Larrain et al. (2021) and I compare a treatment group of firms that completed an IPO with a control group of firms that filed for an IPO but later withdrew. I estimate the regression using the following equation:

$$Y_{it} = \mu_i + \theta_t + \alpha \mathrm{IPO}_{it} + \delta_{cy} + \partial_{qt} + u_{it} \quad (4.1)$$

where *i* represents the firm, *t* represents time, *c* represents the IPO cohort, *y* the calendar year, and *q* the quarter of the IPO. The independent variables include the *number of publications*, *collaborations*, *paper forward citations*, *paper appliedness* and *basicness*. The *IPO* variable is the difference-in-difference estimator, and equal to 1 for the treatment group in the post treatment periods. The firm fixed effect is represented by μ_i, the time fixed effect is represented by θ_t, Treatment is irreversible, meaning that once a unit is treated, it will remain treated indefinitely, even in future periods.

I stack all observations centring at the time of the IPO and consider five years before and after treatment to ensure that the panel is balanced, and the two groups are comparable. The IPO year is considered as the first treatment period. The panel is balanced in event-time but not in calendar-time. Next, I estimate a Two Way Fixed Effects (TWFE) model controlling for cohort-by-calendar year fixed effects δ_{cy}. Firm-by-cohort fixed effects are not necessary because firms do not belong to multiple cohorts. As in Larrain et al. (2021) I control also for the quarter in which the IPO is completed, with a fixed effect that captures the interaction between a quarter dummy and a post treatment dummy (∂_{qt}).[46]

The stacked difference-in-difference allows to overcome the usual problems of TWFE regressions, such as differential timing, and heterogeneity in treatment. This method can be implemented with either a static or dynamic specification and employs an estimator calculated using Ordinary Least Squares (OLS) or Pseudo Poisson Maximum Likelihood (PPML). Essentially, the stacked difference-in-difference method involves estimating separate regressions for each cohort and then weighting the estimates across them. It is important to note that the parallel trends assumption must hold within each stacked cohort.

I exclude firms that were acquired within three years after their IPO to eliminate the influence of factors other than the IPO on the firms' publishing activity.

I consider the entire population of firms that have undergone an IPO and have at least one publication or patent in the five years preceding the IPO. I chose this sampling because I am interested in studying changes in publications behaviour of firms that were already active

[46] Larrain et al. (2021) accounted for the calendar-month effect during the post-IPO-decision period. The rationale behind this control variable is that firms going public in January may have a different impact on innovation compared to those going public, for instance, in November. Similarly, I grouped the IPOs in quarters, and I controlled for quarter effect for the post-IPO-decision period ∂_{qt}.

before the IPO, rather than those that become innovative after the IPO. I acknowledge that both treated and untreated firms exhibit some anticipation behaviour, using publications and patents to signal their value before the IPO. However, the timing of this behaviour is known, and my model accounts for it.[47]

To further treat endogeneity issues, I implement an instrumental variable approach as in Bernstein (2015) and Larrain et al. (2021). The decision of withdrawing from an IPO is endogenous. However, the decision is highly influenced by the average stock market returns around the IPO. If the conditions are not ideal for an IPO, the firm can decide to withdraw its decision and postpone the IPO when market conditions will be more favourable.[48] Firms decide to withdraw because it is costly to wait for the natural expiration of the IPO filing after 270 days of the last amendment of the IPO filing given that in the meantime they cannot disclose new information to investors or banks, or issue private placement (Bernstein 2015; Lerner 1994). Furthermore, good prior returns reflect the sentiment of the investors (Cornelli, Goldreich, and Ljungqvist 2006; Derrien 2005; Larrain et al. 2021) and increase the likelihood of a successful IPO, given that the surplus gained from going public is positively associated with the value of comparable firms (Edelen and Kadlec 2005).

For these reasons, I instrument the IPO decision with the average market returns in the two months post IPO filing. The first stage is calculated as follows.

$$IPO \quad = market\ returns * post_{it} + \mu_i + \theta_t + \delta_{cy} + \partial_{qt} + u_{it} \tag{4.2}$$

The stock returns are interacted with *post* because the returns during the post-filing period have an impact only on the outcomes that follow the filing and not on those that occurred before. The exclusion restriction holds if the sole way in which the market returns before an IPO impact firms' publication behaviour is through the IPO decision. The relevance of the instrument can be seen in the first stages in Table 4.4.

[47] The results should predominantly be seen as a local effect around the IPO threshold disclosure strategy, and not generalizable for the entire population of (public) firms.

[48] For further explanation on why firms decide to withdraw their filing see Bernstein (2015) and Larrain et al. (2021).

4.4 RESULTS

Table 4.2 shows the Average Treatment effect on the Treated (ATT)[49] of IPO on the number of publications and collaborations. The first and third columns present the difference-in-difference estimates obtained via OLS, while the second and fourth columns via PPML, that is the most used specification when dealing with count dependent variables. The coefficients are positive and statistically significant. Going public leads to a 43% increase in publications and 48% in collaborations.

Table 4.2: ATT of IPO on number of publications and collaborations

	# Publications		# Collaborations	
	(1)	(2)	(3)	(4)
	OLS	PPML	OLS	PPML
IPO	0.913***	0.356**	0.540**	0.390**
	(0.271)	(0.141)	(0.233)	(0.189)
Observations	8,964	5,852	8,964	4,804
R-squared	0.861	0.790	0.857	0.760
Firm FE	Yes	Yes	Yes	Yes
Relative Year FE	Yes	Yes	Yes	Yes
Cohort*Time FE	Yes	Yes	Yes	Yes
Quarter*Post FE	Yes	Yes	Yes	Yes

Note: This table presents the difference-in-difference results of the impact of IPO on the number of publications and collaborations. The sample extends until 2010 in order to provide a five-year post-IPO observation period for the outcome variable. The sample is conditional on having at least one publication or patent in the 5 years before the IPO. Standard errors in parentheses are clustered at the firm level. In the PPML model the pseudo R-squared is displayed. Confidence intervals are *** p<0.01, ** p<0.05, * p<0.1.

Figure 4.2 shows the results of an event study regression estimated using the following formula:

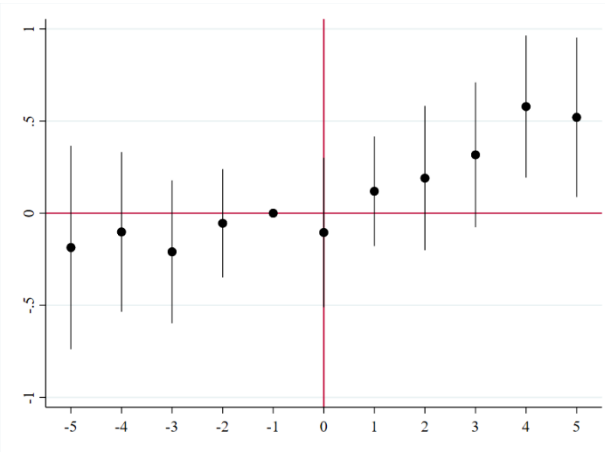
[49] The Average Treatment effect on the Treated (ATT) is the population mean treatment effect for the units assigned to treatment (Cunningham 2021). Formally the ATT corresponds to $ATT = E[Y_i^1|D_i = 1] - E[Y_i^0|D_i = 1]$. $Y_i^1$and Y_i^1 and $Y_i^0$are the two potential outcomes for individual *i*, in case of treatment and non-treatment. The two potential outcomes are conditional on $D_i = 1$ if unit *i* is actually treated, and $D_i = 0$ otherwise. The real value of the ATT is not possible to know because it requires to know the two potential outcomes of unit *i*, however, it is possible to estimate it imposing some conditions. In the case of a difference-in-difference specification the assumptions are that treatment is independent of potential outcomes, the two groups have parallel trends in outcome, stable unit treatment value assumption (the treatment is homogeneous across units, no treatment externalities) and that the units outcomes are parallel in absence of treatment (Cunningham 2021)

$$Y_{it} = \mu_i + \theta_t + \sum_{\tau=-5}^{-2} \alpha_\tau IPO_{i\tau} + \sum_{\tau=0}^{5} \beta_\tau IPO_{i\tau} + \delta_{cy} + \partial_{qt} + u_{it} \quad (4.3)$$

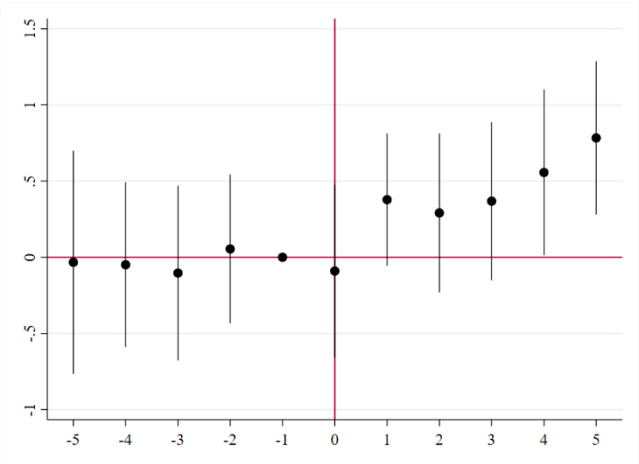
Year minus one is the omitted category. The pre-IPO period goes from year -5 to year -1, while the post period from year 0 to year 5. The effect becomes statistically significant starting from time period t+4, with a 95% confidence interval.

Figure 4.2: IPO impact on publications and collaborations, event study

a) # Publications



b) # Collaborations



Note: Same sample of Table 4.2. Figure a and b extend until 2010 in order to provide a five-year post-IPO observation period for the outcome variable. Conditional on having at least one publication or patent in the 5 years before the IPO. Calendar year dummies are included. The coefficients are obtained from the regression $Y_{it} = \mu_i + \theta_t + \sum_{\tau=-5}^{-1} \alpha_\tau IPO_{i\tau} + \sum_{\tau=1}^{5} \beta_\tau IPO_{i\tau} + \delta_{cy} + \partial_{qt} + u_{it}$ via PPML. Standard errors in parentheses are clustered at the firm level. Confidence intervals are p<0.05.

Table 4.3 shows the ATT of IPO on publication forward citations, appliedness and basicness. Publication forward citations, appliedness and basicness coefficients are negative but statistically insignificant. In an event study regression, all coefficients have the same sign but are not significant (Appendix Figure D 1).

Table 4.3: ATT of IPO on patent impact, publication citations, appliedness, and basicness

VARIABLES	(1) Pub citations	(2) Appliedness	(3) Basicness
IPO	-0.159	-0.0579	-0.0182
	(0.438)	(0.107)	(0.0360)
Observations	2,590	2,009	1,339
R-squared	0.244	0.482	0.395
Firm FE	Yes	Yes	Yes
Relative Year FE	Yes	Yes	Yes
Cohort*Time FE	Yes	Yes	Yes
Quarter*Post FE	Yes	Yes	Yes

Note: This table presents the OLS results for the impact of IPO on publication forward citations, appliedness and basicness. The sample extends until 2010 in order to provide a five-year post-IPO observation period for the outcome variable. The sample is conditional on having at least one publication or patent in the 5 years before the IPO. Standard errors in parentheses are clustered at the firm level. Confidence intervals are *** p<0.01, ** p<0.05, * p<0.1.

Table 4.4 introduces the instrumental variable specification for the number of publications and collaborations. Columns 1 and 3 show the first stages, which show that the market returns in the two months following the IPO filing have a positive and significant impact on the decision to go public. The magnitude of the coefficient and the high F-statistics suggest that the relevance condition is satisfied. Columns 2 and 4 show the second stage results. All coefficients are positive and statistically significant, consistent with the OLS specification. however, the coefficients are greater than those obtained through OLS, suggesting that the OLS coefficients represent the lower bound of the causal effect of IPO on corporate science.

Table 4.4: Instrumental variable regression on publication and collaboration number.

	(1) First Stage	(2) # Publications	(3) First Stage	(4) # Collaborations
IPO		10.35***		7.701***
		(2.606)		(1.683)
Market returns	0.413***		0.413***	
	(0.036)		(0.036)	
	128.62		128.62	
Observations	8,210	8,210	8,210	8,210
Firm FE	Yes	Yes	Yes	Yes
Relative Year FE	Yes	Yes	Yes	Yes
Cohort*Time FE	Yes	Yes	Yes	Yes
Quarter*Post FE	Yes	Yes	Yes	Yes

Note: This table presents the results for the instrumental variable specification. Columns 1 and 3 present the first stages. Columns 2 and 4 present the 2SLS estimations. The sample extends until 2010 in order to provide a five-year post-IPO observation period for the outcome variable. The sample is conditional on having at least one publication or patent in the 5 years before the IPO. Standard errors in parentheses are clustered at the firm level. Confidence intervals are *** p<0.01, ** p<0.05, * p<0.1.

Table 4.5 introduces the instrumental variable specification for publication citations, appliedness and basicness. Columns 1, 3 and 5 show the first stages, which again show that the market returns in the two months following the IPO filing have a positive and significant impact on the decision to go public. The magnitude of the coefficient and the high F-statistics is lower but still statistically significant and relevant. Columns 2,4, and 6 show the second stage results. All coefficients are not statistically significant.

Appendix Table D 1 and D 2 show the results suing as alternative instrument the average market returns only in the company's book buillding phase and not for the two months post IPO filing. The results are in line with the baseline results.

Table 4.5: Instrumental variable regression on publication forward citations, appliedness and basicness

	(1) First Stage	(2) Pub cit	(3) First Stage	(4) Appliedness	(5) First Stage	(6) Basicness
IPO		-7.401		-1.031		0.388
		(5.836)		(0.734)		(0.253)
Market returns	0.287***		0.324***		0.316***	
	(0.055)		(0.061)		(0.063)	
F statistic	26.87		28.58		24.90	
Observations	2,616	2,616	2,029	2,029	1,889	1,889
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Relative Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort*Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Quarter*Post FE	Yes	Yes	Yes	Yes	Yes	Yes

Note: This table presents the results for the instrumental variable specification. Columns 1, 3 and 5 present the first stages. Columns 2, 4 and 6 present the 2SLS estimations. The sample extends until 2010 in order to provide a five-year post-IPO observation period for the outcome variable. The sample is conditional on having at least one publication or patent in the 5 years before the IPO. Standard errors in parentheses are clustered at the firm level. Confidence intervals are *** p<0.01, ** p<0.05, * p<0.1.

To sum up, the main results highlight that IPOs have a positive impact on the number of publications and collaborations. No effect is found on the number of papers forward citations, basicness, or appliedness. In the next section I will explore some of the potential mechanisms that may explain those results.

4.5 MECHANISMS

In this section I explore the mechanisms that may be behind the positive impact of IPO on corporate science. The increase of publications and collaborations may be due to higher productivity of the scientists already working for the company (intensive margin), or due to an expansion of the lab that results in new scientists publishing (extensive margin). Furthermore, additional capital raised through an IPO may be correlated with positive outcomes in science and innovation.

First, I look at the intensive margin testing for changes in incumbent scientists' publication numbers, forward citations, appliedness and basicness. Second, I test whether the number of new scientists and inventors increases after the IPO, hinting at potential turnover. Last, I test

whether additional capital raised through an IPO is correlated with positive outcomes in science and innovation.

4.5.1 Intensive and extensive margin

In this section I test the impact of IPO at the scientist level. As noted by Bernstein (2015) an IPO may lead to inventors leaving a company, causing a decline in patent quality. To explore this possibility, I disambiguate scientists and scientist-inventors using a string similarity match on scientists' names before and after the IPO. I define as unique individuals those scientists that appear in the same company before and after the IPO. The margin of error is minimal because it is highly unlikely to find two individuals with the same first and last name in the same company. Using this approach, I can only identify the scientists that stay in the same company and not those who move. The final sample consists of 6699 scientists, 4,939 inventors and 715 scientists-inventors from 274 IPO firms and 67 withdrawn firms.

Testing the impact of IPO on individuals' productivity raises fewer concerns about endogeneity of treatment because it can be assumed that the IPO decision is almost exogenous for the scientist/inventor, as they are unlikely to have a significant influence on the decision to go public. The regression is estimated as follows:

$$Y_{it} = \mu_i + \theta_t + \alpha \mathrm{IPO}_{it} + \delta_{cy} + \partial_{qt} + \gamma_j + u_{it} \quad (4.4)$$

where *i* represents the inventor, *t* represents time, *c* represents the IPO cohort, *y* the calendar year, *q* the quarter of the IPO, and *j* the firm. The independent variables include the *number of publications*, *collaborations*, *publications forward citations* and paper *appliedness*. The *IPO* variable is the difference-in-difference estimator, and equal to 1 for the treatment group in the post treatment periods. The model is estimated as equation (4.1) with stacked observations centred at the time of the IPO. I consider five years before and after treatment to ensure that the panel is balanced, and the two groups are comparable. The coefficient of interest is estimated via TWFE with individual fixed effects μ_i, time fixed effects θ_t, and cohort-by-calendar year fixed effects δ_{cy}. Furthermore, I control for firm fixed effects (γ_j) and quarter*post FE (∂_{qt}).

Table 4.6 shows the impact of IPO on the productivity of scientists and scientists-inventors. Columns 1 and 3 show the OLS results for the number of publications and collaborations for the group of scientists. Columns 2 and 4 show the 2SLS results. Columns 5 and 7 show the OLS results for the number of publications and collaborations for the group of scientist-inventors. Columns 6 and 8 show the 2SLS results. I find statistically significant coefficients

only in columns 1 and 2 for the effect on the number of publications. The coefficient is positive in the OLS specification, but negative in the IV one. Going public reduces the number of papers published by a scientist by 0.792. All the other coefficients are not statistically significant.

Table 4.7 shows the impact of IPO on publication citations and appliedness of scientists and scientists-inventors. Columns 1 and 3 show the OLS results for publication citations and appliedness for the group of scientists. Columns 2 and 4 show the 2SLS results. Columns 5 and 7 show the OLS results for the number of publication citations and appliedness for the group of scientists-inventors. Columns 6 and 8 show the 2SLS results. The only statistically significant coefficient is the one in the OLS specification in column 1. However, the coefficient becomes negative and not statistically significant in the IV specification in column 2.

Table 4.8 shows the impact of IPO on the number of new scientists publishing after the IPO. New scientists are measured as individuals who have not previously appeared on publications and patents before the IPO. Both columns 1 and 2 show positive and statistically significant coefficients. In other words, firms that go public have 0.44 more new scientists and 0.78 more new scientists-inventors than firms that did not. Despite a smaller sample size, the coefficient for scientists-inventors is still positive and statistically significant.

At least in quantitative terms, IPOs have a negative impact on scientists' productivity. However, considering that the impact of IPO on publications at the firm level is positive, it is reasonable to assume that the effect is driven by the attraction of new scientists and inventors to join the firms' labs.

Table 4.6: Impact of IPO on publications and collaborations at the scientist level

	Scientists				Scientists-inventors			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	# Publications OLS	# Publications 2SLS	# Collaborations 2SLS	# Collaborations 2SLS	# Publications OLS	# Publications 2SLS	# Collaborations 2SLS	# Collaborations 2SLS
IPO	0.0474**	-0.792**	-0.118	4.774	0.0422	-0.542	0.275	7.232
	(0.0237)	(0.344)	(0.179)	(9.418)	(0.103)	(0.367)	(0.402)	(11.11)
Observations	33,233	33,211	4,091	4,091	5,258	5,258	1,302	1,302
R-squared	0.355		0.598		0.416		0.640	
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Relative Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort*Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Quarter*Post FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: Scientists are the individuals who published at least one publication in the five years before an IPO. Scientists-inventors are the individuals that published at least one publication and patent in the five years before the IPO. The sample extends from 1996 to 2010. Standard errors in parentheses are clustered at the firm level.

Table 4.7: Impact of IPO on publication impact and appliedness at the scientist level

	Scientists				Scientists-inventors			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Paper cit OLS	Paper cit 2SLS	Appliedness OLS	Appliedness 2SLS	Paper cit OLS	Paper cit 2SLS	Appliedness OLS	Appliedness 2SLS
IPO	1.461**	-1.953	0.0716	0.855	1.116	0.244	0.0433	1.027
	(0.722)	(15.84)	(0.0730)	(1.038)	(1.100)	(14.81)	(0.126)	(2.347)
Observations	4,091	4,091	2,658	2,658	1,302	1,302	862	862
R-squared	0.488		0.674		0.492		0.679	
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Relative Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort*Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Quarter*Post FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: Scientists are the individuals who published at least one publication in the five years before an IPO. Scientists-inventors are the individuals that published at least one publications and patent in the five years before the IPO. The sample extends until 2010 from 1996 to 2010. Standard errors in parentheses are clustered at the firm level.

Table 4.8: Impact of IPO on the number of unique scientists and scientists-inventors

	(1) # unique scientists	(2) # unique scientists-inventors
IPO	0.437*	0.784***
	(0.253)	(0.291)
Observations	612	204
Firm FE	Yes	Yes
Relative Year FE	Yes	Yes

Note: This table presents the difference-in-difference results of the impact of IPO on the number of unique scientists and scientists-inventors. Columns 1 and 2 are restricted to 2010 and to firms with at least one scientist before the IPO. Standard errors in parentheses are clustered at the firm level.

4.5.2 Additional capital through the IPO

Another plausible mechanism is that the capital raised at the IPO serves to finance more science and innovation. To test this, I collect information on the capital collected at the IPO. I estimate the capital raised at the IPO multiplying the number of shares issued by the average price per share paid for them. Firms that withdrew are assumed to have capital raised set to zero. I this way I can test how different intensity in treatment affects scientific and innovation output. Recent literature on difference-in-difference with continuous treatment (Callaway, Goodman-Bacon, and Sant'Anna 2021) points out that the assumptions have to be stronger than the difference-in-difference with binary treatment. In addition to the parallel trends assumption, it has to be assumed also homogeneous causal responses across groups. In this specification, however I do not aim at testing the causal impact, but only to give additional support to the positive impact of IPO on corporate science found in the previous section.

The models are estimated as follows:

$$Y_{it} = \mu_i + \theta_t + \alpha Capital\ raised_{it} + \delta_{cy} + \partial_{qt} + u_{it} \quad (4.5)$$

where Y_{it} is the number of publications or patents, i represents the firm, t represents time, c represents the IPO cohort, y the calendar year, and q the quarter of the IPO. $Capital\ raised_{it}$ is the capital raised at the IPO. $Capital\ raised_{it}$ is built as a difference-in-difference indicator, so equal to zero for the treated firms in the pre ipo period and equal to the capital raised afterwards. It is always set to 0 for the untreated firms. μ_i and θ_t are firm and year and fixed effects, while δ_{cy} and ∂_{qt} the cohort*calendar year FE and quarter*post FE.

Table 4.9 shows the regression results. All columns present a negative and statistically significant coefficients. However, the coefficients are close to zero. The capital raised at the IPO is calculated in millions, so an increase in one million in capital raised would lead to a 0.00005 increase in publications and 0.0001 in patents. Therefore, these results show that the amount of money raised is not relevant for the post IPO performance. In other words, companies do not invest more in corporate science and innovation the higher the capital raised at the IPO.

Table 4.9: Cash raised and number of publications, patents and R&D expenses

	(1) # Publications PPML	(2) # Patents PPML
Capital raised	-4.92e-05**	-9.98e-05***
	(2.31e-05)	(3.17e-05)
Observations	5,578	9,455
R-squared		
Firm FE	Yes	Yes
Relative Year FE	Yes	Yes
Cohort*Time FE	Yes	Yes
Quarter*Post FE	Yes	Yes

Note: This table shows the regression results of capital raised at the IPO on the number of publications and patents. Capital raised is calculated in millions. Firms that withdraw have capital raised set to 0. Firms that complete an IPO have capital raised set to 0 before treatment, and equal to the capital raised post treatment. Columns 1-2 presents a Pseudo Poisson Maximum Likelihood regression with firm, relative year, cohort*year, and IPO month*post FE. Standard errors in parentheses are clustered at the firm level. Confidence intervals are *** p<0.01, ** p<0.05, * p<0.1.

