A Multilevel Exploration of Science Dynamics with Computational and Quantitative Techniques

Carlo Debernardi

UiA University of Agder

Doctoral Dissertations at the University of Agder 518

A Multilevel Exploration of Science Dynamics with Computational and Quantitative Techniques

Carlo Debernardi

A Multilevel Exploration of Science Dynamics with Computational and Quantitative Techniques

Doctoral dissertation for the degree of Philosophiae Doctor (Ph.D.)

University of Milan / University of Agder

Faculty of Political, Economic and Social Sciences / Faculty of Social Sciences

2025

Doctoral Dissertations at the University of Agder 518

ISSN: 1504-9272

ISBN: 978-82-8427-236-8

Printed by MAKE!Graphics, Kristiansand

Acknowledgements

A PhD is never a straight and easy path. Especially so when the challenges of academia pile up with those of personal life. Mine is no exception, and I acknowledge the fact that I could at least reach the end thanks to the support of many.

I wish to thank Marco Seeber for his continuous help and guidance, as well as Gabriele Ballarino and Flaminio Squazzoni for the insights, suggestions, and supervision. I am grateful for the time and effort they invested in supporting me and for the many opportunities they gave me access to.

I shared a long part of the journey, and much fun, with many other people. Even though Covid allowed us to meet for the first time only at the end of the first year, I wish to thank my fellow PhDs of the Behave Lab, Francesco Renzini, Fabio Torreggiani, Esteban Muñoz. Along with them, I also want to thank all the other members of the Behave in Milan, for the insightful and supportive research environment.

While spending my year in Kristiansand, I also got to know, and enjoy time with, many people that I wish to thank. Among them are Tobias Hofelich, Julie Pihlmann, Kasper Nome, Luiz de Andrade, Laszlo Bugyi, Magnus Bragdø, Johan Andersen, Frans Malmborg, Marina Rabinovich and Aleksandar Avramovic.

My experience in Norway, along with many other activities, would not have been possible with the constant support of academics and administrators of both the Unimi and UiA PhD programmes. Thus, I also wish to thank Gabriele Ballarino, Marco Guerci, Licia Papavero, Jarle Trondal, Stefan Gänzle and Cecilie Mawdsley.

I am also grateful to Eugenio Petrovich and Alberto Baccini for the opportunity to keep working on the quantitative study of science as a collaborator of an exciting research project.

Finally, I wish to thank my housemates in Gimleveien, for making me live a really Norwegian experience and for appreciating my Italian obsession with food. A very deep thank you also to my family and friends, and – most importantly – to Giorgia and Coco, for the little thing we are building together.

Summary

While contemporary societies are increasingly reliant on science, our understanding of how exactly the scientific community works is still limited. The research work has experienced tremendous changes during the past few decades, and it is still evolving both as a result of endogenous dynamics and explicit science policies. For relatively easy to measure dimensions, like the number of publications and citations, we have solid evidence of strategic adaptation to incentives by individual scholars – e.g. self-citations. However, we know very little regarding the more elusive – yet central – dimension of the research topics, i.e. what scholars choose to investigate. Are there general patterns of specialisation? Do the many filters of academia – i.e., peer-review, recruitment, competitive funding, etc. – have an impact on the research topic distribution? Can these factors affect the breadth of individual research agendas, by widening or narrowing down the freedom of choosing the research questions? This dissertation aims to investigate such questions by using quantitative techniques – borrowing recently developed language modeling tools – to measure research topics starting from the textual content of scientific publications. The dissertation is composed of a Part I, introducing theory and methods, and a Part II, presenting three independent empirical studies. The first is a global overview of the climate change literature, exploring country-level specialisation patterns. The second is an investigation of the role of research topics in determining the success rate of ERC proposals, thus exploring the role of institutional filters on the research topic distribution. Finally, the third study examines a sample of scholars across four disciplinary areas and four countries, tracking how their individual research agendas changed over their careers, in relation to their academic seniority, mobility and scientific collaboration.

Sommario (IT)

Sebbene le società contemporanee facciano sempre più affidamento sulla scienza, la nostra comprensione di come funziona esattamente la comunità scientifica è ancora limitata. Il lavoro di ricerca ha subito enormi cambiamenti negli ultimi decenni ed è ancora in evoluzione sia come risultato di dinamiche endogene che di politiche scientifiche esplicite. Per dimensioni relativamente facili da misurare, come il numero di pubblicazioni e citazioni, abbiamo prove concrete di adattamento strategico agli incentivi da parte dei singoli studiosi – ad es. self-citations. Sappiamo però molto poco riguardo alla dimensione più sfuggente – ma centrale – dei temi di ricerca, ovvero ciò che gli studiosi scelgono di indagare. Esistono pattern generali di specializzazione? I numerosi filtri del mondo accademico – ovvero peer-review, reclutamento, finanziamenti competitivi, ecc. – hanno un impatto sulla distribuzione degli argomenti di ricerca? Questi fattori possono influenzare l'ampiezza delle agende di ricerca individuali, allargando o restringendo la libertà di scelta delle domande di ricerca? Questa tesi si propone di indagare tali questioni utilizzando tecniche quantitative – impiegando modelli di linguaggio di recente sviluppo – per misurare gli argomenti di ricerca a partire dal contenuto testuale delle pubblicazioni scientifiche. La tesi è composta da una Parte I, che introduce teoria e metodi, e una Parte II, che presenta tre studi empirici indipendenti. Il primo è una panoramica globale della letteratura sul cambiamento climatico, che esplora i modelli di specializzazione a livello nazionale. Il secondo è un'indagine sul ruolo dei temi di ricerca nel determinare il tasso di successo delle proposte ERC, esplorando così il ruolo dei filtri istituzionali sulla distribuzione dei temi di ricerca. Infine, il terzo studio esamina un campione di studiosi di quattro aree disciplinari e quattro paesi, monitorando come i loro programmi di ricerca individuali sono cambiati nel corso delle loro carriere, in relazione alla loro anzianità accademica, mobilità e collaborazione scientifica

Summary (NO)

