



## NETWORK AND INFORMATION-THEORETIC STUDIES ON THE EFFECTS OF RESEARCH ENVIRONMENT IN SCIENTIFIC CAREERS: GEOGRAPHY, PROMINENCE AND GENDER

Lluís Danús Amengual

**ADVERTIMENT.** L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

**ADVERTENCIA.** El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

**WARNING.** Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.

# Network and information-theoretic studies on the effects of research environment in scientific careers: geography, prominence and gender

Lluís Danús Amengual



Universitat Rovira i Virgili

Doctoral Thesis

2023

UNIVERSITAT ROVIRA I VIRGILI

NETWORK AND INFORMATION-THEORETIC STUDIES ON THE EFFECTS OF RESEARCH ENVIRONMENT IN  
SCIENTIFIC CAREERS: GEOGRAPHY, PROMINENCE AND GENDER

Lluís Danús Amengual



**UNIVERSITAT  
ROVIRA i VIRGILI**

**Network and information-theoretic studies on the  
effects of research environment in scientific  
careers: geography, prominence and gender**

**Lluís Danús Amengual**

Doctoral thesis  
supervised by:

Dr. Marta Sales Pardo  
Dr. Roger Guimerà Manrique

**Department of Chemical Engineering**

Tarragona  
2023



UNIVERSITAT ROVIRA I VIRGILI

NETWORK AND INFORMATION-THEORETIC STUDIES ON THE EFFECTS OF RESEARCH ENVIRONMENT IN  
SCIENTIFIC CAREERS: GEOGRAPHY, PROMINENCE AND GENDER

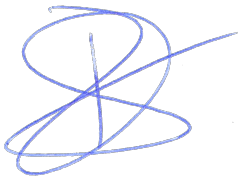
Lluís Danús Amengual

# Declaration

WE STATE that the present study, entitled “Network and information-theoretic studies on the effects of research environment in scientific careers: geography, prominence and gender”, presented by Lluís Danús Amengual for the award of the degree of Doctor, has been carried out under our supervision at the Department of Chemical Engineering of this university, and that it fulfils all the requirements to be eligible for the International Doctorate Award.

Tarragona, April 12th, 2023

Doctoral Thesis Supervisor/s



Dr. Roger Guimerà Manrique



Dra. Marta Sales Pardo

UNIVERSITAT ROVIRA I VIRGILI

NETWORK AND INFORMATION-THEORETIC STUDIES ON THE EFFECTS OF RESEARCH ENVIRONMENT IN  
SCIENTIFIC CAREERS: GEOGRAPHY, PROMINENCE AND GENDER

Lluís Danús Amengual

# Abstract

Science is the result of the coordinated effort of many individuals, and as such it is a complex social system. Consequently, its features depend on the interactions between many actors and its evolution cannot be explained from the behavior of the different actors composing it and thus, its study is extremely challenging. Understanding the way science is done and which factors condition its development is crucial from different points of view. Scientifically, understanding the factors driving disrupting ideas or team performance will allow researchers to take action on the laboratory composition or their research agenda. From a social point of view, having grounded information about the factors driving the evolution of science will help policy makers to better distribute the funds or elaborate policies to overcome possible inequalities among others.

The goal of this work is to gain understanding of some of the social factors driving science nowadays, and to do so by using state of the art statistical inference and machine learning tools without losing sight on the interpretability of the results. In this thesis, we will study different aspects of science namely, geographical differences in the collaboration networks, factors driving prominence and the effect of research environments and gender in academia. As these are social problems, we will use network science as a baseline discipline since it has proven useful for the study of a variety of social interactions and the study of science itself. Network science bases its study on the mathematical objects called networks or graphs, which are composed by actors (nodes) and their interactions (represented by links or edges). Particularly we will make use of two types of networks,

simple undirected multigraphs for the study of the collaboration networks, and bipartite word-document networks to study the evolution of research topics along the career of a researcher.

The use of a networks approach offers the advantage of using some of the methods developed to study network structure, such as the Stochastic Block Models (SBM). SBMs are a family of generative models that propose a partition of the network nodes into groups where the probability of two nodes being connected depends uniquely on the group they belong to, which ends up being a natural representation of real social networks. Alongside with SBMs, we will combine different techniques from machine learning as well as information theory concepts to aid in the advance in some of the aforementioned problems.

With this combination of techniques we will address some of the questions raised at the beginning. In a first place, we will study how the conditions that exist in different geographical environments (such as those in North America and Europe) shape the collaboration strategies of prominent researchers. To that aim, we will construct networks of collaboration between researchers based on the coauthoring of scientific publications. In this way, identifying different groups will imply that different collaboration strategies are at play. We found that despite there being no differences in terms of publications and number of collaborators, a more nuanced analysis of the network reveals highly affiliation-polarized groups of researchers, with different collaboration structures for different affiliations. In general, prominent scientists that are based in North America, tend to collaborate less with other prominent researchers while those based in Europe, create highly-collaborative clusters with each other. Interestingly, this is not translated into a greater impact. Although collaboration increases the number of citations of scientific publications, it seems that North America-based researchers take most profit of their collaboration strategy, increasing the impact more noticeably than Europe based researchers do. Exploring some of the mechanisms of this phenomenon, we observed how despite collaboration always being beneficial when compared to not collaborating, excessive repetition tends to decrease the impact of research on average, implying that highly collaborative clusters might end up producing less impactful

research.

Secondly, we explored the conditions for prominence of the aforementioned set of prominent researchers. Analysing some of the factors of their early career highlights how when compared to the average researcher, these particular researchers stand out in terms of several metrics such as researching at a top 25 university, publishing in top ranked journals or having collaborated with other prominent researchers. Also, we observe differences between prominent researchers when disaggregating them into quartiles according to the number of citations accrued during the first 5 years of career, with the ones in the top quartile having an even higher share of these characteristics. This advantage is not a temporary hot streak but a sustained advantage over time. Indeed, by looking at the quartile researchers belong to after 20 years of career, we show a vast majority of researchers who were at the top two quartiles remained in the same position, suggesting that the initial advantage (or disadvantage) is hard to overcome. Then we examined the predictive power of different factors at play, showing how without taking into account measures directly related to the number of citations during the first five years, we are able to accurately assign the quartile at 20 years of career in most of the cases, with collaborating with other prominent researchers and publishing at top ranked journals being the most predictive ones.

Finally, we study how the effect of joining a new department shapes the research portfolio for early-career researchers in chemical engineering. First, we start by constructing a word-document network with all the publication history in two wide-scope journals of the field and fitting this network with a SBM in order to classify words into different topical groups. Once we can classify words into topics we proceed with the analysis of publications of two different sets of researchers. The first set is composed by a group of early-career authors who were offered to join a department, and some of them accepted the offer and the other declined. This allows us to establish a comparison group to observe the effect of joining a department in terms of research topics. We compare the distribution over topics of researchers before and after the offer was made, to the one of the established researchers in the department (those that were already there before

the hiring process started). The results show how those who joined the department start accruing new departmental collaborations and that without including new papers in collaboration, most of them approach to the departmental topics indicating that there's an influence that goes beyond direct collaboration. The fact that some researchers, diverge from the department despite having joined it, prompts out the question of this effect being equal to all researchers. To address this, the second set comprises a group of authors who all joined a department in any of the top 34 chemical engineering institutions in North America and Europe. The results show how despite having similar number of new collaborators and publishing at a similar rate with other colleagues of the department, female researchers that converge towards the department represents a 34% of the total female incorporation in front of a 52% for male new faculties. Motivated by this, we further explore the nature of the collaborations between males and females in the faculty. We show how when entering in a faculty, while male researchers collaborate equally with both genders, female new faculty tend to collaborate less with their female colleagues, with this effect being more pronounced when the collaboration implies senior female researchers.

Through these three problems, network science and statistical inference have proven to be to provide valuable insights when studying the evolution of science and how it is produced. In addition, the results suggest possible interventions in some of the differences and inequalities identified. Finally, it opens the door to studying the scientific field from different perspectives and research lines.

# Agraïments

En aquest treball xerram de sistemes complexes, aquells el comportament dels quals no es pot entendre a partir del comportament dels individus si no que hem de mirar-ne el conjunt. De la mateixa manera, el camí que m'ha portat fins aquí és fruit de l'interacció de moltes peces fonamentals.

En primer lloc, agrair a na Marta i en Roger per haver-me donat aquesta gran oportunitat tot i venir d'un àmbit força diferent. Durant aquests anys han compartit un coneixement d'un valor incalculable, consells i suport quan les coses no sortien com un inicialment havia planejat.

Agrair també als membres del SEESLab, passats i presents: Sergio, Ignasi, Oscar, Lluç, Anegelo, Manuel, Oriol, Teresa, Maribel per les discussions als group meetings i als dinars. També m'agradaria donar les gràcies a l'Alejandro qui m'ha ajudat estalviar incomptables hores de feina. Mencionar també la important feina realitzada pel personal administratiu de la universitat. En aquest sentit agrair a la Núria i a la Susi la seva tasca.

Mai podré donar les gràcies a bastament a la Rocío, per ser la millor companya de viatge que un podria desitjar i per la seva paciència infinita.

Finalment, agrair a mons pares i a sa predina Antònia, tot l'esforç que han fet per ajudar-me a arribar fins on estic i també pels valors inculcats, tan importants com el coneixement adquirit durant aquests anys. També a la meva germana, una de les persones més fortes que conec, qui creu més en mi que jo mateix.



UNIVERSITAT ROVIRA I VIRGILI

NETWORK AND INFORMATION-THEORETIC STUDIES ON THE EFFECTS OF RESEARCH ENVIRONMENT IN  
SCIENTIFIC CAREERS: GEOGRAPHY, PROMINENCE AND GENDER

Lluís Danús Amengual

# Contents

<b>1</b>	<b>Introduction</b>	<b>13</b>
<b>2</b>	<b>Infering community structure from data and information theory</b>	<b>23</b>
2.1	Inferential or descriptive approaches . . . . .	23
2.1.1	Bayesian formulation of SBMs . . . . .	25
2.1.2	Selection of the best model . . . . .	26
2.1.3	Connection with information theory . . . . .	27
2.2	Information entropy to measure distances between distributions . . . . .	28
<b>3</b>	<b>Geographical differences in scientific collaboration</b>	<b>31</b>
3.1	Prominent researchers across fields and continents . . . . .	33
3.1.1	Affiliation does not imply scientific output differences	35
3.2	Collaboration network structure unveils geographical polarization of structural roles . . . . .	38
3.3	Affiliation shapes the structure of collaboration patterns . .	43
3.4	The implications of repeated collaborations among prominent scientists . . . . .	47
3.5	Discussion . . . . .	51
<b>4</b>	<b>Prominence and early-career factors</b>	<b>55</b>
4.1	Four common factors in prominent researchers . . . . .	58

CONTENTS	12
<hr/>	
4.2 Early-career factors are connected to research impact . . .	63
4.2.1 Early-career performance is a strong indicator of performance throughout later career stages . . . . .	68
4.3 Factors driving citations and h-index in researchers' early career . . . . .	73
4.4 Discussion . . . . .	77
<b>5 Departmental influence, gender and the selection of research agenda</b>	<b>83</b>
5.1 Creating a background of a scientific field . . . . .	86
5.1.1 Topic models with SBM . . . . .	86
5.1.2 Removing the stopwords from the corpus . . . . .	88
5.1.3 Measuring topical distances between researchers: Jensen-Shannon distance . . . . .	90
5.2 The effect of a new research environment . . . . .	92
5.3 Gender differences in the adaptation to the new faculty position	96
5.4 Career age biased collaborations . . . . .	102
5.5 Discussion . . . . .	104
<b>6 Conclusions and final remarks</b>	<b>107</b>
<b>A Networks and names of researchers for all the fields</b>	<b>121</b>

# Chapter 1

## Introduction

### Science as a social complex system

Societies are systems composed by many interacting individuals whose macroscopic properties cannot be explained by the actions of each component, thus constituting a complex system. People (included some scientists) tend to conceive science as a more rational and independent system within society, but nothing furthest from the truth. Science, because it is composed by social individuals, inherits the properties of the society it is embedded within, displaying its flaws and strenghts applied to the specific context of the scientific community.

Importantly, science comprises many ? actors besides individual researchers, such as institutions or funding bodies. Indeed, although the need of funding varies across fields, it is necessary to procure salaries for the scientific team, recruit new members, buy equipment, attending to international conferences to position the group's research and connection with their peers or paying publication costs in top-tier journals. This, in the end, implies that science is not just about research but also politics. Soft skills such as the ability to establish connections within the institutional structure will be crucial when applying for positions to promote or acquire funding, since on equal terms with a competitor, the hiring committee may be inclined

to choose the one whose name and research is known to them. Funding can be public if it comes from institutions like the government or private if it is from foundations. This highlights the importance for the scientist to know how to tackle different audiences. The committees in charge to decide whether to grant a project, might or might not have specific knowledge on the project's field of research or even not scientific knowledge at all and the skill to communicate the benefits of its research to a broader audience will determine if the project is funded. Of course not only the skills of the researcher will be responsible of the funding received, macroeconomics will also play a role. During periods of financial or global crises, like the 2009 housing crisis or the recent COVID-19 pandemic, global scientific funds might be affected. For instance, in the case of the former, frugal governments cut back the funding for science since they needed to show society how they spent money on things that had an instantaneous effect. Interestingly in this line of acting towards the public opinion, the COVID-19 pandemic had the reverse effect, with society realizing the importance of investing in science, governments raced to increase the funding on R&D gaining in this way a few votes.

Between this constant economic struggle that science faces and the pressure to publish as much as possible, different behaviors emerge when designing the operational structure of a research group. Some research groups are highly hierarchical creating a top-to-down structure where the ideas flow from the PI to the lower levels and rarely backwards, which has potential effects on creativity, motivation of the researchers in training and disruptcy (1). These economic constraints will also affect the size of the teams for a project, the less resources the smaller the team, which will fundamentally affect the type of science that they can perform (2).

Public opinion is also a known driver of societal changes and, since science is a social system, it also can be a driver of scientific change. The most prominent case might be the so-called "War of currents" between Thomas A. Edison and Nikola Tesla. Although Tesla's option was better, Edison, aware of how public opinion could play a decisive role in the success of such a business venture, deceived people with grotesque animal executions using alternating current into believing that it was dangerous. He was temporar-

ily successful in this endeavor, until reality imposed to the fireworks.

In the same way as public opinion, popularity also plays a role in science. There are *rock stars of science*, those scientists with whom everyone wants to collaborate, such as Paul Erdős. Erdős was a brilliant (and somewhat eccentric) Hungarian mathematician, who instead of having an established location, spent most of his career sleeping on his colleagues' couches. He would knock on a colleague's door and say "my mind is open" and after they had written a few papers together, he would go to the next collaborator's house. Collaborating with him was such an honor, that several mathematicians even defined a measure for how many degrees of separation (in terms of scientific work) there were between them, the Erdős number. So, given that there are some scientists more popular than others, we might expect to observe differences between scientists regarding their collaborators, which might lead also to a differences in the variety of research topics (new collaborators bring new ideas) or the total number of publications (share the work load will reduce the amount of time dedicated by article) which ultimately can have influence on the author's impact.

With all these constraints and conventions that rule how science is done, the complexity in the understanding its evolution emerges naturally and important questions arise from there. Are there differences between wealthiest and developing countries? If so, are there measurable consequences of these inequalities? Do the same rules apply to a different kinds of scientists, or are there systematically privileged individuals? Do social problems like race and gender discrimination, have an impact in science? To answer these questions, scientists have a powerful tool at their disposal, the scientific method.

## The scientific method to understand science

Scientists devote most of their time to explain the reality that surrounds us, from matter's atomic structure to cosmic scales. To unveil the mechanism ruling a phenomenon they formulate a plausible hypothesis to explain it, then (when possible) perform experiments to corroborate the hypothesis

and finally compare the output of such experiments with the observed data. This process is the *scientific method* and has been used for centuries as a way to explain nature's mechanisms in the most accurate way possible. Then, it is natural to think if we can use it to explain phenomena from physics, chemistry, biology and so, we should be able to apply the same method to explain science and the factors driving its evolution.

Such a goal is not new. Charles Babbage, a 19th century english mathematician, wrote his "*Reflections on the Decline of Science in England, and on Some of its Causes*" (3). In this work, he analyzed the decreased production and quality of English science. To that aim, he examined the data available at that time about the performing of the educational system, how selective where english scientific institutions in order to accept new members, the investment per capita of the country in scientific areas as compared with their rival nations and also the perception of the average citizen towards science.

Despite this early attempts to analyze science and scientists, it was not until the 1960s when scientists of different areas started to apply a rigorous methodology to examine the bibliometric data that was already available at that time giving birth to the so called *Science of science* or *Metascience*. Some of the most renamed scientometricians were Eugene Garfield, father of the citation indexing and the journal impact factor (4), Robert K. Merton, who coined the term *Matthew effect* (5) in reference to the mechanism by which authors or papers that are already highly cited are more likely to be cited again creating a cascade of citations or Margaret W. Rossiter, who introduced the *Matilda effect* (6) as a contraposition to this Matthew effect where women are less likely to be recognized for the same work than their male colleagues. Among all of them, we can affirm with little doubt, that the most impactful scientometrist has been Derek J. De Solla Price. In his collected lectures *Little science, Big Science* (7), Price analyzes the transition from little science to big science with the exponential growth in science production during the last century which he summarizes in the aphorism "*80 to 90 percent of all the scientists that have ever lived are alive now*". He also covered the emergence of supranational colleges of research thanks to the increasing mobility of academics and collaboration

in scientific publications, as well as the economic and political aspects of science and the relation between the growth of the scientific workforce in developed countries and its increased cost.

With the steady growth of scientific publications (2.5 million per year and increasing) fuelled by the wider offer of online-only journals, the increased scientific workforce, the need to publish in order to obtain funding and other factors, digitalization of bibliometric data and the development of new analysis techniques aided by more powerful computational resources are crucial to analyze science at larger scales than ever. Currently it is possible to access to the publication, citation, authorship and content of millions of scientific articles from all times which has boosted the possibilities to study science gathering scientists from different disciplines, as the study of a multi-faceted complex system requires. From sociologists that contribute with their knowledge of the social processes ruling human interactions, to physicists and mathematicians, whose abstraction and training allows the development of new numerical models to unravel emergent phenomena in many-body systems. Thanks to this junction of knowledge, we observed how funding strategies are related to different impact output (), how the resources and human biases lead to gender inequalities in science production and impact (8) or the broadening and shrinking of scientific cognitive extent (9).

## Network science

Networks or graphs are versatile mathematical objects that can be used to represent real world complex systems. In this representation, the elements or actors of a system are represented by nodes/vertices and their connections are represented as links/edges. This simple yet powerful representation of a system allows us to apply refined mathematical tools that have been developed during many decades in the graph theory framework. Indeed, *Solutio problematis ad geometriam situs pertinentis*, considered as the first work on graph theory by Leonhard Euler, dates back to the 18th century. This work proposed a solution for the *Seven birdges of Königsberg*



problem. Königsberg was a prussian city at both sides of the Pregel river with two fluvial islands with seven bridges to connect all the parts of the city. The mathematical problem was to devise a path to visit all the parts of the city crossing just once by each bridge returning to the initial point. Euler modeled the city as a network (Fig. 1.1) where vertices were the different parts of the city and the edges were the bridges connecting them, proving that such a path was not possible since for such a path to exist, all vertices must have an even number of edges connecting them.

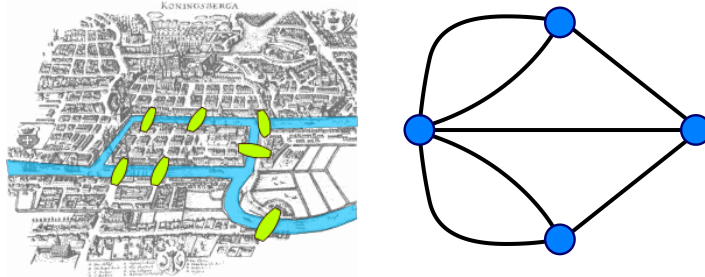


Figure 1.1: **Map of 18th century Königsberg and graph model of the city.** To the left we can observe a map of the city of Königsberg formed by to mainland, two islands and seven bridges. To the right the graph representation, where nodes are the different pieces of land and the edges represent the bridges, with nodes connected as many times as bridges connect two pieces of land.

Moving away from thought experiments, network science has also proven its usefulness solving real problems when combined also with other disciplines. For instance, Gustav Kirchoff combined graph theory with physics to analyze the conservation of current and voltages in electric circuit grids, formalizing this knowledge in the laws that carry his name. With the increase of the computational and algorithmic capacity, network science has been used to tackle a wide variety of areas like the study of air transportation networks (10), urban mobility grids (11), protein interaction networks (12) and also the study of social interaction networks (13). This is

due to the fact that beyond the simple initial formulation of a graph, we can increasingly add layers of complexity to its structure ending in objects with more and more characteristics. To name a few, we can find directed graphs where the direction of the edge between two nodes is important like in the case of links between websites, bipartite networks with more than one type of node and would be representative of plant-pollinator networks and network-based recommendation systems or multilayer graphs consisting on different interconnected graphs representative of different social systems.

Since networks have been used to study social systems and that we can consider science as a social complex system, one might wonder how can we apply this to the study of science. In fact, De Solla Price, who we have already mentioned, used network analysis back in the 60s to study the citation patterns between papers. He constructed a citation network where nodes are scientific papers and they are connected to other papers if there's a citation between them like in Fig.1.2. With this analysis, he conjectured about the "immediacy factor" that quantified the likelihood of citation between two papers according to their proximity in the publication year. He found that most papers are more likely to cite work that lies up to 7 years back of their publication and less likely to cite older papers, pointing that most of the work that is being published relies on findings of their contemporaries.

Focusing more on the social aspects of science at the community level, recent research analyzing the patterns of coauthorship and citation has found that those more centralized scientific communities (those with a highly connected cluster of authors) are more prone to propagate less replicable results than their more decentralized counterparts (15). In this direction, other authors have also analyzed how scientific ideas spread through the community network as a diffusion process with new ideas spreading and gaining popularity in networks with well connected clusters of acquaintances than in the more sparse ones, reflecting the importance of the ego-network structure of a researcher (16). At the individual team level, researchers have also found how the hierarchy as well as the diversity of a team in terms of gender, ethnicity and socioeconomical status affect the later performance and impact () due to the convergence of points of view

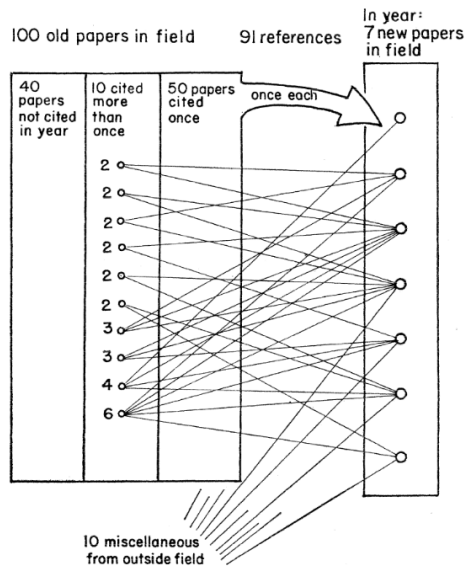


Figure 1.2: **Conceptual citation network from "Networks of scientific papers"** Adapted from (14)

fruit of different vital experiences.

In light of these findings, it is clear that understanding that different economical and sociocultural processes have an impact on how research is produced and on its impact is crucial in order to improve the quality and impact of research. Indeed, a better comprehension of these effects would aid policy makers and institutions to take action and correct possible flaws in the system improving the development of the scientific community as a whole.

This thesis will explore this direction. Using network science as the baseline framework, we will address how different geographical locations, funding agencies and research cultures promote different types of publication strategies ending up in distinct collaboration networks that affect the connectivity of the community and the research impact. Furthermore, we

will also analyze how the research environment shapes the research landscape of early-career researchers and also the role that gender and academic age play in this process.

## Scope of this thesis

The main purpose of this work is to study at which extent some societal biases are reproduced in science. How the different cultural economical and institutional differences shape the way that researchers in geographically distant locations collaborate with each other. Also how scientific prominence is defined and how this definition affects the way we evaluate the careers of scientists along the years. Finally, we address how early career faculty adapt their topics of research when joining a new department, specially how female researchers, a traditionally underrepresented group, evolve. To that aim, this thesis will be divided into several chapters.

Chapter 2 is more methodological; contains an introduction to the techniques used in the posterior chapters. Mainly we will discuss about a type of generative models called Stochastic Block Models (SBM) which will classify the nodes in our networks into blocks or communities according to different structural roles and connection patterns. The flexibility of this framework will be helpful approaching different types of networks that will be used in this work, namely unipartite and bipartite networks. We will also explore concepts from information theory such as the entropy of natural text in the identification of stopwords, the description length to find the most plausible model or the different metrics of distances between distributions.

In chapter 3, we will apply SBM to the collaboration networks of prominent researchers of different fields. Here, we will consider the collaboration networks (in the form of coauthorship networks) that prominent researchers of different fields based in North America and Europe establish with their colleagues. We show how developing research under different economical and academic cultures, such as those in North America and Europe, leads to distinctive collaboration patterns that are consistent across fields with researchers based In Europe engaging in more densely connected communities

than those composed by researchers based in North America. In addition, we will analyze the impact of this different collaborative behaviours in the impact of the research output as well as if the consistency of this behaviors and effects over time.

With the set of prominent researchers from the previous study, chapter 4 will explore the metrics and the definition of prominence in science. We will analyze simultaneously different factors during the early career of a prominent researcher in the aim to predict the main drivers of their excellence. We will show how not only those researchers have an initial advantage, but this privileged start is persistent during the first 20 years of their career. Finally, we will also explore the predictive power of this early metrics when it comes to determine future performance of these researchers by using machine learning tools.

Finally in chapter 5, we study the new collaborations and the research topics of early career faculty after joining a department. Our hypothesis is that are more ductile career stages, such as the initial ones, the topics of research are prone to be affected by the entrance into a new research community like an academic department. We find how the rise of new collaborations and the gender of researchers influences the research topics of these early stage faculty. Firstly, we compare the research topics of authors who joined a department and their counterpart authors who did not, and observe how years after, the research topics of authors who joined the department are more similar to the departmental line of research than the topics of those who declined to join. Then, analyzing a second set of authors who all joined the new department, we focus on the gender differences in this research topics shift and observe how most female researchers, despite having the same number of collaborators, do not converge towards the department and also how they are more likely to collaborate with researchers of different gender than with other females.

## Chapter 2

# Infering community structure from data and information theory

### 2.1 Inferential or descriptive approaches

As stated in the previous chapter, besides the natural choice of using networks to describe interactions between researchers, there are some perks of using them as a representation of a social system. Probably one of the most notorious one, is to use mathematical tools developed in the context of graph theory to study the structure of graphs. The myriad of methods in network science that address the problem of community detection can be divided in two main categories, the descriptive and the inferential ones. The first type of methods consists in analyze the characteristics of a network, like the number of edges between a given set of nodes, and propose a community partition from there. The main issue with this approach is that the output communities carry no explanatory power of the mechanisms that created the network but just describe the output, which can lead to detect communities even in networks originated from a random process (overfitting). Some of the methods belonging to this category are

*modularity maximization* and *RankPage*. On the other hand, inferential methods, generate partitions of nodes and evaluate how likely are these partitions to have generated the observed network. The main advantage of inferential methods is that we can extract knowledge of them, since we have an explanation about the formation process of the network, becoming the preferred option if we want to explain some phenomena. Despite the drawbacks of descriptive methods over inferential ones, they are faster and in some context even a better option to inferential ones. A more exhaustive discussion about whether to use one or the other approach can be consulted in (17).