4.6 IPOS AND INNOVATION

In the previous section I showed that IPOs have a positive impact on scientific publications and collaborations. My results, obtained using publications as an indicator of innovative activity, show an opposite trend to those of Bernstein (2015) and Larrain et al. (2021) which use patents. I want to understand whether this difference is solely due to variations in indicators, suggesting that IPOs have a positive impact on basic research but a negative impact on downstream activities. Alternatively, it could also be due to differences in sampling or methodology, which arise from the need to adapt the techniques used for patents to my data.

In this regard, I will replicate those analyses by studying the impact of IPO on innovation. The selected patents are USPTO, Canadian and WIPO published patents. *Patents citations* refers to the number of citations a patent receives within three years of its publication. To normalize this variable, I divide it by the expected number of citations that patents in the same year and in the same four-digit IPC class would receive.

Table 4.10 shows the impact of IPO on the number of patents and patent forward citations. Columns 1 and 2 show positive and statistically significant coefficients. The PPML coefficient is equivalent to a 71% increase in the number of patents after an IPO. Column 3 shows a positive but nonsignificant coefficient.

Table 4.10: ATT of IPO on number of patents and patent forward citations.

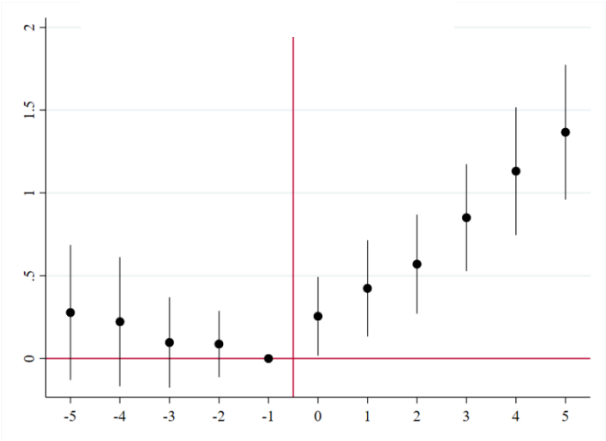
	# Patents		Patent Cit
	(1)	(2)	(3)
	OLS	PPML	OLS
IPO	6.971***	0.537***	0.114
	(1.844)	(0.186)	(0.100)
Observations	6,692	6,289	4,472
R-squared	0.519		0.194
Firm FE	Yes	Yes	Yes
Relative Year FE	Yes	Yes	Yes
Cohort*Time FE	Yes	Yes	Yes
Quarter*Post FE	Yes	Yes	Yes

Note: This table presents the difference-in-difference results of the impact of IPO on the number of patents and forward citations. The sample extends until 2014 in order to provide a five-year post-IPO observation period for the outcome variable. The sample is conditional on having at least one patent in the 5 years before the IPO. Standard errors in parentheses are clustered at the firm level. Confidence intervals are *** p<0.01, ** p<0.05, * p<0.1.

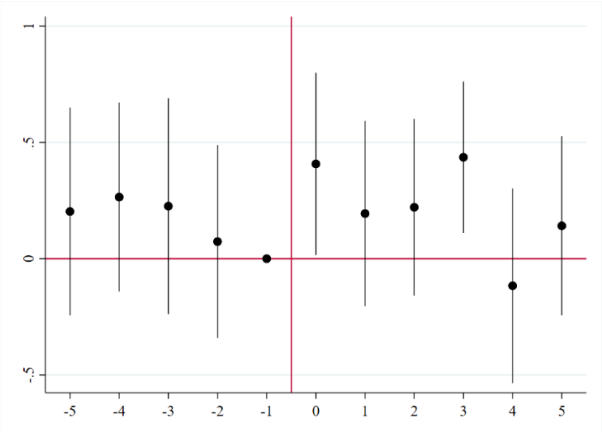
Figure 4.3 continues the analysis showing an event study as per equation (4.3). Table a) shows the results for the number of patents. All coefficients are positive and statistically significant and increasing in t. Table b) shows the results for patent forward citations. The coefficients are positive and statistically significant only at time 0 and 3.

Figure 4.3: IPO impact on patent number and citations, event study

a) # Patents



b) Patent Citations



Note: Same sample of Table 4.10. Figure a*)* and b*)* extend until 2014 in order to provide a five-year post-IPO observation period for the outcome variable. *The sample is c*onditional on having at least one patent in the 5 years before the IPO. Calendar year dummies are included. The coefficients are obtained from the regression $Y_{it} = \mu_i + \theta_t + \sum_{\tau=-5}^{-1} \alpha_\tau IPO_{i\tau} + \sum_{\tau=1}^{5} \beta_\tau IPO_{i\tau} + \delta_{cy} + \partial_{qt} + u_{it}$ via PPML. Standard errors in parentheses are clustered at the firm level. Confidence intervals are p<0.05.

Table 4.11 shows the IV regression. Columns 1 and 3 confirm a positive relation between stock market returns and likelihood of IPO completion. Columns 2 and 4 show the 2SLS results. The coefficient for the number of patents is positive and statistically significant, while positive and insignificant for patent citations. On average, going public leads to an increase of around 25 patents in the 5 years following an IPO.

Table 4.11: Instrumental variable regression on patent number and forward citations

	(1) First stage	(2) # Patents	(3) First stage	(4) Patent cit
IPO		25.05**		0.661
		(9.725)		(0.591)
market returns	0.491***		0.603***	
	(0.042)		(0.062)	
F statistic	134.88		95.70	
Observations	7,783	7,783	4,205	4,205
Firm FE	Yes	Yes	Yes	Yes
Relative Year FE	Yes	Yes	Yes	Yes
Cohort*Time FE	Yes	Yes	Yes	Yes
Quarter*Post FE	Yes	Yes	Yes	Yes

Note: This table presents the results for the instrumental variable specification. Columns 1 and 3 present the first stages. Columns 2 and 4 present the 2SLS estimations. The sample extends until 2014 in order to provide a five-year post-IPO observation period for the outcome variable. The sample is conditional on having at least one patent in the 5 years before the IPO. Standard errors in parentheses are clustered at the firm level. Confidence intervals are *** p<0.01, ** p<0.05, * p<0.1.

The results for patents align with the results presented in the previous section regarding scientific publications. This indicates that IPOs have a positive impact on both publications and patents, while they do not affect the number of forward citations. This evidence is in contrast with Bernstein (2015) and Larrain et al. (2021). Bernstein (2015) finds a positive but non-significant coefficient for the number of patents, and a negative and statistically significant coefficient for the number of forward citations. Larrain et al. (2021) find a negative and non-significant coefficient for granted patents, which becomes significant in countries with high anti-self-dealing and high disclosure requirements.

There are two possible explanations for why I cannot replicate the previous findings. First, concerning Bernstein (2015), the time horizon differs. It is possible that more recent IPOs may differ from those in the 80s and 90s. Second, measurement error could bias my estimations. Orbis patent coverage, especially for the earlier years in my sample, is unsatisfactory. Second, Larrain et al. (2021) study focuses on Europe and EPO patents, which implies that geographic considerations may have an impact.

Next, I switch my focus to the inventors' dynamics, aiming to determine if the same trends observed among scientists also apply to inventors. Table 4.12 shows the IV results for the number of patents, and forward citations for the inventors staying in the company after the IPO. Similar to the findings of Table 4.6 and Table 4.7 the incumbents' number of patents decreases after the IPO, while no effect can be found on forward citations. At the same time, as Table 4.13 shows, the number of new inventors after the IPO increases, exactly as in the case of scientists. This evidence suggests that inventors and scientists follow similar dynamics, and that new inventors join the company compensating for the decline in productivity of the incumbents.

Table 4.12: IPO and innovation at the inventor level

	(1)	(2)	(3)	(4)
	# Patents OLS	# Patents 2SLS	Patent Cit OLS	Patent cit 2SLS
IPO	0.116***	-4.838**	-0.177	-5.228
	(0.0297)	(2.044)	(0.158)	(4.285)
Observations	44,236	44,173	7,115	7,105
R-squared	0.288		0.345	
Individual FE	Yes	Yes	Yes	Yes
Relative Year FE	Yes	Yes	Yes	Yes
Cohort*Time FE	Yes	Yes	Yes	Yes
Quarter*Post FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

Note: Inventors are the individuals who published at least one patent in the five years before an IPO. The sample is restricted from 1996 to 2014. Standard errors in parentheses are clustered at the firm level. Confidence intervals are *** p<0.01, ** p<0.05, * p<0.1.

Table 4.13: Impact of IPO on the number of unique inventors.

	(1) # unique inventors
IPO	0.421**
	(0.183)
Observations	1,502
Firm FE	Yes
Relative Year FE	Yes

Note: This table presents the difference-in-difference results of the impact of IPO on the number of unique inventors. Column 1 is restricted to 2014 and to firm with at least one inventor before the IPO. Standard errors in parentheses are clustered at the firm level. Confidence intervals are *** p<0.01, ** p<0.05, * p<0.1.

4.7 CONCLUSIONS

IPOs play a crucial role in shaping the innovation and scientific output of science-intensive firms. In this study, I examine the population of companies that undertook an IPO in the US from 1996 to 2010 with at least one publication or patent and estimate the causal impact of going public on corporate science. Contrary to recent literature, my findings reveal a positive effect of IPO on the number of publications and collaborations. However, I do not find any effect on publication forward citations, appliedness or basicness. Further evidence on the mobility of scientists and the capital raised at the IPO supports the main results. The effect appears to be driven by the new scientists and inventors who join the company after the IPO.

This chapter presents some limitations that I will address in a future version of this draft. First, there are concerns about the fact that firms that withdraw from an IPO are a proper comparable group. Withdrawing from an IPO may represent a signal of weakness and affect the firms' performance and contribute to overestimate the impact of IPO on scientific research. To mitigate these concerns, I will test the impact of IPO using and alterative sample of private firms that present similar characteristics to the firms that go public. Second, I need to thoroughly test the relevance and the exclusion restriction of my instrument. To address the first issue, I will show that historically the highest IPO completion coincides with years with favourable market conditions. To address the second issue I will run two placebo regressions as in Bernstein (2015), testing if market returns in the years preceding and following the IPO decision have a direct impact on the companies scientific output. For the exclusion restriction to be respected the market returns should show no correlation. Last, I will assess the parallel

trends assumption by conducting a placebo regression, analysing the impact of a simulated treatment in period t-3 on the firms' scientific output.

My results suggest areas for further research. First, I only consider IPOs and no other types of exits. The impact on science and innovation may be influenced by any type of exit strategy, not just IPO. A sample of firms that have undergone other types of exits, such as mergers or acquisitions, could provide additional insights into the mechanisms behind these effects. Second, my analysis does not distinguish between different types of science-intensive firms, and future research could examine whether the impact of IPOs on science and innovation varies across industries or company size. Third, with current data I cannot track intra-firm mobility and the likelihood of staying or leaving after an IPO. Last, the IPO prospectuses could be further used to match the "use of proceeds sections" directly with scientific publications and patents to test if companies genuinely invest in the areas announced at the moment of the IPO.

Conclusions

In this thesis, I explored the changing nature of corporate R&D in the US and contributed to the existing literature in the following ways.

First, as explained in Chapter 1, I presented an original dataset, which is presently the largest dataset available on corporate science in the US, as it includes not only large firms but also small and medium ones. I obtained it from the merging of multiple data sources containing firms' financial information (Orbis), their scientific publications (Web of Science), their distance to the publication-patent frontier (Marx and Fuegi 2020), and patents (Orbis and PATSTAT), from 1980 to 2014. In the same chapter, I presented an overview of general trends in corporate science, which suggests that corporate science has been indeed declining, as suggested by authoritative sources such as Tijssen (2004) and Arora, Belenzon, and Patacconi (2018), but only after 2000. Besides, contrary to the overall decline, business companies have persisted in publishing in biological sciences and medical sciences, witness the increase in the share of corporate publications until 2013. This contrasts with fields like physics, engineering, computer sciences, and chemistry, where corporate publications' shares declined. Besides using the dataset in the chapters of my thesis, I plan to make it available to other researchers. To stay within the legal constraints imposed by Clarivate Analytics and Bureau Van Dijk, it will have to be an anonymized version of my data with selected information.

In Chapter 2, I provided a systematic picture of patterns of university-industry collaborations in the US, as measured by papers co-authored by academic and business scientists. Collaborations have increased over time, due to the increasing "burden of knowledge" faced by R&D-intensive firms (where by burden of knowledge I mean the historical accumulation of scientific notions and skills, well beyond any individual's absorption capacity and leading to an increasing division of labour (Jones 2009)). I found that university-industry collaborations have increased in most industrial and scientific sectors. Specifically, collaborations increased from 2% of the total number of publications (including universities, government, nfp, etc.) in 1980 to 6% in 2013. Simultaneously, firms reduced the amount of research they perform alone. In 1980, about 6% of publications were not co-authored with other institutions, while in 2013, less than 2%. Next, I tested the impact of an increase in the burden of knowledge, measured through references' age, on the likelihood of firms collaborating with

universities. This evidence suggests a positive relationship between the burden of knowledge and collaborations. I also found that this relationship is mediated by firm size, as there is an inverted U-shaped relationship between speed and collaboration. For the same level of burden of knowledge, the likelihood of collaboration increases as size increases, until around 2670 employees. After this threshold, the likelihood of collaboration decreases as size increases. Robustness checks with alternative measures of the burden of knowledge support this evidence.

In Chapter 3, I tested if corporate science is becoming more applied and less basic, contributing to the literature on the decline of corporate science that remarks a shift of companies' in-house basic research toward short term results and commercialisation (Lim 2004; Tijssen 2004; Arora, Belenzon, and Patacconi 2018; Krieger et al. 2021). Based on Stokes' (1997) taxonomy, I measured the appliedness and basicness of scientific research with separate indicators under no assumption of a trade-off between the two. I showed that, over the time period examined, corporate science has become more applied and less basic, thus moving away from what Stokes called the Pasteur's Quadrant towards the Edison's quadrant. Still, some field differences persist. All the life sciences, computer science, geosciences and physics have experienced an increase in appliedness, but only biological sciences, computer sciences, engineering and physics saw a decline in basicness. University-industry collaborations follow a trend similar to that of purely corporate publications but much less pronounced, suggesting that collaborative science may also be moving away from Pasteur's quadrant, but at a slower pace. Last, I found no difference when comparing firms of different size or age.

In Chapter 4, I tested the causal impact of IPOs on corporate science. Many authors have found that the disclosure requirements and pressure from the shareholders coming with going public may affect negatively affect companies' innovation outcomes (Aggarwal and Hsu 2014; Bernstein 2015; Gao, Hsu, and Li 2018; Markovitch, Steckel, and Yeung 2005; Moorman et al. 2012; Wies and Moorman 2015; Wu 2012). However, most evidence concerns patents, while none is available on companies' scientific publications. Scientific publications are important when studying corporate R&D strategies, as they signal more long-term scientific research, which might otherwise remain unnoticed when solely relying on patents. Using data from the population of US companies with at least one patent or publication that underwent an IPO from 1996 to 2010, I found a positive impact of IPO on corporate science, measured as scientific publications and collaborations. Firms' increased access to capital and the inflow of new scientists likely drive the effect. I find no effect on the forward citations of both publications and collaborations, appliedness, and basicness.

This thesis presents some limitations that can be addressed in future research. Some concern the whole thesis, others specific chapters.

Tracing ownership structure across 30 years is difficult and requires much effort, especially for medium and small enterprises. I carefully checked the ownership structure of the largest publishers. However, some imprecision in the ownership of the smaller entities is inevitable. On the one hand, working with large datasets allows one to tackle research questions more systematically. On the other hand, it becomes impossible to check manually all the data collected.

I had to exclude conference proceedings from my econometric analyses because of their limited coverage from the 1990s to the early 2000s in Web of Science. This choice may be problematic because engineering and computer science researchers publish extensively in conference proceedings or open-source databases like ArXiv (Boyack, Klavans, and Börner 2005; Frank et al. 2019; Lin et al. 2020). Missing these publications might underestimate the companies' contributions to science.

Publications are a partial indicator of corporate science because they measure only companies that engage in scientific research and decide to disclose their results in scientific publications. Disclosure is a sensitive issue for companies, as competitors could potentially exploit their research. For this reason, many companies, despite relevant R&D spending, innovate without publishing a single paper (Lim 2004). In this thesis, I do not address this issue directly, studying the incentives and the propensity of firms publishing.

The companies that published the largest number of publications are large multinational corporations with R&D laboratories spread all around the globe. Focussing only on publications with US addresses, I can better analyse the dynamics of US corporate science. However, I cannot observe if companies outsourced or transferred their R&D elsewhere. Collecting worldwide publications comes with the tradeoff of lower data accuracy and higher computational effort. A possible compromise is to focus on a smaller sample of large companies and their subsidiaries and observe how their corporate science changes geographically.

As for the chapter-specific limitations, they are the following.

In Chapter 2, I defined scientific fields using WoS classification and calculated the speed of scientific progress for multi-fields papers as the arithmetic average of the individual fields' speed. However, different levels of aggregation are possible and can drive the results in other

directions. Furthermore, Orbis is probably not the best database to identify small science-intensive companies, given that its coverage prioritises large and medium-sized companies.

Regarding Chapter 3 results, I could not measure appliedness and basicness in absolute terms but only in relative ones, comparing companies and university-industry collaborations to university publications. Other metrics of appliedness and basicness could address this issue in the future.

In Chapter 4, I explored the impact of IPO on corporate science. However, an IPO is only one of the potential companies' exit strategies. A company can choose between alternative exits such as IPO, M&A, or leverage buyout. The distinct effects of these diverse exit strategies on corporate science remain unexplored.

This thesis opens to future research avenues. Many scholarly contributions circle around the general idea that university-industry collaborations are replacing in-house corporate research, including mine. However, no causal relationship has been established yet. Following the footsteps of Chapters 2 and 3, it would be interesting to fill this gap, testing if in-house research and collaborations with universities are complements or substitutes.

Next, startups and spinoffs are important for the innovation system, especially in the context of open innovation (Spender et al. 2017). It would be interesting to explore more in-depth the nature of those firms and their contribution to American science. Despite an attempt to identify university spinoffs in Chapter 2, my proxy only covers the period from 2008 to 2013, potentially missing some companies due to Orbis' lower coverage for small businesses. Therefore, undetected small science-intensive startups and spinoffs might drive university-industry collaborations. In this scenario, the role of corporate science would lose even more importance, and universities would acquire a more central role.

In this thesis, I calculated the publications' distance from the science-technology frontier using the backward citations network. However, I do not take into consideration the patent network of backward citations. The distance metric presented in Chapter 3 could be used to calculate patents' distance from the science-technology frontier, interpreting proximity to the frontier as more science-intensiveness. This metric would be particularly relevant to analyse the science intensiveness of companies that do not engage in scientific publications, and thus do not send any signals regarding their scientific research.

There is evidence on scientists mobility (Franzoni, Scellato, and Stephan 2012; Vaccario, Verginer, and Schweitzer 2021; Verginer and Riccaboni 2021), however there is little evidence

on corporate scientists. It could be interesting to study more in detail the movements of scientists and their research trajectories. Specifically, how scientists' output changes after moving from one institution to another. Some examples are scientists transitioning from industry to academia and vice-versa or changing institutions.

Last, most of the literature attention, including the one of this thesis, is on US scientific research. Few exceptions focus on other nations like UK, Germany, and Italy (Calvert and Patel 2003; Abramo et al. 2009; Krieger et al. 2021). The same methodology used to create the WoS-Orbis database can be used to replicate my analysis in other countries. The technical limitation would be harmonising the algorithm to accommodate different country-specific incorporation types, languages, and abbreviations. It would be interesting to replicate the results for Europe, Japan, or raising scientific powerhouses like China and India.

POLICY IMPLICATIONS

Business investments in R&D are important for all the innovation systems, as businesses constitute the primary R&D performers (NSF 2022). Concerning the US, there was a widespread concern about companies changing their R&D strategies and disengaging from scientific research. These concerns have been expressed by Arora, Belenzon, and Patacconi's (2018) influential paper, which has drawn significant media attention in the US (Xie 2014; Porter 2015). These concerns began to alleviate toward the end of the 2010s, especially after substantial investments in basic research by pharmaceutical and biomedical companies (Suresh and Bradway 2016; Mervis 2017; Arora et al. 2019), as well as the increasing research in artificial intelligence (Hartmann and Henkel 2020; Haddad 2023). Furthermore, recommendations have emerged advocating for an open science approach to address the new challenges in science and technology, both from government agencies (Lander 2021; National Science and Technology Council 2022) and scholars (Gold 2021).

My findings align with Arora's and coauthors, indicating a shift towards less basic and more applied research, particularly in corporate publications, but not as prominently in collaborations with universities. Rather than interpreting this as a decline in corporate science, this evidence suggests a shift in its structure. Companies are transitioning from a vertically integrated model of corporate R&D to a more dynamic system involving collaborations with universities and the acquisition of small, innovative startups (Coombs and Georghiou 2002; Arora and Belenzon 2023).

The shift of companies towards more applied science is not inherently problematic for society, as long as there is a collective effort to continue investing in basic research. Firstly, companies continue to engage in less applied and more basic science through collaborations with universities. Therefore, it may be more beneficial to support collaborations rather than advocating for more corporate science as such. Second, the level of industry involvement in scientific research varies based on the maturity of the sector. While certain industries like chemistry and semiconductors have experienced a decline, emerging sectors such as biotechnologies, pharmaceuticals, and, more recently, artificial intelligence witness substantial corporate involvement.

Conversely, if universities disengage from basic science, it could potentially create more problems. This thesis shows that universities continue to prioritise basic research more than companies. However, although our findings provide evidence that corporate science is increasingly applied and less basic, it remains uncertain whether this trend is exclusive to corporate science or if university science is also undergoing a shift towards greater applied emphasis, albeit at a different pace. A reduction in private sector investments in basic research necessitates a compensatory increase in public sector contributions to maximise social profit (Nelson 1959). However, being scientific discoveries serendipitous, it may require a higher level of investments, possibly unsustainable if private companies are underinvesting.

Furthermore, a new worry is arising for established technological leaders. Countries such as India and China are increasingly investing in science and in its technological transfer, including in traditional US-dominated sectors (Stone 2012; Brainard and Normile 2022). The CHIPS and Science Act, enacted on August 9, 2022, tries to face the challenge by unlocking $280 billion in new funding to boost domestic semiconductors R&D and production in the United States. Companies like Micron, Qualcomm, and GlobalFoundries are making multibillion-dollar investments in chip manufacturing (McKinsey 2022). This act aligns with the goal of investing in established, traditional industries that appear to have lost momentum, including those I found to have declining publication trends (e.g., AT&T, Dupont, and General Electric).

We should recognise the evolving landscape of US innovation and not crystallise on old paradigms like the vertically integrated R&D lab. By investing in emerging industries and increasing government funding in strategic sectors like semiconductors, we are heading towards the right direction. Therefore, it is necessary to adopt a more systemic approach that

supports industry-university-government collaborations to ensure that long-term investments in basic research are not overlooked.

A. Appendix Chapter 1

This technical appendix serves to describe in detail the matching algorithm introduced in Chapter 1. The set of corporate science publications results from merging Orbis by Bureau Van Dijk (BvD), a company database, and Web of Science (WoS) by Clarivate Analytics, a bibliographic repository. Data from the two sources were matched at company level, so to produce a database on corporate science, with information on both companies' publications and financial records. Figure A 1 summarizes the matching exercise's workflow.

First, I prepared other inputs such as a list containing the words belonging to the English dictionary and a table containing the abbreviations used in WoS.