Mens moderne samfunn i økende grad er avhengige av vitenskap, er vår forståelse av hvordan det vitenskapelige samfunnet fungerer fortsatt begrenset. Forskningsarbeidet har opplevd enorme endringer i løpet av de siste tiårene, og det utvikler seg fortsatt både som et resultat av endogen dynamikk og eksplisitt vitenskapelig politikk. For relativt enkle å måle dimensjoner, som antall publikasjoner og siteringer, har vi solide bevis på strategisk tilpasning til insentiver fra individuelle forskere – f.eks. selvsiteringer. Imidlertid vet vi svært lite om den mer unnvikende – men likevel sentrale – dimensjonen ved forskningstemaene, det vil si hva forskere velger å undersøke. Er det generelle spesialiserings mønstre? Har de mange filtrene i akademia – det vil si fagfellevurdering, rekruttering, konkurransefinansiering osv. – innvirkning på forskningsemnets fordeling? Kan disse faktorene påvirke bredden i individuelle forskningsagendaer, ved å utvide eller innsnevre friheten til å velge forskningsspørsmålene? Denne avhandlingen tar sikte på å undersøke slike spørsmål ved å bruke kvantitative teknikker – ved å låne nylig utviklede språkmodelleringsverktøy – for å måle forskningstemaer med utgangspunkt i tekstinnholdet i vitenskapelige publikasjoner. Avhandlingen er satt sammen av en del I, som introduserer teori og metoder, og en del II, som presenterer tre uavhengige empiriske studier. Den første er en global oversikt over litteraturen om klimaendringer, som utforsker spesialiseringsmønstre på landnivå. Den andre er en undersøkelse av rollen til forskningsemner i å bestemme suksessraten til ERC-forslag, og dermed utforske rollen til institusjonelle filtre på forskningsemnefordelingen. Til slutt undersøker den tredje studien et utvalg av forskere på tvers av fire disiplinære områder og fire land, og sporer hvordan deres individuelle forskningsagendaer endret seg i løpet av deres karrierer, i forhold til deres akademiske ansiennitet, mobilitet og vitenskapelige samarbeid.

Table of contents

Part I

Introduction

Contemporary societies are increasingly reliant on scientific knowledge. Ranging from large scale challenges like climate change and the COVID-19 pandemic to more specific or local applications, science directly or indirectly permeates most aspects of our lives. The ability of the scientific community to fulfill these important functions depends on a fundamental interplay between exploration – i.e., discovering new knowledge frontiers by exploring new research programmes and topics – and exploitation – i.e., developing incrementally and consolidating the knowledge space on the basis of established results (Kuhn, 1959; March, 1991). The balance between these two modes of inquiry depends on a mix of norms, incentives, and evaluative procedures that are partly the outcome of self-organised, unplanned processes, partly the effect of external regulations with explicitly stated goals. It is thus of paramount importance for the debate on trust and credibility of science to understand how scientists define their own research agenda and how institutional agencies can encapsulate socially shared evaluation standards that help scientists to promote exploration and discoveries, while coordinating collectively to promote reliable knowledge.

The growing importance and pervasiveness of science – coupled with public expenditure and accountability concerns – gave rise to policies linking publication and citational metrics to individual career prospects (Broucker & De Wit, 2015; Osterloh, 2010) and institutional funding (Hicks, 2012). Said policies, progressively radiating from the UK to most other countries, dramatically altered the structure of incentives in academia, contributing – among many other things – to the phenomenon of exponential growth in the number of scientific articles. The well known dictum "publish or perish" – attested at least since the late 20s (Case, 1928) – has become such a deep part of the culture of the academic community that even a card game has adopted this name[1].

Bornmann et al. (2021) estimated the overall growth rate of the number of publications in the last three centuries to be around 4% – with a doubling time of

[1] https://www.nature.com/articles/d41586-024-02511-5

roughly 17 years – and evidenced a steeper 5% annual growth starting around the 50s. In recent decades however, the volume and growth of the scientific literature have started to raise concerns, representing a serious challenge for scholars, practitioners, research institutions and funding agencies (Hanson et al., 2023). With countless articles, papers, and journals being published daily, it is increasingly difficult for any scholar to keep track of this vast sea of information and find relevant and significant contributions in their respective fields. Furthermore, the complex and interdisciplinary nature of today's research makes traditional signals, e.g., disciplinary journals, poorly useful to map the publication landscape, while selective attention by scholars poses serious limitations to research. This context can lead to information overload, where scholars may struggle to identify the most pertinent and reliable sources amidst the abundance of available literature. As a result, valuable research findings may go unnoticed or remain underutilized, impeding the progress of scientific inquiry and innovation.

Furthermore, the exponential growth of scientific publications – coupled with the pressure to publish even more – has implications for peer review and quality control within the academic community (Hanson et al., 2023). With an increasing number of submissions to academic journals and conferences, the peer review process may be overwhelmed, leading to several consequences. Firstly, an increase in the length of the review process, with potential consequences on the academic career, especially of the early career researchers (ECRs). From the point of view of journal editors, this process might entail an increasing difficulty in finding reviewers, which can lead to a mismatch in the expertise among authors and reviewers themselves and to an imbalance in the review workload, with a minority of reviewers carrying out a growing share of the work. This can potentially compromise the rigor and thoroughness of the evaluation, as well as its credibility. This scenario also opens up opportunities for exploitative actors like predatory publishers, able to offer fast a publication process without any rigorous peer review, while financially benefitting from open access priorities.

Taken as a whole, these dynamics are altering the balance of incentives and conditions of the research work. Indeed, a growing body of literature is documenting strategic processes of individual adaptation to this shift.

At the individual level, pressure to publish incentivizes a "corner cutting" approach, and ultimately open scientific misconduct, in a spectrum that ranges from bad statistical practices to outright fabrication of data and results (Edwards & Roy, 2017). The last decade has seen an increased awareness of this issue – with blogs like Retraction Watch[2] or post-publication peer review platforms like PubPeer[3] providing tools to highlight potential cases of misconduct – but more systematic ways of addressing this challenge are still missing.

In addition, these kinds of research misconduct are not the only adaptation strategies. For instance, an anomalous increase in the rate of self-citations has been linked to the introduction of recruitment policies based on individual bibliometric track-records (Baccini & Petrovich, 2023; Peroni et al., 2020; Seeber et al., 2019).

However, shifting incentives can also have a more subtle impact on the scientific community. Most of the quantitative empirical evidence regarding the effects of the changing incentives in academia focuses on strategic choices regarding easily measured outcomes, such as the number of publications, the choice of venues or citational patterns. Conversely, it is much more complicated to draw clear conclusions on the processes of research topics choice. Qualitative research and theoretical elaborations suggest that, for example, the transition from block funding models to competitive grant funding may discourage risky and innovative research (Whitley et al., 2018). Others suggest that the ability to invest time and resources in the development of new lines of research requires a relative stability in terms of career, funds and incentives (Laudel, 2017; Laudel & Bielick, 2018). In short, this line of research suggests that the balance between exploration and exploitation, which is instrumental for scientific progress, has been greatly affected by a context where exploitation seems favoured against exploration. However, the quantitative evidence in this sense is very scattered, also due to the lack of a clear and widely shared methodology for identifying and measuring research topics.