Here, since we want to extract meaningful partitions of researchers into communities, we will use an inferential approach based on Stochastic Block Models (SBM), as it is one of the most accurate tools to unveil the large-scale structure of complex networks. Initially, SBMs were developed by sociologists (18) in order to overcome several limitations regarding the replicability and generalization of the models studying social networks at that time, as a result of the combination of stochastic models and block models.

The main idea behind SBMs is that in any social network actors (nodes) fulfill different structural roles in it i.e. they have distinct connectivity patterns, and this allows us to group them into different blocks or communities. Specifically, SBMs assume that nodes belonging to the same community are statistically equivalent since the probability of being connected to another node is dependent only on their community membership, allowing to all effects to interchange any node in a given group by another of the same group. A typical example of the group structure that can be found would be the following: We can have communities formed by those who barely interact with other communities but they are connected within its own, those who are connected equally to all other communities or those who are well connected between them and with a selected set of other communities. Notice that no assumption is made in regard to the composition of these groups or what defines them, a clear difference with respect to heuristic approaches which typically make stronger assumptions. For instance, modularity maximization assumes that a community would be a set of nodes that are more connected between them than with other groups, being prone to identify

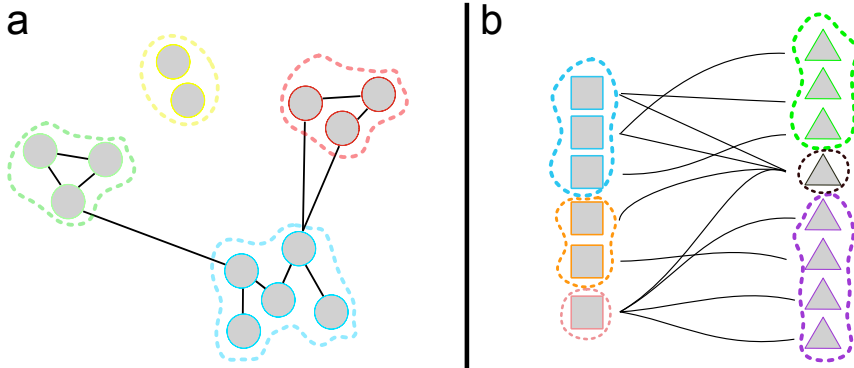


Figure 2.1: **SBMS in different types of networks** **a)** Community structure in a network with simple and undirected links. **b)** Community structure for a bipartite network i.e. a network with two different types of nodes

patterns even in networks generated from random processes (overfitting) or to not identify groups of nodes of similar behavior (underfitting).

Since the main purpose of this work is not to develop community inference tools but to apply them to the studies that we will discuss in later chapters, here we will briefly introduce the Bayesian formulation used in (19).

### 2.1.1 Bayesian formulation of SBMs

Starting from this SBMs framework, we would explore the Bayesian approach to the model selection problem. We consider that there exists a space of models  $\mathcal{M}$  from which we would select a given model  $M$  and examine the plausibility for such a model to have originated our observed network data  $D$  according to the Bayes theorem:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)} \quad (2.1)$$

where  $P(M|D)$  is called the *posterior distribution*,  $P(D)$  is our *evidence* or the probability of the observed data,  $P(M)$  is the probability of the



selected model to be the correct one without any information (also called *prior*) and finally,  $P(D|M)$  is the probability of our selected model to have generated the observed data, which is called the *likelihood*.

Despite its simplicity, Bayes theorem has powerful implications. As one can deduce from the equation, in the absence of knowledge about the system, one can assume a uniform prior distribution  $P(M)$  where each model is equally probable. Then, alongside new observations of our data, this prior need to be updated, since now we have more information about the system, leading us to a better comprehension of which model is the most plausible one.

### 2.1.2 Selection of the best model

Bayes rule in Eq. 2.1 needs to be adjusted by selecting different sets of constants (or parameters) in order to obtain the model that best fits the data. To do so, the prior distribution in Eq. 2.1 can be rewritted as follows:

$$P(M|D) = \frac{1}{P(D)} \int_{\Omega} P(D|M, \Omega)P(\Omega|M)P(M)d\Omega \quad (2.2)$$

where now we have the marginalization of the prior over all the possible values  $\Omega$  of parameters. As one can observe, the predictive power of the posterior will be mostly affected by our prior distributions  $P(\Omega|M)$  and  $P(M)$  and by modifying them we can improve the plausibility of our model.

Focusing on the probability of the model  $P(M)$ , we want to consider some corrections about our previous assumptions. We initially considered that all models are equally probable, but in most cases this leads us to the undesired result of having practically the same number of groups as nodes i.e. small groups consisting of a few nodes, favoring more complex models instead of more simple ones. This can be overcome by sampling the probability of our model from its own prior, *nesting* the distributions. At the same time, this correction introduces a bias towards models with groups of the same size that would be unrealistic in most of real-case scenarios, which is solved by sampling also group sizes from a non-informative distribution.

The final correction to be made, is a correction that overwrites one of the assumptions of the original SBM, the statistical equivalence of nodes in a group. Again, we face that in real-case situations, actors belonging to the same community, are not necessarily equally connected. Therefore, allowing nodes in a given community have different degrees (known as *degree correction*) allows us to find a richer and more plausible variety of models.

### 2.1.3 Connection with information theory

But how should we identify the most plausible model among all models? To address this question we must turn to information theory, more specifically, to the concept of entropy. The interpretation of the concept of entropy has fostered intense debates among physicists along the years. Initially entropy was defined as the heat lost during an spontaneous irreversible process. Years later, Ludwig Boltzmann proposed an statistical interpretation of the entropy, as the probability distribution of the possible microstates of a thermodynamic system. With this as a basis, Claude E. Shannon proposed the information-theoretical interpretation of entropy, defining it as the uncertainty we have about a given event. Formally it is expressed as:

$$H(x) = - \sum_i p(x_i) \cdot \log p(x_i) \quad (2.3)$$

where  $p(x_i)$  is the probability of  $(x_i)$  to happen. This equation gives us a lower bound to the amount of information that we need to completely describe a given event i.e. the length of the most efficient description. In this particular case, since we are taking a base 2 logarithm, the amount of information would be measured in bits. As an example, we can consider the popular game *20 questions* where player 1 has to guess in 20 questions the word that player 2 is thinking of. For simplicity, we will consider a set of three words  $\mathcal{M} = \{tree, car, orchid\}$ . If player 1 takes a more reckless and brave approach directly asking if the word is *car* and fails, he would need an additional question in order to correctly guess the hidden word. On the other hand, if player 1 chooses to play it wiser, he can ask player 2: *is it a living being?* If it is, he would need an additional question, but if it is not, he

would already have the answer with just a single question, giving on average a lower bound of 1.5 questions to win the game. Indeed, if we consider Eq. 2.3 it comes out that  $H(x) = \frac{1}{3} \log(\frac{1}{3}) + \frac{1}{3} \log(\frac{1}{3}) + \frac{1}{3} \log(\frac{1}{3}) = 1.58$ .

If 2.3 gives us the lower bound to the length of the most efficient way to describe a system it would also set a lower bound to our model:

$$\log P(M, \Omega|D) = -\frac{1}{P(D)} \log \int_{\Omega} P(D|M, \Omega)P(\Omega|M)P(M)d\Omega \quad (2.4)$$

where the term inside the logarithm is also called the *description length*, which is the most plausible model to describe our data. To illustrate this, lets see a real-world example by looking at Tables 3.1 which corresponds to the description length (in natural units or nats) at different levels of refinement of the SBM when modelling the collaboration networks that we will analyze in the next chapter. As one can observe, the classic SBM is the one carrying a more complex (long) explanation of the network in each case. Then, along introducing more refined features such as the *degree-correction* and the nested priors for the  $P(M)$  the description length decreases giving the lowest value for all the networks when all the corrections mentioned in this chapter are included.

## 2.2 Information entropy to measure distances between distributions

As we have seen considering data as events happening with a certain probability, allows us to define a lower bound to how surprising is the observation of those events. What happens when instead of isolated events we want to have information about what can we expect from a complete set?

With the Bayesian framework in mind lets consider how we can measure the information gained when we update our beliefs. Let's consider an observed distribution  $p$  and a theoretical distribution drawn from a model,  $q$ . Then difference in the number of bits needed to describe our data could be written as:

$$\Delta H = \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x) \quad (2.5)$$

where the first term is the entropy corresponding to  $p$  being originated from  $p$  and the second term corresponds to the entropy of the expected value of  $p$  being drawn from our model distribution  $q$ . This, general is expressed as:

$$D_{KL}(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \quad (2.6)$$

which is called the *Kullback-Leibler divergence*.

UNIVERSITAT ROVIRA I VIRGILI

NETWORK AND INFORMATION-THEORETIC STUDIES ON THE EFFECTS OF RESEARCH ENVIRONMENT IN  
SCIENTIFIC CAREERS: GEOGRAPHY, PROMINENCE AND GENDER

Lluís Danús Amengual

## Chapter 3

# Geographical differences in scientific collaboration

Science is a social endeavor that progresses through the concerted effort of many individuals, who exchange ideas and interact through intricate collaboration networks (20–23). Due to the increasing complexity involved in the most pressing problems in science and society, and the advantage of diverse groups at solving complex tasks (24, 25), the role of these collaboration networks is becoming more and more important to achieve scientific excellence and advance research fields (26, 27). Additionally, the structure of collaboration networks affects the quality and scope of research outcomes in different ways, some of which have been well described. In particular, networks with more recurrent collaborations (number of publications co-authored by a pair of researchers) have been linked to research with lower impact (23, 28). Also, centralized scientific communities (those with a highly connected cluster in which the same group of scientists repeatedly co-author articles) are more likely to propagate non-replicable claims, and vice versa for decentralized communities with less overlap in co-authorship and more diverse methods (15). The structure of the collaboration network also has an impact on the career of researchers. For example, network structure is predictive of who produces groundbreaking ideas, and who wins scientific

prizes and awards (29).

At the same time, there is mounting evidence that different research environments, such as different funding and institutional arrangements or different scientific and political cultures, leave measurable fingerprints in collaboration networks (2, 30–32). For example, we know that resource-intensive fields (such as astrophysics or high energy physics) typically have collaborations involving a large number of researchers (more than 100) and, therefore, denser collaboration networks (23, 33). Resource demands also result in gender imbalance: women tend to be excluded from resource-intensive fields that require large collaborations (for example, genomics versus plant sciences in biology) and therefore end up working in smaller teams (8).

Since funding and institutional arrangements and scientific and political cultures differ across regions in the world in systematic ways, we surmise that collaboration networks should also differ systematically across regions, independently of other factors such as research field. Additionally, because of the effect of collaboration network structure on research outcomes, we expect to observe systematic differences in the impact of research produced in North America and Europe. Such differences have indeed been observed (34, 35); we explore whether they are affected by collaboration-related factors.

In this chapter, we address the lack of comparative studies on collaboration networks across regions by collecting data on field-specific collaboration networks for eight different fields and classifying prominent researchers based on their institutional affiliation in one of these two regions. Then we will construct the collaboration network of this prominent researchers on the basis of their coauthorships and analyze the community structure of the collaboration network for the whole field and examine their differences in behavior. Finally, we will examine if there is a measurable effect of this different collaborative structures in the impact of their research.

### 3.1 Prominent researchers across fields and continents

We start by collecting data on the scientific collaboration networks between roughly 100 prominent researchers in eight different scientific fields: genetics, development economics, cognitive psychology, philosophy of science, network science, metabolomics, network ecology, and social inequalities in health. We focus on prominent researchers for two main reasons. First, elite researchers are responsible for much of the impact and research focus in any field (5, 36–39). Second, since they also receive a disproportionate share of the funding in their field, they are more likely to be sensitive to institutional arrangements, scientific cultures and funding strategies.

We choose these eight fields because they provide a broad scope of fields spanning across the natural and social sciences and they are sufficiently small and well-defined for prominent researchers to collaborate with one another, while being sufficiently established to have a consistent track record of collaborations between prominent researchers, and of the impact of these collaborations. Moreover, these fields are diverse in terms of the topics covered, their scientific cultures, and how established they are. In particular, the first four fields have longer traditions, whereas the latter are relatively young and have evolved for shorter times. Finally, three of these fields have a majority of prominent researchers based in Europe whereas five have a majority based in North America, with overall 40% of researchers based in Europe and 60% in North America (Fig. 3.8).

For constructing the lists of prominent researchers, we used the following procedures: The list for social inequalities in health was previously collected by another research group in (40). For the four more established fields in our analysis (genetics, development economics, philosophy of science, and cognitive psychology) we selected the 100 researchers with the highest H-index in their field using Google Scholar in January 2021. We then confirmed our initial list using Scopus' citation and H-index data. To ensure that all researchers commonly viewed as the most influential were included in the top 100 lists, we checked common rankings of the most



influential researchers for each of these fields. We incorporated the few top researchers in these ranked lists, who were not already among the top researchers according to Google Scholar. As these four fields have a longer tradition than the other four fields, we only included researchers with publications between 1960 and 2021.

For the four younger fields, for those with well defined conferences and scientific societies (network science and metabolomics) we identified the main conferences (NetSci, NetSciX and CompleNet, for network science; and events of the Metabolomics Society for metabolomics) and societies (Network Science Society and Metabolomics Society), and considered all researchers who gave talks, are in scientific committees and scientific boards, and received awards in these venues. The authors in this list were identified in the Scopus database and ranked by their H-index. We used H-index to rank because it offers balance between the number of publications and the number of citations received by those publications, thus usually being interpreted as a more appropriate measure of the 'quality' of a researcher (41). Additionally, since we analyze quantities related to number of articles and citations, choosing researchers by either of these criteria could lead to selection effects.

For the field of network ecology, we assembled the initial list by querying the Scopus database using a series of keywords (Ecologi\* Network\*, Food Web\*, Environment\* Network\*, Trophic\* Network\*, Trophi\* Web\*) and focusing on top interdisciplinary and ecology journals. We then refined and ranked the list using the H-index, as before.

In all cases, we excluded the few researchers (a total of 6%) not based in Europe or North America, or who did not collaborate with any other prominent researchers in the network. We also checked manually that all researchers in the network really have a significant body of work in the field, and excluded a few scientists that are prominent in other fields and have only made a small contribution to the field under consideration.

Note that we consider all the publications of authors that are prominent in each field, including publications in other fields. This is because we are interested in all the collaborations between these researchers. Additionally, we assign each prominent researcher to their main current affilia-

tion, although some of them have developed parts of their careers in North America and Europe. The latter group is mostly composed by European researchers moving to North America; often, those researchers are still placed in groups with other European researchers, probably due to a maintenance of the preexisting collaboration network. Notice that these researchers are labeled as North-American; if we labeled them as Europeans, the polarization and the differences in collaboration patterns between Europe and North America would be even higher than the ones we report.

The full list with the names of the prominent researchers are provided in Appendix Figs. S??-S??. Note that our criteria guarantee that all scientists in the network are prominent, although different criteria may result in somewhat different prominent researchers. Overall, all 100 researchers identified in each field are among the most highly cited and influential researchers with the highest H-index in their given field. We validated the data set by using an alternative method based on a Scopus search by keywords. The overlap with the network identified here, in network science for example, was 90% and all results in the study remained the same.

### 3.1.1 Affiliation does not imply scientific output differences

Collaboration patterns, and their outcomes in terms of publications, do not appear, at first glance, to be vastly different for prominent researchers in North America and Europe (Fig. 3.1). In both cases, we observe large variability in the total number of collaborators and the total number of publications of prominent researchers. Because of this variability, in what follows we consider the logarithm of the number of collaborations, the number of collaborators and the impact of publications (42). As expected, we observe that the number of collaborators grows with the number of publications; but we observe no consistent significant differences between Europe and North-America (Fig. 3.1A-H) (except in the case of network ecology,  $p = 0.02$ , and genetics,  $p = 0.01$ ).

While prominent researchers in Europe have significantly larger collaborations in network ecology, development economics, genetics and cognitive psychology, the differences are not significant in the other four fields

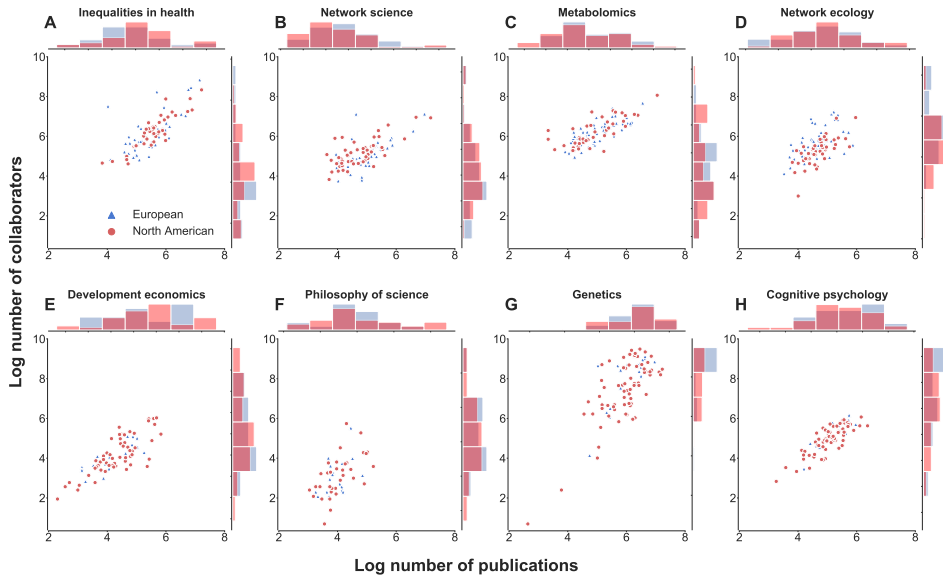


Figure 3.1: **Total number of collaborators and publications for prominent researchers (A) - (H)** Logarithm of the total number of collaborators as a function of the logarithm of the total number of publications for each prominent researcher in: (A) inequalities in health, (B) network science, (C) metabolomics, (D) network ecology, (E) development economics, (F) philosophy of science, (G) genetics and (H) cognitive psychology. Red circles and blue triangles correspond to prominent researchers based in North-America and Europe, respectively. We test whether the points are distributed differently using the 2D Kolmogorov-Smirnov statistic (43), and calculate the significance by resampling the researchers' affiliations. At the 5% confidence level, we can only reject the null hypothesis (that both subsets are drawn from the same distribution) in the case of network ecology.

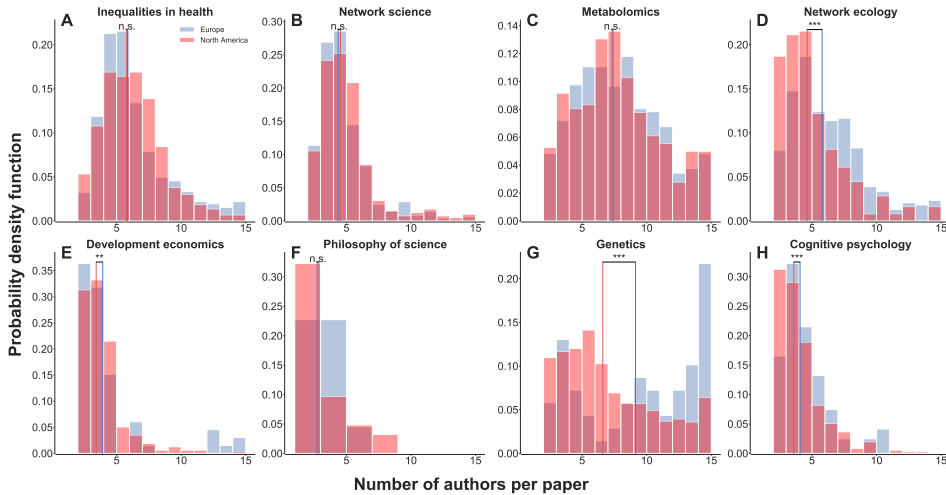


Figure 3.2: **Team size distribution for prominent researchers (A)-(H)** Distribution of the logarithm of the number of authors per paper for papers with only Europe-based prominent researchers (blue) and only North America-based prominent researchers (blue). The vertical lines indicate the mean logarithmic number of authors per paper for each subset of researchers. Stars indicate statistical significance at different levels (\*\*\*: 1%, \*\*: 5%, \*: 10%, n.s.: not significant).

(although collaborations in network science and philosophy of science are slightly larger for prominent researchers in North America). In general, fields like genetics and metabolomics have larger numbers of authors per paper (Fig. 3.2) in part because they often require access to laboratories, large-scale resources, sophisticated instruments etc. that an individual researcher does not generally possess. This is not often the case in fields like philosophy of science and development economics.

## 3.2 Collaboration network structure unveils geographical polarization of structural roles

Despite the lack of systematic differences among geographical groups of prominent researchers in terms of the total number of publications and collaborations, a more nuanced analysis of the structure of the collaboration network between prominent researchers (that is, excluding their collaborators who are not prominent) reveals systematic and consistent differences between North America and Europe. We start by constructing a network where each researcher (node) is connected to other researchers by as many edges as coauthorships they share. Figure 3.3 shows the network of one of the fields under study and there the size of the nodes (researchers) represents the betweenness centrality of the researcher in the collaboration network, which indicates how important is one node in terms of connecting shortest paths between other vertices i.e. in allowing information spreading to the whole network.

Once the network is constructed, we model it using a degree-corrected hierarchical (19) stochastic block model (hSBM) (44–46) (Fig. A.7 and Figs. 3.5–3.6). As mentioned in chapter 2, with this approach, researchers in the same group occupy a similar position in the network and thus play a similar role (44). Unlike other methods to identify groups, roles and/or positions in networks, our approach (Bayesian maximum a posterior, or, equivalently, minimum description length (Tables ??-??); Eq.2.3) guarantees that the partition of the network into groups is the most parsimonious.

We observe that the groups we obtain are markedly polarized in their composition (Fig. A.7B,C), with some groups containing mostly researchers in Europe and others containing mostly researchers in North America, meaning that researchers with the same structural role are typically based in the same continent. To quantify the affiliation imbalance of the groups identified by the hSBM, we defined group polarization  $g_p$  as follows. For each researcher  $i$  in a group of prominent researchers, we calculated the fraction of others in the group that belong to the same continent as  $i$ . Then, the mean group polarization  $g_p$  is calculated as a mean over all researchers

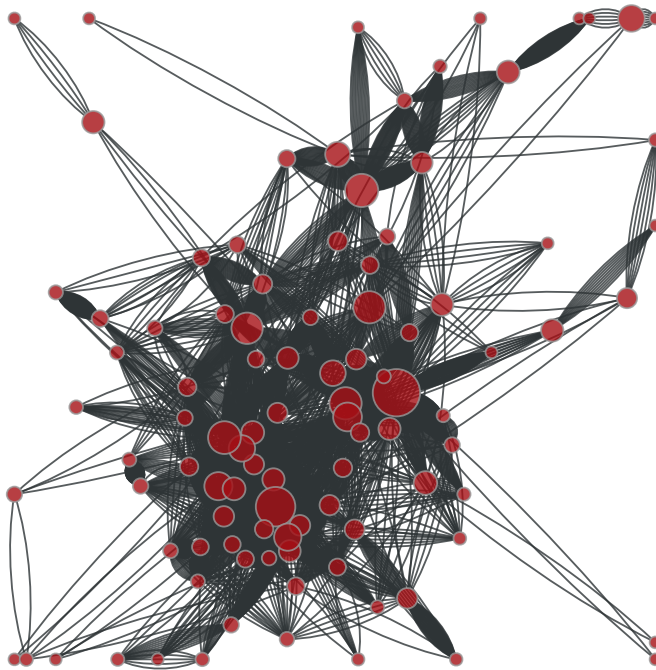


Figure 3.3: **Collaborations in a field:** Example of the collaboration network for the field of *Social inequalities in health*. Each one of the nodes is a researcher and nodes are connected as many times as they appear as coauthors in publications.

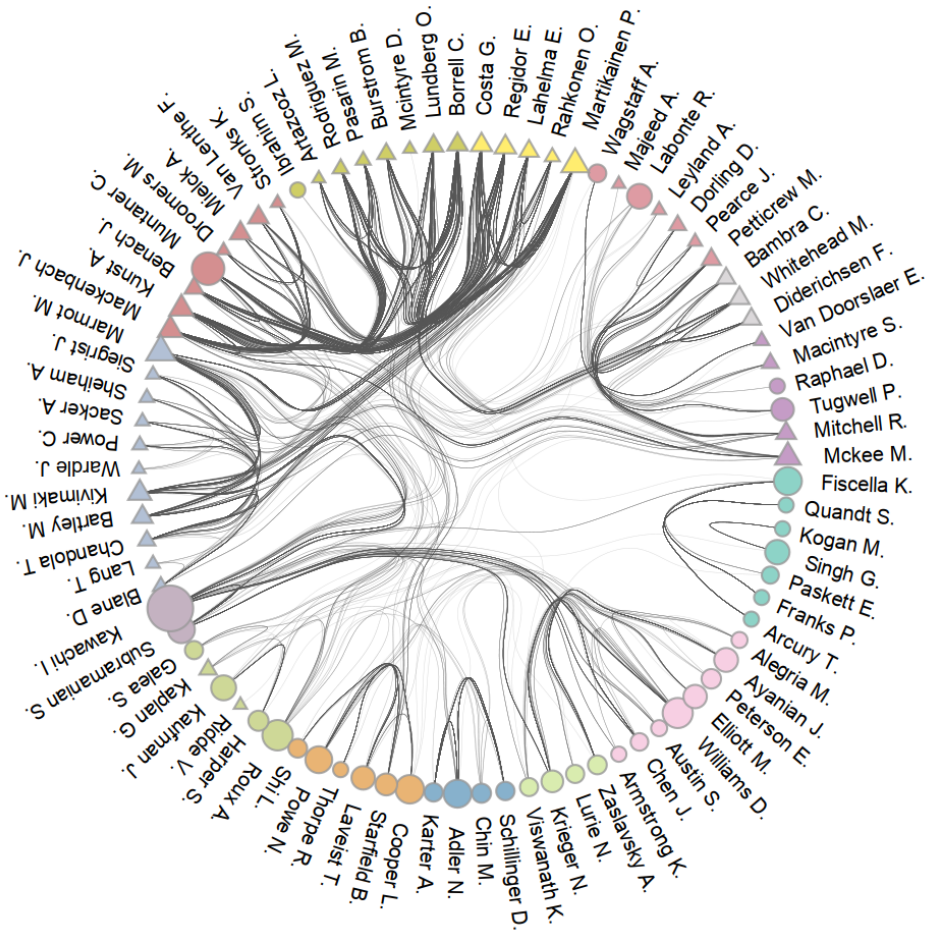


Figure 3.4: **Community structure:** Collaboration network and best fit of the hierarchical stochastic block model (hSBM) for the network in Fig. 3.3. Each node in the network represents a prominent researcher, and each edge represents a different collaboration (coauthored paper) between a pair of researchers. Prominent researchers in North America and Europe are represented as circles and triangles, respectively. Different colors correspond to the groups identified by the hSBM, so that nodes with the same color have a similar collaboration pattern with other researchers and therefore fulfill a similar structural role in the collaboration network. Node size represents the betweenness centrality of the researcher in the network.