Second, I cleaned the Orbis inputs. I lowercased and removed trailing spaces in the company names and city, then I standardized the zip codes to a 5 digits zip code. The matching algorithm compares word by word the strings from Orbis and Wos, thus I tokenised (divided word by word) the strings in Orbis as in Table A 9. The company names were abbreviated according to the WoS abbreviations in order to have a structure coherent to the WoS names.

Third, I selected the right WoS inputs. I started from all the articles published on papers and proceeding in the US for the period 1980-2014. Each unique identifier was linked to an affiliation name and an address. I filtered the dataset removing universities and the biggest governmental agencies.

Last, I loaded an index file (Table A 19) that sets up the algorithm structure. It allows to load chunks of 1000 rows at the time, giving the opportunity of running multiple matches at the same time.

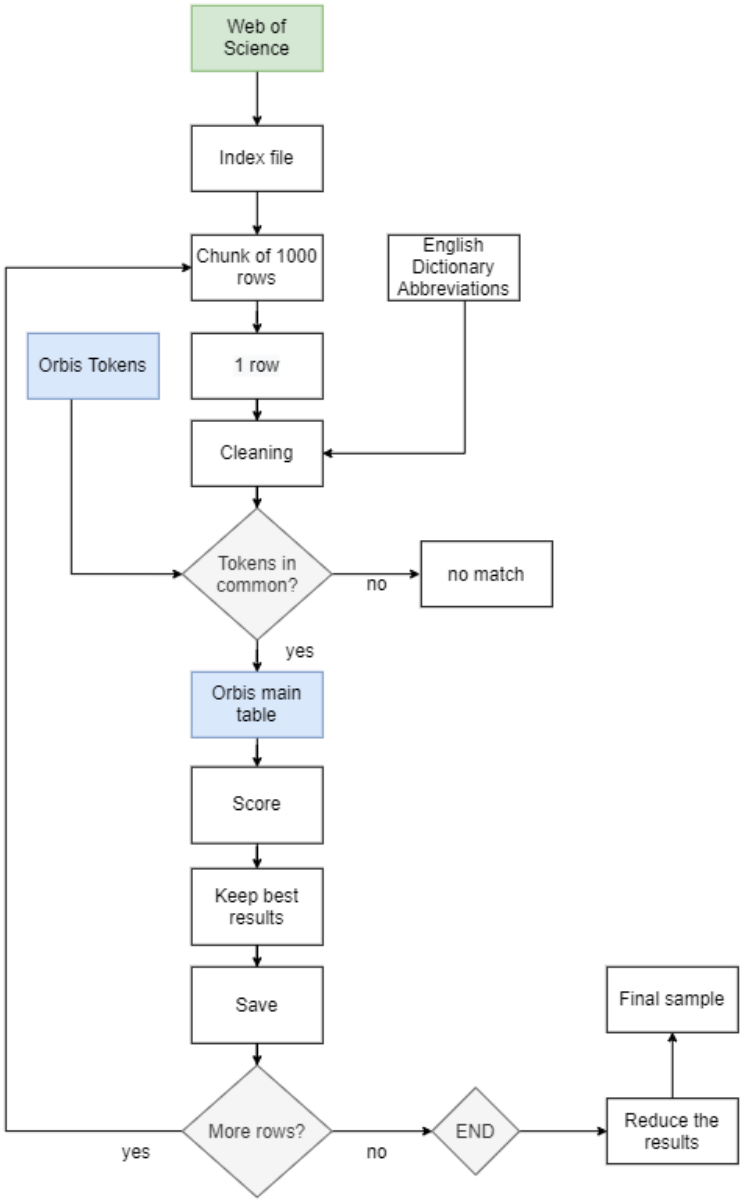
At the beginning of the matching algorithm, I loaded the clean Orbis main table (Table A 13), the tokens (Table A 9), the English dictionary, the abbreviations (Section A.1.3), and a 1000 rows WoS chunk.

Proceeding in the following steps row by row for each WoS row, I:

1. Cleaned the affiliation name, city, and zip code. Names were abbreviated according to Section A.1.3
2. Created a bin (subset) containing the company names in Orbis that share at least one token, up to a maximum of 400k rows.

3. Scored the rows in the bin using fuzzy scores (Levenshtein Distance similarity scores), geography scores (city and zip in common) and dictionary scores (non-dictionary words in common)
4. Filtered the results keeping only the ones with highest fuzzy score, best geography score and best non dictionary score.

Figure A 1: Matching algorithm flowchart



Note: This figure shows the matching algorithm flowchart. The main data sources are WoS, Orbis affiliations and other inputs as stopwords, abbreviations and dictionary words list (to identify non-dictionary words). First a chunk of 1000 rows is read from WoS. Starting from the first row (n=1), I clean the input name and retrieve from Orbis all the affiliations with at least one word in common. Afterwards the potential matches are scored, and I select only the best matches. I save the final results in a csv and I proceed with matching another affiliation if n<1000, otherwise the match concludes.

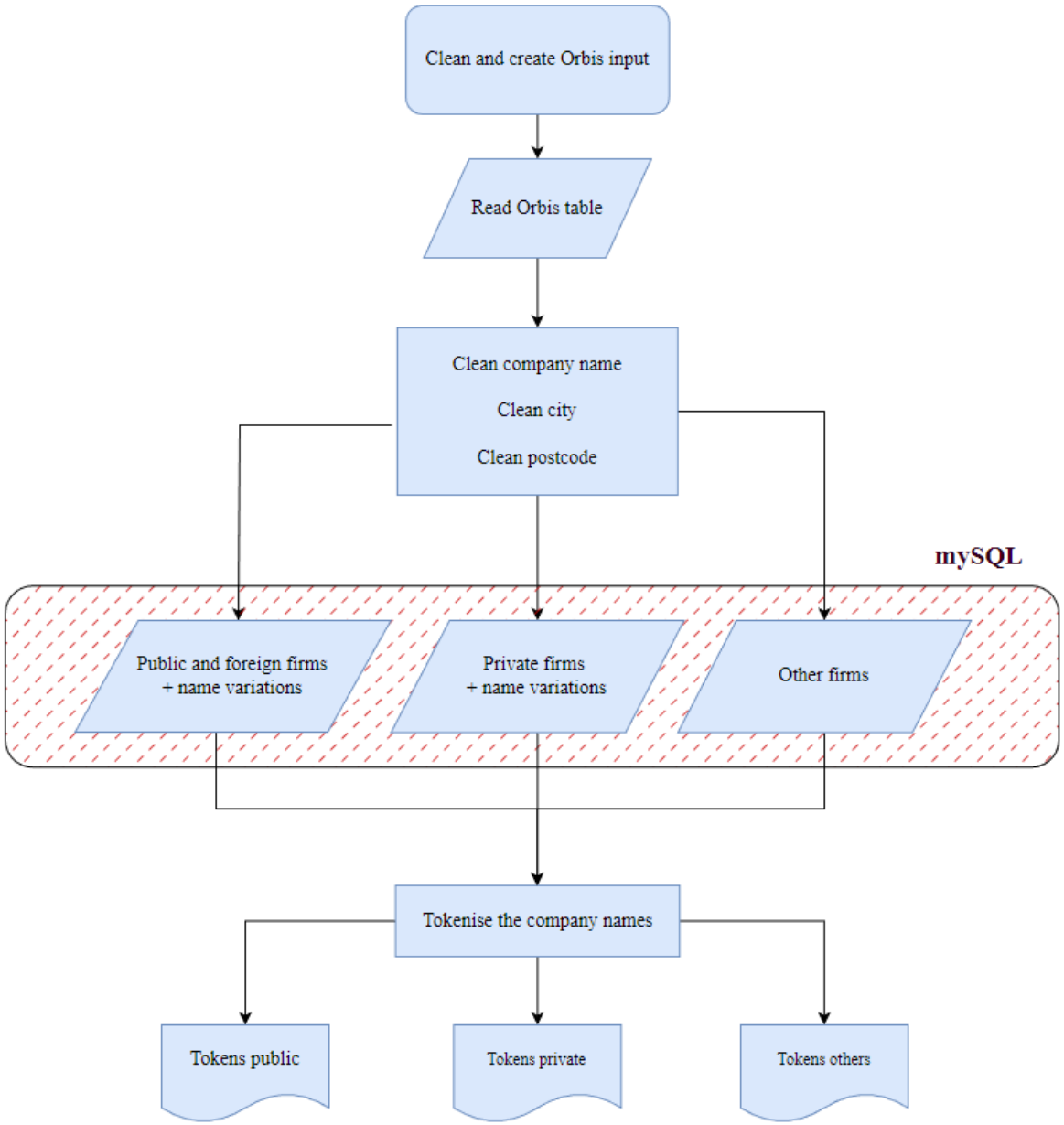
The algorithm produces two output files, for the matched and unmatched names. Post-match I further reduced the results ending up with a single match for each row in WoS, and divided the affiliations according to their type (firms, educational institutions, not for profit organisations, medical centres and clinics, governmental agencies).

The major challenge of the matching exercise was posed by the need to harmonize companies' names, that may appear in different variations (ex: IBM, IBM corp, International Business Machines etc.), both within each source and across them. In Sections A.1 and A.2 I will introduce separately Orbis and Web of Science, then in Section A.3 I will describe in detail the matching process. Section A.4 presents how I selected the final match candidate.[50] Section A.5 shows the diagnostics of the matching algorithm.

[50] Refer to footnote 8 for acknowledgments regarding the previous work I inherited and upon which I based the algorithm.

A.1.ORBIS

Figure A 2: Orbis cleaning workflow



Note: This chart presents schematically the Orbis cleaning operations

Orbis is one of the most comprehensive datasets on private companies. It contains information about more than **375 million large, medium, and small companies** across the globe. Other commonly used datasets like Worldscope database by Thomson Financial and Compustat by S&P Global Market Intelligence, contain mostly information on large publicly listed companies. Orbis is the best option for the purposes of my analysis because in the United States,

medium and small enterprises constitute a core part of the economic system (Kalemli-Ozcan et al. 2015).

BvD has been selling Orbis since 2005. During this time, they put together a data release every year, which compiled all the information for firms they had gathered that year, including information from previous years.

Until October 2019, Swinburne University subscribed to the “Orbis Historic” data service, which provided access to all these yearly releases. At last, 13 different releases are available from 2005 to 2017. Each proving one year's release of the full Orbis data. In other words: the 2005 data server allow to download information in Orbis as of 2005, the 2006 as of 2006, and so on until 2017.

The data servers/data disks from 2005-2011 had a bit of a different user interface than the data servers from 2012 onward. For this reason, BvD calls the 2005-2011 data servers/data disks “Orbis Classic”, and the 2012-onward data servers/data disks “Orbis Neo”. These are not really two different things, but just a subset of Orbis Historic. I am using Orbis Historic both for the better coverage and the possibility to have longer series of data. In addition, I can better follow the changes of ownership and location overtime. The variables available in Orbis Historic are listed in Table A1. When data are missing, I integrated Orbis Historic and Neo with the latest online version of Orbis.

Table A 1: Orbis Historic variables

Category	Variable Name	Type	Downloaded As	Internal Orbis Code
Identifiers	BvDID number		One-to-one	BVDID
	Postcode		One-to-one	POSTCODE
	City		One-to-one	CITY
	Previous company name	multi	Many-to-one	PREVNAME
	Name change date	multi	Many-to-one	NAMECHDT
	Also known as name	multi	Many-to-one	AKANAME
	Country		One-to-one	COUNTRY
	Country ISO		One-to-one	CTRYISO
Others	Company Size		One-to-one	CATEGORY_OF_COMPANY
	Status	multi	Many-to-one	STATUS
	NAICS 2012 Primary code(s)		One-to-one	NAICSPCODE2012
	Last available year		One-to-one	LASTYEAR
	Consolidation code		One-to-one	CONSCODE
	Website	multi	Many-to-one	WEBSITE
	Status date	multi	Many-to-one	STATUSDATE
Financials	Research & Development expenses	yearly	One-to-one	RD
	Export revenue	yearly	One-to-one	EXPT
	Material costs	yearly	One-to-one	MATE
	Cost of employees	yearly	One-to-one	STAF
	Added value	yearly	One-to-one	AV
	Cash and cash equivalent	yearly	One-to-one	CASH
	Sales	yearly	One-to-one	TURN
	Cost of goods sold	yearly	One-to-one	COST
	Fixed assests	yearly	One-to-one	FIAS
	Operating turnover (US$000)	yearly	One-to-one	OPRE
	Employment (person)	yearly	One-to-one	EMPL
	Total assets (US$000) -	yearly	One-to-one	TOAS
Ownership	Global Ultimate Owner - BvDID		Many-to-one	-9105
	Global Ultimate Owner - name		Many-to-one	-9100
	Global Ultimate Owner - ISO code		Many-to-one	-9102
	Subsidiaries - BvDID	multi	Many-to-one	-9305
	Subsidiaries - name	multi	Many-to-one	-9300
	Subsidiaries - ISO code	multi	Many-to-one	-9302
	Domestic Ultimate Owner - name		Many-to-one	-9100
	Domestic Ultimate Owner - BvDID		Many-to-one	-9105
	Immediate parent name		Many-to-one	-9001
	Immediate parent id		Many-to-one	-9006
Patents	Publication identifier	multi	Many-to-one	PUBLICATION
	Application number	multi	Many-to-one	APPLICATION_NUMBER
	Priority number	multi	Many-to-one	PRIORITY_NUMBERS
	Also published as	multi	Many-to-one	ALSO_PUBLISHED_AS
Directors	Full name	multi	Many-to-one	HEADER_FullNameOriginal
	Unique contact identifier	multi	Many-to-one	HEADER_IdDirector
	Appointment date	multi	Many-to-one	BeginningNominationDate
	Resignation date	multi	Many-to-one	EndExpirationDate
	Type of position	multi	Many-to-one	BoardMnemonic
	Board committee or department	multi	Many-to-one	DepartmentFromHierCodeFall2009
	Level of responsibility	multi	Many-to-one	LevelFromHierCodeFall2009

Note: This table presents the variable available in Orbis Historic. The main variables can be divided in 6 groups: identifiers, Others, Financials, Ownership, Patents, and Directors.

A.1.1.Merge different server Years

Orbis Historic allows to have a more complete set of information about US entities. Every data server, however, may present some differences from the others. Firms might have a different: name (J & B importers inc., J&B importers inc., J & B importers), location (firm that relocates, ZIP +4 or ZIP code), ownership (Wyeth that becomes Pfizer after its acquisition in 2009), financials (different data provider or corrected figures). The raw data present 115,439,478 rows and 85,406,826 unique identifiers.

The first cleaning operations were:

1. Cleaning city. After having lowercased and removed punctuation, I abbreviated some words to make them consistent to the ones recorded in WoS. These abbreviations are the following:

Table A 2: WoS abbreviations

pk	park	*se*	southeast
ft	fort	*sq*	square
n	north	*st*	street
s	south	*st*	saint
so	southern	*rd*	road
mw	northwest	*rd*	roads
no	northern	*w*	west
ne	northeast	*e*	east
sw	southwest	*jct*	junction

Note : This table presents the abbreviations used in WoS (*italics*) and their expanded version.

2. Cleaning zip code. The server years 2005-2014 record a five digits zip code while 2015-2017 often record the long ZIP +4 version[51]. WoS records only five digits zip codes, so I standardised all of them to a five digits zip code.
3. Cleaning company names. First, I lowercased, removed punctuation[52] and trailing spaces. Orbis has then an inconsistent behaviour of whitespaces across servers. The same firm may be recorded as “A T & T” or “AT&T”, “T A J chemicals” or “TAJ

[51] A ZIP+4 Code is a five digits code plus four additional numbers to identify blocks. Group of apartments, post-office boxes, units.

[52] !"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~

chemicals", thus I merged together up to four single letters. Second, I removed common words that are not relevant for the matching, commonly called "stopwords".

4. Abbreviating the names according to name-abbreviation pairs (ex: research-res). A more detailed explanation follows at Section A.1.3.

   Last, I dropped all the duplicate rows in my sample grouping by company, name and zip code.

Table A 3: Stopwords

inc	corp	of	associate
incorporated	limited	and	associates
llc	ltd	partnership	co
company	llp	partnerships	gmbh
corporation	pc	enterprise	
lp	the	enterprises	

Note: Stopwords used to clean company names. I used the most common incorporation types and uninformative English words (and, of, the)

A.1.2.Private and public firms

For simplicity I divided the firms in Orbis in 3 subsets. The first one contains public, formerly public companies, and foreign subsidiaries. This set contains the largest firms in Orbis, and the ones most like to publish more. The second subset contains private companies' Ultimate owners. The last subset contains all the firms that do not fit the previous two subsets and do not have any ownership information.

I enriched the previous sets of companies including all the name and address variations of the companies in Orbis. Starting from the Domestic Ultimate owner in the case of public and private companies, and the Global ultimate owner in case of foreign subsidiaries. I considered as name and address variations all those branches and subsidiaries that have a similar name to ultimate owner. I last used a conversion table provided by BvD to convert old identifier into new ones.[53] I did not perform this procedure for the last group of companies because no ownership structure is available.

[53] Frequently companies in Orbis Historic have an identifier in the format US-123-456 which are converted into US136547L in Orbis Neo. The two identifiers do not have anything in common. Initially I tried to group companies with same name and same address as a unique entity- However this approach was interfering with the

A.1.3.Abbreviations

WoS, differently from Orbis, abbreviates the names of his affiliations. Official WoS sources do not provide a comprehensive list of abbreviations, so I integrated It with abbreviations I detected manually, reaching a total of 409 abbreviations. In addition, most of the words are composite like: microelectronics, geophysics and thermodynamics and considering just the root word is not enough. Consider for example the pair “Elect” - “Electronic”. When abbreviating "Electronic" to "Elect," I account for all possible prefixes (e.g., "micro-electronic") and suffixes (e.g., "electronic-s"). As general rule the prefixes remain unchanged and the suffixes are omitted. The result will thus be as follows.

Before the abbreviation:

Table A 4: Orbis table before abbreviating

Orbis	**WoS**	**Fuzzy Score**
Cabot microelect	Cabot microelectronics	84
westinghouse elect	westinghouse electric	92

After the abbreviation:

Table A 5: Orbis table after abbreviating

Orbis	**WoS**	**Fuzzy Score**
Cabot microelect	Cabot microelect	100
westinghouse elect	westinghouse elect	100

In some cases, the technique just presented fails because some words are subsets of others and have different abbreviations (see Table A 6). For example, international is most often abbreviated as “int”, but national may be abbreviated “ntl”. These specific cases are treated differently, without taking into account prefixes and suffixes as before.

Table A 6: Exceptions in abbreviations

ntl	national	*lab*	laboratory
int	international	*lab*	labs
lab	laboratories		

ownership structure and grouping together firms that should have remained separate. Taking a more conservative approach I only convert old identifiers to new ones using the table provided by BvD, even if not comprehensive.

All these tables are saved as dictionaries, making easy to access to all the name-abbreviation pairs. The following tables show the complete dictionary of abbreviations.

Table A 7: Abbreviations dictionary

absorpt	absorption	*aviat*	aviations
acoust	acoustic	*behav*	behavior
acoust	acoustics	*behav*	behaviors
acoust	acoustical	*behav*	behavioral
adm	administration	*bion*	bionical
adv	advanced	*bion*	bionic
adv	advance	*bion*	bionics
aerosp	aerospace	*bldg*	building
agcy	agency	*bldgs*	buildings
agr	agriculture	*blvd*	boulevard
amer	american	*bot*	botanic
anal	analysis	*bot*	botanical
analges	analgesic	*bot*	botanics
analges	analgesics	*bros*	brothers
analyt	analytic	*bur*	bureau
analyt	analytics	*calif*	california
analyt	analytical	*canc*	cancer
anat	anatomy	*carbonizat*	carbonization
anat	anatomic	*catalyt*	catalytical
anat	anatomical	*catalyt*	catalytic
anat	anatomics	*catalyt*	catalytics
anim	animal	*ceram*	ceramic
anim	animals	*ceram*	ceramics
antiinfect	antiinfective	*champ*	champion
antiinfect	antiinfectives	*champ*	champions
apparat	apparatus	*chem*	chemicals
appl	applied	*chem*	chemical
applicat	application	*chem*	chemistry
applicat	applications	*chiropract*	chiropractics
assoc	association	*chiropract*	chiropractic
assoc	associations	*clin*	clinic
astron	astronomy	*co*	company
atmospher	atmospheric	*co*	corporation
atmospher	atmospherics	*coll*	college
atom	atomic	*collaborat*	collaborative
atom	atomics	*combust*	combustion
automat	automation	*comm*	committee
automat	automations	*commercializat*	commercialization
ave	avenue	*commun*	communications

commun	communication	*diffract*	diffraction
comp	computing	*diffract*	diffractions
comp	computer	*dimensional*	dimens
computat	computational	*dis*	diseases
conf	conference	*dis*	disease
conservat	conservation	*dispers*	dispersion
constellat	constellation	*dispers*	dispersions
consultat	consultations	*dist*	districts
consultat	consultation	*dist*	district
convers	conversion	*div*	divisions
convers	conversions	*div*	division
cosmet	cosmetics	*dynam*	dynamic
cosmet	cosmetical	*dynam*	dynamics
cosmet	cosmetic	*econ*	economics
cpd	compound	*econ*	economics
cpds	compounds	*econ*	economical
cpl	corporal	*elect*	electron
creat	creative	*elect*	electronic
creat	creation	*elect*	electronics
creat	creations	*elect*	electronical
ctr	center	*elect*	electric
ctr	centers	*elect*	electrics
cty	county	*elect*	electrical
cycl	cyclic	*elect*	electr
cycl	cyclical	*energet*	energetic
cycl	cyclics	*energet*	energetics
cytol	cytology	*engn*	engineering
dakocytomat	dakocytomation	*engn*	engineerings
def	defense	*engn*	engineer
degradat	degradation	*engn*	engineers
dent	dental	*engn*	engines
dent	dental	*engn*	engine
dept	department	*environm*	environment
depts	departments	*environm*	environmental
dermsurg	dermsurgery	*estab*	establishment
dev	development	*estab*	establishmenta
diabet	diabetes	*explorat*	exploration
diag	diagnosis	*explorat*	explorative
diagnost	diagnostic	*explorat*	exploration
diagnost	diagnostics	*expt*	experiment

expt	experiments	*immunol*	immunology
exptl	experimental	*inc*	incorporated
extens	extensions	*ind*	industry
extens	extension	*ind*	industrial
fac	faculty	*ind*	industries
facil	facility	*infect*	infectuous
fdn	foundation	*infirm*	infirmary
fed	federal	*informat*	information
fertilizat	fertilization	*innovat*	innovative
forens	forensic	*innovat*	innovation
fus	fusion	*innovat*	innovations
fus	fusions	*inspect*	inspection
gen	general	*inspect*	inspections
gen	genic	*inst*	institute
gen	genics	*int*	international
generat	generation	*integr*	integrity
generat	generations	*interferometr*	interferometric
genet	genetic	*interferometr*	interferometrics
genet	genetics	*isl*	island
genet	genetical	*kinet*	kinetic
genom	genomic	*linguist*	linguistic
genom	genomics	*log*	logic
geodes	geodesic	*ltd*	limited
geodes	geodesics	*lyophilizat*	lyophilization
geol	geologic	*magnet*	magnetic
geol	geological	*mat*	material
govt	government	*mat*	materials
grad	graduate	*math*	mathematic
graph	graphic	*math*	mathematical
graph	graphics	*math*	mathematics
graph	graphical	*mech*	mechanic
grp	group	*mech*	mechanics
gynu	gynuity	*mech*	mechanical
hist	history	*med*	medic
hist	historical	*med*	medicine
histol	histology	*med*	medical
hlth	health	*med*	medicinal
hosp	hospital	*mem*	memorial
hyg	hygiene	*met*	metal
immunizat	immunization	*met*	metals

metab	metabolic	*ol*	ological
metab	metabolics	*olymp*	olympic
metab	metabolical	*opt*	optic
metr	metric	*opt*	optics
metr	metrics	*opt*	optical
mfg	manufacturing	*optimizat*	optimization
mgmt	management	*org*	organization
mil	military	*orthopaed*	orthopaedics
min	mining	*orthopaed*	orthopaedic
mkt	marketing	*orthoped*	orthopedics
mobilizat	mobilization	*orthoped*	orthopedic
modrnizat	modrnization	*otol*	otolaryngology
modulat	modulation	*paediat*	paediatric
modulat	modulations	*paediat*	paediatrics
mol	molecular	*panasonic*	panason
mol	molec	*pathol*	pathology
mt	mount	*pediat*	pediatric
mt	mountain	*petr*	petroleum
narcot	narcotic	*pharma*	pharm
narcot	narcotic	*pharma*	pharmaceut
nat	nature	*pharma*	pharmaceutical
natl	national	*pharma*	pharmaceuticals
naturalizat	naturalization	*photon*	photonic
nav	navigation	*photovolta*	photovoltaic
ne	northeast	*phys*	physical
neurol	neurology	*phys*	physics
neurol	neurologic	*phys*	physic
neurol	neurological	*plantat*	plantation
no	nothern	*plantat*	plantations
nucl	nuclear	*plast*	plastic
nutr	nutrition	*plast*	plastics
nw	northwest	*populat*	population
obes	obesity	*precis*	precision
observ	observatory	*precis*	precisions
obstet	obstetrics	*predict*	prediction
occupat	occupation	*predict*	predictions
off	office	*preservat*	preservation
og	ography	*proc*	process
ol	ology	*prod*	product
ol	ologic	*prod*	products

prop	propulsion	*stat*	statistical
prot	proteins	*sterilizat*	sterilization
prot	protein	*stn*	station
protect	protection	*struct*	structural
prov	province	*struct*	structure
psychiat	psychiatric	*struct*	structures
psychol	psycholog	*subst*	substance
pty	proprietary	*subst*	substances
pulm	pulmonary	*supercomp*	supercomput
qual	quality	*supercond*	superconduct
radiat	radiation	*supercond*	superconductor
radiat	radiations	*supercond*	superconductors
recreat	recreation	*supercond*	superconducting
recreat	recreational	*surg*	surgery
rehabil	rehabilitation	*synth*	synthesis
remediat	remediation	*syst*	systems
res	research	*syst*	system
resp	respiratory	*technol*	technological
rev	review	*technol*	technologies
revitalizat	revitalization	*technol*	technology
robot	robotic	*tel*	telephone
sanit	sanitary	*temp*	temperature
sch	school	*text*	textile
sci	scientifics	*text*	textiles
sci	scientific	*tronic*	tronics
sci	sciences	*tron*	tronic
sci	science	*therapeut*	therapeutic
secur	security	*therapeut*	therapeutics
semicond	semiconductor	*therapeut*	therapeutical
semicond	semiconductors	*transportat*	transportation
serv	service	*undustrializat*	undustrialization
serv	services	*univ*	university
simulat	simulation	*utilizat*	utilization
simulat	simulations	*vasc*	vascular
simulat	simulation	*vet*	veterinary
simulat	simulations	*vet*	veteran
soc	society	*vet*	veterans
solut	solution	*victimizat*	victimization
solut	solutions	*visualizat*	visualization
spect	spectroscopy	*weap*	weapon
stabilizat	stabilization	*weap*	weapons
stand	standard	*welf*	welfare
starteg	strategic	*zool*	zoologic
stat	statistic	*zool*	zoological
stat	statistics	*zool*	zoology