[2] https://retractionwatch.com/

[3] https://pubpeer.com/

If we look at this same problem from the aggregate point of view of the scientific community rather than from the perspective of the individual scientist, we can conceive scientific institutions – both formal and informal – as a system of 'filters'. This perspective has been proposed, for instance, by Smaldino & McElreath (2016). The authors show that – along with other minimal conditions – pressures to publish in order to progress in the academic career can result in what they call *the natural selection of bad science*, i.e. a spread of bad research practices that does not even require a conscious and strategic effort by the individuals involved.

Indeed, many of the steps in academic life have to do with selection processes – e.g. peer-review, funding, recruitment, the awarding of prestige, etc. – which are either deliberately or spontaneously designed to provide some level of assurance on the quality of final research products. In this framework, we know very little regarding the effect of these filters on the distribution of research topics. For instance, in the context of competitive grants, other things being equal, do proposals have a different probability of being funded based on the topics they deal with? Does the likelihood of having a successful career in academia vary based on the topics that make up an individual's research agenda? Naturally, we would agree on the fact that they would do, but as soon as we try to specify the details of how this mechanism is supposed to work and which research topics are more or less successful, things get seriously complicated. For instance, it is not clear how much the distribution of research topics varies before and after the application of a given institutional filter. One would therefore question, for instance: is there a difference in the distribution of the research topics of submitted manuscripts versus peer-reviewed publications? Or among proposals and funded projects in the context of competitive grants?

In summary, while we have a substantial stream of literature covering the effects of these transformations of research work on, for instance, the number of publications or citations, we still know very little about the potential effects on the choice of research topics. This gap is even more concerning, given the fact that – while citation and publication counts might be relevant for the allocation of prestige – the choice of research topics, in the aggregate, determines *the overall development trajectory of science itself.*

The objective of this work is thus to investigate the hidden, and often overlooked, dimension of the research topics. It does so by taking a multilevel approach – i.e., by discussing the macro, meso, and micro-level factors shaping the individual choice of research topics by scientists. An understanding of these factors is also of outmost importance in the attempt to formulate effective policies in the context of science governance. Indeed, missing information on both the current state of things and the counterfactual one – in which certain policies have been put in place altering the research environment – might results in suboptimal or even detrimental adaptive dynamics.

This aim is pursued by exploring the use of new techniques, such as those developed in the context of Natural Language Processing – i.e. text embedding models like SentenceBERT (Reimers & Gurevych, 2019). These tools can allow us to develop a new point of view, to be combined and compared with more traditional citational analysis, and to place texts – i.e. the codified results of the research work – in an abstract space where distances depend on semantic relations.

This dissertation is divided in two parts. The aim of Part I of the dissertation is to provide the theoretical and methodological background for the articles of Part II, by combining insights from sociology, history, and philosophy of science. Part II consists of three chapters presenting independent studies with empirical contributions to the literature.

In particular, the first Chapter of Part I links theoretical motivations from sociology and philosophy of science for the choice of studying science as a social phenomenon, rather than exclusively focusing on the content of scientific theories. It then moves to the history of science perspective, summarising the complex evolution of formal and informal scientific institutions in the shift from modern to contemporary science. Finally, the Chapter closes with a review of empirical studies on the factors affecting the choice of research topics and the evolution of the research agenda of individual scholars.

The second Chapter of Part I focuses instead on the research methods, with a focus on quantitative techniques for the identification of research topics. It opens with an overview of classical citational analysis – recalling part of the debate on the nature and theoretical issues of using citations as data – to then move the discussion

on text-based methods. First, it discusses older and well established techniques like co-occurrence analysis and statistical models; it then introduces the new techniques and opportunities emerged from recent advances in the Machine Learning literature.

Part I closes by summarising the three independent studies reported in Part II, showing the connections among them, and discussing limitations and future developments.

As anticipated, Part II is divided into three Chapters. The first one analyses the global literature on climate change research providing an example of how computational techniques can be used to map the variety of research topics. It then relates this variety to the specific country-level contexts, underlining how different challenges and opportunities can give more or less prominence to certain research areas.

Chapter 2 considers the case of ERC – the single largest research funding program in the EU – to show how research topics are an important predictor of the success of proposals in the context of competitive grant funding, even in a bottom-up funding scheme, where top-down prioritization of specific research interests is explicitly avoided. The study highlights the role scientific institutions play in shaping what type of research – and regarding what phenomena – is carried out.

Finally, Chapter 3 takes a micro-level perspective by tracking the development of the individual research agendas of scholars across four scientific fields in four countries. It considers aggregate paths from individual careers, and analyses patterns of diversification and specialisation. Furthermore, it does so by proposing relational measures of the topic landscape, thus overcoming certain common limitations of previous studies using research topics as discrete entities.

1. Theoretical framework

1.1 Science as a collective effort

Before going further with the mapping of the dramatic transformations the scientific community has undergone during the past decades – and the attempt at identifying the consequences of said transformations – we briefly revise the main motivations to analyse science with the tools of the social sciences. Indeed, while today the social character of the scientific enterprise is often taken for granted, this was not always the case. A long and venerable tradition of scholars analysed science mostly through the lens of *scientific theories* rather than considering the scientists themselves – i.e., those developing theories. Traditional epistemology has historically maintained an individualistic focus, emphasizing the exploration of theories and the development of methods to either verify or falsify them. Rooted in a broadly internalist perspective (Shapin, 1992), these approaches concentrated on the formal procedures through which knowledge is acquired and justified. In this framework, scientific inquiry is often perceived as a detached pursuit, with a certain degree of independence from any social, cultural and/or political influences.

1.1.1 Science as a social institution

The opposite of the internalist view can be seen as instantiated by the so-called *Strong Programme*. This stream of sociological thought rejected the conventional understanding that "facts" are universally discovered and instead asserted that they are socially *constructed* through a process of negotiation (Latour & Woolgar, 1986). As a consequence, comprehending and elucidating scientists' beliefs, as well as the development of scientific theories, requires to examine various social factors (Bloor, 1984, 1991).

While the radically externalist position of the Strong Programme might be considered extreme, other influential scholars have underscored the significance of social factors – alongside "internal" factors, that relate to the content of scientific knowledge itself – in comprehending the dynamics of scientific development.

Kuhn (1962) introduced the concept of paradigm shifts to explain how scientific communities undergo radical changes in their fundamental beliefs and practices. He argued that scientific progress is not a linear, cumulative process, but rather a series of revolutionary shifts in which dominant paradigms are replaced by new ones. Crucially, Kuhn emphasized the role of social and cultural factors in creating the conditions for these shifts.

In some of the seminal writings of the sociology of science, Merton (1979) highlighted the importance of social factors in explaining how scientific communities function and how scientific knowledge is produced, considering science as a *social institution* rather than as a *type of knowledge*. In addition to his well-known exploration of the scientific ethos – with the introduction of the Mertonian norms – he scrutinized the significance of peer recognition as a collective tool for coordinating and fostering scientific endeavours. Simultaneously, he acknowledged the potential drawbacks of this system, exemplified by phenomena like the 'Matthew effect' (Merton, 1968).