Model	MDL Ineq. health	MDL Netw. sci.	MDL Metabol.	MDL Netw. ecol.
DC-hSBM	<b>3467.19</b>	<b>3819.11</b>	<b>4546.09</b>	<b>3512.45</b>
NDC-hSBM	3491.99	3868.24	4570.06	3562.95
DC-SBM	5960.63	5246.32	6582.81	4850.89
NDC-SBM	6516.09	5530.36	7071.51	5200.47

Model	MDL Dev. econ.	MDL Genetics	MDL Cogn. psy.	MDL Phi. science
DC-hSBM	<b>1495.42</b>	<b>5214.26</b>	<b>1338.41</b>	<b>453.26</b>
NDC-hSBM	1527.16	5222.34	1385.35	460.72
DC-SBM	1893.01	7455.22	1697.17	477.65
NDC-SBM	2008.08	7978.52	1794.72	486.83

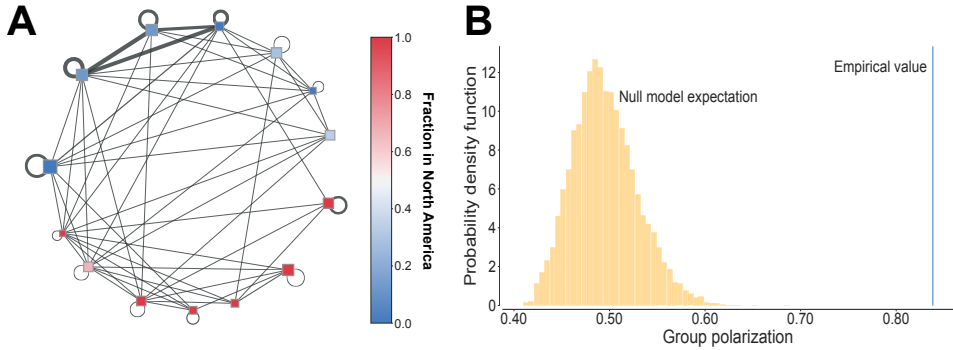
Table 3.1: Minimum description lengths (MDL) achieved by different models for each of the collaboration networks (DC-hSBM: degree-corrected hierarchical stochastic block model; NDC-hSBM: non-degree-corrected hierarchical stochastic block model; DC-SBM: degree-corrected stochastic block model with non-informative priors; NDC-SBM: non-degree-corrected stochastic block model with non-informative priors). The MDL among models (bold) always corresponds to the DC-hSBM model.

in all groups:

$$g_p = \frac{1}{N} \sum_{g=1}^G \sum_{i=1}^{n_g} \frac{c_{ig}}{n_g - 1} \quad (3.1)$$

where  $N$  the number of researchers in the network,  $G$  is the number of groups,  $c_{ig}$  the number of researchers in group  $g$  (other than  $i$ ) belonging to the same continent than node  $i$ , and  $n_g$  the total number of nodes in group  $g$ . Thus, the polarization of the network is  $g_p = 1$  if all groups comprise researchers only from North America or only from Europe, but no group contains researchers from both. In order to assess the statistical significance of the observed polarization, we compare those numbers to the null expectation obtained by resampling researchers' institutional affiliations where we randomly reassigned the affiliations of all researchers in the network and calculated  $g_p$  maintaining the same groups, and repeated this operation many times to obtain the null distribution of  $g_p$ . We find that group polarization is highly significant in all fields except philosophy of science, where the scarcity of collaborations leads to non-significant results (Fig. 2C and Fig. 3.6). This indicates that prominent researchers in





**Figure 3.5: Group dynamics and polarization (A)** Block model of the collaboration network. Each node represents a group of researchers with similar collaboration patterns (that is, a different color in Fig. A.7), with node size representing the number of researchers in the group. The width of the edges represents the number of collaborations between groups, and loops represent collaborations within each group. The color of each node indicates the fraction of researchers in the group that are based in North America, so that dark blue nodes represent groups with mostly Europe-based researchers, and red nodes represent groups with mostly North America-based researchers. **(B)** We define the polarization of a group as the number of same-continent researchers in the group over the random expectation for such number. The vertical line indicates the mean group polarization for the observed collaboration network. We randomize authors' affiliations and calculate the distribution of expected (null) polarization values. The empirical value is well above the null expectation, so that the group structure of the observed network is significantly polarized

North America and Europe fulfill distinct structural roles in collaboration networks between prominent researchers.

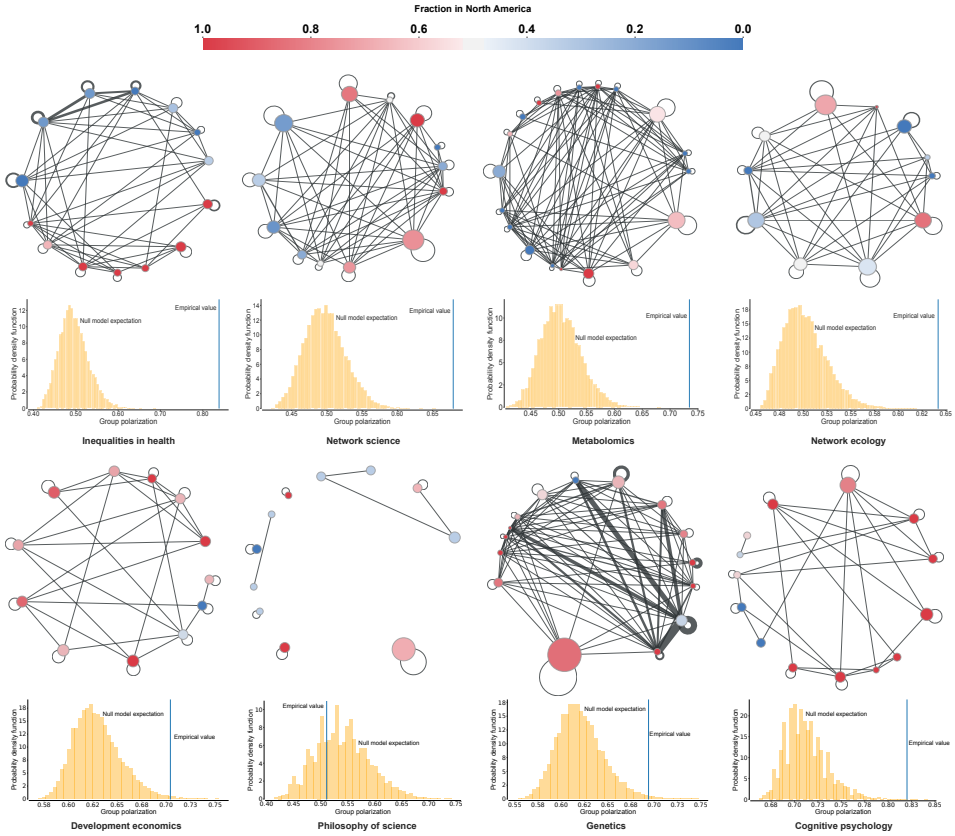
Group polarization could be naively attributed to geographic proximity, that is, to the tendency of researchers based in the same continent to collaborate; indeed, this would lead to polarized groups. However, deeper analysis of the collaboration networks and the corresponding block models (Figs. A.7-3.6) reveals that this is not the only factor at play. Rather, we observe genuinely different collaboration patterns across continents. Groups with more Europe-based researchers tend to have more

within-group and between-group collaborations, whereas groups with more researchers in North America tend to have fewer collaborations altogether. In the following, we quantify these differences directly in the collaboration networks between prominent researchers.

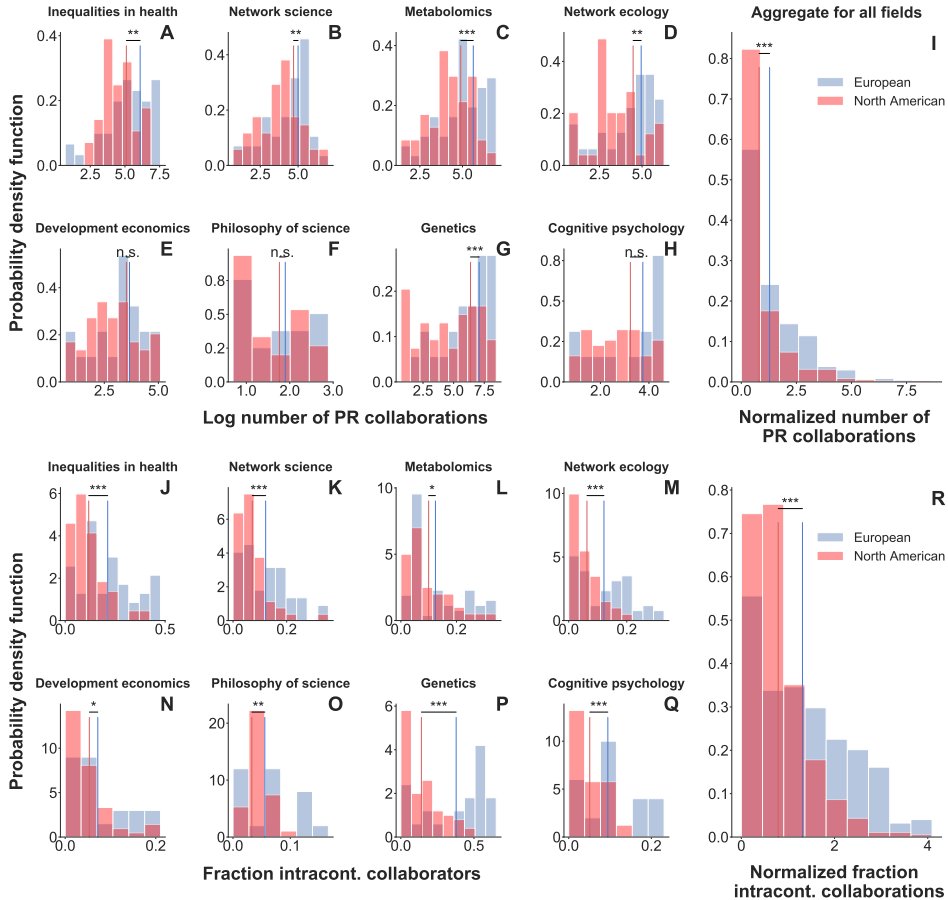
### 3.3 Affiliation shapes the structure of collaboration patterns

First, we measure the total number of collaborations between each researcher and other prominent researchers (Fig. 3.7A-H). When counting collaborations, several repeated collaborations with the same alter prominent researcher are counted separately, so that one collaborator can give rise to several collaborations. Across all fields, we find that the average number of collaborations with other prominent researchers is always higher in Europe than in North America even though in fields with lower collaboration rates among prominent researchers the differences are not statistically significant (Fig. 3.7). When all fields are combined (normalizing each field by its expected number of collaborations), the difference is significant at the 1% level (Fig. 3.7I). Similarly, a significant majority of researchers with above-median number of collaborations with other prominent researchers are based in Europe, whereas the majority of researchers with below-median number of collaborations with other prominent researchers are based in North America (Fig. 3.8). Taken together with the fact that the total number of collaborators does not differ significantly between Europe and North America (Fig. 3.1 and Fig. 3.2), these results indicate that researchers based in North America have a higher tendency to collaborate with non-prominent researchers, whereas in Europe the research elite in a specific field is more tightly knit.

Second, we measure, for each prominent researcher, the fraction of prominent researchers in their continent with which the researcher has collaborated. We call this the *fraction of intracontinental collaborators* (Fig. 3.7J-Q); a value of 0.5 indicates that a prominent researcher has collaborated with half of the prominent researchers in their continent. If we



**Figure 3.6: Group dynamics and polarization for all fields** Block models of the collaboration networks of all fields. As in Fig. 3.5(A), each node represents a group of researchers with similar collaboration patterns, the width of the edges represents the number of collaborations between groups and loops the within-group collaborations, and the color of the nodes the number of researchers in the group that are based in North America. Below each block model, we represent the polarization measured as in Fig. 3.5(B) with the number of same-continent researchers in the group over the random expectation for such number and the vertical line indicates the mean group polarization for the observed collaboration network.



**Figure 3.7: Differences in collaboration patterns between prominent researchers in North America and Europe.** (A)-(H) Number of prominent researcher (PR) collaborations. We plot the distribution of the logarithm of number collaborations for prominent researchers in North America (red) and Europe (blue). The vertical lines indicate the mean log-number of collaborations for each subset. The significance of the difference between the European and North American means was established by resampling researcher affiliations (one sided test). (I) Aggregated distribution for all fields. The log-number of PR collaborations are normalized by the mean in each field so as to make all fields comparable. (J)-(Q) Fraction of intracontinental collaborators, defined as the fraction of prominent researchers in the same continent with which a prominent researcher collaborates. We plot the distribution of the fraction of intracontinental collaborators in North America (red) and Europe (blue). The vertical lines indicate the mean fraction of intracontinental collaborators for each subset. The significance of the difference between the European and North American means was established by reshuffling researcher affiliations (one sided test). (R) Aggregated distribution for all fields. The fractions of intracontinental collaborations in each field are normalized by the mean of the field so as to make all fields comparable. Stars indicate significant differences (\*\*\*: 1%, \*\*: 5%, \*: 10%, n.s.: not significant).

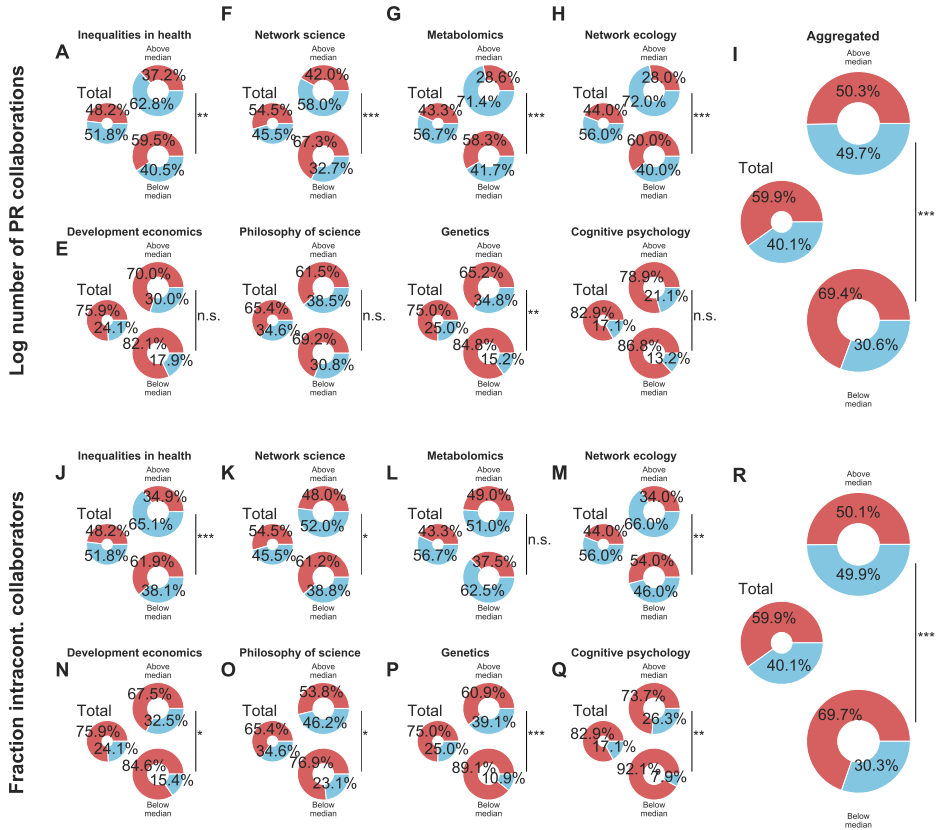


Figure 3.8: Fraction of researchers above the median of the total number of connections (first two rows) and fraction of researchers above/below the median of the fraction of intracontinental collaborators (last two rows). We represent the same metrics as in Fig. 3 as the fraction of researchers that are above and below the median. The significances have been obtained by reshuffling researcher's affiliations and comparing the random median with the empirical value. Stars indicate significant differences (\*\*\*: 1%, \*\*: 5%, \*: 10%, n.s.: not significant).

pool all fields together, we find that the fraction of intracontinental collaborators normalized by field is significantly higher in Europe than in North America at 1% level (Fig. 3.7R). For individual fields, we find that the mean fraction of intracontinental collaborators is always significantly higher in Europe than in North America and that prominent researchers in Europe have significantly above-median intracontinental collaborators for all fields (Fig. 3.8), except for metabolomics.

### 3.4 The implications of repeated collaborations among prominent scientists

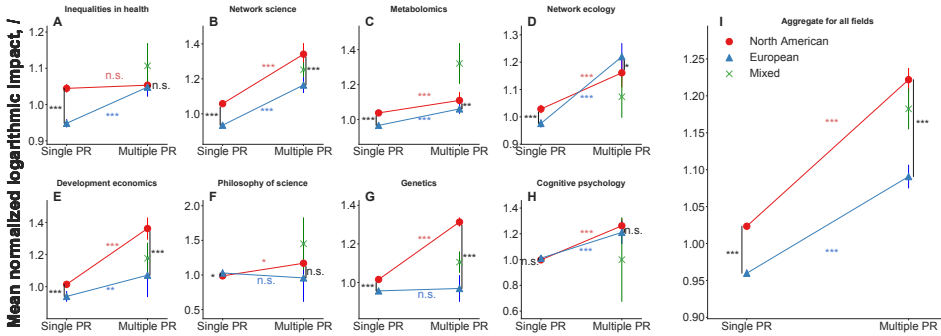
If, as we have shown, collaboration patterns are different across continents, and if collaboration network structure affects research performance (15, 23, 28), then we expect systematic collaboration-dependent differences in impact across continents. To investigate this question, we analyze the impact of publications of prominent researchers, both when they publish with and without other prominent researchers (Figs. 3.9-3.10).

To measure how collaborating with other colleagues affects the research impact we define the normalized logarithmic impact as follows. The normalized logarithmic impact  $I_i$  of a paper  $i$  is the logarithmic number of citations (plus 1)  $\log(k_i + 1)$  of the paper divided by the mean of the logarithmic number of citations (plus 1) of papers with no prominent researcher collaboration in the same publication year

$$I_i = \frac{\log(k_i + 1)}{\langle \log(k + 1) \rangle_{y_i}}. \quad (3.2)$$

Here  $\langle \dots \rangle_y$  is the mean over all papers published in year  $y$  by single prominent researchers (and, possibly, other non-prominent researchers, but not multiple prominent researchers). Comparing with publications in the same year, allows us to avoid the artifact of later collaborations being less impactful just because they have had less time to accrue citations.

We find that, in general, researchers in North America publish significantly more impactful papers than those in Europe when they publish



**Figure 3.9: Impact difference under different collaborative strategies.** Mean normalized logarithmic impact for articles authored by either a single prominent researcher (PR) or multiple PR in: **(A)** inequalities in health, **(B)** network science, **(C)** metabolomics, **(D)** network ecology, **(E)** development economics, **(F)** philosophy of science, **(G)** genetics and **(H)** cognitive psychology. The normalized logarithmic impact  $I_i$  of a paper  $i$  is the logarithmic number of citations (plus 1)  $\log(k_i + 1)$  of the paper divided by the mean of the logarithmic number of citations (plus 1) of papers with no prominent researcher collaboration in the same publication year (Methods). **(I)** Aggregated normalized logarithmic impact for all fields. Stars indicate significant differences (\*\*\*: 1%, \*\*: 5%, \*: 10%, n.s.: not significant). See Extended Data Fig. 5 for the whole distributions of the logarithmic number of citations.

without other prominent researchers in their field (in philosophy of science and cognitive psychology the differences are not significant). Since, as we have seen earlier, prominent researchers in Europe collaborate more with other prominent researchers (Fig. 3.9), this may provide a mechanism to compensate, by means of collaboration, for the lower impact of their work without other prominent researchers.

We also find that collaborating with other prominent researchers increases, by 15% on average across all fields, the impact of publications (differences not significant for Europe-based researchers in philosophy of science and genetics, and North America-based researchers in inequalities in health). The prominent researchers in Europe and North America who benefit the most, in terms of higher publication impact, by collaborating

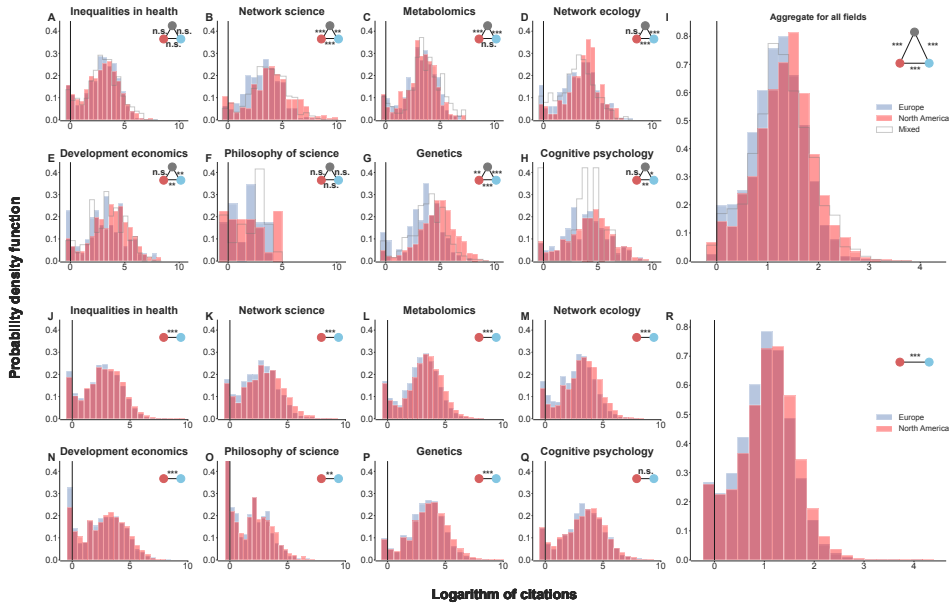


Figure 3.10: Distribution of the logarithmic number of citations for articles with multiple prominent researchers (A-H) and a single prominent researchers (J-Q) For the same publications as in main text Fig. 3.9, we plot the whole distribution of the logarithmic number of citations. Articles with 0 citations are represented left of the vertical axis at 0. The differences between distributions are calculated using the means of the distributions, and their significance is obtained by repeatedly reshuffling researchers' affiliations. The right column represents the aggregated value for all fields in multiple prominent researcher (I) and single prominent researcher (R) publications



with other prominent researchers are those in the fields of network science, with an increase of 25% and 33%, and development economics, with 20% and 30%, respectively. This finding shows that the publishing with prominent researchers is not only beneficial for early career scientists, but, in general, for prominent scientists as well (47).

However, previous results linking collaboration network structure to outcome quality (23, 48) have generally indicated that repetitive collaborations with the same researchers and largely closed collaboration networks (as those observed in Europe) result in lower reproducibility and impact. Given the observed differences in collaboration patterns between continents, we investigate in more depth the effect of repeated collaborations (collaboration number) on the value added by collaboration (Figs. 3.11-3.12). Specifically, we analyze the normalized logarithmic impact for the first two collaborations among each pair of prominent researchers, the third to fifth collaborations, and the sixth collaboration and higher. Although the numbers in each field are small, often leading to non-significant differences, when all fields are pooled together a clear and significant pattern emerges: the more times a collaboration is repeated, the lower the impact (with collaborations between prominent researchers in North America always having higher normalized logarithmic impact). The first two collaborations among prominent researchers increase (on average) the impact with respect to papers with a single prominent researcher by 34% for North America based and 23% for Europe based researchers. For 3-5 (and 6 or more repeated collaborations) the increase in impact is lower: 29% (21%) and 22% (12%) for North America and Europe, respectively. Trends among Europeans and North Americans follow similar patterns within all fields (with Europeans having overall slightly lower impact). Nonetheless the increase and subsequent decrease varies across all fields. This suggests that the nature of the returns to repeated collaborations are also influenced by field-specific features and not just the overall research environment and the number of times researchers collaborate. Note that our results are not contradictory with the finding that super-ties (i.e. scientific collaboration dyads that are sustained over time) in general increase publication impact (49): our results rather suggest that super ties have a discernible positive effect when they

involve only one prominent scientist.

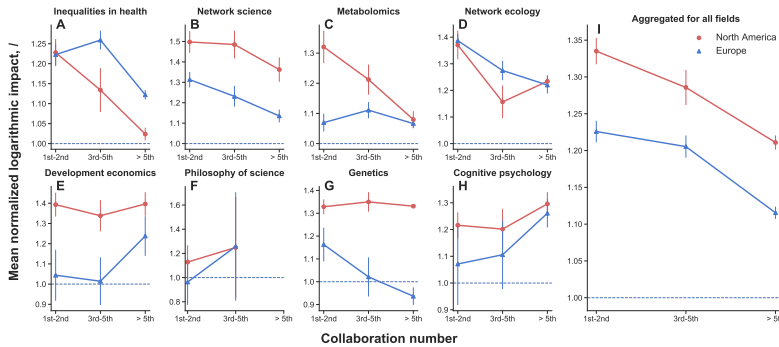


Figure 3.11: **Evolution of impact with repeated collaborations.** Mean normalized logarithmic impact of publications authored by a pair of prominent researchers (PR) as a function of the collaboration number (the number of times two prominent researchers have co-authored a paper: 1-2, 3-5, or >5; Methods). The normalized logarithmic impact  $I_i$  of a paper  $i$  is the logarithmic number of citations (plus 1)  $\log(k_i + 1)$  of the paper divided by the mean of the logarithmic number of citations (plus 1) of papers with no prominent researcher collaboration in the same publication year (Methods). (A) inequalities in health, (B) network science, (C) network ecology, (D) metabolomics, (E) development economics, (F) philosophy of science, (G) genetics and (H) cognitive psychology. (I) Aggregated normalized logarithmic impact for all fields, as a function of the number of collaboration.

### 3.5 Discussion

In studying complex systems like the scientific process or collaboration networks, we are often constrained in precisely measuring causal relations. Here, we surmised that collaboration networks and scientific impact differ systematically across regions, and we found that the empirical evidence indeed supports this hypothesis. This does not prove that the research environments in Europe and North America are directly responsible for the observed differences in collaboration structure (and, indirectly, impact);

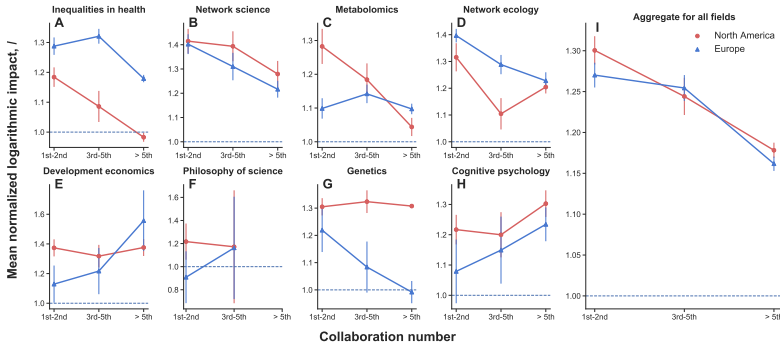


Figure 3.12: **Evolution of impact with repeated collaborations.** We represent the same data as in Fig. 3.11 normalized by each affiliation separately.

but considering that research environment is known to affect collaboration network structure in some cases (2, 30–32), we can conjecture about causal mechanisms that could potentially lead to some of the observed differences.

In Europe, relative scarcity of research funds, collaboration-by-design in framework programs, and the European Commission’s funding schemes can in part account for the larger number of collaborations among Europeans and the formation of a close-knit network of prominent scientists (50, 51). This collaborative strategy has resulted in EU15 competing with the US as the world’s largest scientific producing block in the last decades (34, 35), although East Asia is catching up quickly. Paradoxically, even if collaborative productivity increases, this does not necessarily imply greater impact since largely closed networks of prominent scientists in Europe could result in less original and impactful research (23). Indeed, as illustrated above, the US has systematically been found to be more impactful across scientific fields (34, 52).

Nonetheless, the observation that for Europe-based scientists there is an advantage to collaborating with prominent Europe-based scientists suggests that there might be other mechanisms at play that go beyond funding

agency norms. Europeans for example have shorter average travel distances and live in similar time zones, and North Americans are commonly viewed as slightly more competitive and self-confident in their work (53, 54). Citations, famously referred to by Merton as “pellets of recognition,” contribute to appointment and promotions decisions (39, 55). A growing supply of scientists and a stagnant number of tenured positions in the last three decades has led to greater competition for good jobs among scientists in Europe, vis-à-vis North America (56).