A.1.4.Tokenization

The matching algorithm compares word by word the string from Orbis and the one from Wos. The process of dividing in words a whole string is called tokenisation. In first place I saved the results in the format showed in Table A 8.

Table A 8: Inefficient Orbis tokens table

Token	Bvd_ids
Int	7234162
Business	7234162
Machines	7234162
Gen	5814092
Motors	5814092
Luxury	8922792
Motors	8922792
Int	8922792

Then I collapsed the results to access the Bvd_ids with less computational effort.

Table A 9: Orbis tokens table

Token	Bvd_ids
Int	7234162, 8922792
Motors	5814092, 8922792
Business	7234162
Machines	7234162
Gen	5814092
Luxury	8922792

A.1.5.Financials

Orbis historic servers provide financial information on companies in three different formats, flagged with the following acronyms:

1. AY: absolute year. Downloaded from the server ticking the year on the interface (2004, 2005 etc.).
2. RY: relative year. Downloaded from the interface going backwards from a selected year. For example, RY1 (-1 years), RY2(-2 years) and so on.
3. LAY: latest available year for this company.

Each financial entry in servers 2005-2012 is linked to a closing date, i.e., when the company provided its financial information to BvD. From 2013 onwards, relative years and closing dates are not available.

Table A 10: Orbis financial data from all servers

Bvd_id	Operative revenue	Flag	Closing Date	Data Year	Server year
123	50	RY	1/6/2008	2008	2009
123	100	RY	1/12/2008	2008	2009
123	100	RY	1/12/2008	2008	2010
123	100	RY	1/12/2008	2008	2011
123	110	RY	1/12/2008	2008	2012
123	110	AY	1/12/2008	2008	2012
123	110	LAY	1/12/2008	2008	2012

First, I kept the most recent closing date within each server and drop duplicates regardless the flag. Second, I divided the financials by data year and I kept the most frequent variable for each data year. In the case of bimodal values, I kept the most recent non-blank value. From Table A 10 I obtain Table A 11.

Table A 11: Collapsed financial data

Bvd_id	Operative revenue	Data Year	Server year
123	100	2008	2009

Orbis historic financials are complemented with the latest version of Orbis ("Orbis Neo"). Financials in this version come in bulk and do not present the issue of harmonising different server years. There is some overlapping with Orbis historic in terms of coverage, but there is a consistent number of firms that belong only to one version or the other. In general, Orbis Neo has a better coverage for more recent large publicly listed companies.

A.1.6.Ownership

Orbis provides two different information about ownership: the domestic ultimate owner (DUO) and the global ultimate owner (GUO). The GUO is the entity at the top of the corporate

ownership structure, while the DUO is the highest owning entity in the country.[54] Each server year provides the ownership data relative to the year of the server. Thus, I cannot follow in time in each server the changes in ownership, but I have to compare the data provided by different servers.

I complemented the ownership data with Arora, Belenzon, and Sheer (2021b) open-source data[55] on US public firms with at least one patent and positive R&D spending, and with Orbis M&A, a more recent product of BvD. Residually, all the firms that do not have any ownership information are considered to be independent.

A.1.7.Final sample

Table A 12 shows the final Orbis sample. Note that the same firm may have many correspondents *bvd ids* assigned to a single Ultimate Owner*(UO)*, so the number of *bvd ids* is not an indicator of the number of firms, but of the number of rows.

Table A 12: Final Orbis sample

	Financials	**Ownership**	**Total**
Unique bvd ids	41,134,756	13,683,557	73,484,490

The clean Orbis main table, ready to be used as input in the matching algorithm, will look like the example in Table A 13. *Match id* is the corresponding row number in the table, *bvd id* is the firm identifier and *company name* its name. *UO id* and *UO name* are the id and name of the ultimate owner, *City* and *zip code* contain the geographical information, *Sample* indicates if a firm is public (1), private (2), foreign subsidiary (4) or its type is unknown (3).

[54] For more information about Orbis data consult Bvdinfo.com , accessed 10 June 2020

[55] Duke Innovation & Scientific Enterprises Research Network (DISCERN, 2020), Duke University, accessed 14 June 2022, https://zenodo.org/record/4320782

Table A 13: Clean Orbis main table

Match id	Bvd id	Company name	UO ID	UO name	City	Zip code	Sample
270398	1	pfizer	1	pfizer	port washington	53074	1
270896	2	pfizer	1	pfizer	eden prairie	55344	1
272452	3	pfizer pharma	1	pfizer	memphis	38184	1
274166	4	pfizer res dev ctr	1	pfizer	ann arbor	48105	1
275643	5	pfizer	1	pfizer	coalville	84017	1

Note: This table presents the look of the clean orbis main table, ready to be used as an input of the matching algorithm. *Bvd id* is the firm identifier, and *company name* its name. *Uo id* and *UO name* refer to the company's ultimate owner. City and Zip code are the company's geography information. Sample indicates if the firm is is public (1), private (2), foreign subsidiary (4) or unknown (3).

A.1.8.Upload the database in mySQL

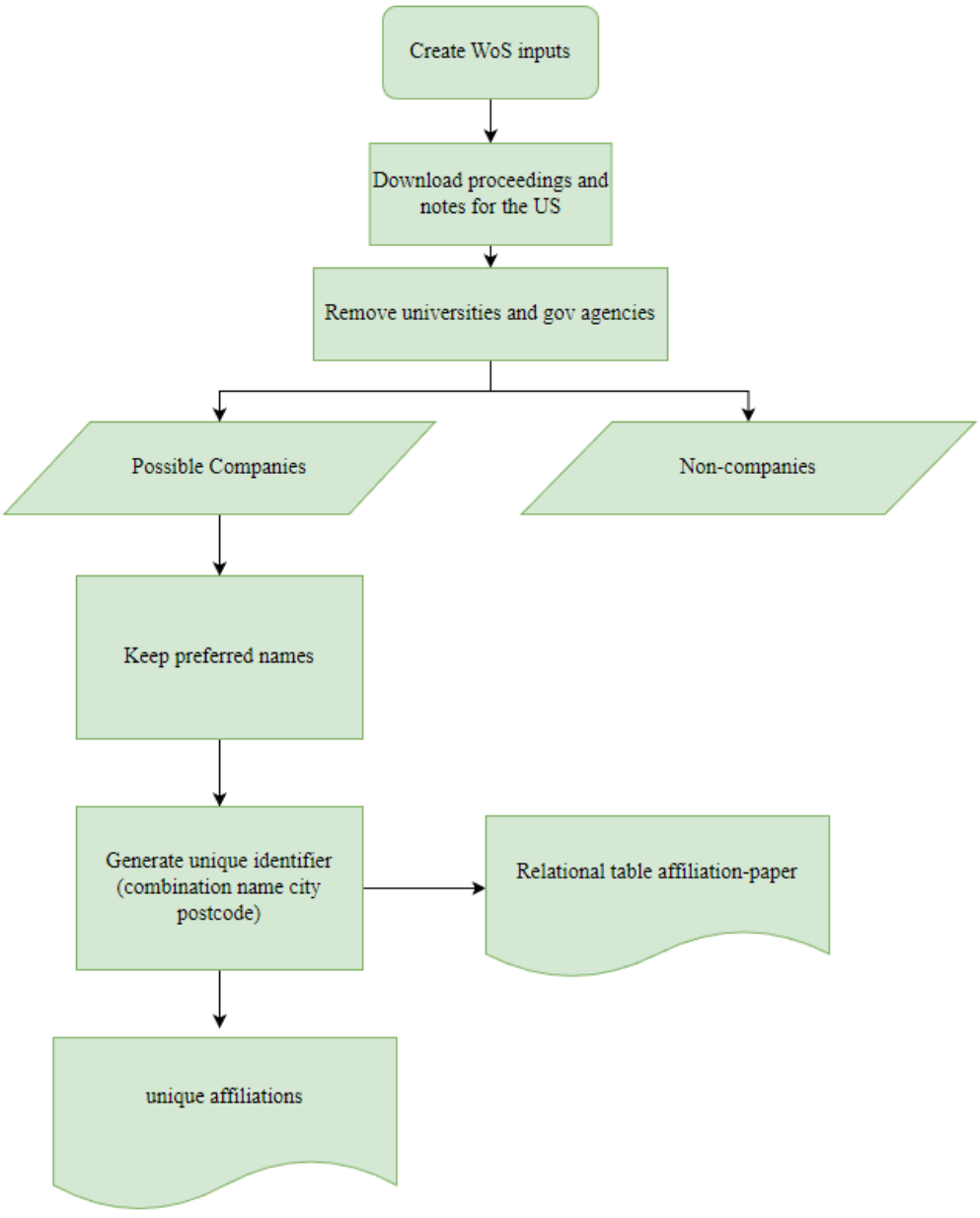
The Orbis table with *bvd id* and address information is finally uploaded on a server in mySQL. This choice allows to avoid loading a heavy 73 million rows file while performing the matching algorithm. *Match id* is set as index to allow a faster loading of the data.

.

A.2.WEB OF SCIENCE

Web of Science is one of the largest global citation database, tracking scholarly journals, books, and proceedings, across disciplines and time from over 1.7 billion cited references from over 159 million records. Specifically, I will use 10,027,418 research articles and proceedings reporting at least an US affiliation.

Figure A 3: WoS cleaning flowchart



Note: This figure shows schematically the WoS cleaning procedure and input preparation. First, I obtained the relevant publications from WoS. Then, I removed the major universities and governmental agencies. Next, I kept the preferred names of the remaining affiliations. Finally, I assigned a unique identifier to all the unique combinations of names, city and zip code. The final inputs are stored in two different tables: one containing only the unique affiliations, and the other containing a reference table that assigns the publications to its affiliation.

Precisely, the fields I selected are:

1. Articles: Reports of research on original works. Includes research papers, features, brief communications, case reports, technical notes, chronology, and full papers that were published in a journal and/or presented at a symposium or conference.
2. Proceedings: Published literature of conferences, symposia, seminars, colloquia, workshops, and conventions in a wide range of disciplines. Generally published in a book of conference proceedings.
3. Notes: A paper that mentions or remarks on a published paper on a specific subject. Generally, finds records dating back to 1996 or before (Clarivate Analytics, 2020).

WoS publications are indexed at the journal level, that is if in that specific year the journal is indexed, all the publications from that journal will be present. The scientific field is assigned to a publication at the journal level through WoS subject categories. So, all the publications from that journal will have the same subject categories assigned. Publication's authors can be linked to their affiliation, but only after 2008.

Figure A 3 shows schematically the cleaning procedures and input preparation.

A.2.1.Preferred names

WoS in the affiliation table provides two different types of names: the main name of the organisation (flagged by "org"): and "sub", a sub name that complements the first one. An example is Auburn University (org), dept phys (sub). The sub names are too generic and do not help getting a better score in the matching, thus I dropped them.

Clarivate analytics, in addition, grouped together the name variations of the biggest publishers, creating an "enhanced organisation" name, signalled by a flag "Y". I perform the matching algorithm keeping both the enhanced-organisation name and the standard WoS name. It is common to find a parent company instead of a name harmonisation in the enhanced-organisation field. Considering only the enhanced-organisation name would lead to inaccuracy in the ownership structure because the parent organisation name can be found even before the actual acquisition of the company, leading to inaccuracy in the ownership structure assigning papers of an independent company to its future acquiror.

Before the cleaning

Table A 14: WoS table before the cleaning operations

Uid	Org_type	Organisation	Pref	Addr_nr
WOS:000338106000061	Org	Auburn Univ		2
WOS:000338106000061	Org	Auburn University	Y	2
WOS:000338106000061	Org	Auburn University System	Y	2
WOS:000338106000061	Sub	Dept Phys		2
WOS:000338106000061	Org	IBM TJ Watson Res Ctr		3
WOS:000338106000061	Org	International Business Machines	Y	3

After the cleaning

Table A 15: WoS table after the cleaning operations

uid	Org_type	Organisation	Pref	Addr_nr
WOS:000338106000061	Org	auburn univ		2
WOS:000338106000061	Org	auburn university	Y	2
WOS:000338106000061	Org	auburn university system	Y	2
WOS:000338106000061	Org	international business machines	Y	3
WOS:000338106000061	Org	ibm tj watson res ctr		3

A.2.2.Flag Universities and government agencies

To reduce the amount of possible matching I start the cleaning procedures flagging and removing the major universities and government agencies using the keywords of Table A 16. It has to be remarked that those keywords are not exhaustive and do not identify all non-corporate entities. A more thorough cleaning procedure is performed post-match and it is presented in Section A.4.2.

Table A 16: Government and University keywords

Government	**University**
American cancer society	College
House us representatives	Faculty
National institutes of health	University
National oceanic atmospheric administration	Institute of technology
Oak ridge national laboratory	School
Research laboratory army	
Research laboratory navy	
United states air force	
United states army	
United states department of energy	
United states department of agriculture	
United states department of defense	
United states department of protection	
United states geological survey	
Us naval academy	
Us bureau + labor/statistics/mines/census	
Us food drug administration	
Us geol survey	
Us military academy	
Us nuclear regulation commission	

Note: This table presents the keywords used to identify government agencies and organisations and universities and research institutes

All the remaining affiliations are potential firms and will be the input of the matching algorithm. I generate a unique identifier (*my_wos_id*) grouping by organisation name, city and zip code, then I drop the duplicates. Tokenisation and abbreviations are performed within the match and not before as in the case of the Orbis data. Before starting the match, I last changed manually some names of big publishers that had name variations that was leading to unsatisfactory results (see Table A 17).

Table A 17: Whole string substitution

Analytica International	Analyt Int	IBM	International Business Machines (IBM)
Ashima	Ashima res	Immunex	Immunex research development
Att	AT&Ts Olymp Efforts	Kaiser permanente	Kaiser Permanente Div Res
Avaya	Avaya Labs Res	Kaiser permanente	No Calif Kaiser Permanente
Bbn tech	Bolt beranek newman	Kaiser permanente	Kaiser Permanente Georgia
Bp Amoco chem	Amoco chem corporation	Labcorp America	Lab Corp Amer
dionex	Dionex Chem Corp	Machine Intelligence Research Labs	MIR Labs
Disney research Pittsburgh	Disney res	Marlow industries	Marlow Ind Inc
Eastman kodak	Kodak res labs	Motorola	Motorola labs
Ei du pont de nemours	Dupont	Mountain whisper light	Mt Whisper Light Stat Consulting
Electrical Geodesics	Elect Geodes	Northrop Grumman	Grumman aerosp
Equity engineering group	Equ Engn Grp	Rockwell automation	Rockwell international science center
Fairchild semiconductor	Fairchild research center	Rockwell international	Rockwell science
Fairway medical technologies	Fairway Med Technol Inc	Sarnoff	David sarnoff research center
General atomics	General Atomics & Affilated Companies	Shell international exploration production	Shell Int Explorat & Prod Inc
General electric capital	GE capital	Sirtris	Sirtris pharmaceuticals
Hitachi San Jose Research Center	San Jose Res Ctr	Spectra phys	Spectra diode labs
Honeywell	Honeywell Labs	Strategy solutions	Strateg Solut Inc
Honeywell international	Honeywell technology center	Xerox	Xerox parc
Hughes aircraft	Hughes stx	Yahoo	Yahoo labs

Note: This table presents the stings that were used to manually modify company names in web of science.

Using again the previous provided example, the final input will take on the following form.

Table A 18: Final WoS input table

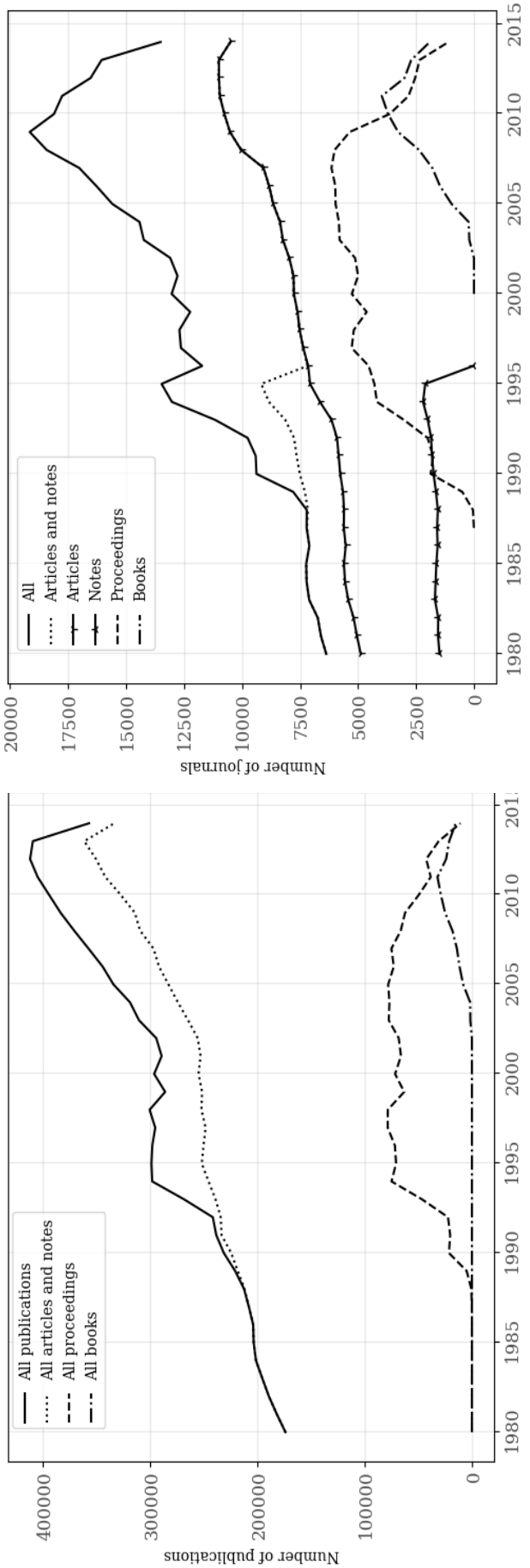
uid	Org_type	Organisation	Pref	Addr_nr
WOS:000338106000061	Org	international business machines	Y	3
WOS:000338106000061	Org	ibm tj watson res ctr		

A.2.3.Coverage

The next two Figures show information on WoS coverage. Figure A 4 shows the number of publications (left) and the number of journals in the whole WoS. Conference proceedings start being indexed in 1990, while books only from 2005. Articles are always indexed since the beginning of the sample in 1980.

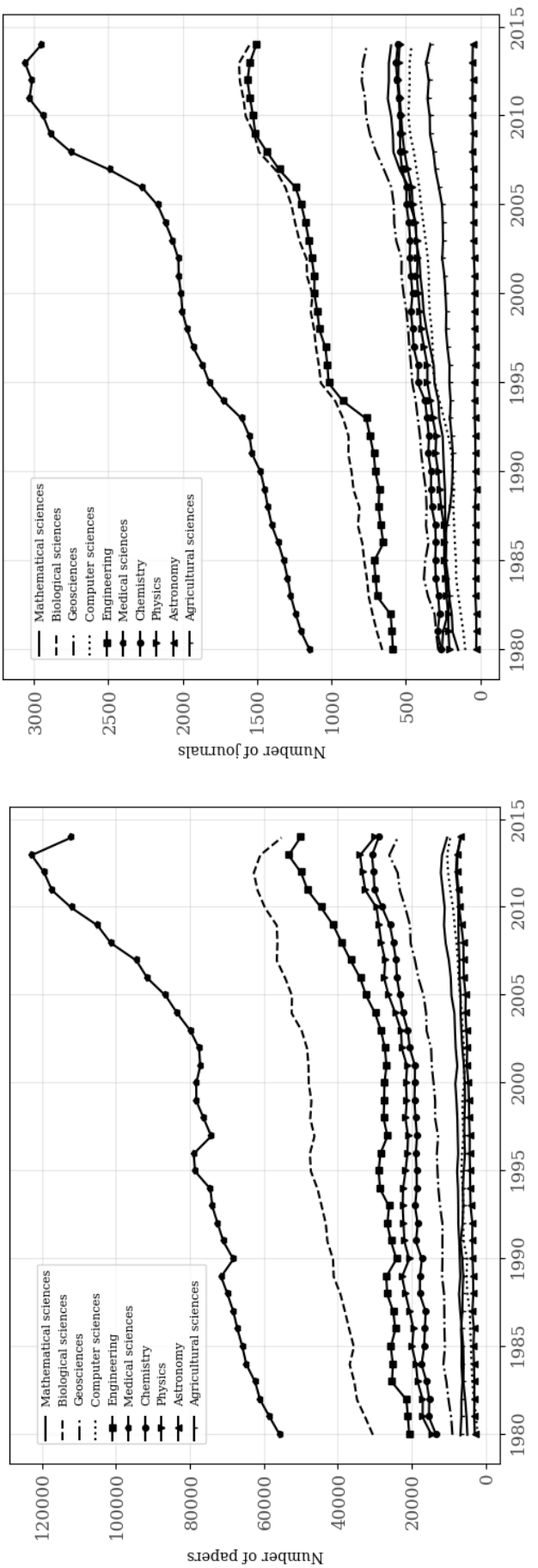
Figure A 5 shows the number of papers (left) and the number of journals (right) excluding books and proceedings. The WoS subject categories are grouped in broad fields using Milojević (2020) classification. As Clarivate analytics itself claims, WoS as better coverage in natural sciences, health sciences and engineering. Medical sciences papers and journals increase more rapidly than the other subjects, followed by biological sciences and engineering.