> Seen in composite […] the reward system of science reinforces and perpetuates the institutional emphasis upon originality. It is in this specific sense that originality can be said to be a major institutional goal of modern science, at times the paramount one, and recognition for originality a derived, but often as heavily emphasized, goal. In the organized competition to contribute to man's scientific knowledge, the race is to the swift, to him who gets there first with his contribution in hand. (Merton, 1979)

1.1.2 Science as an evolutionary process

Another theoretical approach to examine scientific practice is to treat it as a combinatorial process, through which the scientific community explores (the space of) possible solutions to a problem, in search of optimal ones – or at least the ones that are better than those provisionally adopted up to a certain time. Once we assume this point of view, it is natural to wonder about the optimal distribution of effort across theoretical alternatives: i.e., how can we allocate the minimum effort while maximising the probability of getting the right answer?

Surprisingly enough, this question arose very early on in the work of Peirce (1879). This early contribution started a stream of literature in philosophy of science trying to deal with the organization of research and the division of cognitive labour, rather than with specific scientific theories. In his manuscript, Peirce develops a model for the optimal allocation of resources among two or more alternative research efforts, assuming decreasing marginal returns. While this model provides a description of the optimal distribution, it does not consider any possible mechanisms to coordinate individuals in such a way to achieve said distribution. Rather, in the closing Peirce underlines that his model holds only when truth is the final aim of the investigation:

> It is to be remarked that the theory here given rests on the supposition that the object of the investigation is the ascertainment of truth. When an investigation is made for the purpose of attaining personal distinction, the economics of the problem are entirely different. But that seems to be well enough understood by those engaged in that sort of investigation. (Peirce, 1879)

On the contrary, by using a seemingly abstract model, Kitcher (1990) argued – in a Mertonian vein – that self-interested scientists competing for their own recognition can, given some conditions, attain a distribution of the effort yielding in the aggregate better probabilities of success with respect to a community of researchers purely interested in knowledge. This result is driven by the fact that even though a certain research programme could have lower chances of success, it might be rational for an individual to pursue it if they have higher chances of being rewarded due to the overcrowding of the alternatives.

The same reasoning is followed by Strevens (2003), who explicitly linked this mechanism to the *priority rule* in science and argued for the existence of a deep link between two aspects of the scientific endeavour that he named *winner-confers-all* and *winner-takes-all*. The first one refers to the idea that the first researcher or group to find the right answer to a scientific puzzle provides the society with all the benefits of a solution; the second implies that the rewards for the solution are bestowed to said researcher or group.

Taking their move from these abstract models, some authors started to rely on computational modeling – in particular Agent Based Modeling – to relax some of the assumptions entailed by previous works. This resulted in two main branches of models. The first one is more directly concerned with the structure of the scientific community and the way local interaction among researchers shape their convergence to a consensus (Zollman, 2010). The second instead aims at explicitly representing the process of combinatorial search in the space of possible research approaches, the so called *epistemic landscape* (Weisberg & Muldoon, 2009); it also eventually incorporated a more direct representation of the role played by institutions like funding programmes (Avin, 2019).

1.1.3 In search of equilibria

The previous Sections introduced two perspectives on scientific research that may seem drastically different from each other. On the one hand, the tradition of sociology of science views research as a social institution, thus emphasising on the role of individuals in a context mediated by disciplinary ethos, prestige and a variety of other social factors. On the other hand, inspired by computational models of biological evolution, is a set of idealized models that treat the research effort as a combinatorial search problem.

Although these two approaches may seem very distant between each other, they allow us to formulate important questions regarding the internal mechanisms of the scientific community, with particular emphasis on the relationship between the functioning of scientific institutions and their outcomes. As Kitcher (1990) points out, there is no philosopher-monarch – or any other type of centralized institution – capable of directly planning scientific development and allocating research work optimally. Instead, a multitude of factors – including disciplinary norms, editorial practices, recruitment rules, funding schemes, etc. – contribute to guiding the work of relatively autonomous and independent scholars by setting common rules and incentives. These same factors, in turn, are not centrally planned by a single actor, but are the – provisional – result of a long and gradual process of idiosyncratic development:

> Like other institutions, the institution of science has developed an elaborate system for allocating rewards to those who variously live up to

> its norms. Of course, this was not always so. The evolution of this system has been the work of centuries, and it will of course never be finished. (Merton, 1979, p. 297)

1.2 Towards science as an industry: a brief on how we got here

For most of its history, scientific research has been neither a profession nor a job; in most cases, it was self-funded or made possible by some kind of patronage relationship and quite often was a collateral activity of some other profession – like the medical one. Individuals were often able to span wide areas of expertise, managing to contribute meaningfully to the advancement of several of them.

However, by the start of the 19th century, the times of the intellectuals as polymaths was over. Rudwick (1985) characterizes this period as the time of the *gentlemanly specialists*. This term underlines the link with the past social organisation of science – i.e. an amateur activity for the upper classes, able to independently bear the related costs – as well as the increased disciplinary specialization, resulting in the recognition of specific identities – e.g. chemist, geologist, etc. The consolidation of these identities – paired with the desire for and debate around peer recognition – gave rise to an influx of scientific societies: the «Arenas of Gentlemanly Debate» (Rudwick, 1985).

One of the most pervasive institutions of contemporary science – the scientific journal – emerged in this environment often as outlet of said scientific societies, with the earliest examples being in the 17th century. It would, however, be a mistake to conflate the historical format and functions of the journals with the current one (Fyfe et al., 2022). First, scientific publishing was not very formalized, and editors and editorial boards were taking most of the decisions, with external expert judgment – i.e. peer-review – being at least uncommon (Baldwin, 2018; Fyfe et al., 2020). The work of analysing and vetting theories and findings was mainly carried out through public debate among societies' fellows. Second, the literary *genre* of the scientific paper itself was still not strictly codified, and journals used to publish a mix of accounts of new research accounts, book reviews, news, commentaries, educational material and so on, without a clear distinction between the different types of content. Discussing the historical case of the

Philosophical Transactions, Fyfe et al. (2022) notice that it was «[…] as much the ancestor of the abstract or review journal, or of the scientific news and book reviews now found in the front half of Nature, as it is of the peer-reviewed research journal».

Nevertheless, the 19th and early 20th centuries were also a period of slow transition towards science as a profession. A driver of this dynamic was the increasing propensity of the states to fund universities and other research institutions, establishing state-funded professional figures expected to dedicate part of their working time to research. The nature of this transition has been gradual, to the point that – as Shapin (2008) noted:

> By the end of the nineteenth century, the transition from science as a calling to science as a job had at most just begun. […] So at the beginning of the twentieth century the identity of the scientist was radically unstable. To be a scientist was still something of a calling but it was becoming something of a job. (Shapin, 2008, pp. 44–46)

Indeed, the building up of this new professional identity was not entirely uncontroversial, as shown by the fact that the debate around the adoption of *scientist* as an umbrella term for the – at the time – so called *men of science* or *scientific workers* was ongoing on the pages of *Nature* as late as the 1920s (Baldwin, 2015).