In this context, for aspiring Europe-based scientists, co-authorship with prominent scientists might be a dominant and effective social mechanism of professional advancement to secure access to scarce tenured positions (57, 58). On the other hand, in North America the existence of individual soft money for career promotion coupled with less secure and influential tenured positions (59, 60) could lead to permeable networks which are more open to newcomers and with fewer incentives for social closure through collaborations with respect to Europe. In fact, in North America the competition for resources through soft-money positions, prestige of first and last-authorship (to which researchers often renounce in large collaborations), and individual rewards could be a deterrent for prominent scientists to engage in systematic collaborations with other prominent scientists (60). Social stratification is well known to play an important role in student acceptance and hiring inequalities (61–63) and could also play an important role in shaping the collaborations that prominent scientists establish. Further studies based on our findings could examine which forms of social stratification result in differential access to networks of prominent scientists in North America and Europe.

More generally, statistical analysis of network structures linked to impact of the scientific output can be limited in providing a precise causal mechanism given factors that are not easily measurable such as researchers’ personality traits (such as being more competitive and self-confident) and individual motivations towards collaboration (64), and the social norms that shape differences in scientific cultures across continents. Nonetheless, our finding that research that involves several prominent researchers has larger impact, which however wanes in repeated collaborations, holds

across regions and scientific fields can have important implications. On the one hand, the share of research funding allocated to teams (and to repeating teams) may need to be reassessed for existing funding schemes. On the other, early career researchers may need, given different hiring criteria across fields, to strike a balance between work they do by themselves and in collaboration.

## Chapter 4

# Prominence and early-career factors

During the previous chapter we have introduced the concept of prominent researcher, scientists who have achieved excellence in their field and that have been recognized by their peers in form of awards or appearances as keynote speakers in main conferences. But the truth is that the factors behind high-impact science and how this prominence is achieved are important questions that are still not fully understood (38, 65–72). Hiring committees, funding bodies and university departments make decisions by trying to assess which factors will shape the scientific trajectories of researchers often using limited information. In front of this situation one can ask: is it possible to identify a prominent career in its early stage? If so, which are the main factors that we can expect to observe in a future prominent researcher? Can we rely on just bibliometric data? The use of common bibliometric indicators, such as number of publications, journal impact factors and citations, as metrics for assessing research impact has been put into question by some researchers (?). Other metrics such as open access publications and altmetrics have been proposed as complements or alternatives for improving the way we assess research (?). Yet any measure of scientific impact and prominence faces constraints. A necessary step

in identifying ways to evaluate research more fairly is to apply predictive models that help identify inherent biases to science's current incentive and evaluation system. To this end, we comprehensively analyze the careers of prominent scientists to identify to what extent early-career factors predict the success of researchers later on in their career.

Most studies on the drivers of high-impact science focus on the role of an individual factor in isolation, such as the prestige and ranking of researchers' university (73–76), ranking of published papers in journals (42, 77, 78), and collaborations (2, 10, 26–28, 47, 79–82). Total citation counts and h-index of the world's prominent scientists capture only past accomplishments, but not what has driven those achievements. Rarely are there studies conducted to identify the factors driving the production of high-impact research over time, (70, 71, 80, 83, 84) combining the different key factors in a single study to understand the relative importance of each factor (42, 73–77) and studying fields across the natural, behavioural and social sciences simultaneously (38, 81, 82). Here, we do so by conducting a comparative analysis of these key established factors to shed light on how early-career choices and factors shape the path to later become prominent researchers. To this end, we examine four key early-career factors (researchers' university prestige, journal ranking of their top publication, collaboration with other prominent researchers, and overall impact of their early research) that capture the scientific achievements during the first 5 years of the career of the 100 most prominent researchers in eight different fields across science. These include genetics, development economics, cognitive psychology, network science, social inequalities in public health, network ecology, metabolomics, and philosophy of science. We assess how these key factors are related to their h-index later on in their career, while controlling for factors like their geographic location, (34, 35, 85) gender (8) and scientific field (8, 28) (Fig4.1).

The results of this chapter show how top researchers across fields have, in the first five years of their career, an advantage compared to the average researchers – the comparison group – that lasts throughout the rest of their career: they are more likely to research at one of the top 25 ranked universities worldwide, publish a paper in a top 5 ranked journal in their field,

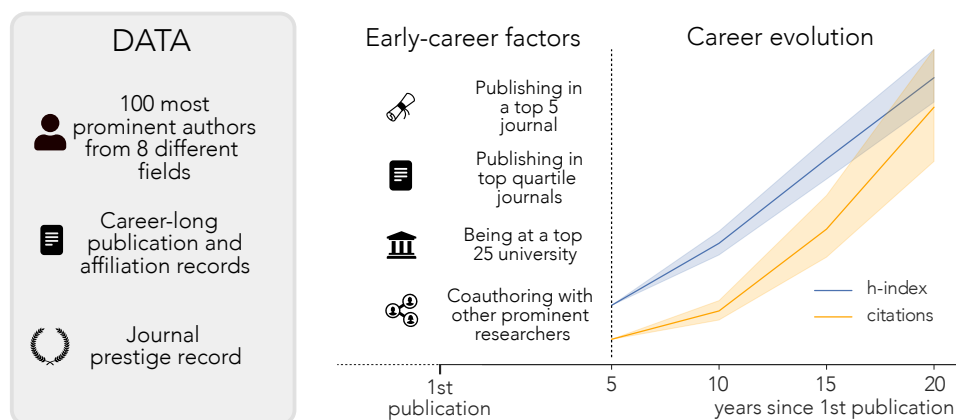


Figure 4.1: **Conceptual map of the study.** We compiled a list of the 800 most prominent scientists in 8 research fields across science. We obtained for each researcher a full publication list, history of citations of the publications as well as their affiliation records over time from Scopus. Using this information, we obtained data on early-career factors (within the first 5 years after their first publication): being at a top 25 university, publishing in a top 5 journal or most papers in Q1 journals within a specific area of knowledge (according to Journal Citation Reports), and coauthoring with other prominent researchers. We then look at the subsequent career of the researchers and measure the evolution of their number of citations and h-index over 5, 10, 15 and 20 years since their first publication.



publish most papers in top quartile journals, and collaborate with other prominent researchers. Indeed, this trend holds for prominent researchers across scientific fields: The prominent researchers at the top of their field early on in their career (compared to their peers) are consistently at the top as their career progresses. Our results highlight how the early steps in a scientific career have a very strong impact on how the researcher will perform in the future. The implications of our findings are vast and can provide young researchers with a means to evaluate their own expected career trajectories. Yet because these four attributes of ultra-successful scientists are highly predictable, the findings also suggest how closed the scientific system already is. The results also point to shortcomings in using the common and highly-influential indicators of success, namely citation and h-index metrics. This is because early career advantages – measured using these metrics – are so strong that it predefines ‘highly-successful scientists’ without further information about the content or social and policy impact of their research.

## 4.1 Four common factors in prominent researchers

Analysing the first 5 years of the academic career (starting at the first publication) of all 100 prominent researchers, across the aforementioned fields, we find that 47% were at a top 25 ranked university, 77% published a paper in a top 5 ranked journal in their field, 59% of their papers were published in top quartile (Q1) journals and 27% co-authored a paper with another prominent researcher in their field.

In order to compare this factors in prominent researchers with the ones of the average researcher globally, we proceed as follows. First, less than 1% of all researchers worldwide – an estimated 0.6% – are at one of the top 25 universities. This share is calculated using UNESCO data on the total number of researchers worldwide at 8,854,288 (86) divided by the total number of researchers (university staff) at the same top 25 universities (using QS World University Rankings) at 56,900. For comparison, the top 25 universities account for 1.8% of the total 1396 universities in the

Times World University Rankings (87). Second, an estimated 3 - 14% of all researchers worldwide have published a paper ranked in the top 5% in their field. This share is calculated by using data on the total number of all publications ranked top 5% in researchers' field at 267,966 publications indexed in Web of Science using the Leiden Ranking (88) divided by the total number of researchers worldwide at 8,854,288 (86) or by the total number of researchers (university staff) at 1,914,149 (87) that results in a 3% (lower bound) or 14% (upper bound) estimate, respectively. Thirdly, about one third of all articles worldwide (upper bound estimate) are published in top quartile journals indexed in Web of Science; (89, 90) and as many individual researchers publish multiple articles in quartile 1 journals it is likely that the share is significantly lower for the average researchers to publish at least half of their papers in quartile 1 journals. Fourthly, about 14% of junior researchers on average have co-authored a paper with a senior researcher between 1990 and 2012 in a global study covering about 1000 journals across the sciences (totalling about 6 million publications), with the shares varying across the fields of biology (15%), physics (13%), chemistry (13%), medicine (16%) and mathematics (6%), including the top three multidisciplinary journals (Nature, Science and PNAS) at about 19% for each journal (57). Fifth, the average h-index using university-level data is estimated at about 27 (median 25) as an upper bound estimate that includes only the top 500 universities (91). The average h-index using all journal-level data from the Scimago Institutions Ranking (92) via Scopus is estimated at about 32 (median 14). Note that both the mean university-level and journal-level h-indexes are upper bound estimates – i.e. higher than the mean researcher-level h-index given that researchers with lower h-indices are not represented in such estimates.

As we can observe, the shares we observed in the prominent researchers subset, are significantly higher than for the average researcher. Indeed, 92% of all prominent researchers had at least one or more of these four features, with the share increasing to at least 95% for those in genetics, development economics, cognitive psychology and metabolomics. Moreover, more than half of all prominent researchers placed a paper within a top 5 ranked journal in their field in the first 5 years, with the highest shares

at 93% for researchers in genetics, 86% in metabolomics and 82% in cognitive psychology (Fig. 4.2). This initial prominence is not only a matter of a 'one-hit wonder' but a consistent characteristic, with the majority of prominent researchers publishing more than a half of their papers in top quartile journals (except for philosophy of science).

A researcher's early institution is also strongly correlated with scientific prominence across a number of fields (73–76). Indeed, we find that over 50% of researchers in development economics, cognitive psychology, and genetics were at one of the top 25 ranked universities worldwide in the first 5 years of their career. However, this is not the case in younger scientific fields such as network science, network ecology or metabolomics, suggesting that the role of institutional prominence seems to be more important in well-established, more traditional fields. Being at a top university is the only feature, among the four early-career features, that illustrates a strong difference between newer and older fields. Another factor that highlights differences between fields, is the collaboration network that prominent researchers establish between them, with network science as the more collaborative field (42% of prominent researchers in this area have co-authored a publication with another prominent researcher) while philosophy of science stands out as the last collaborative with a 17% of their prominent researchers establishing collaboration with their equals.

In terms of geographic differences, we find that prominent European researchers are, in their early career, overall more likely to have top publications and to have been at a top 25 ranked university across all fields (Fig. 4.3), even though North America has a larger concentration of top universities whose graduates occupy the majority of faculty positions in US universities (93). Prominent European researchers are, however, less likely to have co-authored a paper with another of these top 100 researchers in their field, except in development economics and cognitive psychology (85).

In terms of gender differences, our results confirm that the gender gap is even more exacerbated in the scientific elite: females account for 15% of all prominent researchers across fields, ranging from 29% in social inequalities in public health to only 6% in genetics (8), which contrasts with their prominence in the early stage of their careers. During the first five years,

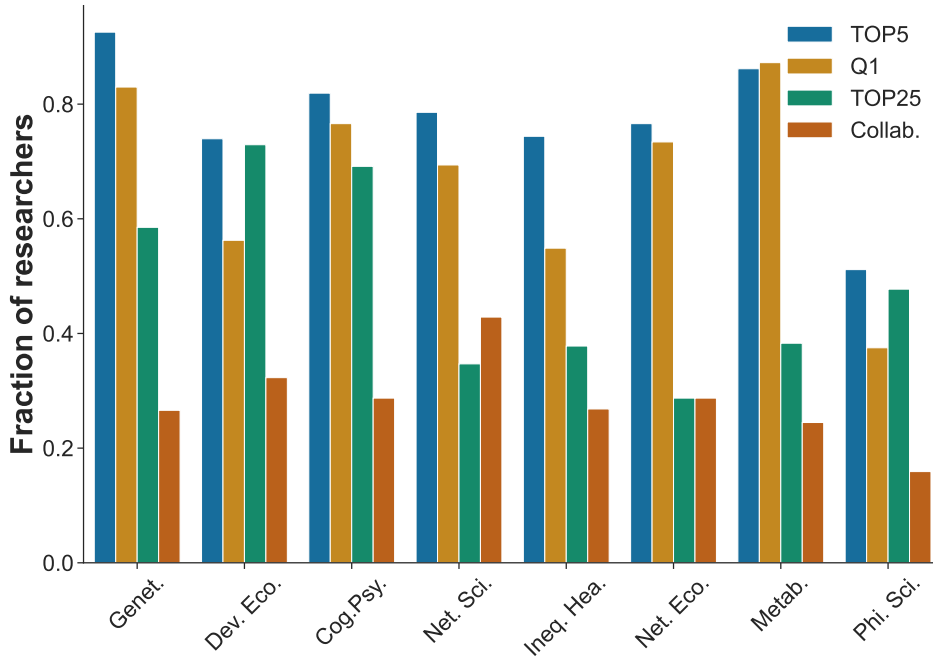


Figure 4.2: **Early-career factors of prominent researchers across fields** Fraction of researchers by field for the four key variables studied: TOP5 represents whether a researcher published in a top 5 ranked journal in their field. Q1 represents whether a researcher published most of their papers in a top quartile journal. TOP25 represents whether a researcher was affiliated to one of the top 25 universities worldwide. Collab represents whether a researcher co-authored a paper with another prominent researchers in their field. To compute all variables we consider publications and institutions for the first 5 years since the first publication.

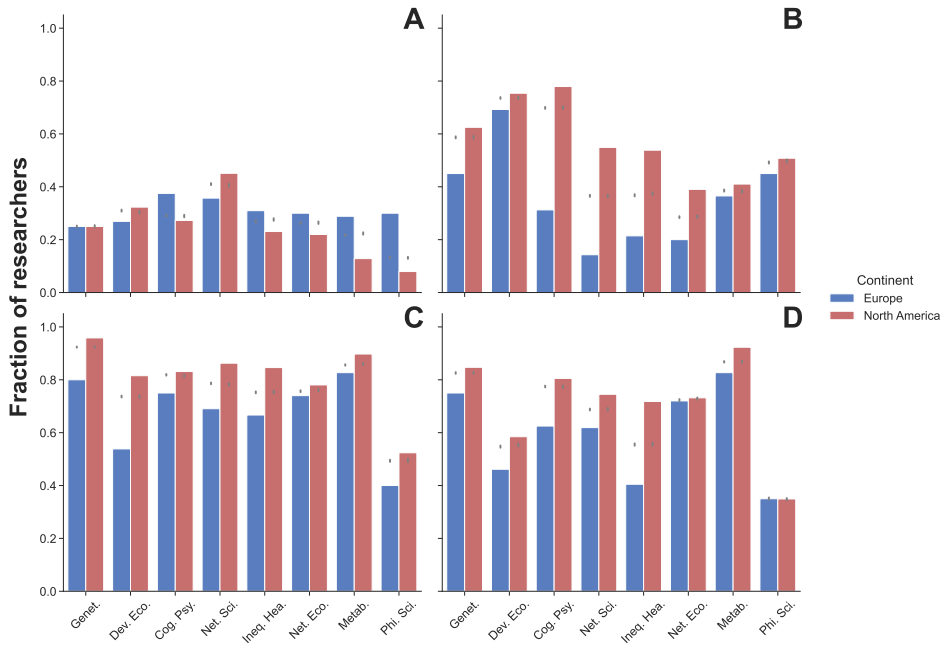


Figure 4.3: **Early-career factors of prominent researchers disaggregated by continental affiliation** Fraction of researchers by field and affiliation who have: (A) Publications with other prominent researchers in their field. (B) Affiliation in one of the top 25 universities. (C) A paper published in a top 5 ranked journal in their field. (D) Most of their papers published in a top quartile journal in their field. Grey points represent the values and the 95% confidence interval expected when randomizing the citation quartiles within each field.

prominent female researchers, have overall, a similar share of papers in top quartile journals as their male counterparts and also they are more likely to have conducted research in a top 25 university and to have collaborated with another prominent researcher (Fig. 4.4)

## 4.2 Early-career factors are connected to research impact

To understand the relationship of early-career factors with early performance, we disaggregate researchers in four quartiles of increasing number of received citations during the first five years (i.e. researchers in quartile 1 (QI) are those with the lowest 25% of citations, while researchers in quartile 4 (QIV) – the top cited quartile – are those with the highest 25% of citations). We find that there is a strong correlation between the four early-career drivers and the impact of research output early on in researchers' career. The fraction of prominent researchers in the top citation quartile in the first five years are, in general, more likely than expected by chance to have any of the four early-career features than other prominent researchers in lower citation quartiles (Fig. 4.5).

**The role of publishing with other prominent researchers.** Collaboration among scientists has been recognised as a source for innovation and creativity leading to increased research impact (10, 26). Our analysis is consistent with these findings: co-authorship is strongly correlated with higher citations across all fields, and the relationship is particularly strong in the natural sciences including genetics and network science (Fig. 4.5).

Remarkably, the effect of co-authoring with prominent researchers is even greater. We find that only 27% of prominent researchers co-authored at least one paper (and overall 11% of their papers) with another prominent researcher in the first 5 years of their career. The papers co-authored by at least two prominent researchers have much higher number of citations than other papers. The effect, intensity and size of collaborations, however, is not homogeneous across geographic locations nor across fields

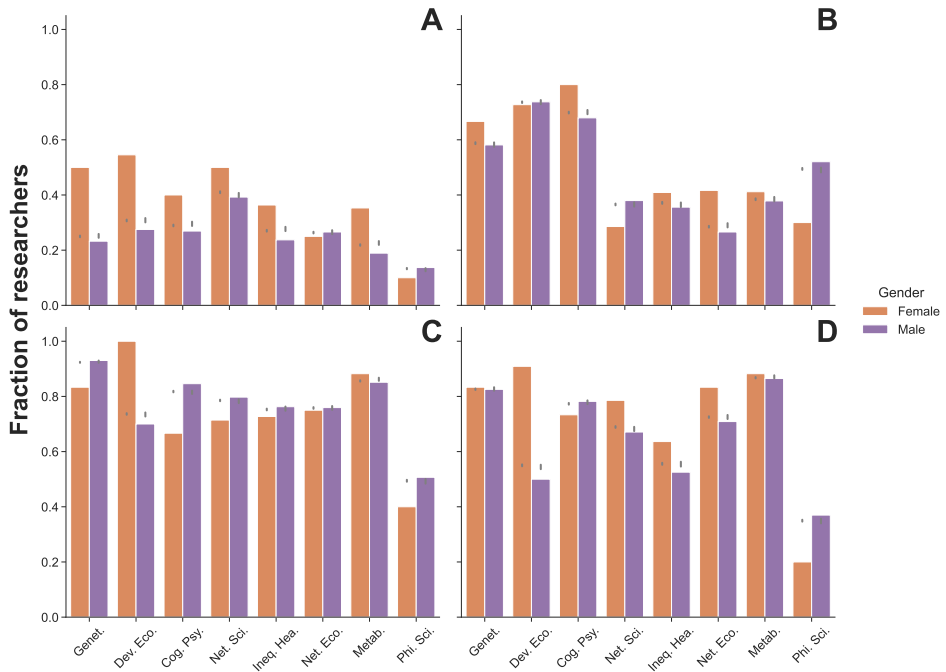


Figure 4.4: **Early-career factors of prominent researchers disaggregated by gender** Fraction of researchers by field and gender who have: **(A)** Publications with other prominent researchers in their field. **(B)** Affiliation in one of the top 25 universities. **(C)** A paper published in a top 5 ranked journal in their field. **(D)** Most of their papers published in a top quartile journal in their field. Grey points represent the values and the 95% confidence interval expected when randomizing the citation quartiles within each field.

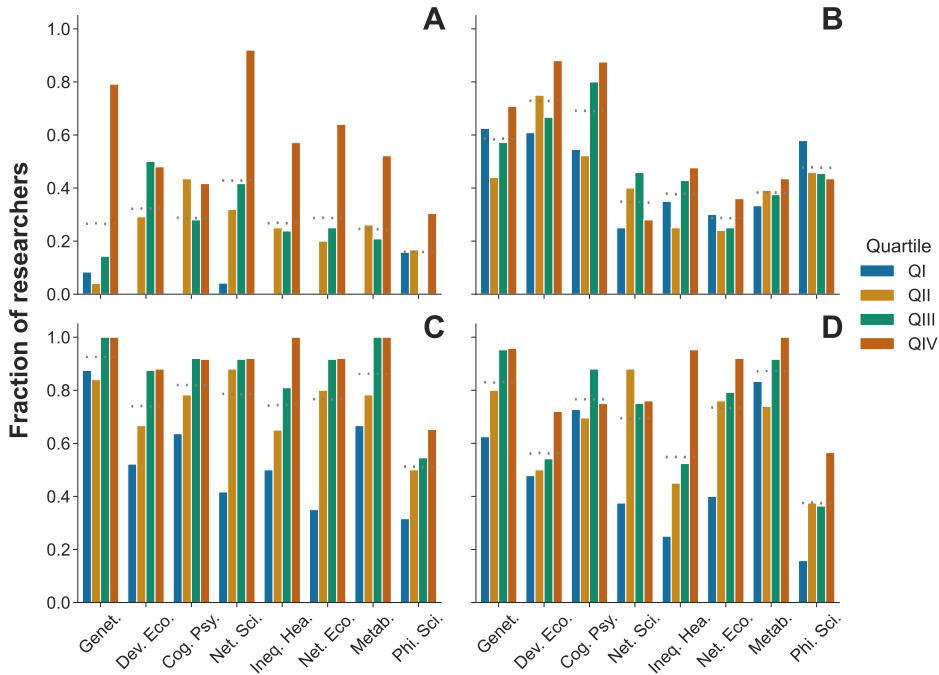
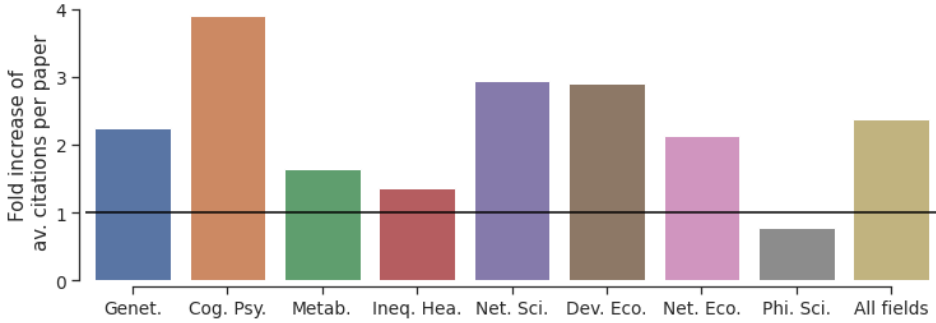


Figure 4.5: **Early-career factors of prominent researchers disaggregated by citation quartiles** Fraction of researchers by field and quartile who have: (A) Publications with other prominent researchers in their field. (B) Affiliation in one of the top 25 universities. (C) A paper published in a top 5 ranked journal in their field. (D) Most of their papers published in a top quartile journal in their field. Grey points represent the values and the 95% confidence interval expected when randomizing the citation quartiles within each field.





**Figure 4.6: Fold increase of average citations per paper in collaboration with other prominent scientists during the first 5 years of researchers' career** For each field, and for all fields combined, each bar shows the ratio between the number of citations per paper, in which a researcher collaborates with other prominent scientists in their field and the citations per paper for those papers not in collaboration with other prominent scientists. The black line indicates no fold increase. Bars above 1 show increased citations in papers with prominent collaboration, whereas bars below 1 indicate no increase. Note that in all fields except for philosophy of science, there is an increase in the number of average citations per paper. Overall, papers published with other prominent scientists during the first 5 years of their career receive over two times more citations than those papers not in collaboration over that period of time.

(Fig. 3.9). Furthermore, the disaggregated data by citation quartiles reveal that researchers in the lowest citation quartile have very low shares of co-authorship in their early career across fields with other prominent researchers in their field compared to an average of 56% for those in the top citation quartile (Fig. 4.5A). This finding suggests that co-authorship with other prominent researchers early on has a very large return across all fields. Indeed, already during the first five years of the career of scientists in our study, papers with other prominent scientists have overall received more than twice the number of citations than those not co-authored with other prominent scientists in their field (Fig. 4.6).

Our findings are thus in line with previous studies that analyzed the advantages of co-authoring with leading researchers in one's field. Working under leading researchers can boost career development through greater ci-

tations and mentorship (57), and provides visibility early on in a scientist's career (47). In fact, junior scientists at less recognised universities are most likely to benefit from co-authorship with leading researchers (47). Young scientists can also apply what they learn from high-impact, established researchers in their own career, (80, 81, 94) providing them with a competitive advantage relative to their peers (95).

**The role of prestige of researchers' institution.** Researchers at top universities have a qualitative advantage with respect to researchers in other institutions. They enjoy a high-quality research environment, generally with access to greater resources. Additionally, researchers at prestigious institutions are sought for collaboration as a way to boost their academic careers (27). Here, we assess the relationship between being at a top university and early-career impact. The share of researchers who have spent part of their early career in such institutions is not homogeneous across fields, with traditional disciplines having much larger shares, as outlined earlier. Not surprisingly, we find that only for these disciplines there is a relationship between early-research impact and being at a top university (Fig. 4.5B).

**The role of publishing in highly-ranked journals.** Publishing in higher impact journals early on can increase impact. Publishing in highly ranked journals, by increasing citations, benefits researchers' career opportunities, increases their prestige and recognition, and helps promotion (77). Getting one's best paper into a top journal increases early-career citations. Nearly all prominent researchers across fields placed their best paper in their early career within a highly ranked journal, which thus appears to be a necessary condition for becoming a prominent researcher. Researchers publishing in highly-ranked journals is strongly correlated with greater early-career impact, more so than just publishing within journals in Q1 (Fig. 4.5C, D).

#### 4.2.1 Early-career performance is a strong indicator of performance throughout later career stages

During the progress of a scientific career, the number of publications increase and so does the number of citations and thus, the h-index of a researcher. This increase is not homogeneous across fields, but rather reflects differences in the rate of publication, collaborations and size for each field (Fig. 4.7). We can observe that although when considering within-field evolution of the h-index over time is a good proxy for the success of the academic career of a given researcher, when comparing across fields, career-age and h-index should be controlled.

To assess whether early-career performance translates into a sustained advantage over time, we analyse the evolution of h-indices and citations over time (Fig. 4.8). To this end, we divide researchers into quartiles based on the normalized h-index and the normalized number of citations at 5, 10, 15 and 20 years since the first publication. We then look at the probability of transition over time between quartiles using the 5-year mark as the reference point. We observe that the initial advantage in the first 5 years is still present at 20 years of researchers' career. Figure 4.8C and F shows that 90% of researchers that started their career in the two top citation quartiles (QIII and QIV) have maintained this prominent position over time. Conversely, we observe the same situation for those scientists who were in the lower two quartiles (QI and QII). Both findings are consistent, whether we look at quartiles defined by h-index (Fig. 4.8 first row) or by citations (Fig. 4.8 second row) and across fields (Figs. 4.9-4.10). Although some fields display higher mobility between lower and upper quartiles, such as in network science and metabolomics, researchers are very unlikely to transition from the top-two to the bottom-two quartiles, suggesting that the fingerprints of that initial advantage are present along the researcher's career.