Figure A 4: Number of publications (left) and journals (right) in WoS by document type, 1980-2014



Note: These plots present the coverage in WoS from 1980 to 2014. In the left plot is displayed the number of publications by document type, while in the right one the number of journals by document type. The document types are Articles and notes, conference proceedings, and books.

Figure A 5: Number of publications (left) and journals (right) by broad fields, 1980-2014

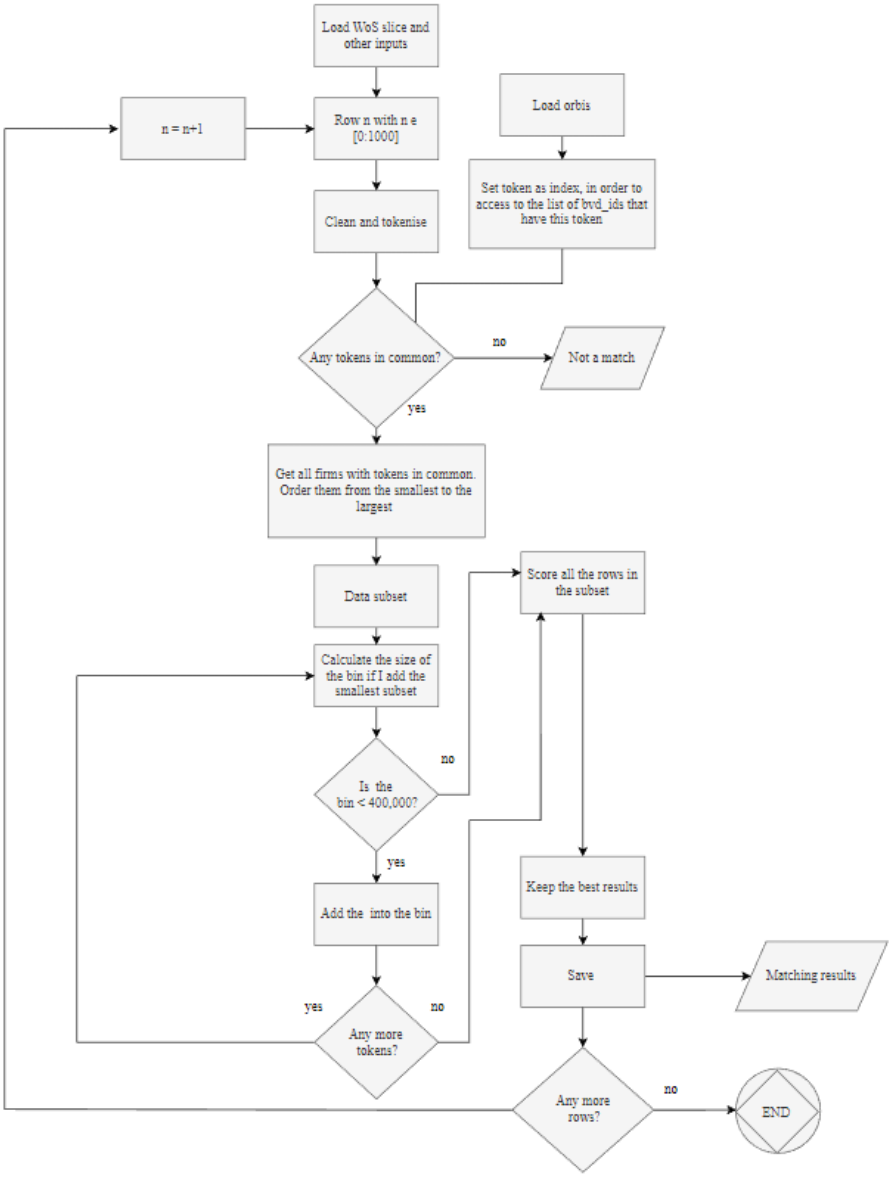


Note: This figure shows the number of publications (left) and journals (right) in WoS. The different markers indicate the broad scientific field the publications or journals belong. The scientific fields are grouped following Milojević (2020).

A.3.MATCHING ALGORITHM

In the following sections I will explain all the steps of the matching algorithm. At the beginning, I loaded the Orbis main table (Table A 13), the tokens (Table A 9), the English dictionary, the abbreviations, and a 1000 rows WoS chunk. I proceeded in the following steps row by row. First, I cleaned the WoS name, city, and zip code. Second, I tokenised it. Third if I have tokens in common with the Orbis Tokens file I created a subset (bin) with all the Orbis rows that have at least one token in common. Last, I scored the rows in the bin and I kept only the best matches.

Figure A 6: Matching algorithm flowchart



Note: This flowchart shows the matching algorithm steps more in detail. The starts loading the inputs (1000 rows slice from WoS, Orbis tokens, language dictionary, abbreviations and stopwords). Then it selects the first row (n=1).After cleaning and tokenising the WoS input it creates the bin using the firms in web of science with at least one word in common. Afterwards the algorithm calculates the matching scores on bins up to 400,000 rows. Last the best results are selected and saved. The procedure is repeated until n=1000.

A.3.1.Input slicing and parallel job structure

I ran the match in three parts. First, I matched my articles on public firms. The affiliations that did not match were then matched with public firms. Finally, the remaining unmatched affiliations were matched with firms that do not have ownership information.

I performed the match in three steps to improve the accuracy of the matches. Branches of large companies do not have ownership information in Orbis. Those branches would match with the branch rather than the parent company if I pool all companies together. Even if theoretically the match is correct, those branches are difficult to handle when using the database for empirical analysis. Many branches (thus large firms) would be systematically labelled as small companies, overestimating their contribution to corporate science.

The WoS input file contains 545,919 affiliations, defined as unique combinations of affiliation name, city, and zip code. I divided those affiliations in 561 csv files containing each 1000 affiliations with the same starting letter, as follows.

Table A 19: Index file

File name	Total entries	Number of slices	Min_row	Max_row
/fred/oz077/../a.csv	63456	64	0	1000
/fred/oz077/../a.csv	63456	64	1000	2000
..				
/fred/oz077/../b.csv	39583	40	0	1000
/fred/oz077/../b.csv	39583	40	1000	2000
..				
/fred/oz077/../c.csv	79555	80	0	1000
/fred/oz077/../c.csv	79555	80	1000	2000

In order to run multiple matches at the same time I used a parallel job structure submitting an array of python files that runs the matching algorithm on any 1000 rows chunk independently. After having loaded a 1000 rows chunk I cleaned the institution name, city and zip code as in Section A.1.1. The Orbis data are already cleaned before the match. Last, I loaded the tokens table with *bvd id* and tokens (Table A 9), the WoS abbreviations, the words contained in the English dictionary, and the stopwords of Table A 3.

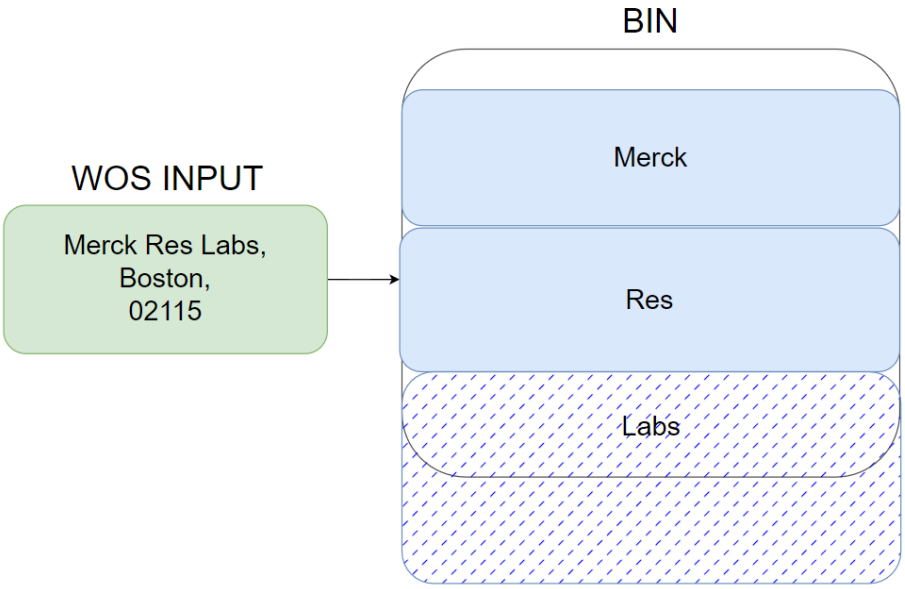
A.3.2.Binning

Grouping data in smaller groups or “bins”, according to certain characteristics, is a technique usually called binning in data science. Without binning, having 800,000 rows in WoS and 73 millions in Orbis, the algorithm would have to score 5.84x10^13 combinations. I chose to score

all the rows that have at least one token in common with the input row. Let's observe the example of Merck Research Laboratories as in Figure A 7. First, I saw how many rows in Orbis have the word merck, then res (abbreviation of research), and last labs (abbreviation of lab, laboratory and laboratories). Second, I ordered those groups from the smallest to the biggest: merck 100,000 rows, res 200,000 rows, labs 300,000 rows. I want to keep into my bin, i.e., the rows that will be scored by the algorithm, maximum 400,000 rows. Thus, I started adding the smallest group: merck. If the size of my bin is still below the 400,000 threshold, I add the second group: res. Again, if there's still space into the bin, I add the next group. In this case the group of labs (300,000) is too big to fit into the bin, so I discard it. Last, my final bin will be composed by all the rows in Orbis that contain the word merck and the word res.

The algorithm has access to the groups with a word in common through the token table (Table A 9). Once I binned, I can recover the Orbis company name, city and zip code from mySQL through the firm identifier linked to the tokens. This choice allows to read only the data necessary for each match, without the need of having a large table always loaded. A match that requires little memory allows to run more processes at the same time without incurring in memory errors. At last, I can keep 40 mySQL connections open at the same time, that means 40 contemporary matches on 40 different 100 rows chunks.

Figure A 7: Simplified visualisation of the binning



Note This figure shows schematically the binning procedure using "Merck Res Labs" as an example. On the left there is the WoS input (green), on the right bin with the Orbis firms containing the word Merck or Res. Firms with the word Labs are excluded because exceed the maximum bin size.

A.3.3.Scores

The results of the matching are stored in a dictionary, that contains all the Wos and Orbis information. The potential matches are then scored using the following scoring system.

Fuzzy score: The Wos and Orbis strings are compared using the python library "fuzzywuzzy". It calculates a figure from 0 to 100 using Levenshtein Distance[56] to test string similarity. 100 is the best possible outcome, 0 the worst.

Geography score: either

- **City score**: If two input strings share the same city, I assign to the city match a score of 1.
- **Zip score**: If two input strings share the same zip code, I assign to the zip match a score of 1.

Non-dictionary score: First, I added all the WoS abbreviations to my English dictionary, so that they do not result as non-dictionary words. Then I added one to the non-dictionary score for every non-dictionary word in common.

A.4.REDUCING THE RESULTS

A.4.1.Scoring

After having scored all the rows in the bin, I created a scoring criterion to select the best match for each affiliation. I ranked all the matching algorithm outcomes with a scale from 1 to 27 (1 the best). The perfect match has a string similarity score equal to 100, same city, same zip code and a non-dictionary word. Afterward, I proceeded ranking the scores using different combinations of string similarity scores, geography, and shared non-dictionary words in the following way. I consider 100% string similarity score, non-dictionary words and shared zip code (2). Then 100% string similarity score, non-dictionary words and shared city (3), 100% string similarity score and shared geography (4), 100% string similarity score and shared zip code (5), 100% string similarity score and shared city (6), 100% string similarity score and non-dictionary words (7). The same mechanism applies to the scoring interval from 99% to 90% (8-13). Intervals 14 and 15 require string similarity from 90% to 99% and shared non-

[56] "**Definition: (1)** The smallest number of insertions, deletions, and substitutions required to change one *string* or *tree* into another. (2) A $\Theta(m \times n)$ algorithm to compute the distance between strings, where m and n are the lengths of the strings" (NIST, 2019)

dictionary words, or just 100% string similarity. From interval 16 to 27, instead, it is required the presence a non-dictionary word and at least city or zip code in common. The scoring proceeds as follows: 89% to 80% (16-18) , 79% to 70% (19-21), , 69% to 60% (22-24), and 59% to 50% (25-27). All the remaining cases are considered as unmatched.

Table A 20: Scores of the matching algorithm outcomes

C	**CZD**	**ZD**	**CD**	**CZ**	**Z**	**C**	**D**	**None**
Fuzzy = 100	1	2	3	4	5	6	7	15
90≤fuzzy<100	8	9	10	11	12	13	14	
80≤fuzzy<90	16	17	18					
70≤fuzzy<80	19	20	21					
60≤fuzzy<70	22	23	24					
50≤fuzzy<60	25	26	27					

C = same city, Z = same zip code, D non-dictionary word

I kept all the scores until 27. The presence of non-dictionary words and common geography are a strong indicator of a possible match and allow to have good precision also in matches with lower similarity scores. Big firms have very recognisable names and may have a lot of sub names, as follows:

Table A 21: Matching example

WoS Name	**City**	**Zip Code**	**Orbis Name**	**City**	**Zip Code**	**Fuzzy Score**	**Score**
acnielsen analyt serv	minneapolis	55426	acnielsen	Minneapolis	55401	60	22

A.4.2.Remove non-corporate entities

Even if I removed universities and the main governmental agencies Orbis still contains other government agencies, universities and not for profit corporations. To identify them I proceed implementing a more thorough keyword search. I complement the keyword search suing Orbis companies' web addresses.

A.4.2.1.Keywords

Most of the words chosen relate specifically to a type of organisation. When keywords are not enough in identifying the non-corporates, I remove manually the biggest publishers.

Table A 22: Government keywords

Government			
DoD	Comm	Hlt syst	State laboratory
FBI	Commis	Hlth system	State labs
Langley Research Center	Commiss	Lib congress	Survey
NCAR	Commission	Medical examiner	Us fda
NCGR	Committee	Metropolitan	Vamc
NIOSH	Congress	Nasa	Vet affairs
NMFS	Coroner	National archives	Veteran
NOAA	Council	Natiobal center	Warfare center
NSF	County	National gallery	Water conservat labs
Administration	Department	National pk	Water dist
Afrl	Dept	National vet service labs	Water management
Agcy	District	Navy	Zool pk
Agency	Division wildlife	Observatory	
Air force	Federal reserve	Police	
Bureau	Fhwa	Publ	
Circuit	Health care system	Public	
City of	Health network	Senate	
Coast guard	Health system	Smithsonian	
comiss	Healthcare system	spawar	
	Highway administration	State lab	

Table A 23: Specific government institutions

Specific institutions	
Ames research	Johnson space flight center
Ames research center	National centers for coastal ocean science noocs
Bronx psychiat center	National radio astronomy observatory
Centers for disease control prevention	New York city poison control center
Cold spring harbor laboratory	New York state off mental health
Connecticut mental health center	New York state mental retardant development disabil
David w taylor naval ship research development center	Rocky mountain research station
Edgewood chem biology center	Sheep experiment station
Forestry experiment station	So reg research center
Framingham heart disease epidemiol study	Supercond super collider labs
Glenn research center	Us fish wildlife service
Goddard space flight center	Usa hcg reference service
Handford engine dev laboratory	

Table A 24: International organisations keywords

International Organisations			
world bank	save children	IMF	Interamer Dev Bank
united nations	unesco	WHO	WWF
unicef	world wildlife fund	UNDP	

Table A 25: Medical centres keywords

Medical centres		
Aids	Mayo clinic	Scripps health
Blood center	Mayo system	Sinai
Cancer center	Med center	Spine center
Childrens	Medical center	Trauma center
Clinic	MGH	Urol
Headache center	Oncology	
Heart center	Pediat oncol group	
Heart center	Prostate	
Hospital	Rehabilitation center	
Jewish	Rheumat	

Table A 26: Specific medical centres

Specific institutions		
Arizona health science center	Group health cooperat Puget sound	Strang cancer prevent center
Eunice kennedy shriver center mental retardant	New York eye ear infirmary	Sw oncol group

Table A 27: Not-for-profit keywords

Not for profit			
Aarp	Cen	Pharmacopeia	Society
Accelerator	Church	PRBO	Supercomp
	Consortium	Primate	Telescope
Afl cio	Draper lab	Program	VA
Alcohol research group	Foundation	RAND	Vanderbilt
Aquarium	Ieee	Red cross	Volunteer
Arema	Institute	RTOG	Volunteers
Association	Library	SAMSI	ymca
Botanic garden	Museum	Scholars	zoo
Botanical gardens	Organisation	Seminary	
Carnegie institute	PATH	SISSA	

Table A 28: Specific not-for-profit

Specific institutions	
Burroughs wellcome	Jackson laboratory
Center strategy international studies	Itamp
Charles f kettering research labs	Microelectr center n Carolina
East west center	Rocky mountain biology labs
Fhi 360	Rti international
Fisheries	Santa Barbara research center
Souther California coastal water research project	Sri international
Geophys labs	The citadel
GIA lab	The scientist
Houston advanced research center	Welcome research labs
Intermountain health care	

Table A 29: Educational institutions keywords

Educational Institutions				
Acad	Iit	Polytechnic	Stanford syncroton	Uab
Assoc	Library	PRBO	Sunarc	Ucla
Caltech	MIT	RAND	Suny	Undp
Center astrophys	NCGR	RTOG	Supercomputing	upmc
Channing labs	Nshe	SAMSI	Texas agrilife	
Educ	Nyu	Scientist	Texas am	
educational	PATH	SISSA	Theol	

Table A 30: Specific educational institutions

Specific institutions			
Becton Dickinson immunocytometry system	Ek shiver center	GIA lab	Stanford syncroton radiat lightsource
Cancer therapy research center	Emory eye center	Labs phys science	Western Carolina center
Center advanced study behaviour science	Forestry science labs	Maryland psychiatric research center	Wright labs
Eastman dental center	Friday harbor labs	Office naval research	

A.4.2.2.Orbis Web addressess

If the web address ends with .org, I assume it is a not-for-profit institution, if .edu a university/educational entity, if .gov a government agency.

A.5.DIAGNOSTICS

In the following section I am going to present some diagnostics about the precision of the algorithm. Precision is defined as the ratio of true positives and total cases identified as positives ($\frac{true\ positives}{true\ positives + false\ positives}$). By true positives I mean here the correct matches, while false positives are matches wrongly produced by the algorithm.

Unfortunately, I am unable to calculate the recall of the algorithm because I do not possess a golden set that allows me to calculate the false negatives. While possible, the creation of a golden set would require significant manual work and the hiring of research assistants to guarantee fairness in the valuation of the false negatives.

The matching algorithm presents problems handling the following scenarios:

1. When an undetected abbreviation does not bin on a Orbis token
2. When an abbreviation is considered a non-dictionary word
3. When WoS strings are long and with many uninformative tokens. Ex: “johnson johnson ctr study pediat psychopathol”
4. When the company name is short ex: “hcs inc.”

A.5.1.Precision

Precision is calculated as the ratio of true positives and the sum of true positives and false positives. I selected a stratified random sample of 100 matched affiliations per scoring interval and I checked manually for true positives and false positives. When affiliation names are identical, and geographical information coincide, the assessment of the truthfulness of the match is self-explanatory. However, in instances where the assessment was less evident, I performed an internet search to verify whether the two affiliations were indeed the same.

Table A 31: Precision in a random sample of 100 affiliations per scoring interval

Score	N affiliations	N publications	% True positives	Precision (cumulative)
1-5	122,846	656,533	100	100%
6-11	102,494	323,643	94.00	97.27%
12-13	14,065	29,489	87%	96.66%
16-21	16,772	43,446	90%	96.23%
22-27	12,746	33,914	67%	94.84%
14-15	74,532	179,889	?	?

Note: This table shows the precision for a random sample of 100 matches. Each row represents a different scoring interval. The second column shows the % of correct matches within each scoring interval, while the last column displays the cumulative precision from score one up to the given scoring interval.

In Table A 31 I show the accuracy of matches within every scoring interval, and the cumulative precision. Matches in the scoring interval 1-5 are almost error free. Precision lowers to 97.27% for the 6-11 interval, to 96.66% in the 12-13 interval, to 96.23% in the 16-21 interval and finally to 94.84 in the 22-27 interval. It is important to remark that the scoring interval 1-11 captures 77.82% of publications and 64.40 % of the affiliations with a precision of 97.27%. Assessing the precision of the scoring interval 14-15 presents a more intricate challenge. Given the absence of shared geographical data, the only method to assess the quality of the match involves a manual internet search for both company names. However, despite implementing this method, occasionally the correctness of the match cannot be assessed with certainty, especially for older or lesser-known firms. While these matches are included into the sample, they necessitate further scrutiny and examination.

B. Appendix Chapter 2

Table B 1: Distribution of subjects

# subjects	#papers
1	36,424,090
2	14,722,774
3	5,335,424
4	1,344,302
5	341,165
6	75,713
7	9,444
8	5,577
9	2,881
10	318

Note: The table shows the distribution of subjects. Almost half of the subjects are composed by only subjects, while most of the subjects have less than 5 subjects.

Table B 2: Impact of Speed on the probability of collaborating, multi field and multi company. LPM

	(1) Collaboration	(2) Collaboration
Speed	0.351*** (0.0120)	0.499*** (0.0170)
Observations	387,007	364,492
R-squared	0.016	0.208
Year FE	Yes	Yes
Firm FE	No	Yes
Sc. Field FE	No	No

Note: This table presents the regression results of equation $Collaboration_{ijft} = \beta_0 + \beta_1 speed_{\ j} + \lambda_f + \mu_t + + u$ introducing different level of fixed effects: year (column 1), year and firm (2). Papers with multiple companies and scientific fields are included.

Table B 3: Impact of IRA and Avg Auth on the probability of collaborating, multi field and multi company. LPM

VARIABLES	(1) Collaboration	(2) Collaboration	(3) Collaboration	(4) Collaboration
IRA	2.433***	2.967***		
	(0.0416)	(0.0592)		
Avg Auth			0.0535***	0.0652***
			(0.000587)	(0.000898)
Observations	387,007	364,492	387,007	364,492
R-squared	0.022	0.212	0.032	0.218
Year FE	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes
Sc. Field FE	No	No	No	No

Note: This table presents the regression results of equation $Collaboration_{ijft} = \beta_0 + \beta_1 IRA_{\ j} + \lambda_f + \mu_t + + u$ and $Collaboration_{ijft} = \beta_0 + \beta_1 Avg\ Auth_{\ j} + \lambda_f + \mu_t + + u$ introducing different level of fixed effects: year (column 1 and 3), year and firm (2 and 4). Papers with multiple companies and scientific fields are included.