The Second World War represented a dramatic turn. The increased availability of resources and the unprecedented scale of research projects had a deep impact on the experience of scientists and the whole public:

> The mobilization of American science during the Second World War – especially in the Manhattan Project and in the construction of radar, but spreading across much of the scientific landscape – propelled a generation of academic scientists into a world that was largely unfamiliar to them: the experience of large-scale organization; of teamwork; of interdisciplinary project-oriented research; of unlimited resources and severely limited time; of close contact with the sorts of people – especially the military and the commercial worlds – they had not known

> much about and, after the end of the war and the beginning of the Cold War, the experience – for some of them – of political power. (Shapin, 2008, pp. 64–65)

The mobilization of science was by no means limited to the US, but the scale and visibility of the phenomenon in the American context makes its dynamics particularly salient. It is not a coincidence that we can observe an explicit strategy of post-war transition in the report to the US President by Vannevar Bush (1945) – at the time Director of the Office of Scientific Research and Development. In this report, he argued for the creation of a central federal agency devoted to funding basic research – later realised in the NSF – and discussed the challenges of the reconversion of research from the war effort. This marked the rise of the so-called *big science* (Price, 1963), i.e., a mode of research based on large-scale collaborations of individuals, often across several institutions, typically sharing very expensive equipment.

The increased availability of resources via both direct and grant funding, however, came at the cost of an increased formalisation and bureaucratization. During the Cold War, both journals and funding agencies – which earlier relied heavily on editor and/or internal staff – began to increasingly rely on peer-review to assess manuscripts and proposals. A practice that started mainly as a way to ease the work of journal editors in their decisions and to allow them to keep up with the increased complexity of submissions – and thus as a way to make the publication process more *efficient* – gradually came to be perceived as vital to the research process itself, as a badge of quality – thus becoming a signal of *effectiveness* of the selection process (Baldwin, 2018).

The same period also saw the rise of a new way of keeping track of the scientific literature: while discipline-specific indexes of publications and citations were already present, only during the late ’50s of the 20th century the systematic collection of these metadata started to be practiced – as a commercial activity – at a large scale and irrespectively of disciplinary boundaries, with the birth of the *Science Citation Index* (Garfield, 1955).

The research and higher education systems are deeply intertwined. After the war the increased funding for research and the increased demand for access to tertiary

education fueled a consistent and steady expansion of both systems that can be observed in historical data about funding, personnel, and enrollment (Snyder, 1993).

The trend of this expansion eventually slowed down, and the parallel rise of the paradigm of *New Public Management* increased the demand for control and accountability on how resources were utilised (Ferlie et al., 1996; Power, 1999). While this shift had various consequences in the wider context of higher education, which are outside the scope of this dissertation, it mainly caused a gradual – and it might be argued partial – replacement of incentives towards a market-like system within academia (Broucker & De Wit, 2015), for which careers and funding availability started to greatly depend on performance, usually measured by bibliometric indicators derived from publication and citations counts (Osterloh, 2010). In many countries, performance evaluation followed different systems – based on peer-review, quantitative indicators, or a mix of both – and this also informed the allocation of national funding to universities and other research institutions (Hicks, 2012).

This – albeit brief and not comprehensive – historical overview is key to set the ground to the overall argument of this work. First, tools and practices related to the academic community gradually evolved over time in an idiosyncratic way – depending on particular social and historical contexts – being standardized and incorporated into formalized procedures by various science stakeholders. In this process, these frameworks and procedures often changed their core functions. For example, peer-review gradually shifted from being an optional support for journal editors to a seal of «*actual science*» (Baldwin, 2018); the systematic collection of citational metadata shifted from a research tool detecting «*association-of-ideas*» (Garfield, 1955) to an instrument of assessment and allocation of resources (Hicks, 2012). Second, even the figure of the scholar underwent a process of parallel specialisation and professionalisation, which mostly reflected a shift in the incentives and norms of the scientific community. This cultural transition has been well captured during its unfolding by Merton (1979, p. 298):

> The echo of these complaints [about unrewarded efforts] still reverberates in the halls of universities and scientific societies, but chiefly with regard

to material rather than honorific rewards. With the growth and professionalization of science, the system of honorific rewards has become diversely elaborated, and apparently at an accelerated rate.

1.3 The choice of research topics

As we have seen, the transition towards science as an industry has implied many transformations in the inner machinery of the scientific community. Among those, the arguably single most important shift is the new centrality of standardised quantitative measures. These measures have started to be relevant both from a formal point of view – e.g. in hiring procedures, national assessment exercises, etc. – and from an informal one; a deep transformation has taken place in the culture of the academic community – it is, for example, common practice to look at the number of publications, citations, or the H-index of a person rather than (or prior to) reading their papers.

Researchers have to find a balance between the exploration of new topics, theories and methods and the exploitation of the already acquired knowledge and skills (Kuhn, 1959; March, 1991). However, the transformations we described – and particularly the increased pressure to publish – are likely to affect this balance.

As previously discussed, the aim of this work is exactly to try and identify the factors affecting such a balance, since the aggregate of individual research choices determines the overall direction of development of science. In the attempt to do so, this Section reviews previous research on the choice of research topics and the formation of individual research agendas.

This literature represents a specific fraction of a larger body of research dealing with science and the scientific community. Even though we focus on just part of this production, the attempt to summarise the overall results of these studies is a challenging one, in particular due to the heterogeneity of approaches. Most fields of science devote part of their attention to reflecting upon their own mechanisms, and thus the study of the scientific community – despite some recent unification efforts (Fortunato et al., 2018) – is theoretically and methodologically fragmented.

As such, multiple lenses would be appropriate to attempt a systematization of the relevant results. Here, we will follow the level of description of the phenomena under investigation, as well as the research methods, in order to organize the discourse. We first discuss research with a broad focus, often investigating field-level transformations. We will then dive deeper into the micro-level, by reviewing the qualitative literature on factors that operate along the scientific career. Finally, we will discuss examples of quantitative studies aimed at estimating the effect of specific institutional policies – i.e., meso-level interventions. It must be noted that this organization of this Section does not imply any order of causation, since these levels interact with each other in an intricate and complex way.