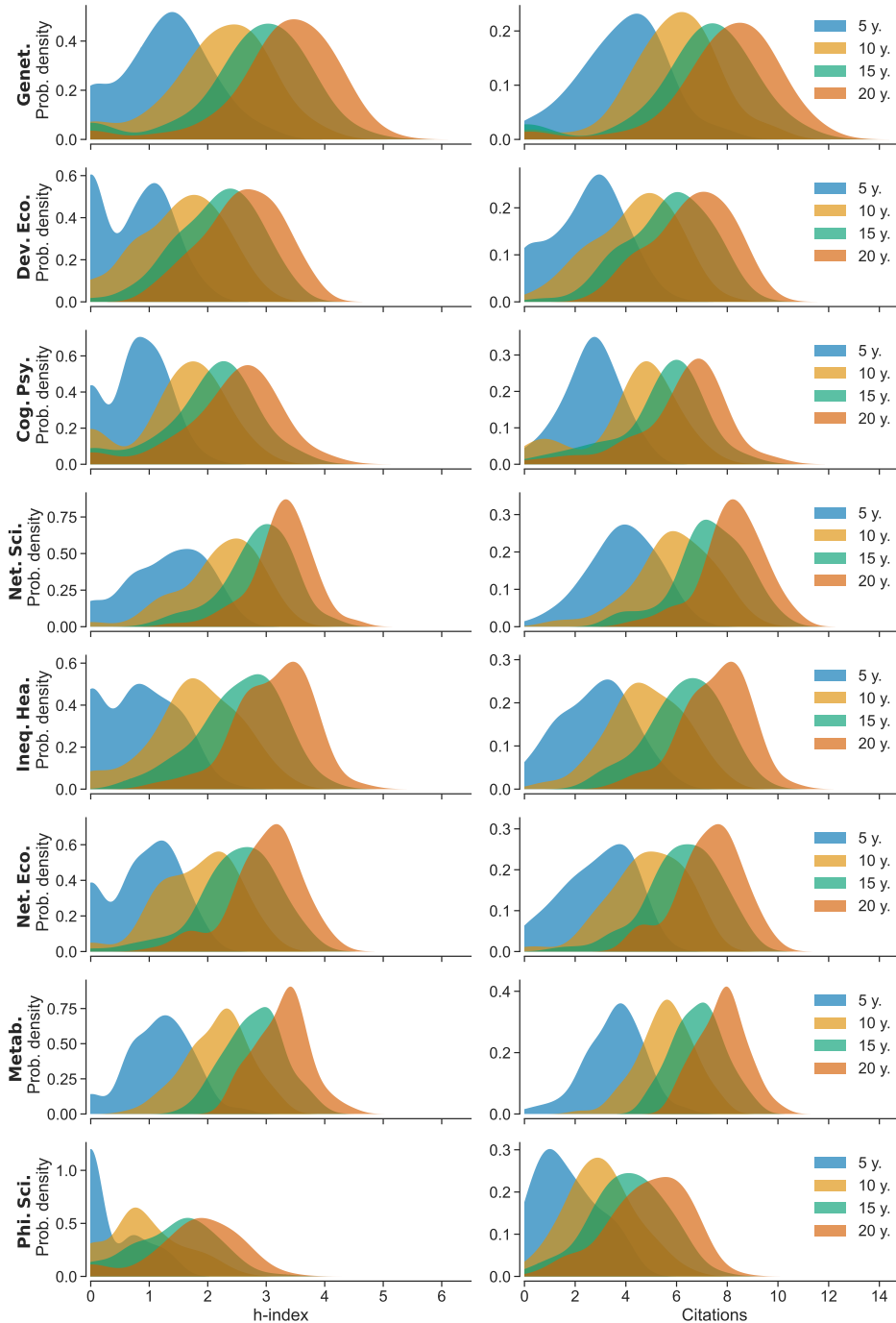


Figure 4.7: Evolution of the number of citations and h-index over time. Distributions of the h-index and the number of citations (logarithmic scale) at 5, 10, 15 and 20 years after the first publication by field.

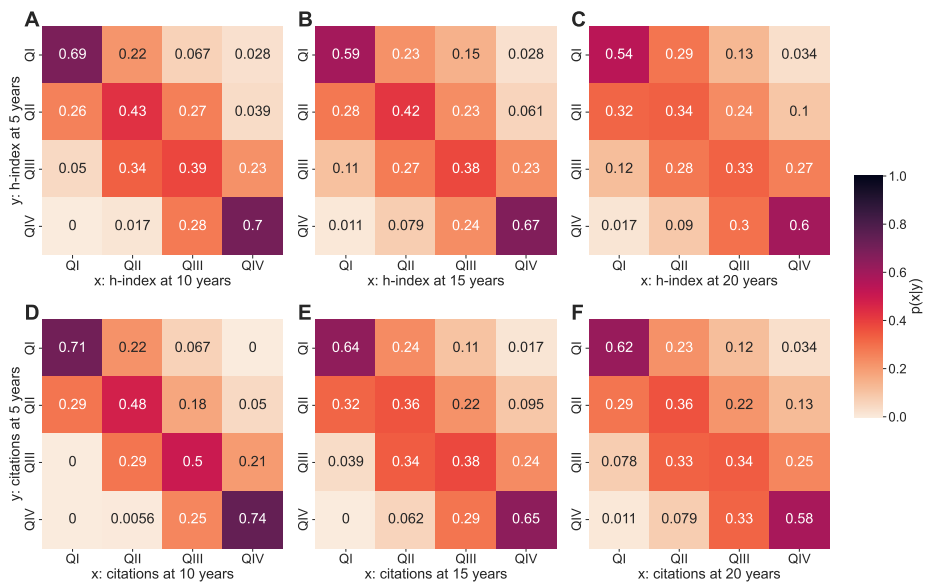
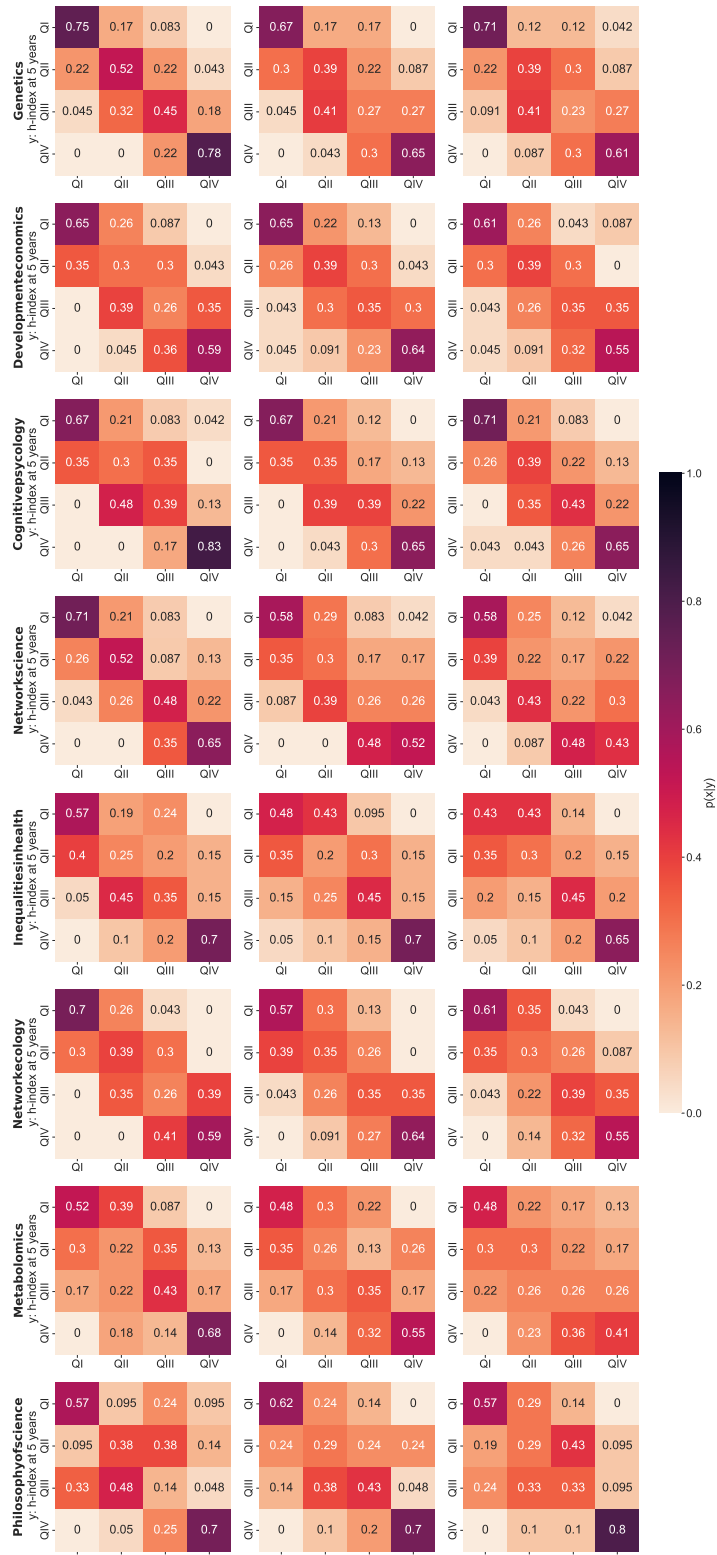


Figure 4.8: **Researchers mobility across quartiles** H-index quartile at five years compared to h-index quartile at 10, 15 and 20 years (first row, panels **A-C**), and citation quartile at five years compared to citation quartile at 10, 15 and 20 years (second row, panels **D-F**). The darker the region, the stronger the coincidence between the quartile at 10, 15 and 20 years relative to the quartile at the first 5 years. The results reflect the aggregated and normalized data for all fields.

Prominence and early-career factors



Prominence and early-career factors



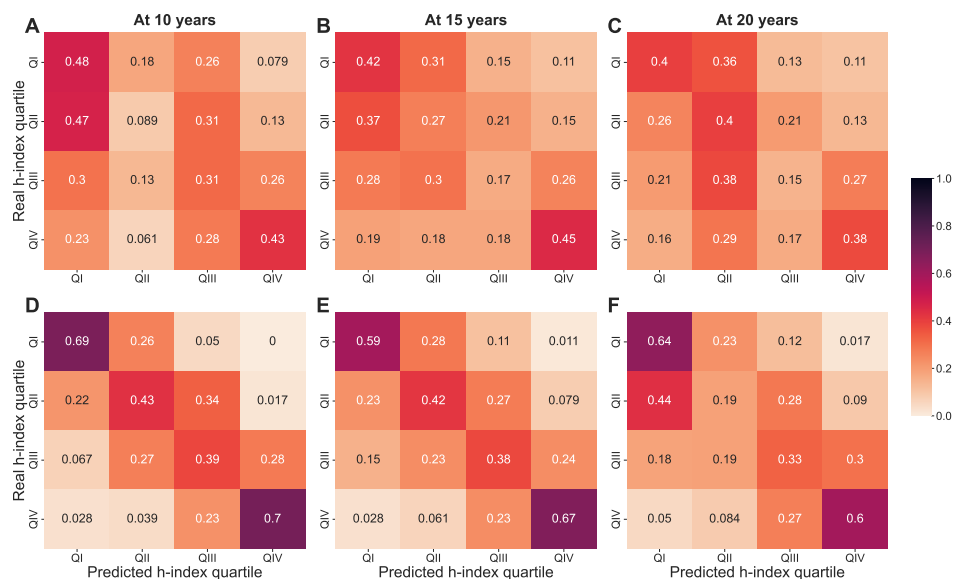
### 4.3 Factors driving citations and h-index in researchers' early career

So far, our results in Figs. 4.5, 4.8, 4.9 and 4.10 show that there is a clear relationship between the impact of research output and key early-career factors. Here, we want to assess the extent to which early-career factors can explain the evolution in citations and h-index over a scientific career. To this end, we train a random forest classifier over two different models: First, an econometric model which considers the same factors we have been assessing during this chapter (affiliation with a top 25 university, publishing at least one paper in a top 5 ranked journal, publishing more than half of the papers in Q1 and collaborating with other prominent researchers) and also controlling for two common background factors that include researchers' geographic location (whether they are based at a university in North America or not) and their gender (whether they are male or not), which are standard control variables applied in economics and the social sciences. In studying the role of researchers in science, these six factors have been assessed previously, yet this is the first study to examine these factors simultaneously by assessing how they shape the success of leading researchers in their early career. Second, we will compare it to a model that will take only the h-index quartile at 5 years.

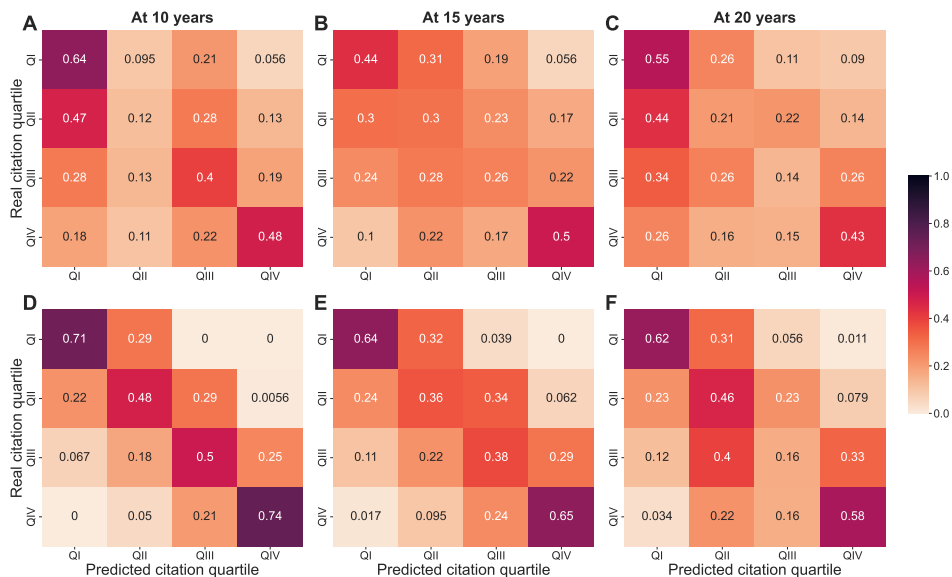
Once we have defined the models, we train the random forest classifier (RFC). The RFC analysis behaves similarly to that of a random forest regressor but produces a categorical output instead of a continuous one (labelling each observed sample into a category). In this sense, the classifier iteratively evaluates several decision trees over different parts of the data and averages the resulting outputs. We evaluated the fitness of the classifier with a 10-fold cross validation: in this procedure the dataset is divided in 10 folds from which one is selected as the test and the others as the training folds iterated several times until each fold has been used as a test, preventing in this way common problems such as over and under fitting.

In Fig. 4.11, we show the prediction results of the two different models, the econometric one (A-C) and the one including the h-index quartile after





**Figure 4.11: Prediction of h-index quartile based on early-career factors**  
 Predicted h-index quartile at five years compared to observed h-index quartile at 10, 15 and 20 years (first, second and third columns). **A-C** illustrate the prediction results with model 2 (which takes into account the four early-career factors as well as the geographic location and gender of researchers; see Methods). **D-F** illustrate the prediction results with model Q5 (which only takes into account the quartile of the first 5 years). The darker the region, the higher the number of researchers that are correctly classified by the algorithm. The results reflect the aggregated and normalized data for all fields.



**Figure 4.12: Prediction of citation quartile based on early-career factors**  
 Predicted h-index quartile at five years compared to observed citation quartile at 10, 15 and 20 years (first, second and third columns). **A-C** illustrate the prediction results with model 2 (which takes into account the four early-career factors as well as the geographic location and gender of researchers; see Methods). **D-F** illustrate the prediction results with model Q5 (which only takes into account the quartile of the first 5 years). The darker the region, the higher the number of researchers that are correctly classified by the algorithm. The results reflect the aggregated and normalized data for all fields.

5 years of the first publication (D-F). Our classification analysis reveals that assessing the h-index quartile at 5 years (Q5), the classifier is more accurate than if we only include the early-career factors. Nonetheless, if we do not include h5, the classifier is still able to correctly predict overall 40% of the researchers that fall into the lowest quartile (QI) and 38% who fall into the top quartile (QIV) at 20 years from the start of their career – significantly higher than the expected 25% for random quartile assignment. These results are consistent also when examining citation quartiles (Fig.4.12).

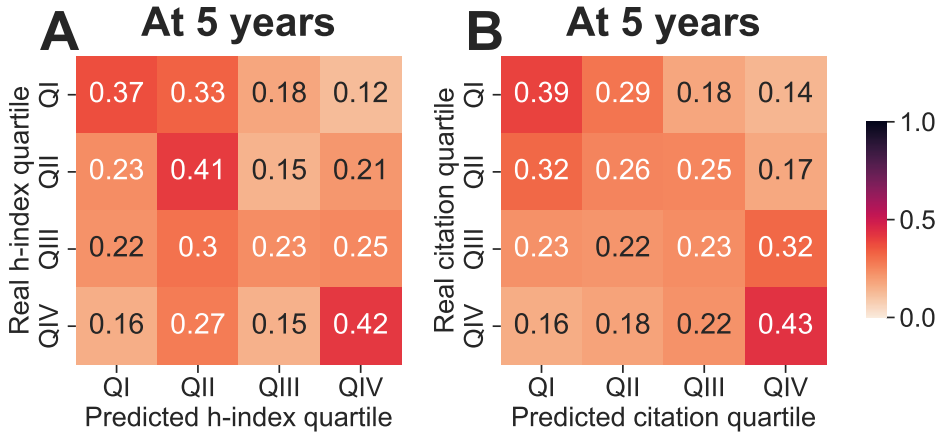


Figure 4.13: **Confusion matrices for predicting h-index and citation quartile with the econometrics model (A)**H-index quartile at 5 years, and **(B)** citation quartile at 5 years. These matrices show, for each row  $Q_i$ , the fraction of researchers in  $Q_i$  at 5 years who are classified as  $Q_I$ ,  $Q_{II}$ ,  $Q_{III}$  and  $Q_{IV}$ .

Our results also show that the early-career factors we study can explain trends in the share of researchers who remain in the same h-index/citation quartile over their career (Figs. 4.11 4.12). We also observe that for both classifiers, missclassification tends to happen between neighboring quartiles, so that the fraction of lower quartile researchers are seldom classified as  $Q_{IV}$  researchers and vice versa. This indicates that early-career features capture a substantial part (but not all) of the information captured by the h-index. Nonetheless, our results show that early-career researchers who are already prominent among their peers are very likely to sustain their advantage 15-20 years later.

In addition, these early career factors are able to capture the initial advantage even during the first 5 years (Fig. 4.13) correctly assigning around 40% of researchers in the upper and bottom quartiles for both, h-index and citation number.

As a final step, we compare the relative importance of the four early-career factors simultaneously by examining the *error permutation increase*,

where we run different instances while permuting the variables and observe how the error of the classification increases. As all features are binary (0 or 1), this facilitates comparing the relative importance of each factor (Fig. 4.14). Collaborating with other prominent researchers stands out as the most important factor across career age and for h-index and citation quartiles, followed by publishing a paper in a top 5 journal. Working at a top 25 university and publishing more than half of one's papers in Q1 journals have less explanatory power; and gender appears to be the less predictive variable. The results show how collaborating with established researchers in the field, stands as the best strategy to secure a position in the scientific elite. These results are consistent with results from the analysis of citations (Fig. 4.13) and also when predicting those quartiles during the first 5 years (Fig.4.15).

## 4.4 Discussion

Our analysis shows that the future success of a researcher is often determined early on in their career. Indeed, we show that as early as 5 years after the first publication, we can already make accurate predictions of whether a researcher is going to be within the top quartile of leading researchers later on.

We find four early career factors that are central drivers for later success across science: working at a highly ranked university, publishing a top 5 journal paper, publishing most papers in top quartile journals and co-authoring with prominent researchers at the early stage of researchers' career. Most importantly, we find a strong positive correlation between citations during the first five years of their career and the probability to have had any of these central early-career features we identify: researchers in the top quartile of citations are more likely than expected to have any of the four key features, whereas researchers in the lowest citation quartile are less likely than expected to have any of these features (but still more likely than the average non-prominent researchers). This finding is very insightful, especially because classification models are able to accurately predict

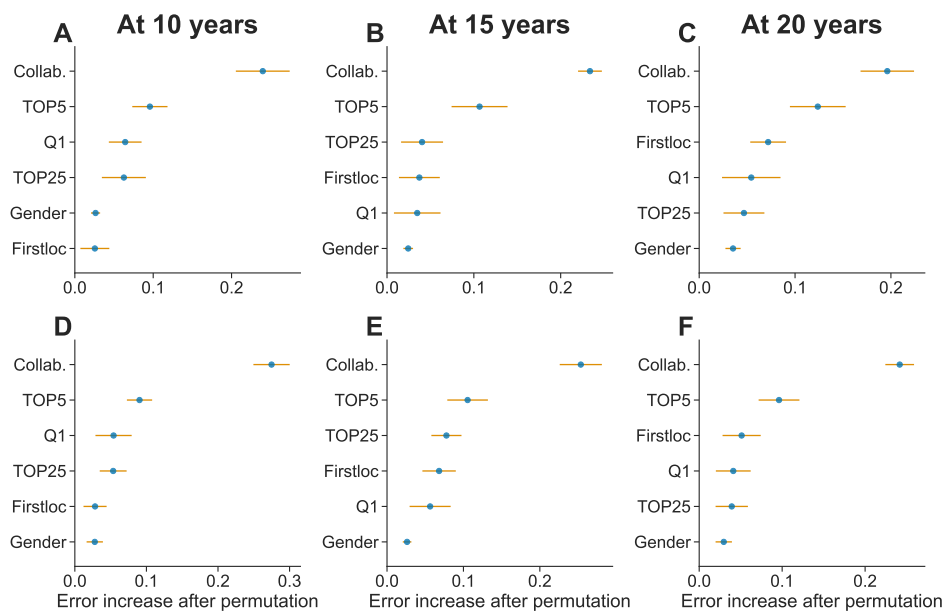


Figure 4.14: **Parameter importance for predicting citation quartile at 10, 15 and 20 years with the econometrics model** The first row (A-C) corresponds to parameter importance for h-index quartile prediction and the second row (D-F) for predicting citation quartile.

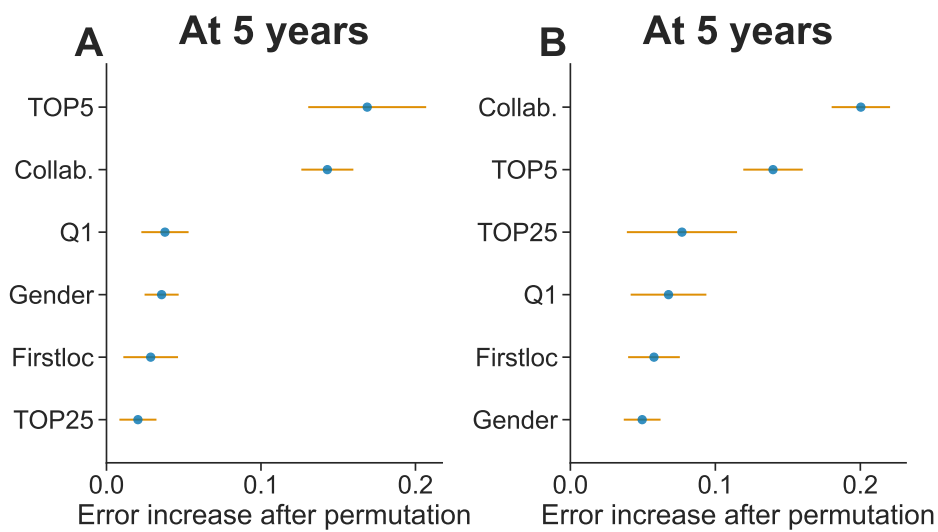


Figure 4.15: **Parameter importance for predicting quartiles at 5 years** Importance for h-index (A) and citation (B) quartile prediction. Parameter importance is an estimate of the increase in prediction error when we remove that feature from the classifier. The greater the error increase, the greater the importance of the parameter.

the citation and h-index quartiles after 10, 15 and 20 years for researchers falling into the top and lowest quartiles: what scientists do early on largely determines their impact later on in their careers.

We also find that in traditional areas of science, being at a top-ranked institution can be an important driver, but in younger disciplines this is less important. This finding is especially interesting in light of recent findings about graduates from top-ranked US universities occupying the majority of faculty positions in the US university ecosystem (93), and raises the question of whether hierarchies in the hiring system pose a threat to innovation and the emergence of new fields of science. Indeed, we also find that in disciplines in which university affiliation is not such an important driver, publishing with other prominent scientists becomes especially important (57).

Our analysis shows that these four key factors are important as a general strategy for young researchers across science and that an early-career jump start gives scientists an advantage that is sustained throughout their career. At the same time, our results suggest that there are also other factors influencing the h-index at 5 years such as individual, more qualitative or psychological traits of researchers (78) or, in relevant cases, the traits of a PhD advisor (94) that have not been considered here. While it can be a limitation, our results also explain that the success of individual researchers cannot be attributed to a single factor but involve a combined set of early-career factors.

Given that these four attributes of ultra-successful scientists are highly predictable, the findings also suggest the scientific system is presently relatively closed. The results also illustrate limitations in using highly-influential metrics of success, such as citations, h-index and JIF. This is because early career advantages on these metrics are so strong that it predefines 'highly-prominent scientists', independent of the content of their research. More generally, the findings point to a needed reform among the scientific community: As some scientists produce good science but are not successful in the 'metrics game', decision makers evaluating the work of researchers should also use additional metrics such as policy and social impact of research, developing new research tools, and the like. Decision makers should

thus not take this as an opportunity to just use citation and h-index metrics to evaluate scientific prominence.





## Chapter 5

# Departamental influence, gender and the selection of research agenda

The topics of scientific research are constantly evolving (*96*): New topics gain attention while others languish in response to external pressures including funding frameworks and societal needs like the recent COVID-19 pandemic (*32, 97*). For individual researchers, the selection of a portfolio of research topics is a critical decision that has direct impact in the evolution of their scientific careers. This is specially true for early career-scientists and young faculty, whose future professional stability hinges upon their early choices (*98*). Despite the importance of this decision we still know very little about which are the factors that affect the selection of research topics by individuals beyond global trends.

Changing and expanding the research portfolio is a common trait of scientific careers of researchers in STEM fields (*99*). However, when selecting possible research topics, individuals need to consider the trade-off between innovation and exploitation of exiting topics: While the former potentially has high rewards but implies high risk, the latter comes with a milder recognition but also implies lower risk (*100*). Another factor that plays a role

in shaping scientific careers is establishing collaborations: Collaboration with top-tier institutions and with prominent researchers increases impact and can help in career promotion (27, 57). However, because collaborative team sizes are largely driven by the amount of resources available, putting an emphasis on collaboration can lead to gender segregation in different fields of study within the same area (95, 101). Indeed, gender is another factor that has a strong impact in the scientific career of individuals: Leaving leaky pipelines apart (102), female faculty publish less and receive less grant money (95, 103, 104), are promoted at a later stage than their males colleagues (104), are more likely to experience issues when discussing authorship (105), and are given less credit for their contributions (106).

Unfortunately, studies looking at the evolution of scientific careers rarely consider the host institution or department as one of the factors playing an important role in the development of early-career faculty. Although formal collaborations, like coauthoring a publication, are the most recognizable way through which researchers share their ideas with each other, departments are collections of people who are exposed to similar scientific influences by attending the same seminars and informal meetings. Indeed, researchers within the same institution are more likely to collaborate (27). These collaborations are typically face-to-face, resulting in lower communication costs (107) and having a larger chance to spark creativity (108). Surprisingly, the effect that research environments have in shaping research careers and whether this effect has gender disparities has not been assessed.

Here, we aim precisely to cover this gap (Fig. 5.1). Our assumption is that departments and research institutions expose researchers to certain research questions and approaches, and become incubators for novel ideas through collaboration among faculty members. To that end we consider two cohorts of early-career researchers in departments of chemical and biological engineering, the engineering discipline with the lowest gender gap (109) and one that covers a broad scope of topics that range from molecular to planetary scales. The first cohort comprises young researchers who were offered an assistant professorship; some of them declined and some accepted. The second cohort, comprises early-career faculty in 34 top chemical engineering departments in Europe and the United States.