Table B 4: Impact of Speed on the probability of collaborating with interaction term, multi field and multi company. LPM

VARIABLES	(1) Collaboration	(2) Collaboration	(3) Collaboration	(4) Collaboration
Speed	0.457***	0.759***	0.605***	
	(0.0135)	(0.0304)	(0.0468)	
Firm Size	-0.0124***	0.00426***		
	(0.000217)	(0.00161)		
Speed*Firm Size		-0.0386***	-0.00883*	4.27e-05
		(0.00365)	(0.00514)	(0.00516)
Observations	322,584	322,584	309,749	309,254
R-squared	0.025	0.025	0.178	0.320
Year FE	Yes	Yes	Yes	Yes
Firm FE	No	No	Yes	Yes
Sc. Field FE	No	No	No	Yes

Note: This table presents the regression results of equation $Collaborations_{ijft} = \beta_0 + \beta_1 speed_{\ j} + \beta_2 \text{Firm size}_f + \beta_3 speed * \text{Firm size}_i + \lambda_f + \mu_t + \gamma_j + u$ introducing different level of fixed effects: year (column 1 and 2), year and firm (3) and year firm and field (4). Papers with multiple companies and scientific fields are included.

Table B 5: Impact of Speed on the probability of collaborating with interaction term, multi field and multi company. SMEs vs Large firms. LPM

	<250 employees			≥250 employees		
	(1) Collaboration	(2) Collaboration	(3) Collaboration	(4) Collaboration	(5) Collaboration	(6) Collaboration
Speed	0.715***	-0.217**		0.838***	1.552***	
	(0.0585)	(0.0947)		(0.0869)	(0.116)	
Firm Size	-0.00979			0.0116***		
	(0.00745)			(0.00384)		
Speed*	-0.0196	0.194***	0.0316	-0.0470***	-0.0972***	-0.0207*
Firm size	(0.0170)	(0.0276)	(0.0268)	(0.00858)	(0.0111)	(0.0108)
Observations	96,676	85,159	84,697	225,908	224,590	224,164
R-squared	0.029	0.411	0.477	0.013	0.096	0.272
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	Yes	No	Yes	Yes
Sc. Field FE	No	No	Yes	No	No	Yes

Note: This table presents the regression results of equation $Collaboration_{ijft} = \beta_0 + \beta_1 speed_{\ j} + \beta_2 \text{Firm size}_f + \beta_3 speed * \text{Firm size}_{\ i} + \lambda_f + \mu_t + \gamma_j + u$ discerning by firms size. Columns 1-2-3 show the results for SMEs while 4-5-6 for large firms. Speed is time invariant and calculated as the average speed from 2000 to 2001. Papers with multiple companies and scientific fields are included.

Table B 6: Impact of Speed on the probability of collaborating with interaction term, multi field and multi company. LPM with quadratic term

VARIABLES	(1) Collaboration	(2) Collaboration
Speed	-0.647***	
	(0.0856)	
Speed*Firm Size	0.422***	0.0337
	(0.0280)	(0.0270)
Speed*Firm Size^2	-0.0287***	-0.00226
	(0.00191)	(0.00184)
Observations	309,749	309,254
R-squared	0.178	0.320
Year FE	Yes	Yes
Firm FE	Yes	Yes
Sc. Field FE	No	Yes

Note: This table shows the results of equation $Collaboration_{ijft} = \beta_0 + \beta_1 speed_j + \beta_2 Firm\ Size_f + \beta_3 Firm\ Size_f^2 + \beta_4 speed_{ij} * Firm\ Size_f + \beta_5 speed_j * Firms\ Size_f^2 + \lambda_f + \gamma_t + \mu_j + u$. This model presents a double interaction, one with *Firm size* and one with *Firm size*^{2}. Papers with multiple companies and scientific fields are included.

Table B 7: Impact of IRA on the probability of collaborating with interaction term, multi field and multi company. SMEs vs Large firms. LPM

	<250 employees			≥250 employees		
	(1) Collaboration	(2) Collaboration	(3) Collaboration	(4) Collaboration	(5) Collaboration	(6) Collaboration
IRA	3.434***	-0.837**		5.066***	6.244***	
	(0.216)	(0.346)		(0.316)	(0.423)	
Firm Size	-0.0163**			0.0202***		
	(0.00643)			(0.00334)		
IRA*	-0.0290	0.912***	0.257***	-0.270***	-0.284***	-0.0347
Firm Size	(0.0610)	(0.0923)	(0.0925)	(0.0311)	(0.0403)	(0.0388)
Observations	96,676	85,159	84,697	225,908	224,590	224,164
R-squared	0.038	0.414	0.477	0.019	0.102	0.272
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	Yes	No	Yes	Yes
Sc. Field FE	No	No	Yes	No	No	Yes

Note: This table presents the regression results of equation $Collaboration_{ijft} = \beta_0 + \beta_1 IRA_j + \beta_2 \text{Firm size}_f + \beta_3 IRA * \text{Firm size}_i + \lambda_f + \mu_t + \gamma_j + u$ discerning by firms size. Columns 1-2-3 show the results for SMEs while 4-5-6 for large firms. Speed is time invariant and calculated as the average speed from 2000 to 2001. Papers with multiple companies and scientific fields are included.

Table B 8: Impact of Avg auth on the probability of collaborating with interaction term, multi field and multi company. SMEs vs Large firms. LPM

	<250 employees			≥250 employees		
	(1) Collaboration	(2) Collaboration	(3) Collaboration	(4) Collaboration	(5) Collaboration	(6) Collaboration
Avg Auth	0.0772***	0.00592		0.121***	0.106***	
	(0.00332)	(0.00537)		(0.00453)	(0.00657)	
Firm Size	-0.00739			0.0293***		
	(0.00464)			(0.00211)		
Avg Auth*	-0.00244**	0.0112***	0.00350**	-0.00777***	-0.00294***	-0.000422
Firm Size	(0.000954)	(0.00149)	(0.00149)	(0.000440)	(0.000614)	(0.000597)
Observations	96,676	85,159	84,697	225,908	224,590	224,164
R-squared	0.053	0.414	0.477	0.023	0.112	0.272
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	Yes	No	Yes	Yes
Sc. Field FE	No	No	Yes	No	No	Yes

Note: This table presents the regression results of equation $Collaboration_{ijft} = \beta_0 + \beta_1 Avg\ Auth_j + \beta_2 \text{Firm size}_f + \beta_3 speed * \text{Firm size}_i + \lambda_f + \mu_t + \gamma_j + u$ discerning by firms size. Columns 1-2-3 show the results for SMEs while 4-5-6 for large firms. Speed is time invariant and calculated as the average speed from 2000 to 2001. Papers with multiple companies and scientific fields are included.

Table B 9: Impact of IRA on the probability of collaborating with interaction term, multi field and multi company. LPM with quadratic term

	(1) Collaboration	(2) Collaboration
IRA	-2.789***	
	(0.371)	
IRA* Firm Size	1.894***	0.170*
	(0.109)	(0.103)
IRA* Firm Size^2	-0.122***	-0.00847
	(0.00710)	(0.00678)
Observations	309,749	309,254
R-squared	0.184	0.320
Year FE	Yes	Yes
Firm FE	Yes	Yes
Sc. Field FE	No	Yes

Note: This table shows the results of equation $Collaboration_{ijft} = \beta_0 + \beta_1 IRA_j + \beta_2 Firm\ Size_f + \beta_3 Firm\ Size_f^2 + \beta_4 IRA_{ij} * Firm\ Size_f + \beta_5 IRA_j * Firm\ Size_f^2 + \lambda_f + \gamma_t + \mu_j + u$. This model presents a double interaction, one with *Firm Size* and one with *Firm Size*^{2}. Papers with multiple companies and scientific fields are included.

Table B 10: Impact of Avg auth on the probability of collaborating with interaction term, multi field and multi company. LPM with quadratic term

VARIABLES	(1) Collaboration	(2) Collaboration
Avg Auth	-0.0353***	
	(0.00553)	
Avg Auth* Firm Size	0.0298***	0.0107***
	(0.00162)	(0.00156)
Avg Auth* Firm Size^2	-0.00178***	-0.000574***
	(0.000105)	(0.000101)
Observations	309,749	309,254
R-squared	0.191	0.320
Year FE	Yes	Yes
Firm FE	Yes	Yes
Sc. Field FE	No	No

Note: This table shows the results of equation $Collaboration_{ijft} = \beta_0 + \beta_1 Avg\ Auth_j + \beta_2 Firm\ Size_f + \beta_3 Firm\ Size_f^2 + \beta_4 Avg\ Auth_{ij} * Firm\ Size_f + \beta_5 Avg\ Auth_j * Firm\ Size_f^2 + \lambda_f + \gamma_t + \mu_j + u$. This model presents a double interaction, one with *bigger* and one with *bigger*^{2}. Papers with multiple companies and scientific fields are included.

C. Appendix Chapter 3

Table C 1: Summary of the metrics

Appliedness	
A&J metric: distance from the dual frontier	If a publication is cited directly by a patent, We consider the patent-publication distance equal to one (D=1). Keeping with A&J's vocabulary, we define the collection of patents and publications at D=1 as the "dual frontier", that if the zone of immediate contact between science and technology. If a publication is cited by a publication that is cited by a patent, we assign distance two (D=2). Any publication cited by other publications along a chain ultimately leading to the dual frontier, that is to a publication directly cited by a patent, will stand at distance D=k from the frontier, where k is the number of publications along the chain. A publication that is never linked to a patent is considered as unlinked (or at infinite distance; D=∞).
Basicness	
Generality: concentration (Rao Stirling index) of subjects of citing publications	$1 - \sum_{i,j=1}^{N_i} (s_{ij} p_i p_j)$ Adaptation of Trajtenberg et al. (1997) metric. Basicness is a Rao Sterling index (that captures diversity) of the publications citing the focal publication. p_i is the proportion of references citing the Science Categories (SC) i among all children papers and s_{ij} is the cosine measure of similarity between SCs i and j. The cosine s_{ij} ensures that SC far from each other have a higher weight. N is the total number of SC of all the citing publications.

Table C 2: Corporate and collaborations appliedness and basicness, 1980, 2014

	Corporate			Collaboration	
	(1) Appliedness	(2) Basicness		(3) Appliedness	(4) Basicness
Corp	0.0300***	0.0075***	Collab	0.0164***	0.0165***
	(0.0007)	(0.0010)		(0.0010)	(0.0014)
Corp*t2	0.0034***	-0.0029**	Collab*t2	0.0031***	0.0029
	(0.0009)	(0.0013)		(0.0012)	(0.0018)
Corp*t3	-0.0008	-0.0092***	Collab*t3	0.0022*	0.0005
	(0.0008)	(0.0013)		(0.0011)	(0.0016)
Corp*t4	0.0002	-0.0046***	Collab*t4	0.0039***	0.0019
	(0.0009)	(0.0013)		(0.0011)	(0.0016)
Corp*t5	0.0033***	-0.0037***	Collab*t5	0.0046***	-0.0001
	(0.0009)	(0.0014)		(0.0011)	(0.0016)
Corp*t6	0.0075***	-0.0144***	Collab*t6	0.0032***	-0.0069***
	(0.0010)	(0.0014)		(0.0010)	(0.0015)
Corp*t7	0.0134***	-0.0298***	Collab*t7	0.0076***	-0.0136***
	(0.0011)	(0.0017)		(0.0011)	(0.0016)
Constant	0.7645***	0.3902***	Constant	0.7659***	0.3902***
	(0.0002)	(0.0003)		(0.0002)	(0.0003)
Observations	6,656,642	6,163,433	Observations	6,656,642	6,163,433
R-squared	0.2713	0.2248	R-squared	0.2707	0.2250
Journal FE	YES	YES	Journal FE	YES	YES

Note: The table shows the regression results of the model $Y_{\text{itj}} = \beta_o + \beta_1 corp_i + \beta_2 \sum_{t=1}^{7} \text{I(year} = \text{t)} + \beta_3 corp * \sum_{t=1}^{7} * \text{I(year} = \text{t)} + \mu_j + u_{itj}$ and $Y_{\text{itj}} = \beta_o + \beta_1 collab_i + \beta_2 \sum_{t=1}^{7} \text{I(year} = \text{t)} + \beta_3 collab * \sum_{t=1}^{7} * \text{I(year} = \text{t)} + \mu_j + u_{itj}$. Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table C 3: Corporate science appliedness by 10 broad fields, 1980-2014

	(1) Agricultural Sciences	(2) Astronomy	(3) Biological Sciences	(4) Chemistry	(5) Computer Sciences	(6) Engineering	(7) Geosciences	(8) Mathematical Sciences	(9) Medical Sciences	(10) Physics
Corp	0.0009	-0.0114*	0.0337***	0.0274***	0.0085**	0.0274***	0.0036	0.0242***	0.0219***	0.0384***
	(0.0047)	(0.0069)	(0.0027)	(0.0015)	(0.0041)	(0.0014)	(0.0028)	(0.0055)	(0.0020)	(0.0014)
Corp*t2	0.0081	-0.0035	0.0086***	0.0025	0.0184***	0.0055***	0.0020	0.0052	0.0037	0.0089***
	(0.0066)	(0.0097)	(0.0033)	(0.0021)	(0.0053)	(0.0018)	(0.0039)	(0.0078)	(0.0026)	(0.0019)
Corp*t3	0.0206***	0.0202*	0.0086***	0.0038*	0.0193***	-0.0003	-0.0032	0.0081	0.0087***	0.0030
	(0.0063)	(0.0105)	(0.0031)	(0.0020)	(0.0052)	(0.0018)	(0.0037)	(0.0078)	(0.0025)	(0.0019)
Corp*t4	0.0153**	0.0164	0.0063**	0.0037*	0.0267***	-0.0016	-0.0018	0.0071	0.0135***	0.0013
	(0.0065)	(0.0108)	(0.0031)	(0.0021)	(0.0051)	(0.0019)	(0.0039)	(0.0081)	(0.0024)	(0.0023)
Corp*t5	0.0196***	0.0133	0.0138***	0.0067***	0.0321***	-0.0048**	-0.0049	-0.0053	0.0147***	-0.0044*
	(0.0070)	(0.0119)	(0.0031)	(0.0021)	(0.0052)	(0.0021)	(0.0042)	(0.0087)	(0.0025)	(0.0025)
Corp*t6	0.0351***	0.0046	0.0145***	0.0010	0.0296***	0.0002	0.0108**	0.0006	0.0163***	0.0089***
	(0.0074)	(0.0139)	(0.0032)	(0.0021)	(0.0057)	(0.0023)	(0.0045)	(0.0094)	(0.0025)	(0.0028)
Corp*t7	0.0409***	0.0752***	0.0211***	-0.0036	0.0280***	0.0197***	0.0497***	0.0048	0.0138***	0.0290***
	(0.0093)	(0.0181)	(0.0035)	(0.0025)	(0.0073)	(0.0030)	(0.0066)	(0.0142)	(0.0027)	(0.0038)
Constant	0.7292***	0.6625***	0.7953***	0.7977***	0.7885***	0.7714***	0.7024***	0.6762***	0.7596***	0.7435***
	(0.0010)	(0.0008)	(0.0004)	(0.0005)	(0.0018)	(0.0006)	(0.0007)	(0.0012)	(0.0003)	(0.0005)
Observations	147,600	134,683	1,327,712	649,569	139,089	796,819	347,407	103,949	2,327,753	679,034
R-squared	0.1961	0.0735	0.2746	0.1879	0.2580	0.2059	0.1682	0.2306	0.1928	0.2593
Journal FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note: The table shows the regression results of the model $Appliedness_{\text{itj}} = \beta_o + \beta_1 corp_i + \beta_2 \sum_{t=1}^{7} \text{I(year} = \text{t)} + \beta_3 corp * \sum_{t=1}^{7} * \text{I(year} = \text{t)} + \mu_j + u_{itj}$.
Every column represents a different scientific field. Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table C 4: Collaborations appliedness by 10 broad fields, 1980-2014

	(1) Agricultural Sciences	(2) Astronomy	(3) Biological Sciences	(4) Chemistry	(5) Computer Sciences	(6) Engineering	(7) Geosciences	(8) Mathematical Sciences	(9) Medical Sciences	(10) Physics
Collab	0.0091	0.0049	0.0185***	0.0049*	0.0081	0.0075***	-0.0031	0.0153***	0.0304***	0.0128***
	(0.0059)	(0.0045)	(0.0028)	(0.0028)	(0.0061)	(0.0023)	(0.0034)	(0.0057)	(0.0023)	(0.0022)
Collab*t2	0.0081	-0.0046	0.0102***	0.0050	0.0042	0.0026	0.0074*	0.0087	0.0056**	-0.0002
	(0.0077)	(0.0060)	(0.0034)	(0.0036)	(0.0075)	(0.0030)	(0.0044)	(0.0075)	(0.0028)	(0.0028)
Collab*t3	-0.0079	-0.0035	0.0084***	0.0021	0.0083	0.0007	0.0090**	0.0052	0.0122***	-0.0025
	(0.0072)	(0.0055)	(0.0031)	(0.0033)	(0.0070)	(0.0027)	(0.0041)	(0.0071)	(0.0026)	(0.0026)
Collab*t4	0.0037	-0.0006	0.0066**	0.0071**	0.0089	0.0008	0.0084**	0.0096	0.0096***	-0.0009
	(0.0069)	(0.0053)	(0.0030)	(0.0032)	(0.0067)	(0.0026)	(0.0039)	(0.0069)	(0.0024)	(0.0026)
Collab*t5	0.0117*	0.0035	0.0053*	0.0021	0.0090	-0.0012	0.0108***	0.0124*	0.0083***	-0.0007
	(0.0066)	(0.0051)	(0.0030)	(0.0032)	(0.0066)	(0.0026)	(0.0038)	(0.0068)	(0.0024)	(0.0026)
Collab*t6	0.0103	0.0021	0.0039	-0.0017	0.0111*	-0.0008	0.0107***	0.0080	0.0042*	-0.0040
	(0.0066)	(0.0049)	(0.0030)	(0.0032)	(0.0066)	(0.0026)	(0.0037)	(0.0070)	(0.0024)	(0.0025)
Collab*t7	0.0100	0.0095*	0.0106***	0.0054*	0.0108	0.0055**	0.0167***	0.0081	0.0017	-0.0004
	(0.0072)	(0.0050)	(0.0030)	(0.0033)	(0.0069)	(0.0027)	(0.0040)	(0.0089)	(0.0024)	(0.0027)
Constant	0.7291***	0.6622***	0.7956***	0.8008***	0.7887***	0.7757***	0.7027***	0.6767***	0.7594***	0.7483***
	(0.0010)	(0.0008)	(0.0004)	(0.0005)	(0.0017)	(0.0006)	(0.0007)	(0.0012)	(0.0003)	(0.0005)
Observations	147,600	134,683	1,327,712	649,569	139,089	796,819	347,407	103,949	2,327,753	679,034
R-squared	0.1963	0.0739	0.2738	0.1849	0.2567	0.2038	0.1682	0.2311	0.1953	0.2551
Journal FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note: The table shows the regression results of the model $Appliedness_{\mathrm{itj}} = \beta_o + \beta_1 collab_i + \beta_2 \sum_{t=1}^{7} \mathrm{I(year = t)} + \beta_3 collab * \sum_{t=1}^{7} * \mathrm{I(year = t)} + \mu_j + u_{itj}$. Every column represents a different scientific field. Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table C 5: Corporate science basicness by 10 broad fields, 1980-2014

	(1) Agricultural Sciences	(2) Astronomy	(3) Biological Sciences	(4) Chemistry	(5) Computer Sciences	(6) Engineering	(7) Geosciences	(8) Mathematical Sciences	(9) Medical Sciences	(10) Physics
Corp	-0.0133	-0.0495***	0.0155***	-0.0082***	0.0183***	0.0052***	-0.0282***	0.0214**	-0.0003	0.0119***
	(0.0084)	(0.0130)	(0.0038)	(0.0021)	(0.0058)	(0.0020)	(0.0051)	(0.0098)	(0.0032)	(0.0019)
Corp*t2	0.0111	0.0625***	0.0031	-0.0022	-0.0233***	-0.0030	-0.0006	0.0123	0.0057	0.0034
	(0.0118)	(0.0185)	(0.0046)	(0.0028)	(0.0076)	(0.0027)	(0.0070)	(0.0140)	(0.0042)	(0.0026)
Corp*t3	0.0110	0.0499**	0.0052	-0.0032	-0.0238***	-0.0105***	0.0168**	-0.0165	0.0038	-0.0086***
	(0.0111)	(0.0200)	(0.0044)	(0.0027)	(0.0074)	(0.0026)	(0.0067)	(0.0141)	(0.0039)	(0.0026)
Corp*t4	-0.0008	0.0521**	-0.0044	0.0096***	-0.0189***	-0.0067**	0.0121*	0.0033	-0.0004	0.0037
	(0.0114)	(0.0208)	(0.0043)	(0.0028)	(0.0073)	(0.0028)	(0.0070)	(0.0143)	(0.0038)	(0.0031)
Corp*t5	0.0000	0.0966***	-0.0060	0.0118***	-0.0282***	-0.0134***	0.0096	-0.0135	0.0052	0.0015
	(0.0121)	(0.0226)	(0.0044)	(0.0029)	(0.0073)	(0.0029)	(0.0074)	(0.0150)	(0.0038)	(0.0033)
Corp*t6	0.0147	0.0880***	-0.0198***	-0.0047	-0.0328***	-0.0216***	0.0094	0.0051	-0.0053	-0.0097***
	(0.0125)	(0.0266)	(0.0045)	(0.0029)	(0.0078)	(0.0032)	(0.0078)	(0.0157)	(0.0038)	(0.0037)
Corp*t7	-0.0333**	0.0555*	-0.0341***	-0.0184***	-0.0543***	-0.0291***	-0.0030	-0.0568**	-0.0203***	-0.0232***
	(0.0155)	(0.0336)	(0.0049)	(0.0033)	(0.0106)	(0.0041)	(0.0114)	(0.0233)	(0.0041)	(0.0050)
Constant	0.3177***	0.2193***	0.3969***	0.3659***	0.4336***	0.3974***	0.3904***	0.3388***	0.4019***	0.3915***
	(0.0017)	(0.0015)	(0.0005)	(0.0007)	(0.0025)	(0.0009)	(0.0012)	(0.0022)	(0.0005)	(0.0007)
Observations	127,068	131,684	1,263,625	610,193	118,573	698,026	322,441	87,498	2,164,367	637,361
R-squared	0.2000	0.3534	0.1794	0.1996	0.2283	0.2089	0.2330	0.3456	0.1696	0.1556
Journal FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note: The table shows the regression results of the model $Basicness_{\text{itj}} = \beta_o + \beta_1 corp_i + \beta_2 \sum_{t=1}^{7} \text{I(year} = \text{t)} + \beta_3 corp * \sum_{t=1}^{7} * \text{I(year} = \text{t)} + \mu_j + u_{itj}$. Every column represents a different scientific field. Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table C 6: Collaborations basicness by 10 broad fields, 1980-2014