1.3.1 Studies on field-level dynamics

Some of the contributions in this area had a more general and exploratory focus. For example, Huang et al. (2022) analyzed data on 25 million papers in the field of computer science, and found that productive and impactful scientists tend to follow the research frontier and diversify their research quite often, while comparatively focusing less on mature topics. Furthermore, they highlighted the tendency of successful scientists to explore new topics early in their career while being more conservative in later stages of their career.

The area of physics attracted much attention especially due to the availability of a relatively standardized classification system of research topics: the Physics and Astronomy Classification Scheme (PACS). For example, Wei et al. (2013) were able to exploit this classification to show a mechanism of preferential attachment of new papers towards larger and growing fields, thus uncovering the propensity of researchers to trace “hot” topics.

By studying publications from the American Physical Society (APS) by over 100,000 authors, Aleta et al. (2019) found that researchers change research topics along their career, but following a strategy that could be described as ‘cautious exploration’: they remain in the same broad area of expertise and move between relatively closely related topics. Additionally, they found a higher level of exploitation in the research agenda of scholars with a shorter career, suggesting the need for time and stability in order to develop an independent research path.

Again studying physicists, Jia et al. (2017) found that changes in research interests – comparing the start and the end of career – follow an exponential distribution, and that the exploration of topics tends to be concentrated in batches rather than being evenly spread in time. In this study, the authors built a random walk model able to account for the three fundamental features for the patterns of development of the individual research agendas: i.e. heterogeneity, recency, and subject proximity.

Building on this contribution, Zeng et al. (2019) underlined how researchers have a narrow distribution of research topics on which they specialize. They found that on average the three main research topics cover more than 70% of a scholar's scientific production. Over time, however, the frequency of switching among these topics has been increasing. They also underlined how this increased probability of switching research topics is correlated with lower productivity in the early career and a higher one later on, while high citation rates are correlated with low topic switching at all career stages.

In an outstanding contribution, Foster et al. (2015) exploited the availability of metadata regarding chemical entities in a *corpus* of over 6 million publications from Medline. They represented the state of chemical knowledge as the time-varying network of co-occurrence among chemicals and proposed five possible strategies to account for scholar's behavior. These included three innovative strategies: i.e., *jump*, *new consolidation*, and *new bridge*; respectively involving a chemical just recently introduced to the network, adding a new link among chemicals in the same cluster, and adding a new link between chemicals in different clusters. Alongside these there were two traditional strategies: i.e., *repeat consolidation* and *repeat bridge*; respectively adding on links that were already present between chemicals in the same knowledge cluster or in different ones. The authors reported a remarkable stability in the share of strategies over time. They also found a very low level of innovative strategies despite a growing potential for innovation given by the increasing number of potentially interlinked chemicals. This pattern is consistent with innovation opportunities being strategically neglected when scholars are trying to maximise citation counts and their bibliometric record. However, if scholars were just citation count maximisers, we would observe even lower levels of innovation. The authors eventually related this

finding with the hypothesis that the potential for prestige and awards – i.e., long-term reputational incentives – might be the driver of higher-than-expected levels of innovation.

1.3.2 Studies on individual careers

The studies reviewed above are good examples of a stream of research with broad goals and based on very large volumes of literature. Interesting insights have also been provided by studies considering smaller samples but exploring more in depth the details of individual careers, often by mixing bibliometric field delineation techniques with qualitative methods like interviews. A structured description of this approach – that allow to track topical changes in the scientific contribution by scholars alongside their biography – is provided by Gläser & Laudel (2015).

For example, Laudel & Gläser (2008) studied a sample of Australian scholars with a mixed methodology employing biographic interviews assisted with individual-level bibliometric information. They argue that we can identify three logically separate careers that not necessarily align with each other: the organisational one is the sequence of different formal positions and contracts linked to a person, the cognitive career represents their contribution to scientific knowledge, while the community career can be understood as a sequence of informal roles – i.e. apprentice, colleague, master and elite scholar. Their study points to the idea that a misalignment of these careers can be an issue especially for early career researchers having to transition from the role of apprentice to developing an independent research path (Laudel & Bielick, 2018). The authors also argue that this development of an autonomous cognitive career is moving towards the postdoctoral phase, and that the pressure resulting from lack of research time, increased levels of competitiveness and fixed-term employments might push ECRs to adapt their research focus to the preferences of senior colleagues and funding agencies, undermining their independence.

The same framework was later extended to a larger sample spanning two fields – molecular biology and history – and countries with different career systems: a chair system (Germany) and two tenure systems, one with strong hierarchies (the Netherlands) and one with flat hierarchies (Australia) (Laudel, 2017). In this case, different formal career systems seemed to produce similar results in terms of the

evolution of individual cognitive careers. The main source of variation is rather the discipline, with strongly codified fields – i.e. with high specialisation and high levels of collaboration – displaying a later emergence of independent lines of research, as opposed to weakly codified fields where individual autonomy rises earlier. The authors claim that the formal positions of the ECRs are a poor predictor of their actual work conditions, and thus are not informative regarding their pattern of research topic choice. What really matters is rather the *protected space* (Whitley, 2000): i.e., «the discretion over the use of needed resources, including their own efforts, to pursue particular problems and approaches before having to produce publishable and collectively valued results. It incorporates authority over the choice and formulation of which topics to study, how to do so, and how to obtain and manage resources» (Whitley & Gläser, 2014). With respect to this dimension, formal relations act merely as a constraint or enabling factor – are thus necessary but not sufficient. The authors, finally, highlight how this perception of stability is limited to a small group of elite scientists, while the majority of ECRs experience limitations to their protected space.

Whitley et al. (2018) again compared researchers from different countries (Germany, Netherlands, Sweden and Switzerland) under the theoretical lens of the protected space, analysing the impact of the increased prominence of grant funding as opposed to block funding. This contribution is though particularly interesting because of a specific focus on the development of three innovation that were key in their respective fields, i.e. i) the Bose-Einstein condensates in physics, ii) the emergence of Evolutionary Developmental Biology, and iii) the introduction of International Large-Scale Student Assessments in educational research. The authors notice how and increased number of funding bodies does not matter *per se*, greater diversity or convergence to mainstream research approaches depend more directly on the variety of goals and objectives pursued by said funding bodies; thus having more funding bodies might, counterintuitively, increase specialisation rather than foster diversification, if they all use the same evaluation criteria. This effect is likely stronger in times of reduced public block funding, because grants become comparatively more important in terms of funding opportunities. Another factor to be considered is the standardisation of project duration and resource packages. Fields and research approaches are highly diverse

in terms of the possibility to employ division of labour and unpack problems into smaller units. An increased standardisation on this front can determine difficulties for emergent approaches that require longer times to develop. Especially in highly technical disciplines lower block funding also means difficulties in maintaining local research infrastructure and personnel, that are essential especially for long-term research projects. Finally, the authors emphasize the inability of policy driven initiatives to stimulate the adoption of new approaches unless these offer clear scientific advantages over the current ones.