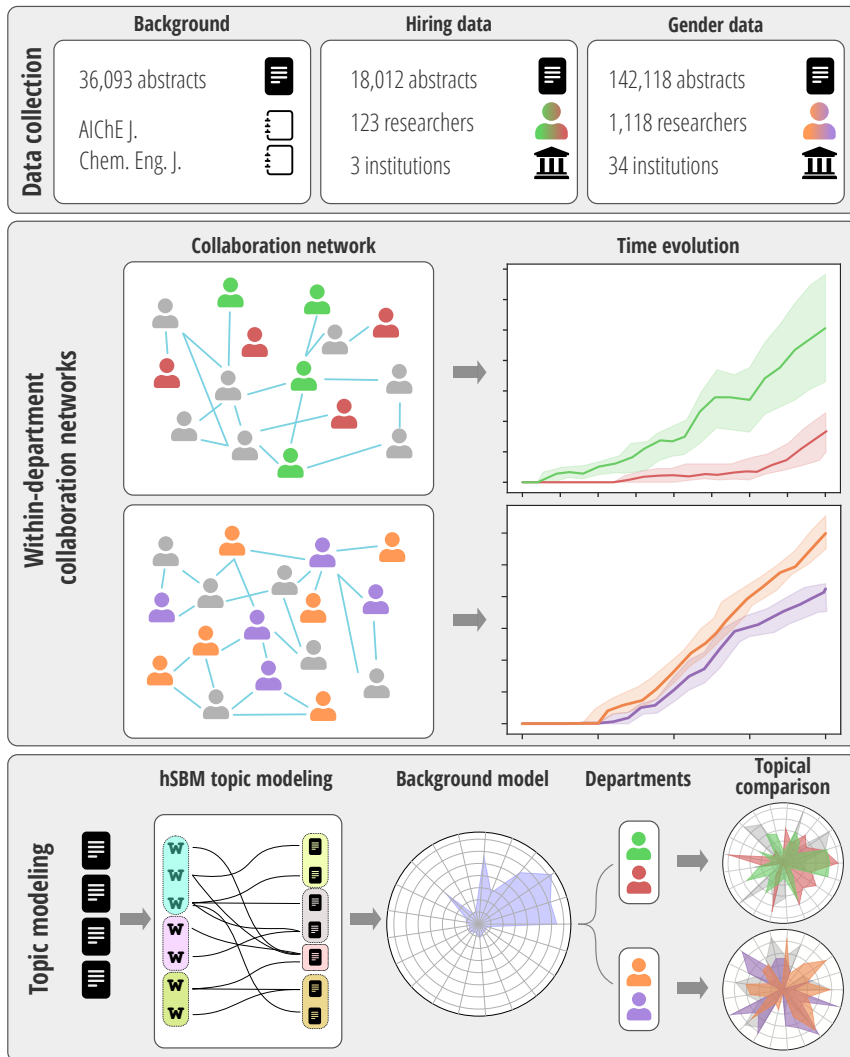


Figure 5.1: Caption

## 5.1 Creating a background of a scientific field

To evaluate the evolution of scientific careers we first need to classify research articles into scientific topics, which implies the construction of such topics from a field-specific background. To that aim, we collected all the publications that appeared in the *American Institute of Chemical Engineers (AIChE)* and in the *Chemical Engineering Journal (CEJ)* since their inception until 2021 comprising a total cohort of 36,093 articles and XXXX words. The main reason of this choice is that in order to classify documents with a broad topical distribution (chemical engineering studies from chemical processing plants to food composition) and thus we need broad-scope journals to build our models from, and also, we need long-record journals in order to capture chemical engineering as a whole and not just recent topics.

### 5.1.1 Topic models with SBM

Once we have collected the publication history of the journals used as a background, we can proceed with the topic model of the field. In particular, as stated in the introduction, we will use a network approach to topic modeling (110) which offers some advantages relevant to our study in front of other techniques. In first place those approaches more common in computer science based in pre-trained models such as BERT, cannot be used here since we are studying a very specific language framework like the scientific one, and also we avoid possible biases on a training set which is unknown to us. On the other hand, widely used inferential approaches such as LDA make unrealistic assumptions about the properties of the text that are mathematically convenient but do not reproduce its real features. By contrast, assuming an agnostic perspective about the number of topics in our corpus or the distribution that words are drawn from, results in simpler and more accurate models even in the cases where synthetic data favors other approaches (110).

With all of this in mind, we start by constructing a bipartite network that represents our coropora. Biparite networks are those networks com-

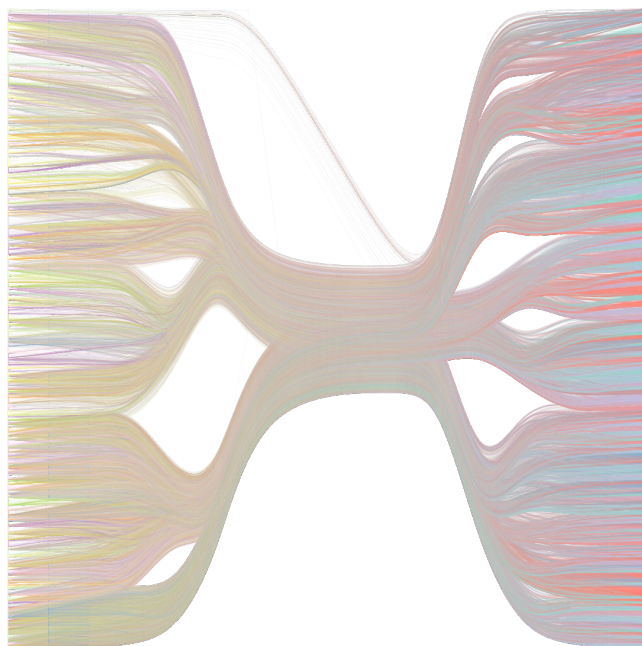


Figure 5.2: **dc-hSBM word-document network fit** Fit for the best model of the word-document network. Left side corresponds to document-nodes and right side to word-nodes. Each different color corresponds to a different group a node is associated with.

posed by two types of nodes that in our case will be word-nodes and document-nodes. In this case the connections are between nodes of different types (nodes from one type connected to nodes of the other type but not between them) and a word-document will be connected if the word appears in the document. Again, as proceeded in Chapters ?? and 3, we will fit the network with a hierarchical degree-corrected stochastic block model (dc-hSBM) since it is the most plausible one (maximizes the Bayesian posterior over partitions) and therefore minimizes the description length and better compresses our data (19, 46).

When analyzing the groups that emerge from the fit of the network, one can observe some groups large in size formed by words that add little value when identifying the topic, like the words *was, I, were...*, which inevitably increase the amount of data to process and also mask the real topics we are aiming for.

### 5.1.2 Removing the stopwords from the corpus

One of the problems that arise when dealing with the processing of natural language is the one of the stopwords. Stopwords are those words that carry no informative power about the text one is dealing with. To illustrate this let's consider the sentence "stellar evolution is the process by which a star changes over the course of time". The words that mostly explain what the sentence is about are: Stellar, evolution, star, time. Then, we have other words like *is, the, by, which, a, of* that can be removed without affecting the content of the sentence, since we can understand that a new phrase like *Stellar evolution star time* is somewhat related to the evolution of stars.

A general approach to identify these words is the so-called "bags of words" approach, in which stopwords are sets of pre-identified words in a given language that are known to be general and not topic-specific. However such an approach comes with some hindrances: Namely bags of words are specific to a given language, there are bags of words for English, Spanish or Chinese for instance, so it's not a flexible approach to the problem and also if we are processing minoritarian languages or dialects it could be difficult to obtain a standardized set of stopwords. In addition, this collection of

words are extracted from a day to day language, so they are not useful when dealing with specific texts.

If one thinks of *Science* as a minority language or a very specific dialect the aforementioned problems seem relevant if we want to extract information from research publications. To overcome this issue, we adopt the information theoretical approach developed in (111). The idea is to calculate the entropy of words according to the expression Eq. 2.3 that we have seen in Chapter 2 using the conditional probability of a document  $d$  to contain the word  $w$  leading to a conditional entropy of the form:

$$H(w|C) = - \sum_d p(d|w) \log p(d|w) \quad (5.1)$$

Which gives us an idea of how a word is distributed. To assess whether a word is informative, we compare the conditional entropy to  $\langle \tilde{H}(w|C) \rangle$ , the expected entropy of a word if we distribute it at random among documents, defining the information content of words as:

$$I(w|C) \equiv \langle \tilde{H}(w|C) \rangle - H(w|C) \quad (5.2)$$

where the first term represents the average random conditional entropy and the second term to the observed entropy. If the information content is zero, means that a word is as informative as it would be at random and that removing them would not affect to the meaning of the sentence.

This approach has immediate advantages relevant for our study:

- It is corpus dependent. It allows us to detect stopwords specific to chemical engineering that are not considered stopwords in a day to day language.
- Reduces the amount of data we have to process: We are not immune to computational resources demands, and reducing the amount of data to process without affecting the information content in it allows to obtain results with less resources and in less time.
- Improves the quality of the models.



### 5.1.3 Measuring topical distances between researchers: Jensen-Shannon distance

Now that stopwords do not interfere in the identification of topical groups of words, we can fit again our word-document network with the dc-hSBM obtaining the most plausible classification of words into topics (Table 5.1). This allows us to define a distribution over topics in each document, the average distribution of topics over all the publication record of a researcher. Since we want to measure if joining a department exerts an attractive force over an incumbent researcher, we need to define a metric of the topical distances between different researchers and the department to compare how far is the researcher before and after joining (or not) a department.

Five most common words					
<b>Topic 0</b>	flow	gas	reactor	transfer	liquid
<b>Topic 1</b>	process	rate	temperature	water	concentration
<b>Topic 2</b>	model	data	membrane	co2	diffusion
<b>Topic 3</b>	adsorption	removal	ph	degradation	capacity
<b>Topic 4</b>	bed	aqueous	selectivity	batch	yield
<b>Topic 5</b>	reaction	surface	catalyst	carbon	activity
<b>Topic 6</b>	filtration	shear	fluids	droplet	surfactant
<b>Topic 7</b>	extraction	fluoride	biosorption	sand	uranium
<b>Topic 8</b>	materials	nanoparticles	composite	metal	photocatalytic
<b>Topic 9</b>	g-1	current	enzyme	anode	bacteria
<b>Topic 10</b>	sites	ion	iron	recovery	biomass
<b>Topic 11</b>	drag	filtered	deformed	tadf	triplet
<b>Topic 12</b>	zeolite	inlet	reactant	outlet	pellets
<b>Topic 13</b>	chitosan	emulsion	imprinted	hgii	electromagnetic
<b>Topic 14</b>	corrosion	inhibitor	benzotriazole	nanocontainers	alkyd
<b>Topic 15</b>	condensation	breakup	cascade	journal	vapour
<b>Topic 16</b>	mesoporus	nitrate	soil	ps	cuii
<b>Topic 17</b>	aggregates	fractal	cbz	nb	quinoline
<b>Topic 18</b>	cofs	strach-based	polymer-grafted	graphene-containing	microdomains
<b>Topic 19</b>	monoliths	n-heptane	adherence	washcoating	nano-structured
<b>Topic 20</b>	mileage	2-ipnhp	sulfur-tolerant	cuo+naohaq	pd/cz

Table 5.1: **Background topics** Five most common words from each topic detected by the dc-hSBM.

To that aim, we compute the Jensen-Shannon distance between two topic distributions  $p$  and  $q$  which is expressed as follows:

$$\text{JSD} = \sqrt{\frac{\text{KL}(p||m) + \text{KL}(q||m)}{2}}, \quad (5.3)$$

where  $m$  is the pointwise mean of  $p$  and  $q$  and KL is the Kullback-Leibler divergence between  $p$  (or  $q$ ) and  $m$  defined in Chapter 2. Using Eq. 5.3 over the KL diverge (which is common in computer science), we avoid divergences when one of the distributions is 0 at the same time that we obtain a quantity that is symmetric as opposed to the KL.

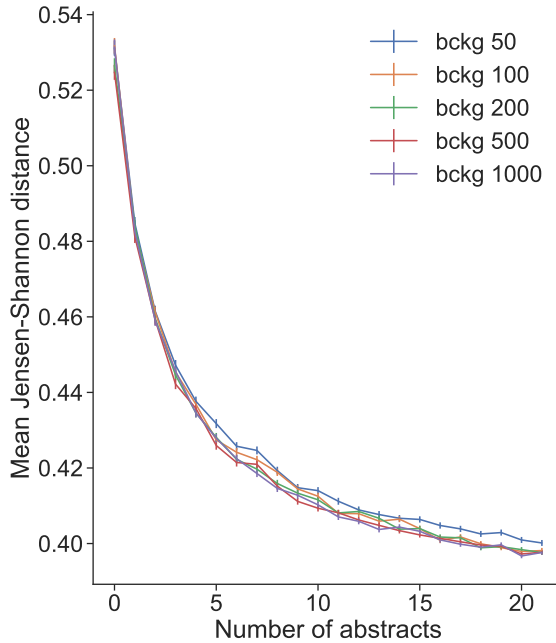


Figure 5.3: **Jensen-Shannon size dependence** Mean JSD between individual researchers and the background varying the number of abstracts and the size of the background.

Nevertheless we need to be cautious with using this as a reference, since it is sensitive to the size of the corpora we are using. To better comprehend

this dependence we examined the variation of the mean JSD between the average distribution of a subset of abstracts and the background. Fig. 5.3 shows how with a fixed background size, the average distance is highly dependent on the number of abstracts that we use to compare. Also that, when big enough, the size of the background does not reflect significant differences at different size. To prevent this and to be able to compare the distances of different researchers no matter their background, in the following sections we will compare random samples of 10 papers from each researcher to the background of their department.

## 5.2 The effect of a new research environment

We start by quantifying the effect of joining a department on the collaborations and research topics of early-career faculty. For this, we analyze the hiring history of three different departments in the United States between 2007 and 2017. Each of the three datasets contains a list of young researchers that were offered an assistant professor position in one of the departments and their response (accepted or declined), as well as the publications of those researchers and of all other faculty in the department, before and after the hiring offer. The candidates who accepted the offer give us the opportunity to analyze the effect of joining the department, in terms of both their number of within-department collaborators and their research topics. The researchers that declined allow us to control for potential confounding effects, such as field-wise shifts in research topics.

As expected, the decision to join a department affects the number of collaborations of a researcher with members of that department. Figure 5.5A shows that, for the three departments considered here, the incorporation to the department increases the number of new collaborators within the department to an average of 2.5 collaborators after 10 years. By contrast, researchers who declined the offers barely collaborate with members of the department after declining.

We surmise that these new collaborations, as well as other forms of scientific socialization within the department, bring the early-career faculty in

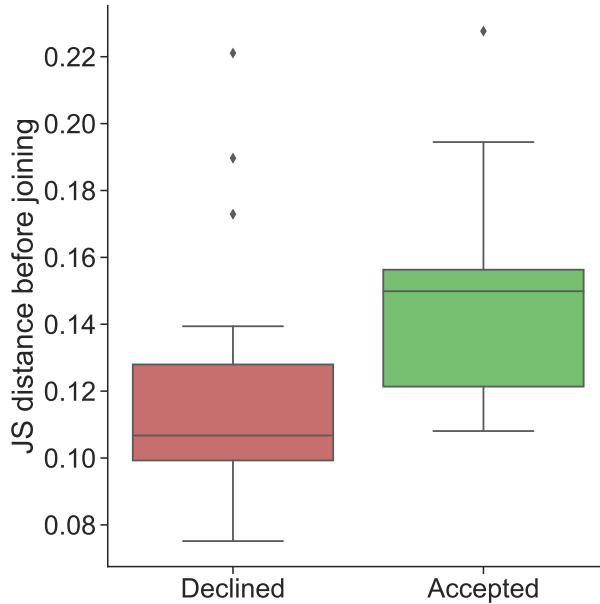
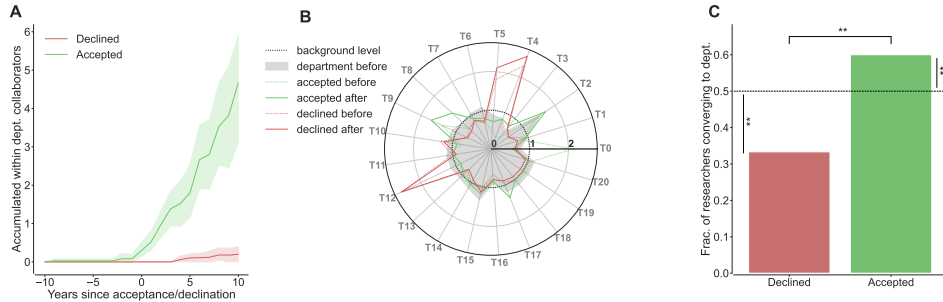


Figure 5.4: **Similarity of researchers to the department before the offer was made.** Jensen-Shannon distance of the corpus of researchers (those who accepted and those who declined) respect to the background of the department before the offer was made.

contact with new methodologies or research questions and ideas. To analyze the influence of the department on the research interests of the newcomers, we analyze the change in research topics by means of topic modeling of article abstracts, before and after the offer to join the department. In particular, we compare the distribution over topics of each researcher with that of the department (Fig.5.5B). We observe that, before joining a department, the early-career faculty who accepted and those who declined were similarly close to the topic distribution of the department (Fig. 5.4). However, after joining a department, early-career faculty tend to shift away from topics that are not popular in the department and towards others that are more prominent, even when we exclude direct collaborations (that is,

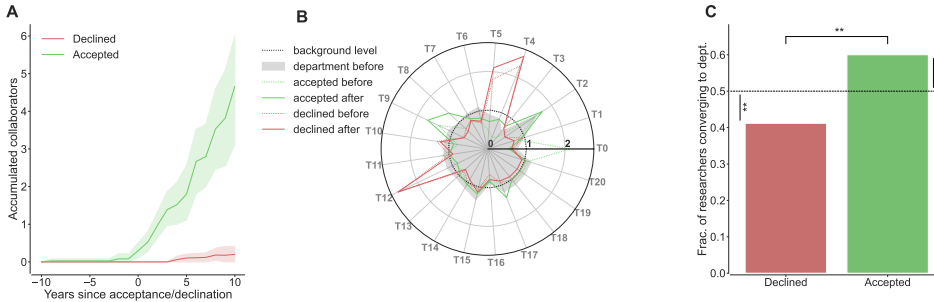


**Figure 5.5: Changes in departmental collaborations and research topics after joining a department.** (A) Accumulated within-department distinct collaborators over time with respect to the year the offer to join the department was made for researchers who accepted (green) and who declined (red). The shadowed area corresponds to the 95% confidence interval. (B) Comparison of topic distributions before (dotted lines) and after (solid lines) the year that the offer to join the department was made for a researcher who accepted and for one who declined. The gray solid area corresponds to the background of the department at the time of the offer. All distributions are normalized with respect to the global distribution of the background dataset. The black dashed line represents the same topic proportion as that of the background. (C) Fraction of researchers whose topic distribution after the offer was made converges to that of the department – researchers who declined the offer, red; researchers who accepted, green. Stars indicate statistical significance obtained from randomizing the accepted and declined labels in our dataset (\*\*\*: 1%, \*\*: 5%, \*: 10%, n.s.: not significant). Note that in the analysis we exclude papers that are in collaboration with department faculty, and, therefore, the effects we report are not a direct consequence of those collaborations.

direct co-authorship) with other department members. This result is in contrast to those who declined, who are less likely to converge towards the topics that are prominent in the department.

To quantify this finding, we compare the topic distribution of researchers and departments using information-theoretic metrics. In particular, we compute the Jensen-Shannon distance between the topic distributions of the early-career faculty (before and after the offer, accepted or declined) and that of the department (before the offer). Again, we exclude from this analysis all publications that involve coauthorship between the early-career faculty and other faculty in the department; therefore, the convergence in

Departmental influence, gender and the selection of research agenda 95



**Figure 5.6: Changes in departmental collaborations and research topics after joining a department.** Same as Fig. 5.5 but restricting to papers published within the 10-year threshold. Stars indicate statistical significance obtained from randomizing the accepted and declined labels in our dataset (\*\*\*: 1%, \*\*: 5%, \*: 10%, n.s.: not significant). Note that in the analysis we exclude papers that are in collaboration with department faculty, and, therefore, the effects we report are not a direct consequence of those collaborations.

topic distribution we observe is not an immediate effect of those collaborations. Researchers converge towards the department if the change in distance with respect to the department is negative (they become closer after acceptance or declination) and diverge from the department if the change is positive. We find that the fraction of accepting early-career faculty who converge towards the department is 60% (Fig.5.5C), while the fraction of declining early-career faculty that converge towards the department is only 32%, both significant at the 5% uncertainty level when compared to the null hypothesis of researchers randomly moving towards or away from the department and therefore converging or diverging with equal probability. The difference between the two fractions is also significant. The results are also consistent if we restrict the topic analysis to publications carried in the 10-year period under study (Fig. 5.6).

### 5.3 Gender differences in the adaptation to the new faculty position

We have established that joining a department has a statistically significant effect on both the collaborations and the research topics of early-career faculty. However, Fig.5.5C also reveals that 40% of the early-career faculty who joined one of the three departments considered in the first cohort did not converge towards it. Therefore, the influence of the department is not equal for everyone. In particular, because studies in the literature have shown stark differences in gender when it comes to scientific output, attribution and authorship (*95, 103, 105, 106*), we further surmise that the research environment can also have a different effect depending on gender.

TU Austin	UC Boulder	Caltech	Carnegie Mellon
Cambridge	Cornell	Delaware	TU Delft
Denmark TU	Eindhoven	EPFL	Georgiatech
ICL	KIT	Leuven	Manchester
McGill	Milano	Minnesota	MIT
NWST	Princeton	PSU	UPenn
Purdue	Standford	Toronto	UCL
UCLA	Wisconsin	Yale	

Table 5.2: The 34 chemical engineering departments in the study

To test this hypothesis, we analyze data from among the 50 most prominent chemical engineering departments according to the QS World University Rankings for Chemical Engineering 2021 (*112*). In particular, we focus on the 34 universities that are based in North America or Europe (Table 5.2). For each one of these departments, we selected the tenure-track researchers (assistant, associate and full professors) and collected their publication and career data. With this information, we constructed the faculty-level collaboration networks and computed the topic distribution of each researcher and the overall department distribution (Fig. 5.7). These data, together with information about gender and the year in which the researcher

Departmental influence, gender and the selection of research agenda 97

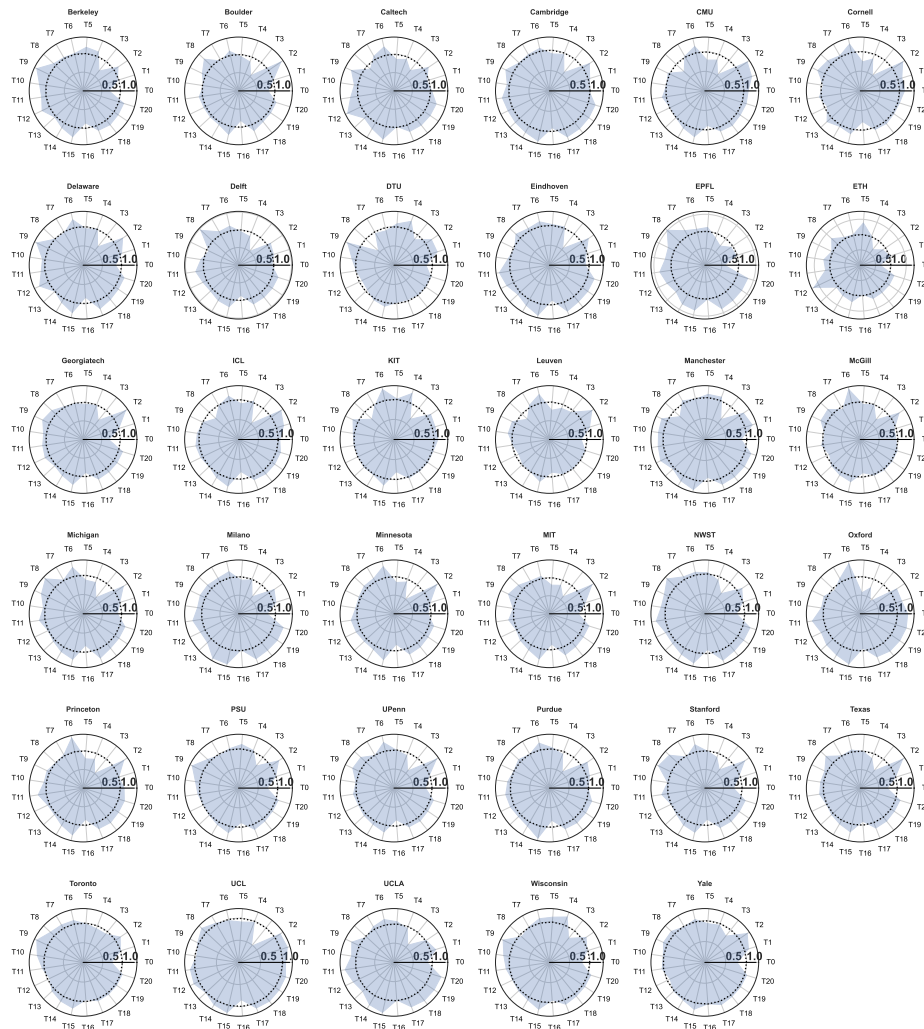


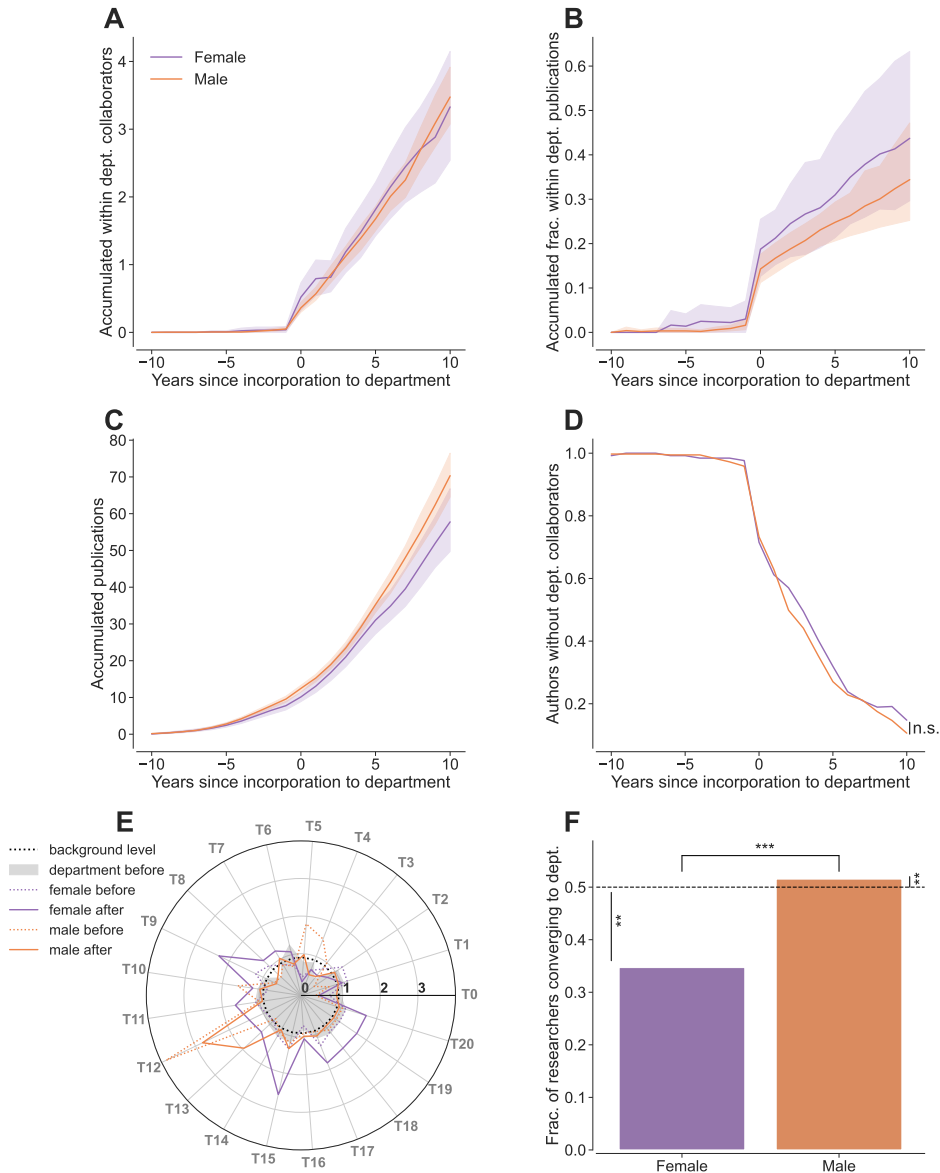
Figure 5.7: **Departmental topic distributions.** Average topic distribution of a department over all 20 topics normalized respect to the background of the field (dashed black line). A value of 1.0 in a topic, means that this topic is as popular in the department as it is in the background of the field.



joined the department, allow us to compare the effects of the department on the early-career faculty. Since we can expect more established researchers to be less permeable to new research topics, even when joining a new department, we focus on researchers who had less than 30 papers published at the time of joining. Other than this restriction, we consider all faculty who joined one of the departments considered after the year 2000.