	(1) Agricultural Sciences	(2) Astronomy	(3) Biological Sciences	(4) Chemistry	(5) Computer Sciences	(6) Engineering	(7) Geosciences	(8) Mathematical Sciences	(9) Medical Sciences	(10) Physics
Collab	-0.0146	0.0056	0.0202***	0.0151***	0.0104	-0.0063*	0.0006	0.0063	0.0284***	0.0190***
	(0.0104)	(0.0085)	(0.0039)	(0.0038)	(0.0087)	(0.0033)	(0.0058)	(0.0102)	(0.0035)	(0.0029)
Collab*t2	0.0106	0.0027	0.0096**	0.0003	0.0031	0.0096**	0.0070	0.0244*	0.0059	-0.0045
	(0.0135)	(0.0112)	(0.0047)	(0.0049)	(0.0106)	(0.0042)	(0.0076)	(0.0132)	(0.0042)	(0.0037)
Collab*t3	0.0111	0.0012	0.0085*	-0.0044	-0.0158	0.0104***	0.0180**	0.0164	0.0059	-0.0080**
	(0.0125)	(0.0103)	(0.0043)	(0.0045)	(0.0100)	(0.0039)	(0.0071)	(0.0125)	(0.0039)	(0.0034)
Collab*t4	0.0268**	0.0123	0.0025	-0.0048	-0.0004	0.0083**	0.0109	0.0120	0.0044	-0.0058*
	(0.0120)	(0.0099)	(0.0042)	(0.0044)	(0.0094)	(0.0037)	(0.0067)	(0.0121)	(0.0037)	(0.0034)
Collab*t5	0.0132	0.0071	-0.0022	-0.0116***	-0.0087	0.0048	0.0093	0.0211*	0.0039	-0.0105***
	(0.0114)	(0.0096)	(0.0041)	(0.0043)	(0.0093)	(0.0037)	(0.0065)	(0.0119)	(0.0037)	(0.0034)
Collab*t6	0.0178	0.0126	-0.0048	-0.0092**	-0.0174*	-0.0026	0.0020	0.0061	-0.0060*	-0.0176***
	(0.0114)	(0.0092)	(0.0041)	(0.0043)	(0.0092)	(0.0036)	(0.0063)	(0.0121)	(0.0036)	(0.0033)
Collab*t7	0.0096	0.0125	-0.0124***	-0.0099**	-0.0200**	-0.0029	0.0006	-0.0127	-0.0211***	-0.0199***
	(0.0123)	(0.0094)	(0.0042)	(0.0044)	(0.0097)	(0.0038)	(0.0068)	(0.0149)	(0.0037)	(0.0036)
Constant	0.3175***	0.2185***	0.3968***	0.3646***	0.4361***	0.3986***	0.3891***	0.3394***	0.4014***	0.3921***
	(0.0017)	(0.0015)	(0.0005)	(0.0007)	(0.0024)	(0.0008)	(0.0012)	(0.0022)	(0.0005)	(0.0007)
Observations	127,068	131,684	1,263,625	610,193	118,573	698,026	322,441	87,498	2,164,367	637,361
R-squared	0.1999	0.3536	0.1798	0.1995	0.2282	0.2088	0.2328	0.3458	0.1705	0.1557
Journal FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note: The table shows the regression results of the model $Basicness_{\mathrm{itj}} = \beta_o + \beta_1 collab_i + \beta_2 \sum_{t=1}^{7} \mathrm{I(year = t)} + \beta_3 collab * \sum_{t=1}^{7} * \mathrm{I(year = t)} + \mu_j + u_{itj}$. Every column represents a different scientific field. Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table C 7: Appliedness and basicness by firm age, 1980-2014

	Young			Old	
	(1) Appliedness	(2) Basicness		(3) Appliedness	(4) Basicness
Young	0.0338***	-0.0040	Old	0.0326***	0.0112***
	(0.0051)	(0.0082)		(0.0007)	(0.0011)
Young*t2	-0.0006	-0.0083	Old*t2	0.0035***	-0.0037**
	(0.0067)	(0.0107)		(0.0010)	(0.0015)
Young*t3	0.0104*	0.0038	Old*t3	-0.0018*	-0.0107***
	(0.0061)	(0.0096)		(0.0009)	(0.0014)
Young*t4	0.0001	0.0133	Old*t4	0.0015	-0.0072***
	(0.0061)	(0.0096)		(0.0010)	(0.0015)
Young*t5	0.0042	0.0048	Old*t5	0.0039***	-0.0056***
	(0.0057)	(0.0090)		(0.0010)	(0.0015)
Young*t6	0.0101*	-0.0132	Old*t6	0.0077***	-0.0175***
	(0.0060)	(0.0093)		(0.0010)	(0.0015)
Young*t7	0.0212***	-0.0163	Old*t7	0.0134***	-0.0338***
	(0.0072)	(0.0110)		(0.0012)	(0.0018)
Constant	0.7634***	0.3898***	Constant	0.7634***	0.3898***
	(0.0002)	(0.0003)		(0.0002)	(0.0003)
Observations	6,149,989	5,695,504	Observations	6,149,989	5,695,504
R-squared	0.2710	0.2244	R-squared	0.2710	0.2244
Journal FE	YES	YES	Journal FE	YES	YES

Note: The table shows the regression results of the model $Y_{itj} = \beta_o + \beta_1 firmage + \beta_2 \sum_{t=1}^{7} \mathrm{I(year = t)} + \beta_3 firmage * \sum_{t=1}^{7} * \mathrm{I(year = t)} + \mu_j + u_{itj}$. Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table C 8: Appliedness and basicness by firm size, 1980-2014

	Small			Large	
	(1) Appliedness	(2) Basicness		(3) Appliedness	(4) Basicness
Small	0.0222***	0.0036	Large	0.0310***	0.0086***
	(0.0039)	(0.0052)		(0.0017)	(0.0022)
Small*t2	0.0178***	-0.0032	Large*t2	0.0050**	0.0017
	(0.0056)	(0.0075)		(0.0024)	(0.0031)
Small*t3	0.0036	-0.0133*	Large*t3	0.0029	0.0012
	(0.0056)	(0.0075)		(0.0024)	(0.0031)
Small*t4	0.0061	-0.0048	Large*t4	0.0045*	-0.0048
	(0.0055)	(0.0074)		(0.0023)	(0.0031)
Small*t5	0.0090*	-0.0108	Large*t5	0.0010	-0.0076**
	(0.0054)	(0.0072)		(0.0024)	(0.0031)
Small*t6	-0.0008	-0.0204***	Large*t6	0.0039	-0.0121***
	(0.0056)	(0.0074)		(0.0024)	(0.0031)
Small*t7	-0.0015	-0.0067	Large*t7	0.0052**	-0.0102***
	(0.0056)	(0.0074)		(0.0024)	(0.0031)
Small*t8	0.0147**	-0.0224***	Large*t8	0.0058**	-0.0160***
	(0.0058)	(0.0077)		(0.0025)	(0.0032)
Small*t9	0.0175***	-0.0229***	Large*t9	0.0049*	-0.0147***
	(0.0058)	(0.0076)		(0.0025)	(0.0032)
Small*t10	0.0234***	-0.0111	Large*t10	0.0105***	-0.0188***
	(0.0058)	(0.0075)		(0.0025)	(0.0032)
Small*t11	0.0168***	-0.0309***	Large*t11	0.0063**	-0.0250***
	(0.0061)	(0.0079)		(0.0026)	(0.0033)
Small*t12	0.0240***	-0.0224***	Large*t12	0.0154***	-0.0294***
	(0.0063)	(0.0081)		(0.0027)	(0.0034)
Small*t13	0.0373***	-0.0113	Large*t13	0.0076***	-0.0359***
	(0.0069)	(0.0089)		(0.0029)	(0.0037)
Small*t14	0.0301***	-0.0403***	Large*t14	0.0119***	-0.0421***
	(0.0084)	(0.0111)		(0.0032)	(0.0043)
Small*t15	0.0148	-0.0698***	Large*t15	-0.0203***	-0.0171**
	(0.0109)	(0.0190)		(0.0041)	(0.0069)
Constant	0.7992***	0.4488***	Constant	0.7992***	0.4488***
	(0.0003)	(0.0004)		(0.0003)	(0.0004)
Observations	3,188,964	3,063,990	Observations	3,188,964	3,063,990
R-squared	0.2845	0.2313	R-squared	0.2845	0.2313
Journal FE	YES	YES	Journal FE	YES	YES

Note: The table shows the regression results of the model $Appliedness_{\text{itj}} = \beta_o + \beta_1 firmsize + \beta_2 \sum_{t=1}^{15} \text{I(year} = \text{t)} + \beta_3 firmsize * \sum_{t=1}^{15} * \text{I(year} = \text{t)} + \mu_j + u_{itj}$ and $Basicness_{\text{itj}} = \beta_o + \beta_1 firmsize + \beta_2 \sum_{t=1}^{7} \text{I(year} = \text{t)} + \beta_3 firmsize * \sum_{t=1}^{7} * \text{I(year} = \text{t)} + \mu_j + u_{itj}$. Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

D. Appendix Chapter 4

Table D 1: Number of papers and collaborations, IV calculated in the book building phase

VARIABLES	(1) First stage	(2) # papers	(3) First stage	(4) # collaborations
IPO		9.994**		6.713***
		(3.924)		(2.484)
Market returns	0.343***		0.343***	
	(0.032)		(0.032)	
F statistic	117.10		117.10	
Observations		5,002		5,002
R-squared				
Firm FE	✓	✓	✓	✓
Relative Year FE	✓	✓	✓	✓
Cohort*Time FE	✓	✓	✓	✓
Quarter*Post FE	✓	✓	✓	✓

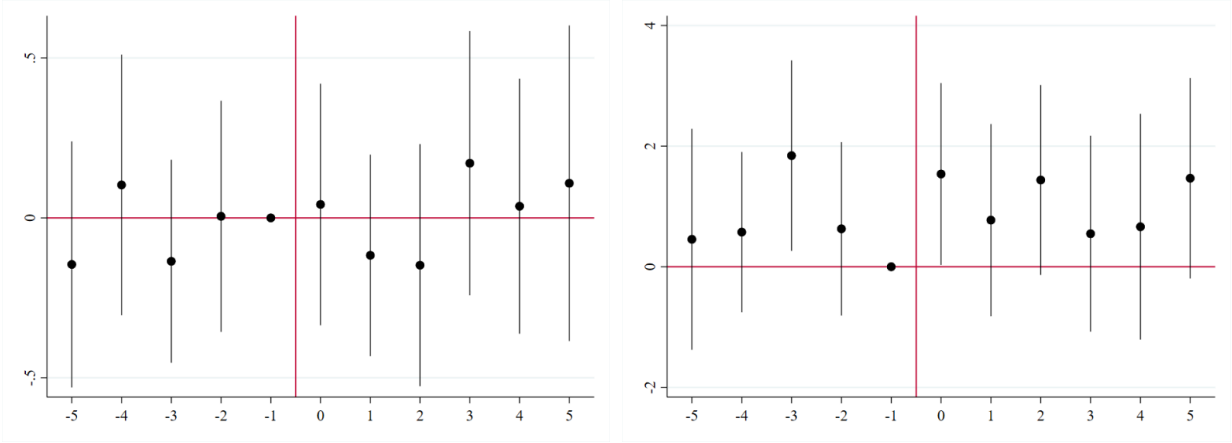
Note: This table shows the regression results for equation $Y_{it} = \mu_i + \theta_t + \alpha IPO_{it} + \delta_{cy} + \partial_{qt} + u_{it}$ with as first stage $IPO = market\ returns * post + \mu_i + \theta_t + \delta_{cy} + \partial_{qt} + u_{it}$. The instrument is the average market returns in the firms' book building phase. The independent variables are the number of publications and collaborations.

Table D 2: Publications forward citations and appliedness, IV calculated in the book building phase

VARIABLES	(1) First stage	(2) Pub Citations	(3) First stage	(4) Appliedness
IPO		-5.947*		0.0308
		(3.495)		(0.454)
Market returns	0.401***		0.445***	
	(0.047)		(0.055)	
F statistic	73.92		65.10	
Observations		2,463		1,914
R-squared				
Firm FE	✓	✓	✓	✓
Relative Year FE	✓	✓	✓	✓
Cohort*Time FE	✓	✓	✓	✓
Quarter*Post FE	✓	✓	✓	✓

Note: This table shows the regression results for equation $Y_{it} = \mu_i + \theta_t + \alpha IPO_{it} + \delta_{cy} + \partial_{qt} + u_{it}$ with $IPO = market\ returns * post + \mu_i + \theta_t + \delta_{cy} + \partial_{qt} + u_{it}$. The instrument is the average market returns in the firms' book building phase. The independent variables are the number of forward citations and appliedness.

Figure D 1: Appliedness (left) and paper citations (right) , event study



Note: This table shows the regression results for equation $Y_{it} = \mu_i + \theta_t + \sum_{\tau=-5}^{-2} \alpha_\tau IPO_{i\tau} + \sum_{\tau=0}^{5} \beta_\tau IPO_{i\tau} + \delta_{cy} + \partial_{qt} + u_{it}$. The independent variables are the number of forward citations and appliedness.

Bibliography

Abramo, Giovanni, Ciriaco Andrea D'Angelo, Flavia Di Costa, and Marco Solazzi. 2009. 'University–Industry Collaboration in Italy: A Bibliometric Examination'. *Technovation* 29 (6–7): 498–507. https://doi.org/10.1016/j.technovation.2008.11.003.

Acharya, Viral, and Zhaoxia Xu. 2017. 'Financial Dependence and Innovation: The Case of Public versus Private Firms'. *Journal of Financial Economics* 124 (2): 223–43. https://doi.org/10.1016/j.jfineco.2016.02.010.

Aggarwal, Vikas A., and David H. Hsu. 2014. 'Entrepreneurial Exits and Innovation'. *Management Science* 60 (4): 867–87. https://doi.org/10.1287/mnsc.2013.1801.

Aghion, Philippe, John Van Reenen, and Luigi Zingales. 2013. 'Innovation and Institutional Ownership'. *American Economic Review* 103 (1): 277–304. https://doi.org/10.1257/aer.103.1.277.

Agrawal, Ajay, Avi Goldfarb, and Florenta Teodoridis. 2016. 'Understanding the Changing Structure of Scientific Inquiry'. *American Economic Journal: Applied Economics* 8 (1): 100–128.

Ahmadpoor, Mohammad, and Benjamin F. Jones. 2017. 'The Dual Frontier: Patented Inventions and Prior Scientific Advance'. *Science* 357 (6351): 583–87. https://doi.org/10.1126/science.aam9527.

Allen, Franklin, and Douglas Gale. 1999. 'Diversity of Opinion and Financing of New Technologies'. *Journal of Financial Intermediation* 8 (1–2): 68–89. https://doi.org/10.1006/jfin.1999.0261.

Almeida, P., J. Hohberger, and P. Parada. 2011. 'Individual Scientific Collaborations and Firm-Level Innovation'. *Industrial and Corporate Change* 20 (6): 1571–99. https://doi.org/10.1093/icc/dtr030.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

Ankrah, Samuel, and Omar AL-Tabbaa. 2015. 'Universities–Industry Collaboration: A Systematic Review'. *Scandinavian Journal of Management* 31 (3): 387–408. https://doi.org/10.1016/j.scaman.2015.02.003.

Arora, Ashish, Sharon Belenzon, Wesley Cohen, and Andrea Patacconi. 2019. 'Companies Persist with Biomedical Papers'. *Nature* 569 (7756): S18–19. https://doi.org/10.1038/d41586-019-01441-x.

Arora, Ashish, Sharon Belenzon, Konstantin Kosenko, Jungkyu Suh, and Yishay Yafeh. 2021. 'The Rise of Scientific Research in Corporate America'. w29260. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w29260.

Arora, Ashish, Sharon Belenzon, and Andrea Patacconi. 2018. 'The Decline of Science in Corporate R&D'. *Strategic Management Journal* 39 (1): 3–32. https://doi.org/10.1002/smj.2693.

———. 2019. 'A Theory of the US Innovation Ecosystem: Evolution and the Social Value of Diversity'. *Industrial and Corporate Change* 28 (2): 289–307. https://doi.org/10.1093/icc/dty067.

Arora, Ashish, Sharon Belenzon, and Lia Sheer. 2021a. 'Knowledge Spillovers and Corporate Investment in Scientific Research'. *American Economic Review* 111 (3): 871–98. https://doi.org/10.1257/aer.20171742.

———. 2021b. ‘Matching Patents to Compustat Firms, 1980–2015: Dynamic Reassignment, Name Changes, and Ownership Structures’. *Research Policy* 50 (5): 104217. https://doi.org/10.1016/j.respol.2021.104217.

Arora, Ashish, Sharon Belenzon, and Jungkyu Suh. 2021. ‘Science and the Market for Technology’. w28534. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w28534.

Asker, John, Joan Farre-Mensa, and Alexander Ljungqvist. 2015. ‘Corporate Investment and Stock Market Listing: A Puzzle?’ *Review of Financial Studies* 28 (2): 342–90. https://doi.org/10.1093/rfs/hhu077.

Astebro, Thomas, Serguey Braguinsky, and Yuheng Ding. 2020. ‘Declining Business Dynamism among Our Best Opportunities: The Role of the Burden of Knowledge’. Working Paper. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w27787.

Atanassov, Julian, Vikram K. Nanda, and Amit Seru. 2007. ‘Finance and Innovation: The Case of Publicly Traded Firms’. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.740045.

Audretsch, David Bruce, and Maksim Belitski. 2021. ‘Knowledge Complexity and Firm Performance: Evidence from the European SMEs’. *Journal of Knowledge Management* 25 (4): 693–713. https://doi.org/10.1108/JKM-03-2020-0178.

Bajgar, Matej, Giuseppe Berlingieri, Sara Calligaris, Chiara Criscuolo, and Jonathan Timmis. 2020. ‘Coverage and Representativeness of Orbis Data’. OECD Science, Technology and Industry Working Papers 2020/06. Vol. 2020/06. OECD Science, Technology and Industry Working Papers. https://doi.org/10.1787/c7bdaa03-en.

Baker, Andrew C., David F. Larcker, and Charles C. Y. Wang. 2022. ‘How Much Should We Trust Staggered Difference-in-Differences Estimates?’ *Journal of Financial Economics* 144 (2): 370–95. https://doi.org/10.1016/j.jfineco.2022.01.004.

Balconi, Margherita, Stefano Brusoni, and Luigi Orsenigo. 2010. ‘In Defence of the Linear Model: An Essay’. *Research Policy* 39 (1): 1–13. https://doi.org/10.1016/j.respol.2009.09.013.

Barton, David K. 2010. ‘History of Monopulse Radar in the US’. *IEEE Aerospace and Electronic Systems Magazine* 25 (3): c1–16. https://doi.org/10.1109/MAES.2010.5464419.

Bellemare, Marc F., Lindsey Novak, and Tara L. Steinmetz. 2015. ‘All in the Family: Explaining the Persistence of Female Genital Cutting in West Africa’. *Journal of Development Economics* 116 (September): 252–65. https://doi.org/10.1016/j.jdeveco.2015.06.001.

Berle, A, and C Means. 1932. *The Modern Corporation and Private Property*.

Bernstein, Shai. 2015. ‘Does Going Public Affect Innovation?’ *The Journal of Finance* 70 (4): 1365–1403. https://doi.org/10.1111/jofi.12275.

Bertrand, Marianne, and Sendhil Mullainathan. 2003. ‘Enjoying the Quiet Life? Corporate Governance and Managerial Preferences’. *Journal of Political Economy* 111 (5): 1043–75. https://doi.org/10.1086/376950.

Bhaskarabhatla, Ajay, and Deepak Hegde. 2014. ‘An Organizational Perspective on Patenting and Open Innovation’. *Organization Science* 25 (6): 1744–63. https://doi.org/10.1287/orsc.2014.0911.

Bhattacharya, Sudipto, and Jay R. Ritter. 1983. ‘Innovation and Communication: Signalling with Partial Disclosure’. *The Review of Economic Studies* 50 (2): 331–46. https://doi.org/10.2307/2297419.

Block, Fred, and Matthew R. Keller. 2009. 'Where Do Innovations Come from? Transformations in the US Economy, 1970–2006'. *Socio-Economic Review* 7 (3): 459–83. https://doi.org/10.1093/ser/mwp013.

Borisov, Alexander, Andrew Ellul, and Merih Sevilir. 2021. 'Access to Public Capital Markets and Employment Growth'. *Journal of Financial Economics* 141 (3): 896–918. https://doi.org/10.1016/j.jfineco.2021.05.036.

Bougrain, Frédéric, and Bernard Haudeville. 2002. 'Innovation, Collaboration and SMEs Internal Research Capacities'. *Research Policy* 31 (5): 735–47. https://doi.org/10.1016/S0048-7333(01)00144-5.

Boyack, Kevin W., Richard Klavans, and Katy Börner. 2005. 'Mapping the Backbone of Science'. *Scientometrics* 64 (3): 351–74. https://doi.org/10.1007/s11192-005-0255-6.

Boyack, Kevin W., Michael Patek, Lyle H. Ungar, Patrick Yoon, and Richard Klavans. 2014. 'Classification of Individual Articles from All of Science by Research Level'. *Journal of Informetrics* 8 (1): 1–12. https://doi.org/10.1016/j.joi.2013.10.005.

Brainard, Jeffrey, and Dennis Normile. 2022. 'China Rises to First Place in Most Cited Papers'. 2022. https://www.science.org/content/article/china-rises-first-place-most-cited-papers.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro H C Sant'Anna. 2021. 'Difference-in-Differences with a Continuous Treatment'.

Calvert, J, and P Patel. 2003. 'University-Industry Research Collaborations in the UK: Bibliometric Trends'. *Science and Public Policy* 30 (2): 85–96. https://doi.org/10.3152/147154303781780597.

Camerani, Roberto, Daniele Rotolo, and Nicola Grassano. 2018. 'Do Firms Publish? A Multi-Sectoral Analysis', 42.

Cameron, A Colin, and Pravin K Trivedi. 2005. 'Microeconometrics : Methods and Applications', 1058.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. 'The Effect of Minimum Wages on Low-Wage Jobs*'. *The Quarterly Journal of Economics* 134 (3): 1405–54. https://doi.org/10.1093/qje/qjz014.

Chacua, Christian. (2020) 2020. 'Sql_mag'. https://github.com/cchacua/sql_mag.

Cohen, Wesley M., and Daniel A. Levinthal. 1989. 'Innovation and Learning: The Two Faces of R & D'. *The Economic Journal* 99 (397): 569–96. https://doi.org/10.2307/2233763.

———. 1990. 'Absorptive Capacity: A New Perspective on Learning and Innovation'. *Administrative Science Quarterly* 35 (1): 128. https://doi.org/10.2307/2393553.

Colombo, Massimo G., Michele Meoli, and Silvio Vismara. 2019. 'Signaling in Science-Based IPOs: The Combined Effect of Affiliation with Prestigious Universities, Underwriters, and Venture Capitalists'. *Journal of Business Venturing* 34 (1): 141–77. https://doi.org/10.1016/j.jbusvent.2018.04.009.

Coombs, Rod, and Luke Georghiou. 2002. *A New" Industrial Ecology"*. American Association for the Advancement of Science.

Cornelli, Francesca, David Goldreich, and Alexander Ljungqvist. 2006. 'Investor Sentiment and Pre-IPO Markets'. *The Journal of Finance* 61 (3): 1187–1216. https://doi.org/10.1111/j.1540-6261.2006.00870.x.

Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. Yale University Press.

Darby, Michael, and Lynne Zucker. 2002. 'Going Public When You Can in Biotechnology'. w8954. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w8954.

Deeds, David L., Dona Decarolis, and Joseph E. Coombs. 1997. 'The Impact of Firmspecific Capabilities on the Amount of Capital Raised in an Initial Public Offering: Evidence

from the Biotechnology Industry'. *Journal of Business Venturing* 12 (1): 31–46. https://doi.org/10.1016/S0883-9026(97)84970-1.

Derrien, François. 2005. 'IPO Pricing in "Hot" Market Conditions: Who Leaves Money on the Table?' *The Journal of Finance* 60 (1): 487–521. https://doi.org/10.1111/j.1540-6261.2005.00736.x.

D'Este, P., F. Guy, and S. Iammarino. 2013. 'Shaping the Formation of University-Industry Research Collaborations: What Type of Proximity Does Really Matter?' *Journal of Economic Geography* 13 (4): 537–58. https://doi.org/10.1093/jeg/lbs010.

Durst, Susanne, and Ingi Runar Edvardsson. 2012. 'Knowledge Management in SMEs: A Literature Review'. *Journal of Knowledge Management* 16 (6): 879–903. https://doi.org/10.1108/13673271211276173.

Edelen, Roger M., and Gregory B. Kadlec. 2005. 'Issuer Surplus and the Partial Adjustment of IPO Prices to Public Information'. *Journal of Financial Economics* 77 (2): 347–73. https://doi.org/10.1016/j.jfineco.2004.05.009.