Horlings & Gurney (2013) attempted a similar analysis – on a sample of condensed matter physicists – by combining a bibliometric analysis with detailed information about individual careers recoverd from CVs. They evidence how a scholar's work can be seen as a set of parallel research paths, and observe how the research immediately following the PhD is usually unconnected to the preceding one. The postdoctoral phase can thus be seen as an expansive and diversifying one, while later in the career the research trails stabilise.

1.3.3 Studies on policy impact

While qualitative inquiries give us a more detailed picture of the evolution of individual research agendas along individuals' careers, quantitatively oriented research has often focused on trying to estimate the effect of specific policies.

For instance, Azoulay et al. (2011) compared two radically different funding programmes in the life sciences, namely grants from the Howard Hughes Medical Institute (HHMI) and the National Institutes of Health (NIH). HHMI allows higher leeway to researchers since it bestows five-year grants – with high chances of renewal –, allows reallocation of resources across projects and is tolerant of early failures, additionally it provides in-depth feedback during the evaluation phase. NIH grants, instead, follow a more standardised cycle of three years with little tolerance of failure when renewal has to be decided, with a focus on clearly defined deliverables and providing limited feedback. The authors compared the career of scholars awarded with grants from these two institutions not only from the point of view of the impact of their work, but also tracking the evolution of their agendas by looking at the novelty and breadth of the keywords they employed to describe it. They conclude that investigators funded by HHMI are more likely to get

involved in new research topics, modify their agenda with a higher degree of diversification and reach a more diverse community by being cited from a broader set of journals.

Analysing Italian scholars across all fields of science, Abramo et al. (2024) tried to estimate the impact of the introduction of the national scientific accreditation (ASN) on the specialisation-diversification trade-off. Since 2012 the ASN is required in order to apply for professorship positions and is awarded by disciplinary-based committees of randomly chosen full professors. Additionally, in some fields the evaluation is also based on the candidates meeting quantitative thresholds of numbers of publications and citations. The authors tracked the distribution of Web of Science (WoS) Subject Categories (SCs) of each scholar's publication over time as a measure of their level of specialisation. They found an overall trend of decreasing specialisation, i.e. research agenda diversity has been increasing over time. They also report specialisation to be positively associated with productivity and gender – with female scholars having more focused research agendas. Collaboration also clearly plays a relevant role in the process, with scholars engaging in more co-authorships having a more diverse agenda, as well as researchers working in cities with relatively few colleagues in their same field of study. As for the impact of the ASN, the authors report the puzzling finding of an increased diversification despite the direction of the individual level incentives and identify an even stronger effect among individuals that would be theoretically more sensitive to said incentives.

Groen-Xu et al. (2023) studied the impact of the Research Excellence Framework (REF), a country-scale periodic evaluation exercise in the United Kingdom. By systematically collecting data on the publications submitted for the evaluation, the authors show evidence of a recurring pattern consistent with strategic choices of projects and publication venues. In fact, they report a spike in the number of publications close to the evaluation deadline. Furthermore, they found articles published close to the deadline to receive less citations, be published in lower-impact journals, have more authors, and have higher probability of retraction.

Considering again the UK context, Madsen & Nielsen (2024) assessed the effectiveness of thematic funding schemes in steering the research interests of

scholars. They longitudinally followed the research agenda of individuals and used community detection on the citational network to identify research topics. By using statistical matching, they are able to find a short term shift in the interests of researchers funded through thematic schemes. However, the effect fades away over time, with scholars going back to their original interests.

By combining bibliometric data with a survey of Japanese authors, Wang et al. (2018) found that projects funded by competitive grants have higher novelty compared to those funded by internal block funds. However, they report this effect to be limited to scholars with high status – i.e., senior male scientists – while internal block funding to be associated with higher novelty for junior and female scholars.

1.3.4 Summary of the literature

Overall, we might conclude that the literature dealing with the choice of research topics is quite fragmented. First, we discussed broadly focused studies that attempted to map field-level dynamics. These works have the advantage of being comprehensive and are able to reveal general patterns of development, however high-level patterns can be compatible with a plurality of micro level mechanisms that require a focus on individuals in order to be disentangled. Furthermore, these studies mostly rely on pre-existing disciplinary labels or entity tagging, which might limit their ability to capture bottom-up and temporal developments. Second, we considered studies of individual scientific careers that mostly employed qualitative or mixed methods. These works provide a compelling theoretical framework, linking the intellectual contributions with the institutional career and the social recognition of the authors. However, they suffer from the inability to systematically generalize to larger samples and contexts. Finally, we considered studies of the impact from specific policies. These contributions can provide a theoretically informed quantification of the effect of such policies but are – by definition – limited in their scope. In addition, they often rely on proxy measures of their outcome variables that are arguably limited, like citation or unique word counts.

In light of these contributions, one possible way forward in the understanding of the dynamics of research topics is to both i) develop new methods to track and

systematically measure research topics and ii) to combine the breadth allowed by quantitative studies of scientific production with the depth of individual-based research designs. The ambition of the present works is thus to contribute to both aims.

1.4 Concluding remarks

Along this first Chapter, we have argued that science is a social enterprise, and scholars are embedded in a system of norms and incentives. This system is the result of both an historical process of evolution and of policies with explicitly stated goals. All these factors contribute to shape the protected space – i.e., the level of autonomy – within which individual scholars develop their research agendas. In particular – as we discussed – the last few decades have seen the introduction of strong incentives towards higher volumes of scientific publication. These incentives had both a direct impact as well as an indirect one, on the culture of the academic community.

There is quite good evidence that scholars adapted to this new environment with strategies related to their publication and citation behavior. However, evidence regarding a possible shift in the balance of exploration and exploitation in much scarcer. This is likely due to also the difficulty in quantifying such a balance, through a systematic tracking of individual research agendas. We can have an intuitive grasp of how a research agenda might look like – for instance as a succession of research topics, possibly related among each other, and mixed along time. However, as often happens, the attempt at formally identifying and measuring this latent dimension is not trivial.

In particular, if we start by operationalizing exploration as the variety or diversity of scientific production, we end up with the problem of quantifying relatedness or similarity among scientific publications.

This issue is a foundational one in the field of scientometrics, where traditional approaches have primarily relied on citation analysis. However, recent advancements in statistics and machine learning have introduced innovative methods that are capable of analyzing textual data. These methods show

considerable promise in capturing the substantive content of publications, thereby offering a more nuanced understanding of their relatedness. The next chapter will review the main methodological approaches in addressing the relatedness measurement and discuss the respective opportunities and limitations.