After joining a department, male and female researchers do not display significant differences in the overall number of departmental collaborators or in the fraction of their publications that involve collaborations with other faculty of the department (Fig. 5.8A-B). There are also similarities in terms of the total number of publications and in the fraction of researchers who have no within-department collaborations. Indeed, Figure 5.8C shows that female researchers have, on average, similar number of publications as their male colleagues; and the fraction of female researchers who have not co-authored any papers with other members of the department after 10 years (15%) is not significantly higher than the fraction of male researchers in the same situation (9.8%, Fig. 5.8D).

Despite these similarities, we do observe gender differences in the convergence toward department topics. At the time of joining the department, female researchers are not further from the department in topic space than are male researchers (Fig. 5.9). However, we observe that female and male researchers evolve differently after joining. A comparison of the topic distributions before and after joining the department (Fig. 5.8E-F) shows that the fraction of male researchers converging towards the department (52%, Fig. 5.8F) is significantly higher than for female researchers (33%). Our results thus show that female researchers behave, in terms of their convergence towards the department topics, similar to the faculty that did not join the departments (Fig. 5.5C). Again, similar to the previous case, the results are consistent limiting the topic analysis to the 10-year period before and after joining the department (Fig. 5.10). This difference between men and women is striking because both groups joined the department in similar conditions and, overall, have similar collaboration patterns with other members of the department.



**Figure 5.8: Gender differences in collaborations and research topics after joining a department.** (A) Accumulated within-department distinct collaborators with respect to the year of incorporation to the department for male (orange) and female (purple) researchers. The shadowed area corresponds to the 95% confidence interval. (B) Fraction of publications that are with departmental collaborators. (C) Accumulated number of publications. (D) Fraction of authors with no collaborators within the department. (E) Comparison of topic distributions before (dotted lines) and 10 years after (solid lines) the year of incorporation to the department for a male and a female researcher. The gray solid area corresponds to the background of the department at that time. All distributions are normalized respect the general distribution of the field. The black dashed line represents the same topic proportion as the background. (F) Fraction of male and female researchers whose topic distributions converges to that of the de-

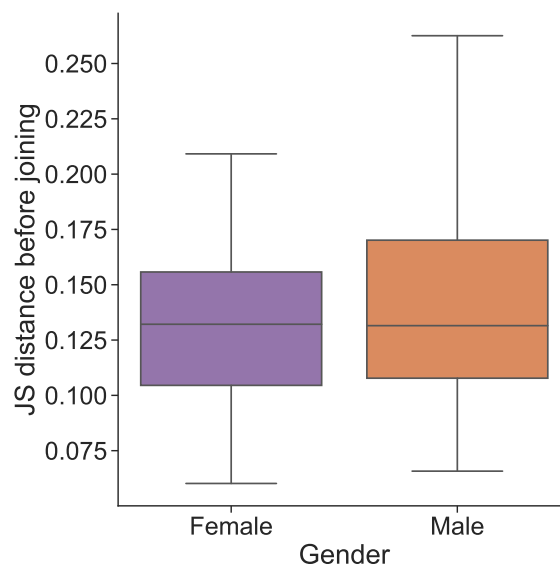


Figure 5.9: **Similarity of researchers to the department before joining the department.** Jensen-Shannon distance of the corpus of researchers (male and female) respect to the background of the department before they joined the department.

Departmental influence, gender and the selection of research agenda 101

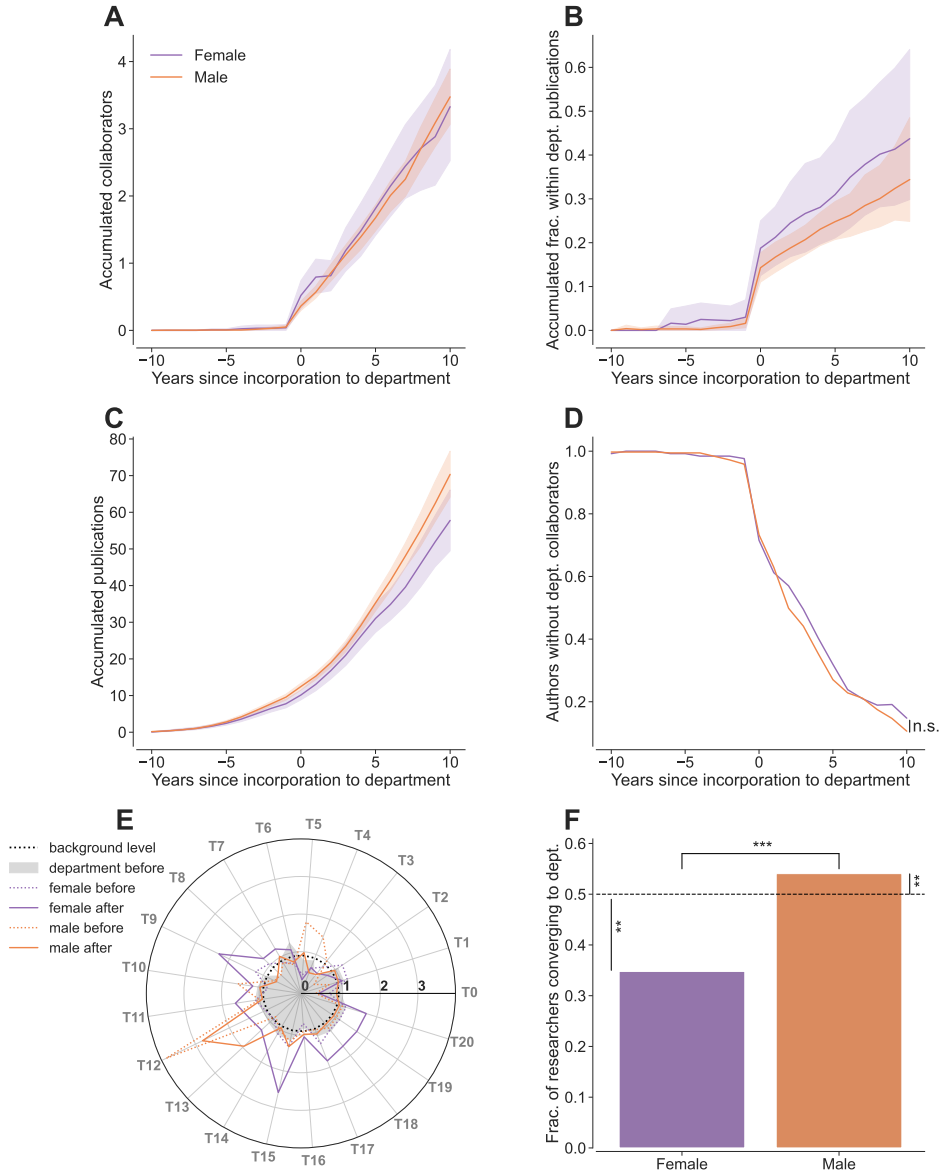


Figure 5.10: **Gender differences in collaborations and research topics after joining a department.** Same as Fig. 5.8 but restricting to papers published within the 10-year threshold. Stars indicate statistical significance obtained from randomizing the accepted and declined labels in our dataset (\*\*\*: 1%, \*\*: 5%, \*: 10%, n.s.: not significant). Note that in the analysis we exclude papers that are in collaboration with department faculty, and, therefore, the effects we report are not a direct consequence of those collaborations.

## 5.4 Career age biased collaborations

We have shown that while departments have an attractive effect on the topic evolution of young faculty, this effect is different on male and female faculty, despite them having similar overall departmental collaborations patterns. This result is in contrast to researchers accepting and declining offers in which the differences in convergence can be explained from the overall differences in collaboration patterns. We surmise that subtle differences in the selection of collaborators might play a role in the gender differences in topic convergence we observe. To explore this possibility, we study separately the evolution of collaborations of early-career faculty member with male and female incumbent members of the department. We observe that, when we take into account the gender of the collaborators, significant differences between male and female early-career faculty appear.

Figure 5.11A-C represents the cumulative fraction of within-department collaborators (that is, the number of collaborators of the early-career faculty divided by the number of collaborators available), aggregated by the gender of the early-career faculty and the gender of the incumbent collaborator. Female and male early-career faculty have indistinguishable fractions of male collaborators, and neither group deviates from the null expectation of selecting collaborators without taking gender into account (Fig. 5.11A). By contrast, there are significant differences in the fraction of female collaborators of early-career faculty (Fig. 5.11B-C): While male researchers have as many female collaborators as one would expect from the null model of gender-unbiased selection of collaborators (except for the last two years), female researchers collaborate with a significantly lower fraction of other female colleagues than expected.

Since the above finding might be a consequence of early-career faculty favoring senior professors within the department and women being under-represented in senior positions (95), we separate collaborations with senior collaborators (those who have a career at least 10 years longer than the early-career faculty member) and junior collaborators (those who have a career at most 10 years longer than the early-career faculty member). Both male and female early-career faculty members (Fig. 5.11D,G) engage in col-

Departmental influence, gender and the selection of research agenda 103

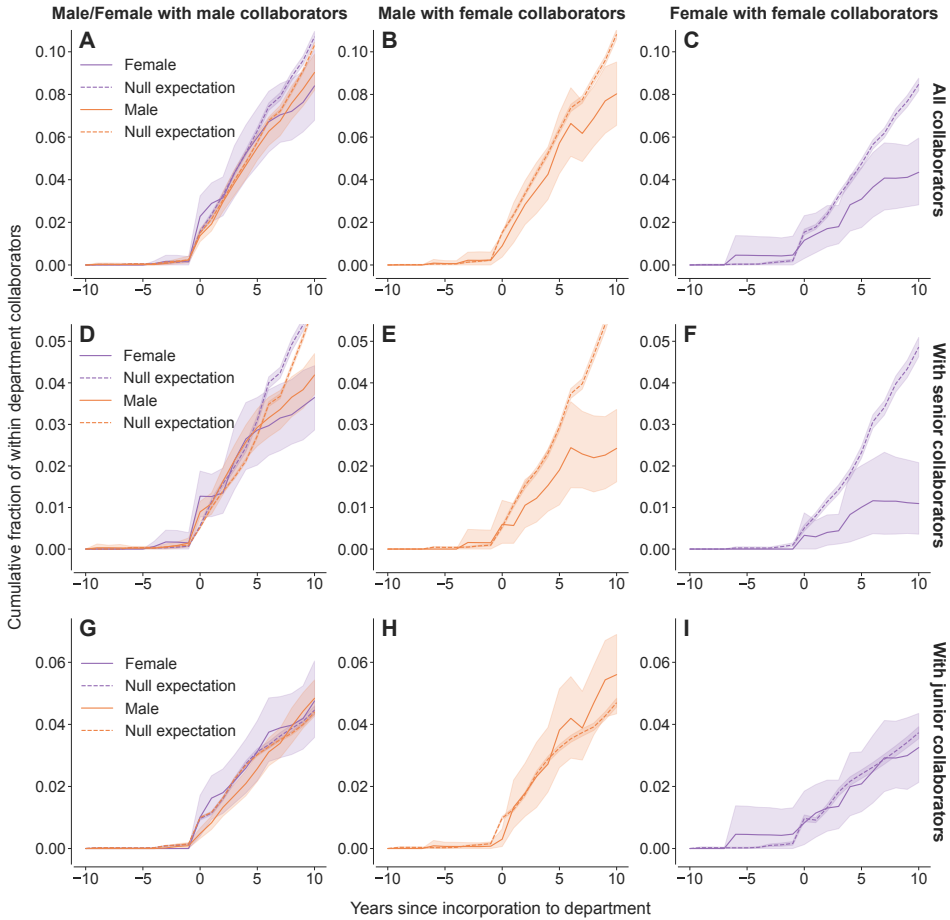


Figure 5.11: **The gendered target of collaborations.** Cumulative fraction of within-department collaborators by gender and career stage (purple represents females and orange represents males). The shadowed area represents the 95% confidence interval. The gray dashed line represents the expected fraction of collaborations if we randomize female and male labels in each department. **(A)** Male and female researchers with male collaborators **(B)** Male researchers with male collaborators. **(C)** Female researchers with female collaborators. **(D)** Male and female researchers with senior male collaborators. **(E)** Male researchers with male senior collaborators. **(F)** Female researchers with female senior collaborators. **(G)** Male and female researchers with junior male collaborators. **(H)** Male researchers with male junior collaborators. **(I)** Female researchers with female junior collaborators.

laboration with the same fraction of junior and senior male researchers and while in the case of collaboration with junior researchers are indistinguishable from what one would expect if collaborations did not have a gender bias, the collaborations with senior members are less than what would be expected, which could be an indicator of the early-career faculty establishing as proper PI's themselves. When engaging with senior female collaborators, male researchers collaborate as much as with other male faculty until their 7th year, after which their average number of senior female collaborators stops increasing. On the other hand, female early-career faculty collaborate significantly less than expected with senior female researchers almost from the first year after joining the department. These results differ from what we observe in the case of collaborating with junior incumbents. In this case, male early-career faculty collaborate as expected until the 7th year, when they start to collaborate more with junior female faculty. Female early-career faculty collaborate as expected with junior female incumbents.

## 5.5 Discussion

Our results show that working environments and in particular hiring departments exert a subtle yet quantifiable attractive force in the way young faculty shape their portfolio of scientific topics. Our results also show that, while constant contact with department colleagues can lead to collaborations that affect the topic selection process, the convergence toward department topics is apparent even when we do not take direct collaborations into account. This finding unmistakably points to the fact that the local research environment permeates into the shaping of academic careers beyond strict paper co-authorship. This finding is surprising, because departments typically do not actively seek to affect researchers portfolios (quite on the contrary, top departments would expect their researchers to pursue independent, fruitful research careers), but they inevitably do have an impact on topic choices.

Our results also clearly show that female and male young faculty do not respond to departmental pressures in the same way. While women seem

to have overall the same collaboration patterns as that of men, we observe stark differences in the way they converge towards department topics and in the way they select collaborators within the department: Female faculty tend to move away from department topics and collaborate less with female colleagues, in particular with senior female colleagues. These findings are both unexpected and worrisome, and they point toward yet another gender difference in academia with unknown causes and unknown consequences down the road.

In recent years, we have witnessed structural changes in many departments (95, 104), with an increase in the fraction of female faculty, and yet large gender differences still exist in terms of promotion and credit: Female researchers are typically given less credit for their contributions to research (106), and female faculty typically get promoted at later stages in their careers (104). These observations could suggest that women might look for different topic portfolios and different collaboration strategies as mechanisms to receive credit that could lead to future promotion, a mechanism male faculty might feel less pressured to use.

Studies from female academics in the 1990s already discussed that young women going into the academic track did not necessarily find more support in departments with a larger presence of women faculty, and that some women felt the need to stand out on their own rather than clustering with other female faculty (102, 113). It is unclear if female faculty currently feel the same way as in the 1990's, and whether or not there are idiosyncratic differences that could explain the differences in collaboration we observe. The obvious questions moving forward are therefore what are the reasons for these differences and what are the consequences of these differences? As many hiring departments make efforts and put policies in place to help reduce the gender gap in STEM departments, we cannot forget that hiring young female academics is not the endpoint but the start to breach the gender gap in academia. The department is the research environment where young faculty flourish, and right now this environment is not gender blind.





## Chapter 6

# Conclusions and final remarks

The main purpose of this work was to explore the useful insights of using network science as an approach to study science from the complex systems point of view. This particular choice was motivated by the long-history of networks applied to different problems and particularly applied to the study of social systems like the relations between scientist themselves, and also by the availability of a wide variety of methods already developed in this area that will help us to approach this issue from different angles.

In this direction, we started by discussing about inferential and descriptive methods, being the former the best option for our case, since we want to extract useful information about the mechanisms that give rise to the communities found in our networks. Thereafter, we introduced one of this inferential methods, namely the Stochastic Block Models (SBM), initially developed to aid in the study of social networks, as a flexible framework that can be adapted to the study of networks from different natures. Finally, we introduced how different variations of the classical SBM lead to a better explanatory models of our data and how to select the most plausible one among all the possibles. After this brief introduction to inferential methods and SBMs, we proceed to apply them to the study of science as a

social complex system in two different areas.

The first question that we have explored is that if different geographical, institutional and funding environments give rise to different collaboration patterns between prominent researchers, who we conjecture that are more sensitive to long-lasting research cultures. We found that, despite there are no systematic differences in term of collaborators or publications nor in the size of the scientific teams they engage, a more profound analysis of the networks of collaboration with an hSBM revealed some interesting behaviours. The groups found by our model were highly polarized in terms of affiliation, indicating that prominent researchers in North America and Europe represent different structural roles when comes to collaboration. Specifically, we found that Europeans establish more denser collaboration networks with their colleagues than prominent researchers in North America do, in terms of the total number of collaboration and collaborators. Finally this difference in collaboration patterns seem to influence the impact of researchers. Despite that collaboration increases the impact of research for both groups, North Americans take more advantage the synergy of collaboration by increasing the normalized logarithmic impact in XX% on average in front of the % for the Europeans, which can be related to the fact that repeated collaboration (that Europeans engage more habitually) implies a decline in this effect, although it's still more beneficial than no collaborating.

In light of this results, it is clear that despite not indicating causal relationships between funding mechanisms or institutions and the collaborative pattern, different allocations lead to different behaviors in terms of conducting research. Research policies by the European Comission in the last decades, led the EU15 to be one of the most important scientific production hubs worldwide competing with North America. Despite this success our results suggest that some those policies might be improved to cover the gap that still exists between both scientific powers, for instance by increasing the availability of soft money for individual early-career researchers rather than the collaboration by design scheme. This study can benefit from future additions and perspectives. How the mechanisms in the different affiliations create the different collaboration patterns remain elusive and exploring the funding available for each one of this researchers

can add explanatory value to our results. Also there is the question that when we talk about collaboration, we are referring to collaboration within the field and there might be collaboration with prominent researchers from other fields that lie beyond our focus. Finally, exploring how this results translate to the context of other research environments such as Asia (with the rise of China as a scientific power) or South America is a quite challenging yet exciting question.

Following the initial question of this work, we assessed if one can identify prominence by examining early-career factors of prominent researchers. We examined the accomplishments of prominent researchers during their early career (first five years) and found how there are some key-factors (publishing in a top 5 ranked journal, publishing more than half of their articles in top quartile journals, collaborating with other prominent researchers and being placed at a top 25 university) where they are overrepresented when compared with the average scientific workforce. Also, we observed how even if only the most prominent researchers are considered, there are still differences between them instead of being an homogeneous group. Indeed, if we disaggregate prominent researchers according to citation quartiles, we observed that those in the higher citation quartile have, in general, higher shares of these key-factors. Our results also show how this initial advantage during the first 5 years is echoed during the whole career of the researcher, since almost 90% of the researchers placed in the higher two quartiles, remain in those quartiles 20 years after the first publication, suggesting that there are some drivers of prominence. Finally, to test if the early key-factors are informative of later prominence, we tested them with a model that also included gender and geographical location and another that taken into account the h-index in the first 5 years. The results of the random forest classification showed that including the h-index turns out to be the best predictor of the citation and h-index quartiles at 20 years, correctly predicting 70%, but that even removing this and taking into account the four factors (plus gender and geographical location) correctly assigns the citation and h-index quartile for 44% of prominent researchers.

From this analysis we can observe how despite that factors accrued during the initial stages of a scientific career are important in predicting

later success, there are other factors at play which might not be captured by measurable metrics which rises a cautionary tale to policy makers and hiring committees that rely on bibliometrics in order to select possible candidates. While citations, h-index and so can capture a part of a researcher's talent, there are other factors which might not prompt at first sight and should be considered individually given the case. Future lines of work derived from this results might explore different sets of researchers to overcome the survivor bias, incorporating those who had a high share of these key factors but did not reach the top of the scientific field or even abandoned the scientific career. Other interesting lines of research might center on the use of different metrics to define prominence, such as altmetrics which measure the social impact of research or to explore the traits of prominence by different continents including those rising scientific powers like South America or East-Asia.

The third question addressed in this thesis was if new research environments have an effect shaping the research portfolio of early-career faculty who join a department. To that aim, we studied two different cohorts, a first one formed by researchers who were offered a position in a given department and their response, and a second one formed by researchers who joined a department in some of the top 50 chemical engineering institutions in Europe or North America. With the help of hSBM we modelled a background of topics for the field of chemical engineering which we used to extract the topical distributions of both, departments and researchers. Analyzing the first cohort, allowed us to examine the effect of joining a department versus not doing it. As expected after joining a department, the number of new collaborators increases, and the knowledge transaction of this collaborations echoes in the publications without other members of the department. Most of the new members of the department approach those topics popular prior to their entrance while for those who did not join we do not observe this effect. The second cohort, allows us to examine if this effect has the same implications for all researchers who join a department. Our results show how despite having a similar number of new departmental collaborators and publishing at a similar rate, while for male early-career faculty this effect is observed with more than 50% of them

converging towards the department, only a roughly 33% of their female colleagues converge towards this topics. Finally in order to explore if different collaboration patterns might explain this differences, we have observed how female early-career faculty have well-differentiated choices of collaborators, engaging equally with other junior early-career faculty regardless of their gender, but collaborating less than what we would expect with senior female faculty.

These results show how despite departments not seeking to provoke a change in the research portfolio of early-career faculty, they exert a change in the research topics of newcomers beyond direct collaborations with other members of the department and that his effect does not apply equally to male and female researchers, which is unexpected in view of the similar engaging with the departmental colleagues. Also, the fact that when entering a department they lack of collaboration with other female researchers is worrisome, since the lack of a strong core of same-gender researchers might affect the retention of this new workforce in the academia. Our study also opens future research possibilities. Since most of the female researchers that we can currently observe have incorporated to the scientific workforce in recent years, exploring how their careers develop in terms of impact and new topics of research, and compare groups according their current collaborative patterns (whether their collaborated with senior colleagues or not), will allow to assess the effects of this career choices. In order to unveil possible mechanisms driving this patterns, it could be interesting to obtain the funding information and determine if there is a relation between the funds or the prestige of a researcher determines whether if the incumbent scientists are more prone to engage into a collaboration with them. Finally, the data used in our study captures only a part of the influence that a research environment can have through direct collaboration, while a probably more subtle influence, like the informal mentoring during departmental meetings can have, and explore in this direction could broaden our understanding of the mechanisms at play in this effect.

During the development of this thesis, network science has proven to be a flexible approach to analyze how science is produced. From the same mathematical model we have studied different geographical collaboration

patterns and networks formed by documents and words, allowing us to explore different problems from novel points of view, that will be also benefited and complemented by future development of new network models exploiting the capacities of this relatively simple yet powerful tool.

# Bibliography

1. E. M. Anicich, R. I. Swaab, A. D. Galinsky, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 1338 (2015).
2. L. Wu, D. Wang, J. A. Evans, *Nature* **566**, 378 (2019).
3. C. Babbage, *Reflections on the Decline of Science in England: And on Some of Its Causes* (B. Fellowes, 1830).
4. E. Garfield, *Science* **178**, 471 (1972).
5. R. K. Merton, *Science* **159**, 56 (1968).
6. M. W. Rossiter, *Social Studies of Science* **23**, 325 (1993).
7. D. J. de Solla Price, *Little Science, Big Science* (Columbia University Press, 1963).
8. X. H. T. Zeng, *et al.*, *PLoS. Biol.* **14**, 1 (2016).
9. S. Milojević, *J. Informetr.* **9**, 962 (2015).
10. R. Guimerà, B. Uzzi, J. Spiro, A. LAN, *Science* **308**, 697 (2005).
11. W. Pan, G. Ghoshal, C. Krumme, M. Cebrian, A. Pentland, *Nat. Commun.* **4**, 1961 (2013).
12. A. Vazquez, A. Flammini, A. Maritan, A. Vespignani, *Nat. Biotechnol.* **21**, 697 (2003).



13. A. M. Sadri, S. Hasan, S. V. Ukkusuri, J. E. Suarez Lopez, *Soc. Netw. Anal. Min.* **8**, 56 (2018).
14. D. J. de Solla Price, *Science* **149**, 510 (1965).
15. V. Danchev, A. Rzhetsky, J. A. Evans, *eLife* **8**, e43094 (2019).
16. A. Montanari, A. Saberi, *Proceedings of the National Academy of Sciences* **107**, 20196 (2010).
17. T. P. Peixoto, Descriptive vs. inferential community detection in networks: pitfalls, myths, and half-truths (2021).
18. P. W. Holland, K. B. Laskey, S. Leinhardt, *Soc. Networks* **5**, 109 (1983).
19. T. P. Peixoto, *Phys. Rev. X* **4**, 011047 (2014).
20. D. J. de Solla Price, *Little Science, Big Science... and Beyond* (Columbia Univ. Press, 1963).
21. R. K. Merton, *The Sociology of Science* (Univ. of Chicago Press, Chicago, 1973).
22. M. E. J. Newman, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 404 (2001).
23. R. Guimerà, B. Uzzi, J. Spiro, L. Amaral, *Science* **308**, 697 (2005).
24. A. Almaatouq, M. Alsobay, M. Yin, D. J. Watts, *Proc. Natl. Acad. Sci. U.S.A.* **118** (2021).
25. T. Carletti, A. Guarino, A. Guazzini, F. Stefanelli, *JASSS* **23**, 4 (2020).
26. S. Wuchty, B. F. Jones, B. Uzzi, *Science* **316**, 1036 (2007).
27. B. F. Jones, S. Wuchty, B. Uzzi, *Science* **322**, 1259 (2008).
28. H. F. Chan, A. S. Önder, B. Torgler, *Scientometrics* **106**, 509 (2016).

29. Y. Ma, B. Uzzi, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 12608 (2018).
30. S. Srivastava, M. Banaji, *Am. Sociol. Rev.* **76**, 207 (2011).
31. B. Uzzi, S. Mukherjee, M. Stringer, B. Jones, *Science* **342**, 468 (2013).
32. L. Bromham, R. Dinnage, X. Hua, *Nature* **534**, 684 (2016).
33. C. S. Wagner, T. A. Whetsell, L. Leydesdorff., *Scientometrics* **110**, 1633 (2017).
34. D. A. King, *Nature* **430**, 311 (2004).
35. B. Lepori, A. Geuna, A. Mira, *PLoS ONE* **14**, 1 (2019).
36. A. M. Petersen, W.-S. Jung, J.-S. Yang, H. E. Stanley, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18 (2011).
37. A. Barabási, C. Song, D. Wang, *Nature* **491**, 1476 (2012).
38. J. Li, Y. Yin, S. Fortunato, D. Wang, *J. R. Soc. Interface* **17**, 20200135 (2020).
39. J. P. A. Ioannidis, J. Baas, R. Klavans, K. W. Boyack, *PLoS Biol.* **17**, 1 (2019).
40. L. Bouchard, M. Albertini, R. Batista, J. D. Montigny, *Soc. Sci. Med.* **141**, 100 (2015).
41. J. E. Hirsch, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 16569 (2005).
42. M. Stringer, M. Sales-Pardo, L. Amaral, *PLoS ONE* **3**, e1683 (2008).
43. W. H. Press, S. A. Teukolsky, *Comput. Phys.* **2**, 74 (1988).
44. H. C. White, S. A. Boorman, R. L. Breiger, *Am. J. Sociol.* **81**, 730 (1976).
45. K. Nowicki, T. A. B. Snijders, *J. Am. Stat. Assoc.* **96**, 1077 (2001).