Egghe, L. 2010. 'A Model Showing the Increase in Time of the Average and Median Reference Age and the Decrease in Time of the Price Index'. *Scientometrics* 82 (2): 243–48. https://doi.org/10.1007/s11192-009-0057-3.

Etzkowitz, Henry, and Loet Leydesdorff. 1997. 'Universities and the Global Knowledge Economy: A Triple Helix of University-Industry Relations'. SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=3404823.

Fabrizio, Kira R. 2009. 'Absorptive Capacity and the Search for Innovation'. *Research Policy* 38 (2): 255–67. https://doi.org/10.1016/j.respol.2008.10.023.

Ferreira, Daniel, Gustavo Manso, and Andre C. Silva. 2014. 'Incentives to Innovate and the Decision to Go Public or Private'. *Review of Financial Studies* 27 (1): 256–300.

Fisk, Jb, Hd Hagstrum, and P.L. Hartman. 1946. 'The Magnetron as a Generator of Centimeter Waves'. Nokia Bell Labs. 1946. https://www.bell-labs.com/institute/publications/bstj25-2-167/.

Fleming, Lee, and Olav Sorenson. 2004. 'Science as a Map in Technological Search'. *Strategic Management Journal* 25 (8–9): 909–28. https://doi.org/10.1002/smj.384.

Frank, Morgan R., Dashun Wang, Manuel Cebrian, and Iyad Rahwan. 2019. 'The Evolution of Citation Graphs in Artificial Intelligence Research'. *Nature Machine Intelligence* 1 (2): 79–85. https://doi.org/10.1038/s42256-019-0024-5.

Franzoni, Chiara, Giuseppe Scellato, and Paula Stephan. 2012. 'Foreign-Born Scientists: Mobility Patterns for 16 Countries'. *Nature Biotechnology* 30 (12): 1250–53. https://doi.org/10.1038/nbt.2449.

Fukugawa, Nobuya. 2022. 'Effects of the Quality of Science on the Initial Public Offering of University Spinoffs: Evidence from Japan'. *Scientometrics*, June. https://doi.org/10.1007/s11192-022-04433-3.

Gao, Huasheng, Po-Hsuan Hsu, and Kai Li. 2018. 'Innovation Strategy of Private Firms'. *Journal of Financial and Quantitative Analysis* 53 (1): 1–32. https://doi.org/10.1017/S0022109017001119.

Gertner, Jon. 2012. *The Idea Factory: Bell Labs and the Great Age of American Innovation*. Penguin.

Godin, Benoît. 2003. 'Measuring Science: Is There "Basic Research" without Statistics?' *Social Science Information* 42 (1): 57–90. https://doi.org/10.1177/0539018403042001795.

———. 2006. 'The Linear Model of Innovation: The Historical Construction of an Analytical Framework'. *Science, Technology, & Human Values* 31 (6): 639–67. https://doi.org/10.1177/0162243906291865.

Gomila, Robin. 20200924. ‘Logistic or Linear? Estimating Causal Effects of Experimental Treatments on Binary Outcomes Using Regression Analysis.’ *Journal of Experimental Psychology: General* 150 (4): 700. https://doi.org/10.1037/xge0000920.

Greene, William. 2004. ‘Fixed Effects and Bias Due to the Incidental Parameters Problem in the Tobit Model’. *Econometric Reviews* 23 (2): 125–47. https://doi.org/10.1081/ETC-120039606.

Grindley, Peter C., and David J. Teece. 1997. ‘Managing Intellectual Capital: LICENSING AND CROSS-LICENSING IN SEMICONDUCTORS AND ELECTRONICS’. *California Management Review* 39 (2): 8–41. https://doi.org/10.2307/41165885.

Guo, Re-Jin, Baruch Lev, and Nan Zhou. 2004. ‘Competitive Costs of Disclosure by Biotech IPOs’. *Journal of Accounting Research* 42 (2): 319–55. https://doi.org/10.1111/j.1475-679X.2004.00140.x.

Hall, Bronwyn H., and Josh Lerner. 2010. ‘The Financing of R&D and Innovation’. In *Handbook of the Economics of Innovation*, 1:609–39. Elsevier. https://doi.org/10.1016/S0169-7218(10)01014-2.

Hamilton, K. 2003. ‘Subfield and Level Classification of Journals’. CHI Report No. 2012-R.

Hartmann, Philipp, and Joachim Henkel. 2020. ‘The Rise of Corporate Science in AI: Data as a Strategic Resource’. *Academy of Management Discoveries* 6 (3): 359–81. https://doi.org/10.5465/amd.2019.0043.

He, Jie (Jack), and Xuan Tian. 2013. ‘The Dark Side of Analyst Coverage: The Case of Innovation’. *Journal of Financial Economics* 109 (3): 856–78. https://doi.org/10.1016/j.jfineco.2013.04.001.

Hellevik, Ottar. 2009. ‘Linear versus Logistic Regression When the Dependent Variable Is a Dichotomy’. *Quality & Quantity* 43 (1): 59–74. https://doi.org/10.1007/s11135-007-9077-3.

Hicks, Diana. 1995. ‘Published Papers, Tacit Competencies and Corporate Management of the Public/Private Character of Knowledge’. *Industrial and Corporate Change* 4 (2): 401–24. https://doi.org/10.1093/icc/4.2.401.

Hiltzik, Michael A. 1999. *Dealers of Lightning: Xerox PARC and the Dawn of the Computer Age*. HarperCollins Publishers.

Horrace, William C., and Ronald L. Oaxaca. 2006. ‘Results on the Bias and Inconsistency of Ordinary Least Squares for the Linear Probability Model’. *Economics Letters* 90 (3): 321–27. https://doi.org/10.1016/j.econlet.2005.08.024.

Hounshell, David A., and John Kenly Smith Jr. 1988. *Science and Corporate Strategy: Du Pont R and D, 1902-1980*. Cambridge University Press.

Howells, J. 2000. *Research and Technology Outsourcing and Systems of Innovation In: Metcalfe, J.S., Miles, I. (Eds.), Innovation Systems in the Service Economy, Measurement and Case Study Analysis*. Springer Science & Business Media.

Jensen, C, and H Meckling. 1976. ‘Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure’. *Journal of Financial Economics*.

Jones, Benjamin F. 2009. ‘The Burden of Knowledge and the “Death of the Renaissance Man”: Is Innovation Getting Harder?’ *REVIEW OF ECONOMIC STUDIES*, 35.

Kalemli-Ozcan, Sebnem, Bent E. Sørensen, Carolina Villegas-Sanchez, Vadym Volosovych, and Sevcan Yesiltas. 2015. ‘How to Construct Nationally Representative Firm Level Data from the ORBIS Global Database’. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2660191.

———. 2019. ‘How to Construct Nationally Representative Firm Level Data from the Orbis Global Database: New Facts on SMEs and Aggregate Implications for Industry Concentration’.

Katz, J.Sylvan, and Ben R. Martin. 1997. 'What Is Research Collaboration?' *Research Policy* 26 (1): 1–18. https://doi.org/10.1016/S0048-7333(96)00917-1.

Kim, Linsu. 1998. 'Crisis Construction and Organizational Learning: Capability Building in Catching-up at Hyundai Motor'. *Organization Science* 9 (4): 506–21. https://doi.org/10.1287/orsc.9.4.506.

Krieger, Bastian, Maikel Pellens, Knut Blind, Sonia Gruber, and Torben Schubert. 2021. 'Are Firms Withdrawing from Basic Research? An Analysis of Firm-Level Publication Behaviour in Germany'. *Scientometrics*, October. https://doi.org/10.1007/s11192-021-04147-y.

Larivière, Vincent, Éric Archambault, and Yves Gingras. 2008. 'Long-Term Variations in the Aging of Scientific Literature: From Exponential Growth to Steady-State Science (1900–2004)'. *Journal of the American Society for Information Science and Technology* 59 (2): 288–96. https://doi.org/10.1002/asi.20744.

Larrain, Borja, Gordon Phillips, Giorgo Sertsios, and Francisco Urzùa. 2021. 'The Effects of Going Public on Firm Performance and Commercialization Strategy: Evidence from International IPOs'. https://www.nber.org/system/files/working_papers/w29219/w29219.pdf.

Latour, Bruno. 1993. *The Pasteurization of France*. Harvard University Press.

Lazonick, William, and Mary O'Sullivan. 2000. 'Maximizing Shareholder Value: A New Ideology for Corporate Governance'. *Economy and Society* 29 (1): 13–35. https://doi.org/10.1080/030851400360541.

Lerner, Joshua. 1994. 'Venture Capitalists and the Decision to Go Public'. *Journal of Financial Economics* 35 (3): 293–316. https://doi.org/10.1016/0304-405X(94)90035-3.

Levenshtein, Vladimir I. 1966. 'Binary Codes Capable of Correcting Deletions, Insertions, and Reversals'. In *Soviet Physics Doklady*, 10:707–10. Soviet Union.

Lim, Kwanghui. 2004. 'The Relationship between Research and Innovation in the Semiconductor and Pharmaceutical Industries (1981–1997)'. *Research Policy* 33 (2): 287–321. https://doi.org/10.1016/j.respol.2003.08.001.

Lin, Jialiang, Yao Yu, Yu Zhou, Zhiyang Zhou, and Xiaodong Shi. 2020. 'How Many Preprints Have Actually Been Printed and Why: A Case Study of Computer Science Preprints on arXiv'. *Scientometrics* 124 (1): 555–74. https://doi.org/10.1007/s11192-020-03430-8.

Lojek, Bo. 2006. *History of Semiconductor Engineering*. New York: Springer.

López-Martínez, R. E., E. Medellín, A. P. Scanlon, and J. L. Solleiro. 1994. 'Motivations and Obstacles to University Industry Cooperation (UIC): A Mexican Case'. *R&D Management* 24 (1): 017–030. https://doi.org/10.1111/j.1467-9310.1994.tb00844.x.

Maksimovic, Vojislav, and Pegaret Pichler. 2001. 'Technological Innovation and Initial Public Offerings'. *The Review of Financial Studies* 14 (2): 459–94. https://doi.org/10.1093/rfs/14.2.459.

Markovitch, Dmitri G., Joel H. Steckel, and Bernard Yeung. 2005. 'Using Capital Markets as Market Intelligence: Evidence from the Pharmaceutical Industry'. *Management Science* 51 (10): 1467–80. https://doi.org/10.1287/mnsc.1050.0401.

Martín-Martín, Alberto, Mike Thelwall, Enrique Orduna-Malea, and Emilio Delgado López-Cózar. 2021. 'Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A Multidisciplinary Comparison of Coverage via Citations'. *Scientometrics* 126 (1): 871–906. https://doi.org/10.1007/s11192-020-03690-4.

Marx, Matt, and Aaron Fuegi. 2020. 'Reliance on Science: Worldwide Front-page Patent Citations to Scientific Articles'. *Strategic Management Journal*, April, smj.3145. https://doi.org/10.1002/smj.3145.

Melin, G., and O. Persson. 1996. 'Studying Research Collaboration Using Co-Authorships'. *Scientometrics* 36 (3): 363–77. https://doi.org/10.1007/BF02129600.

Merton, Robert K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press.

Milojević, Staša. 2012. 'How Are Academic Age, Productivity and Collaboration Related to Citing Behavior of Researchers?' *PLOS ONE* 7 (11): e49176. https://doi.org/10.1371/journal.pone.0049176.

———. 2020. 'Practical Method to Reclassify Web of Science Articles into Unique Subject Categories and Broad Disciplines'. *Quantitative Science Studies* 1 (1): 183–206. https://doi.org/10.1162/qss_a_00014.

Mizik, Natalie, and Robert Jacobson. 2003. 'Trading off between Value Creation and Value Appropriation: The Financial Implications of Shifts in Strategic Emphasis'. *Journal of Marketing* 67 (1): 63–76. https://doi.org/10.1509/jmkg.67.1.63.18595.

Mongeon, Philippe, and Adèle Paul-Hus. 2016. 'The Journal Coverage of Web of Science and Scopus: A Comparative Analysis'. *Scientometrics* 106 (1): 213–28. https://doi.org/10.1007/s11192-015-1765-5.

Moorman, Christine, Simone Wies, Natalie Mizik, and Fredrika J. Spencer. 2012. 'Firm Innovation and the Ratchet Effect Among Consumer Packaged Goods Firms'. *Marketing Science* 31 (6): 934–51. https://doi.org/10.1287/mksc.1120.0737.

Moretti, Enrico. 2021. 'The Effect of High-Tech Clusters on the Productivity of Top Inventors'. *American Economic Review* 111 (10): 3328–75. https://doi.org/10.1257/aer.20191277.

Motohashi, Kazuyuki. 2005. 'University–Industry Collaborations in Japan: The Role of New Technology-Based Firms in Transforming the National Innovation System'. *Research Policy* 34 (5): 583–94. https://doi.org/10.1016/j.respol.2005.03.001.

Mowery, David C. 1983. 'The Relationship between Intrafirm and Contractual Forms of Industrial Research in American Manufacturing, 1900–1940'. *Explorations in Economic History* 20 (4): 351–74. https://doi.org/10.1016/0014-4983(83)90024-4.

———. 2009. 'Plus ca Change: Industrial R&D in the "Third Industrial Revolution"'. *Industrial and Corporate Change* 18 (1): 1–50. https://doi.org/10.1093/icc/dtn049.

Mowery, David C., and Joanne E. Oxley. 1995. 'Inward Technology Transfer and Competitiveness: The Role of National Innovation Systems'. *Cambridge Journal of Economics*, February. https://doi.org/10.1093/oxfordjournals.cje.a035310.

Murray, Fiona, and Scott Stern. 2006. 'When Ideas Are Not Free: The Impact of Patents on Scientific Research', 37.

Muscio, Alessandro. 2007. 'THE IMPACT OF ABSORPTIVE CAPACITY ON SMEs' COLLABORATION'. *Economics of Innovation and New Technology* 16 (8): 653–68. https://doi.org/10.1080/10438590600983994.

Narula, Rajneesh. 2004. 'R&D Collaboration by SMEs: New Opportunities and Limitations in the Face of Globalisation'. *Technovation* 24 (2): 153–61. https://doi.org/10.1016/S0166-4972(02)00045-7.

Onida, Fabrizio, and Franco Malerba. 1989. 'R&D Cooperation between Industry, Universities and Research Organizations in Europe'. *Technovation* 9 (2): 137–95. https://doi.org/10.1016/0166-4972(89)90055-2.

Park, Han Woo, and Loet Leydesdorff. 2010. 'Longitudinal Trends in Networks of University–Industry–Government Relations in South Korea: The Role of Programmatic Incentives'. *Research Policy* 39 (5): 640–49. https://doi.org/10.1016/j.respol.2010.02.009.

Pisano, G. P. 2010. 'The Evolution of Science-Based Business: Innovating How We Innovate'. *Industrial and Corporate Change* 19 (2): 465–82. https://doi.org/10.1093/icc/dtq013.

Polidoro, Francisco, and Matt Theeke. 2012. 'Getting Competition Down to a Science: The Effects of Technological Competition on Firms' Scientific Publications'. *Organization Science* 23 (4): 1135–53. https://doi.org/10.1287/orsc.1110.0684.

Porter, Alan L., and Ismael Rafols. 2009. 'Is Science Becoming More Interdisciplinary? Measuring and Mapping Six Research Fields over Time'. *Scientometrics* 81 (3): 719–45. https://doi.org/10.1007/s11192-008-2197-2.

Price, Derek De Solla. 1976. 'A General Theory of Bibliometric and Other Cumulative Advantage Processes'. *Journal of the American Society for Information Science* 27 (5): 292–306. https://doi.org/10.1002/asi.4630270505.

Rafols, Ismael, Jarno Hoekman, Josh Siepel, Paul Nightingale, Michael M. Hopkins, Alice O'Hare, and Antonio Perianes-Rodriguez. 2012. 'Big Pharma, Little Science? A Bibliometric Perspective on Big Pharma's R&D Decline'. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2012878.

Rajan, Raghuram G. 2012. 'The Corporation in Finance', 64.

Rajan, Raghuram, and Luigi Zingales. 1996. 'Financial Dependence and Growth'. w5758. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w5758.

Reich, Leonard S. 2002. *The Making of American Industrial Research: Science and Business at GE and Bell, 1876-1926*. Cambridge University Press.

Roach, Michael, and Henry Sauermann. 2010. 'A Taste for Science? PhD Scientists' Academic Orientation and Self-Selection into Research Careers in Industry'. *Research Policy* 39 (3): 422–34. https://doi.org/10.1016/j.respol.2010.01.004.

Rosenberg, Nathan. 1990. 'Why Do Firms Do Basic Research (with Their Own Money)?' In *Studies On Science And The Innovation Process: Selected Works of Nathan Rosenberg*, 225–34. World Scientific.

———. 1994. *Exploring the Black Box: Technology, Economics, and History*. Cambridge University Press.

Rosenberg, Nathan, and Richard R. Nelson. 1994. 'American Universities and Technical Advance in Industry'. *Research Policy* 23 (3): 323–48. https://doi.org/10.1016/0048-7333(94)90042-6.

Rotolo, Daniele, Roberto Camerani, Nicola Grassano, and Ben R. Martin. 2022. 'Why Do Firms Publish? A Systematic Literature Review and a Conceptual Framework'. *Research Policy* 51 (10): 104606. https://doi.org/10.1016/j.respol.2022.104606.

Santoro, Michael D., and Alok K. Chakrabarti. 2002. 'Firm Size and Technology Centrality in Industry–University Interactions'. *Research Policy* 31 (7): 1163–80. https://doi.org/10.1016/S0048-7333(01)00190-1.

Savage, Neil. 2018. 'Collaboration Is the Key to Cancer Research'. *Nature* 556 (7700): S1–3. https://doi.org/10.1038/d41586-018-04164-7.

Science. 2021. 'GlaxoSmithKline Focuses on Immunology through Collaboration'. Science | AAAS. 19 March 2021. https://www.sciencemag.org/features/2021/03/glaxosmithkline-focuses-immunology-through-collaboration.

Simeth, Markus, and Michele Cincera. 2016. 'Corporate Science, Innovation and Firm Value', 24.

Simeth, Markus, and Stephane Lhuillery. 2015. 'How Do Firms Develop Capabilities for Scientific Disclosure?' *Research Policy* 44 (7): 1283–95. https://doi.org/10.1016/j.respol.2015.04.005.

Singh, Vivek Kumar, Prashasti Singh, Mousumi Karmakar, Jacqueline Leta, and Philipp Mayr. 2021. 'The Journal Coverage of Web of Science, Scopus and Dimensions: A

Comparative Analysis'. *Scientometrics* 126 (6): 5113–42. https://doi.org/10.1007/s11192-021-03948-5.

Smith, John Kenly Jr. 1990. 'The Scientific Tradition in American Industrial Research'. *Technology and Culture* 31 (1): 121–31. https://doi.org/10.1353/tech.1990.0098.

Sonnenwald, Diane H. 2007. 'Scientific Collaboration'. *Annual Review of Information Science and Technology* 41 (1): 643–81. https://doi.org/10.1002/aris.2007.1440410121.

Spender, John-Christopher, Vincenzo Corvello, Michele Grimaldi, and Pierluigi Rippa. 2017. 'Startups and Open Innovation: A Review of the Literature'. *European Journal of Innovation Management* 20 (1): 4–30. https://doi.org/10.1108/EJIM-12-2015-0131.

Spiegel, Matthew I., and Heather Tookes. 2008. 'Dynamic Competition, Innovation and Strategic Financing'. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1161239.

Stern, Scott. 2004. 'Do Scientists Pay to Be Scientists?' *Management Science* 50 (6): 835–53. https://doi.org/10.1287/mnsc.1040.0241.

Stokes, Donald E. 1997. *Pasteur's Quadrant: Basic Science and Technological Innovation*. Brookings Institution Press.

Stone, Richard. 2012. 'India Rising'. *Science* 335 (6071): 904–10. https://doi.org/10.1126/science.335.6071.904.

The Economist. 2007. 'Out of the Dusty Labs'. *The Economist*, 1 March 2007. https://www.economist.com/briefing/2007/03/01/out-of-the-dusty-labs.

Tijssen, Robert J.W. 2004. 'Is the Commercialisation of Scientific Research Affecting the Production of Public Knowledge?' *Research Policy* 33 (5): 709–33. https://doi.org/10.1016/j.respol.2003.11.002.

Trajtenberg, Manuel, Rebecca Henderson, and Adam Jaffe. 1997. 'University Versus Corporate Patents: A Window On The Basicness Of Invention'. *Economics of Innovation and New Technology* 5 (1): 19–50. https://doi.org/10.1080/10438599700000006.

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. 'Atypical Combinations and Scientific Impact' 342: 6.

Vaccario, Giacomo, Luca Verginer, and Frank Schweitzer. 2021. 'Reproducing Scientists' Mobility: A Data-Driven Model'. *Scientific Reports* 11 (1): 10733. https://doi.org/10.1038/s41598-021-90281-9.

Varma, Roli. 2000. 'Changing Research Cultures in U. S. Industry'. *Science, Technology, & Human Values* 25 (4): 395–416.

Verginer, Luca, and Massimo Riccaboni. 2021. 'Talent Goes to Global Cities: The World Network of Scientists' Mobility'. *Research Policy* 50 (1): 104127. https://doi.org/10.1016/j.respol.2020.104127.

Vismara, Silvio. 2014. 'Patents, R&D Investments and Post-IPO Strategies'. *Review of Managerial Science* 8 (3): 419–35. https://doi.org/10.1007/s11846-013-0113-5.

Wang, Jian, Bart Thijs, and Wolfgang Glänzel. 2015. 'Interdisciplinarity and Impact: Distinct Effects of Variety, Balance, and Disparity'. Edited by Neil R. Smalheiser. *PLOS ONE* 10 (5): e0127298. https://doi.org/10.1371/journal.pone.0127298.

West, Joel, Ammon Salter, Wim Vanhaverbeke, and Henry Chesbrough. 2014. 'Open Innovation: The Next Decade'. *Research Policy* 43 (June): 805–11. https://doi.org/10.1016/j.respol.2014.03.001.

Wies, Simone, and Christine Moorman. 2015. 'Going Public: How Stock Market Listing Changes Firm Innovation Behavior'. *Journal of Marketing Research* 52 (5): 694–709. https://doi.org/10.1509/jmr.13.0289.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data MIT Press*. Vol. 108.

Wu, Geraldine A. 2012. ‘The Effect of Going Public on Innovative Productivity and Exploratory Search’. *Organization Science* 23 (4): 928–50. https://doi.org/10.1287/orsc.1110.0676.

Yujian, Li, and Liu Bo. 2007. ‘A Normalized Levenshtein Distance Metric’. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29 (6): 1091–95. https://doi.org/10.1109/TPAMI.2007.1078.

Zahra, Shaker A, and Gerard George. 2002. ‘Absorptive Capacity: A Review, Reconceptualization, and Extension’, 20.

Zahra, Shaker A., Aseem Kaul, and María Teresa Bolívar-Ramos. 2018. ‘Why Corporate Science Commercialization Fails: Integrating Diverse Perspectives’. *Academy of Management Perspectives* 32 (1): 156–76. https://doi.org/10.5465/amp.2016.0132.

Zhu, Nibing, Chang Liu, and Zhilin Yang. 2021. ‘Team Size, Research Variety, and Research Performance: Do Coauthors’ Coauthors Matter?’ *Journal of Informetrics* 15 (4): 101205. https://doi.org/10.1016/j.joi.2021.101205.

Zucker, Lynne, and Michael Darby. 1998. ‘Capturing Technological Opportunity via Japan’s Star Scientists: Evidence from Japanese Firms’ Biotech Patents and Products’. w6360. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w6360.

Zucker, Lynne G., Michael R. Darby, and Jeff S. Armstrong. 2002. ‘Commercializing Knowledge: University Science, Knowledge Capture, and Firm Performance in Biotechnology’. *Management Science* 48 (1): 138–53.