46. R. Guimerà, M. Sales-Pardo, *Proc. Natl. Acad. Sci. U. S. A.* **106**, 22073 (2009).
47. W. Li, T. Aste, F. Caccioli, G. Livan, *Nat. Commun.* **10**, 5170 (2019).
48. J. Hoekman, T. Scherngell, K. Frenken, R. Tijssen, *J. Econ. Geogr.* **13**, 23 (2012).
49. A. M. Petersen, *Proc. Natl. Acad. Sci. U.S.A.* **112**, E4671 (2015).
50. B. Hoenig, *Europe's new scientific elite: Social mechanisms of science in the European research area* (Taylor & Francis, 2017).
51. A. Olechnicka, A. Ploszaj, D. Celińska-Janowicz, *The geography of scientific collaboration* (Taylor & Francis, 2019).
52. J. Tu, *Scientometrics* **118**, 587 (2019).
53. H. C. Triandis, *The handbook of culture and psychology* (2001).
54. R. Nisbett, *The Geography of Thought: How Asians and Westerners Think Differently— and why* (Nicholas Brealey, 2003).
55. A. Baccini, G. De Nicolao, E. Petrovich, *PLoS ONE* **14**, 1 (2019).
56. B. Maher, M. Sureda Anfres, *Nature* **538**, 444 (2016).
57. V. Sekara, *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 12603 (2018).
58. A. Calvó-Armengol, M. O. Jackson, *Am. Econ. Rev.* **94**, 426 (2004).
59. J. N. Parker, C. Lortie, S. Allesina, *Scientometrics* **85**, 129 (2010).
60. D. D. Beaver, *Scientometrics* **52**, 365 (2001).
61. D. S. Massey, C. Z. Charles, G. Lundy, M. J. Fischer, *The source of the river: The social origins of freshmen at America's selective colleges and universities* (Princeton University Press, 2003).

62. S. Bowles, H. Gintis, M. O. Groves, *Unequal chances: Family background and economic success*. (Princeton Univeristy Press, 2008).
63. A. Clauset, S. Arbesman, D. B. Larremore, *Sci. Adv.* **1** (2015).
64. T. Boyer-Kassem, C. Mayo-Wilson, M. Weisberg, *Scientific collaboration and collective knowledge: New essays* (Oxford University Press, 2017).
65. S. Fortunato, *et al.*, *Science* **359** (2018).
66. A. Clauset, D. B. Larremore, R. Sinatra, *Science* **355**, 477 (2017).
67. P. Azoulay, *et al.*, *Science* **361**, 1194 (2018).
68. J. A. Evans, J. G. Foster, *Science* **331**, 721 (2011).
69. A. Zeng, *et al.*, *Phys. Rep.* **714-715**, 1 (2017).
70. D. E. Acuna, S. Allesina, K. P. Kording, *Nature* **489**, 201 (2012).
71. R. Sinatra, D. Wang, P. Deville, C. Song, A.-L. Barabási, *Science* **354** (2016).
72. D. Wang, C. Song, A.-L. Barabási, *Science* **342**, 127 (2013).
73. E. M. Schlagberger, L. Bornmann, J. Bauer, *Scientometrics* **109**, 723 (2016).
74. H. F. Chan, B. Torgler, *Scientometrics* **102**, 847 (2015).
75. J. P. Ioannidis, *et al.*, *BMC Med.* **5**, 30 (2007).
76. N. Amara, R. Landry, N. Halilem, *Scientometrics* **103**, 489 (2015).
77. E. C. McKiernan, *et al.*, *eLife* **8** (2019).
78. J. A. G. Moreira, X. H. T. Zeng, L. A. N. Amaral, *PLoS ONE* **10** (2015).

79. M. E. J. Newman, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 404 (2001).
80. Y. Bu, *et al.*, *Proc. Assoc. Inf. Sci.* **55**, 29 (2018).
81. J. F. Liénard, T. Achakulvisut, D. E. Acuna, S. V. David, *Nat. Commun.* **9**, 4840 (2018).
82. T. Amjad, *et al.*, *J. Informetr.* **11**, 307 (2017).
83. D. K. Simonton, *Psychol. Rev.* **104**, 66 (1997).
84. Y. Dong, R. A. Johnson, N. V. Chawla, *IEEE Trans. Big Data* **2**, 18 (2016).
85. L. Danus, C. Muntaner, A. Krauss, M. Sales-Pardo, R. Guimera, Differences in collaboration structures and impact among prominent researchers in europe and north america (2022).
86. Statistics and resources | 2021 Science Report.
87. World University Rankings (2019).
88. C. f. S. a. T. Studies (CWTS), CWTS Leiden Ranking.
89. R. Miranda, E. Garcia-Carpintero, *Scientometrics* **121**, 479 (2019).
90. W. Liu, G. Hu, M. Gu, *Scientometrics* **106**, 1273 (2016).
91. M. Huang, *Online Inf. Rev.* **36**, 534 (2012).
92. SJR : Scientific Journal Rankings.
93. K. H. Wapman, S. Zhang, A. Clauset, D. B. Larremore, *Nature* **610**, 120 (2022).
94. Y. Ma, S. Mukherjee, B. Uzzi, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 14077 (2020).
95. J. Duch, *et al.*, *PLoS ONE* **7**, e51332 (2012).

96. T. Kuhn, M. Perc, D. Helbing, *Phys. Rev. X* **4** (2014).
97. C. Wagner, X. Cai, Y. Zhang, C. Fry, *PLoS ONE* **17**, e0261624 (2022).
98. S. Milojević, F. Radicchi, J. P. Walsh, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 12616 (2018).
99. F. Battiston, *et al.*, *Nat. Rev. Phys.* **1**, 89 (2019).
100. A. Aleta, S. Meloni, N. Perra, Y. Moreno, *EPJ Data Sci.* **8**, 27 (2019).
101. X. H. T. Zeng, *et al.*, *PLoS Biol.* **14**, 1 (2016).
102. H. Etzkowitz, C. Kemelgor, B. Uzzi, *Athena Unbound: The Advancement of Women in Science and Technology* (Cambridge University Press, 2000).
103. V. Larivière, C. Ni, Y. Gingras, B. Cronin, C. R. Sugimoto, *Nature* **504**, 211 (2013).
104. H. Boekhout, I. van der Weijden, L. Waltman, Gender differences in scientific careers: A large-scale bibliometric analysis (2021).
105. C. Ni, E. Smith, H. Yuan, V. Larivière, C. R. Sugimoto, *Sci. Adv.* **7**, eabe4639 (2021).
106. M. B. Ross, *et al.*, *Nature* p. in press (2022).
107. J. N. Cummings, S. Kiesler, *Soc. Stud. Sci.* **35**, 703 (2005).
108. E.-A. Horvát, B. Uzzi, *Nature* **605** (2022).
109. G. Ghiasi, V. Larivière, C. R. Sugimoto, *PLoS ONE* **10**, 1 (2016).
110. M. Gerlach, T. P. Peixoto, E. G. Altmann, *Sci. Adv.* **4**, eaaq1360 (2018).
111. M. Gerlach, H. Shi, L. A. N. Amaral, *Nat. Mach. Intell.* **1**, 606 (2019).

BIBLIOGRAPHY

---

120

112. QS World University Rankings for Chemical Engineering 2021 (2021).
113. H. Etzkowitz, C. Kemelgor, M. Neuschatz, B. Uzzi, A. J., *Science* **266**, 51 (1994).

## Appendix A

# Networks and names of researchers for all the fields



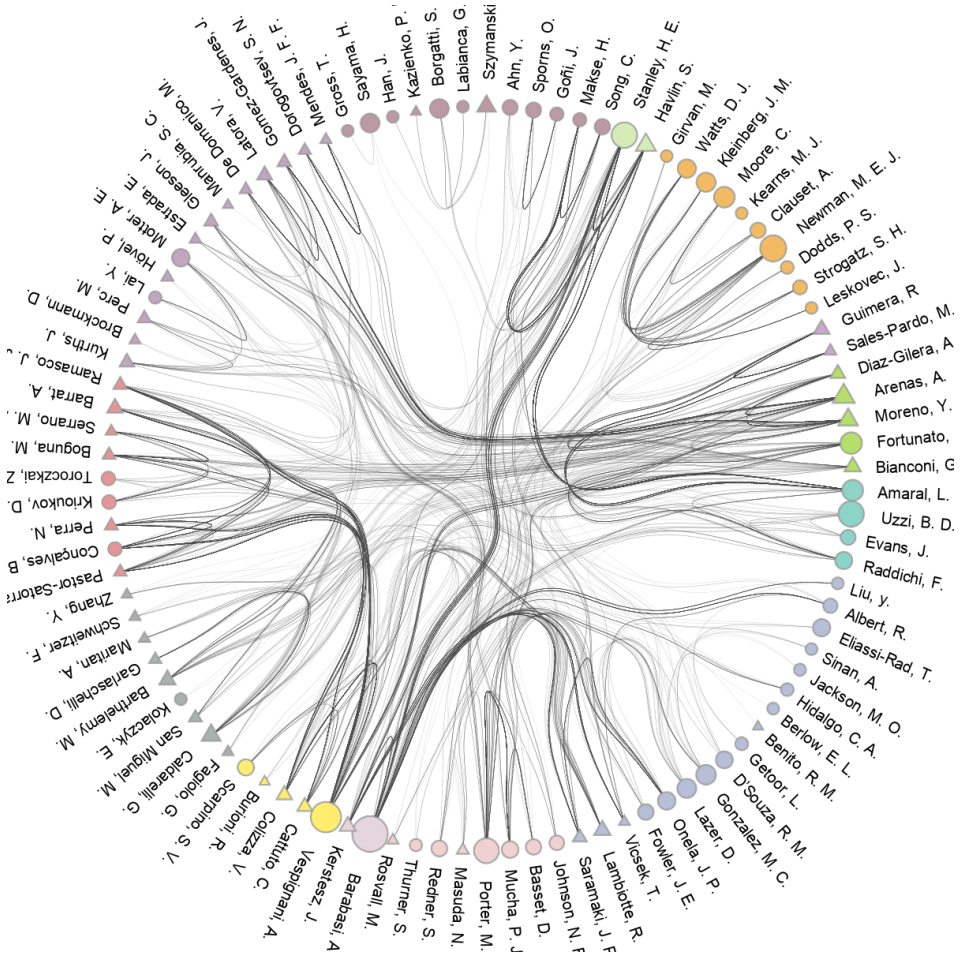


Figure A.1: **Community structure for Network Science:** Each node in the network represents a prominent researcher, and each edge represents a different collaboration (coauthored paper) between a pair of researchers. Prominent researchers in North America and Europe are represented as circles and triangles, respectively. Different colors correspond to the groups identified by the hSBM, so that nodes with the same color have a similar collaboration pattern with other researchers and therefore fulfill a similar structural role in the collaboration network. Node size represents the betweenness centrality of the researcher in the network.

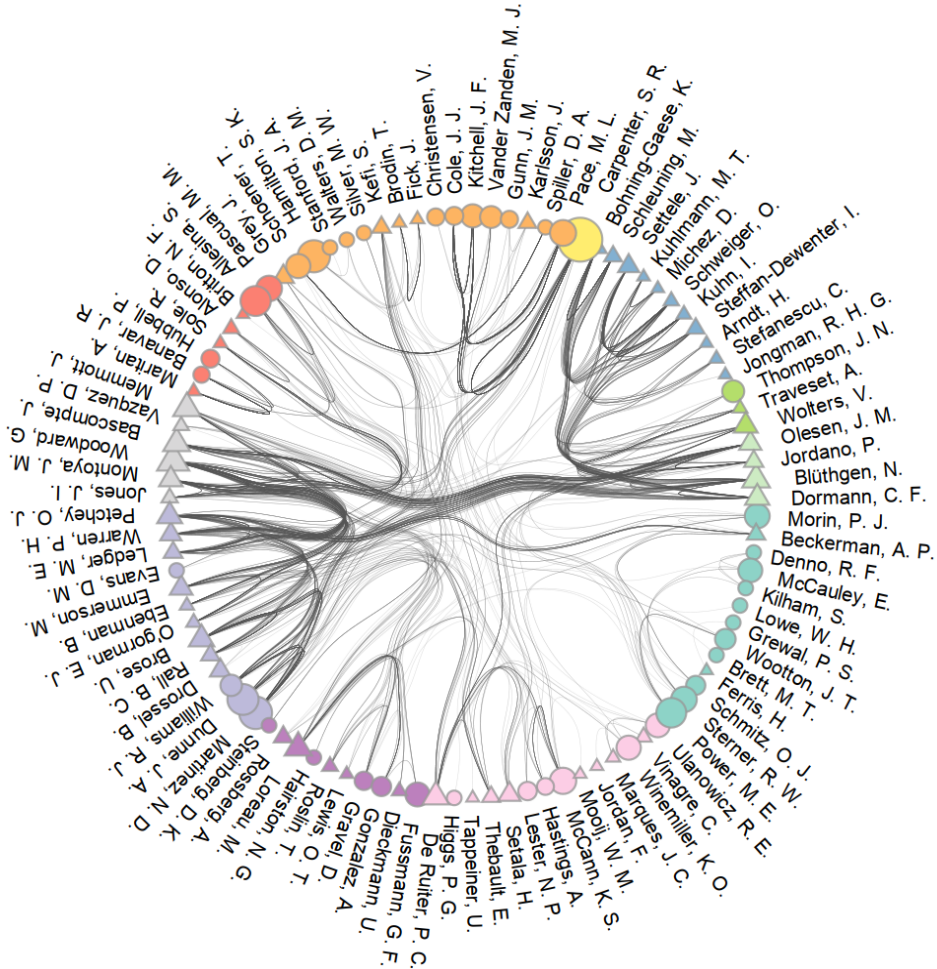


Figure A.2: **Community structure for Network Ecology:** Each node in the network represents a prominent researcher, and each edge represents a different collaboration (coauthored paper) between a pair of researchers. Prominent researchers in North America and Europe are represented as circles and triangles, respectively. Different colors correspond to the groups identified by the hSBM, so that nodes with the same color have a similar collaboration pattern with other researchers and therefore fulfill a similar structural role in the collaboration network. Node size represents the betweenness centrality of the researcher in the network.

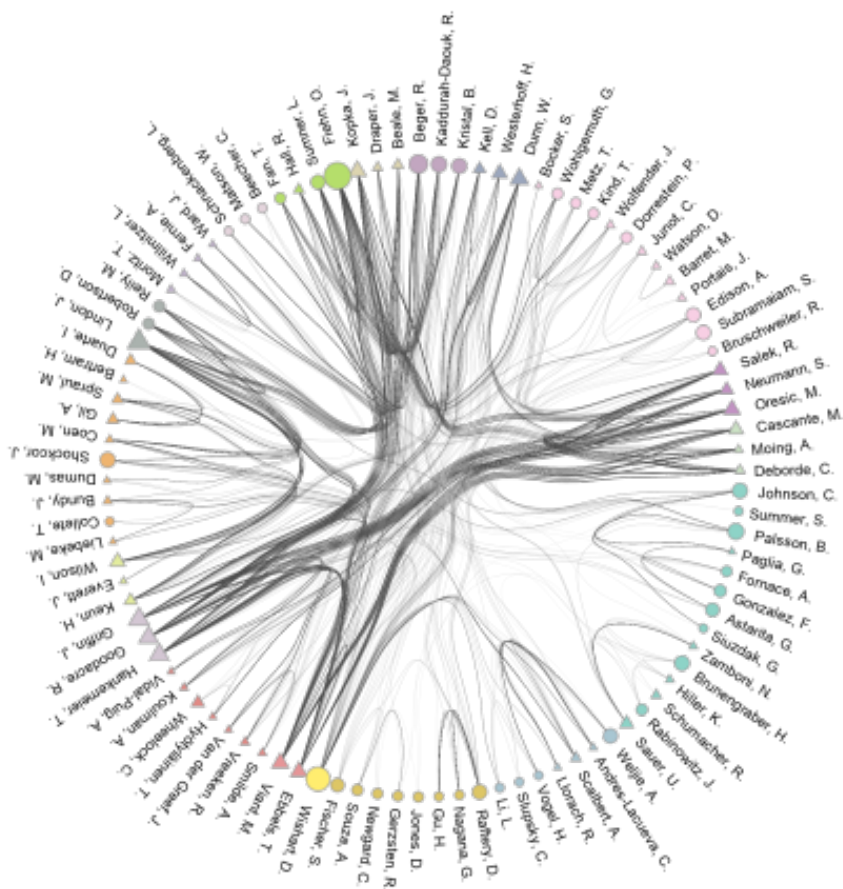


Figure A.3: **Community structure for Metabolomics:** Each node in the network represents a prominent researcher, and each edge represents a different collaboration (coauthored paper) between a pair of researchers. Prominent researchers in North America and Europe are represented as circles and triangles, respectively. Different colors correspond to the groups identified by the hSBM, so that nodes with the same color have a similar collaboration pattern with other researchers and therefore fulfill a similar structural role in the collaboration network. Node size represents the betweenness centrality of the researcher in the network.

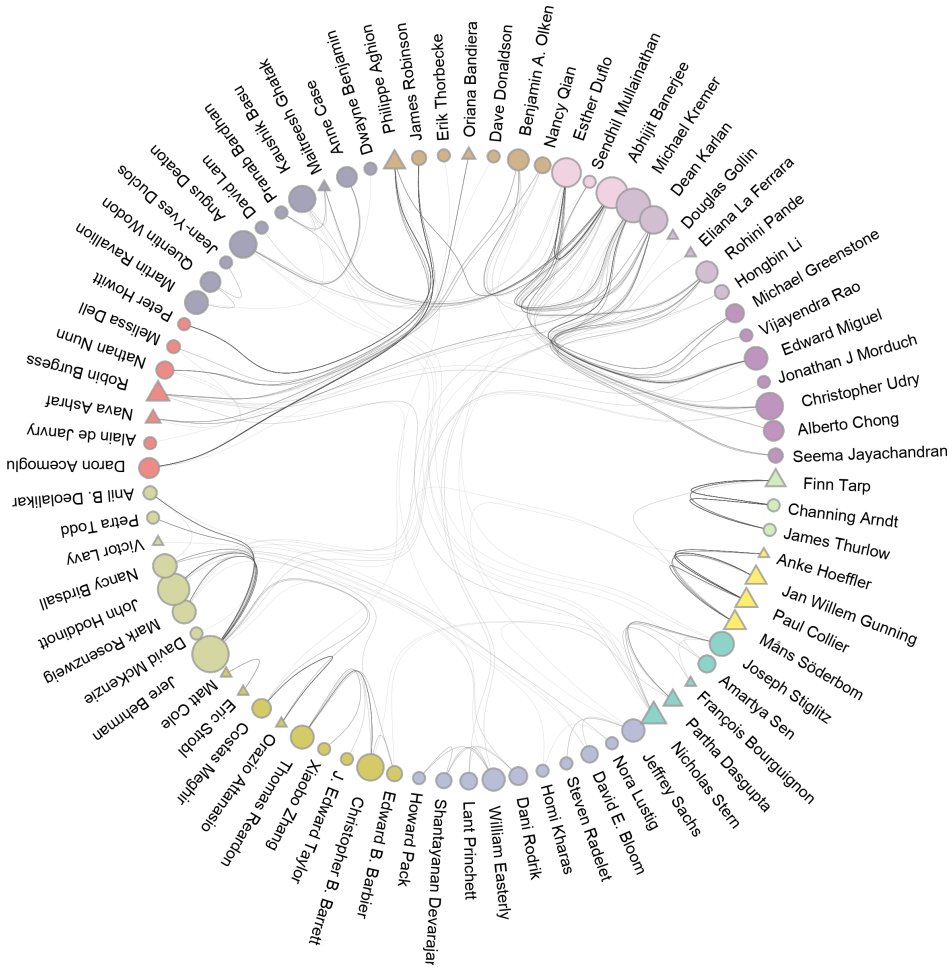


Figure A.4: **Community structure for Development Economics:** Each node in the network represents a prominent researcher, and each edge represents a different collaboration (coauthored paper) between a pair of researchers. Prominent researchers in North America and Europe are represented as circles and triangles, respectively. Different colors correspond to the groups identified by the hSBM, so that nodes with the same color have a similar collaboration pattern with other researchers and therefore fulfill a similar structural role in the collaboration network. Node size represents the betweenness centrality of the researcher in the network.





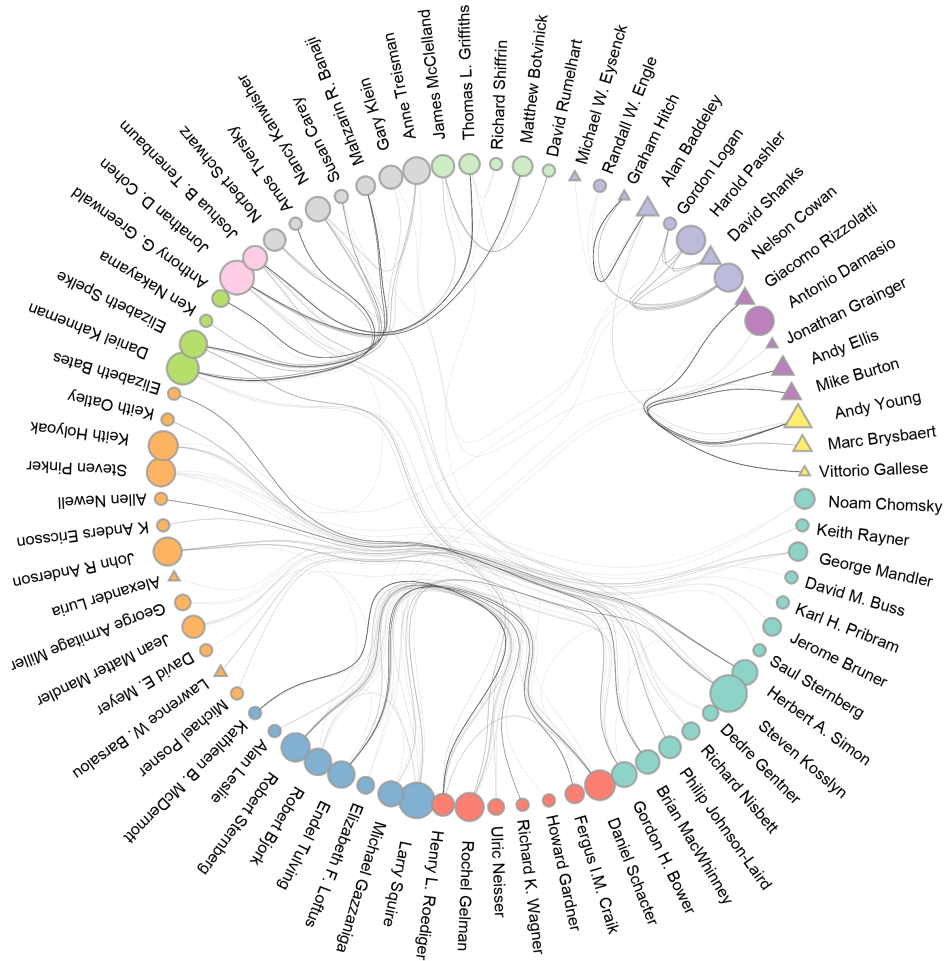


Figure A.6: **Community structure for Cognitive Psychology:** Each node in the network represents a prominent researcher, and each edge represents a different collaboration (coauthored paper) between a pair of researchers. Prominent researchers in North America and Europe are represented as circles and triangles, respectively. Different colors correspond to the groups identified by the hSBM, so that nodes with the same color have a similar collaboration pattern with other researchers and therefore fulfill a similar structural role in the collaboration network. Node size represents the betweenness centrality of the researcher in the network.

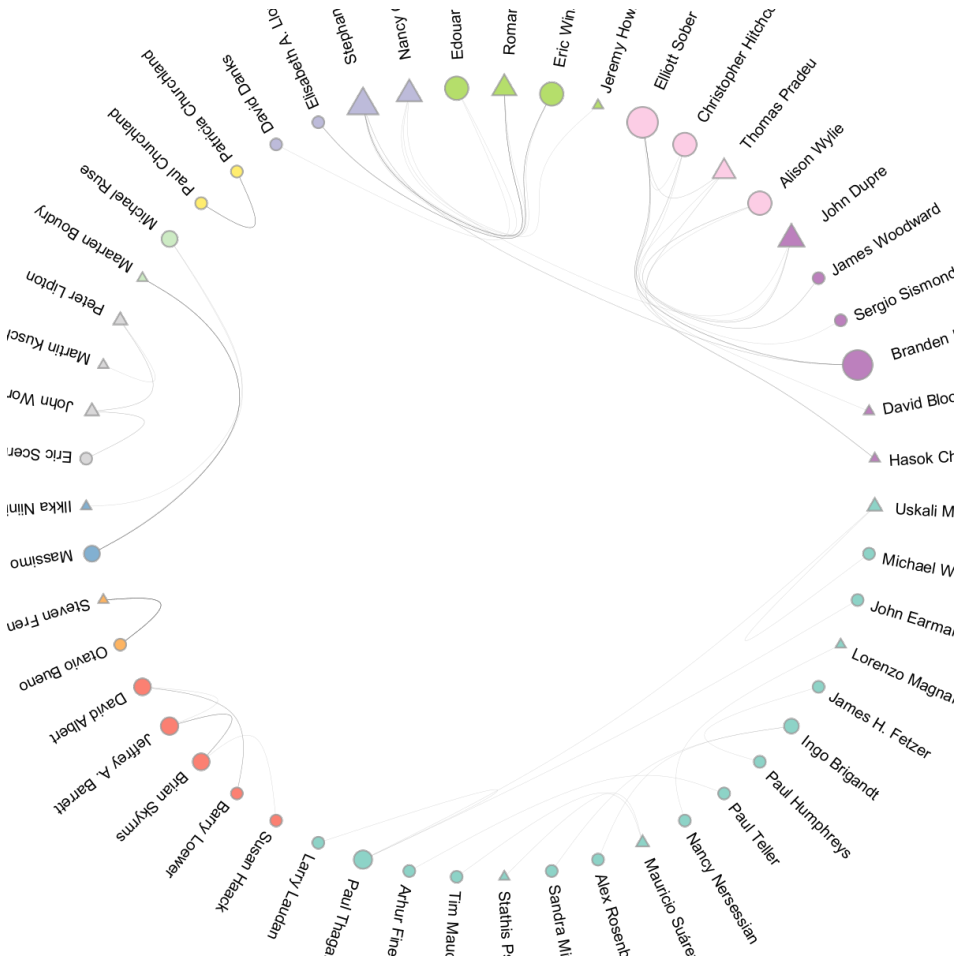


Figure A.7: **Community structure for Philosophy of Science:** Each node in the network represents a prominent researcher, and each edge represents a different collaboration (coauthored paper) between a pair of researchers. Prominent researchers in North America and Europe are represented as circles and triangles, respectively. Different colors correspond to the groups identified by the hSBM, so that nodes with the same color have a similar collaboration pattern with other researchers and therefore fulfill a similar structural role in the collaboration network. Node size represents the betweenness centrality of the researcher in the network.

UNIVERSITAT ROVIRA I VIRGILI

NETWORK AND INFORMATION-THEORETIC STUDIES ON THE EFFECTS OF RESEARCH ENVIRONMENT IN  
SCIENTIFIC CAREERS: GEOGRAPHY, PROMINENCE AND GENDER

Lluís Danús Amengual