2. Methodology

2.1 The identification of research topics

An established point within sociology of science is that the space in which scientific research unfolds is not the one of disciplines. Those are instead relevant in more institutionalised activities like teaching or recruiting. Scientific research displays high levels of specialisation, and within individual disciplines – or often across multiple disciplines – we can find niches related to particular theories, methods and phenomena. This rather more elusive and difficult to delineate dimension involves research specialties (Chubin, 1976). This section aims to provide a general overlook of the development of theory and methods for the definition and quantitative measurement of this latent dimension, jointly combining sociology of science and bibliometrics.

2.1.1 Citational analysis

The idea of systematically collecting and classifying the body of human knowledge has deep historical roots. It was perhaps with the same kind of aspiration that in the 1950s Eugene Garfield imagined the creation of a citation index, which had to be independent of any disciplinary boundaries and congruent with the evolving structure of the scientific literature (Garfield, 1955). In his vision – later embodied in the founding of the Science Citation Index – the systematic collection of citation data would have provided a better entry point to the exploration of the literature, facilitated the communication among scholars, curbed the uncritical use of fraudulent or inaccurate data and sources, and enabled the historical analysis of fields of research; all through the tracing of *association-of-ideas*, the concrete form of which could be identified as citations.

The idea that citations allow us to track the gradual process of knowledge development – the very base of the entire field of bibliometrics – is summarized in a quite evocative way by Cronin (1981):

> Metaphorically speaking, citations are frozen footprints on the landscape of scholarly achievement; footprints which bear witness to the passage of

> ideas. From footprints it is possible to deduce direction; from the configuration and depth of the imprints it should be possible to construct a picture of those who have passed by, whilst the distribution and variety furnish clues as to whether the advance was orderly and purposive.

The direct citation relationship, however, encodes relational information between publications in a very sparse manner. This led very early to the development of derivative measures based on co-occurrence that helped to capture similarity between publications. The two best known and most widely used measures of this type are bibliographic coupling (Kessler, 1963a, 1963b) and co-citation (Marshakova, 1973; Small, 1973).

The former is based on counting shared references between two publications. The idea is that similar publications have an overlap in the sources used. The similarity relationship thus estimated solely depends on the reference list of the two publications considered, consequently, this measure is invariant over time.

The second measure is based on the count of citations shared by the two considered publications: i.e. by how many other publications cite both one and the other. It is clear that in this case the similarity relationship depends on the total body of publications considered and may vary as additional publications are added over time.

Of course, there are variants of both measures adapted to specific use cases through, for example, context-appropriate normalization (Eck & Waltman, 2009). From these basic relationships between individual publications it is then possible to derive relationships referring to aggregate levels of all kinds: for instance between authors (White & Griffith, 1981), journals (McCain, 1991), etc.

This wealth of different citational relationships is meant to capture the semantic similarity – or relatedness – among the objects of analysis, usually documents. Quite often though the end goal of identifying research topics is a discrete categorisation of entities into groups.

This is in general attained by either using a clustering or a community detection algorithm. In the former case, each observation is considered to be represented by a vector of features, and the procedure aims at grouping together entities that are

similar with respect to said features. In the latter, entities are assumed to be linked to each other in a network structure, and the algorithm is used to create a classification that predominantly preserves links among dense areas of the network. There is, however, a large number of different algorithms that could be applied in each case. This variety, combined with the one of relatedness measures and preprocessing procedures, represents one of the biggest challenges in a topic identification endeavour.

Systematic comparisons of the possible alternatives are still uncommon – even though a dedicated special issue of *Scientometrics* provides a virtuous example (Gläser et al., 2017) – and the task is even more complex given the complete lack of ground truth – i.e., in no case we have a "true" classification to compare against the results of each method. Far from being just a technical implementation detail, this problem entails the very definition of "research topic", and requires both a more explicit theoretical grounding and to «avoid normative myopia that recommends the 'best' approach» (Gläser et al., 2017).

2.1.2 Assumptions of citational analysis

Citational approaches have certain – often implicit – underlying assumptions about the function and meaning of citations. The main ones can be summarized as follows:

1. Citations are the currency of the scientific community and are used to recognize the importance of previous contributions (citation counts are thus a proxy – however limited – for quality or peer recognition).
2. Scientific publications can in some way represent the ideas they contain.
3. The body of work cited in a publication is informative about its content.

Albeit coherent, these assumptions were subject to criticism. From the early days of quantitative citation analysis, there was a debate surrounding its limitations and the need for a stronger theoretical grounding, i.e., an explicit *theory of citing* rather than an implicit one (Mulkay, 1974).

Gilbert (1977) for example refused the idea that scientific articles would be neutral reports of the research work carried out. He suggested that they would instead be fundamentally rhetorical devices aimed at persuading the audience. Under this

lens, citations are not simply a way to resolve priority claims or recognize contributions. Rather, they are employed to strengthen the authors' argument by referring to authoritative papers, signal allegiance to a specific part of the scientific community or underline the novelty of the contribution; all of this in ways that might not directly reflect the credibility or even relevance of the cited literature.

> Hence, authors preparing papers will tend to cite the 'important and correct' papers, may cite 'erroneous' papers in order to challenge them and will avoid citing the 'trivial', and 'irrelevant' ones. Indeed, respected papers may be cited in order to shine in their reflected glory even if they do not seem closely related to the substantive content of the report. (Gilbert, 1977)

Another example of such criticism is a review of the seven main problems of citational analysis, which quite bleakly concluded as follows: «whether or not, and in what ways, citations can be used as data remains unclear» (MacRoberts & MacRoberts, 1989).

Cronin (1984) provided a wide ranging discussion of the arguments on the two – sometimes partially overlapping – fronts of the critics and practitioners of citational analysis. He proposed a partial synthesis and called for a higher level of awareness of the theoretical implications of the methodology, while recognizing that «there is no single, all-embracing theory of citation. […] the appropriateness of the various analytical perspectives is situationally or contextually dependent».

> Kuhn has argued that there are two sorts of history of science, the internalist and the externalist, and it may be that the same can loosely be said to apply to citation. That is to say: most citation analyses have been internalist rather than externalist in character, in that they have concentrated on quantities and frequency distributions rather than on the contexts within which, and processes by which, authors employ citations. Citation analysis requires the sort of firm and fruitful contact between these two approaches which Kuhn feels is needed in the history of science. (Cronin, 1984)

Several decades later, Cronin's call for a theoretically grounded synthesis of the two perspectives is still open, as no single comprehensive theory of citing is explicitly accepted by a majority of the scholars involved in scientometric research. Research *practice*, however, has implicitly recognised some of the critiques to naive citational analysis and has incorporated a more complex account of citational behaviour. It is the case, for instance, of research on strategic adaptation to incentives in the form of metric gaming and self-citation (e.g., Baccini & Petrovich, 2023; Peroni et al., 2020; Seeber et al., 2019).

2.1.3 Text-based approaches

Distancing themselves from the problems and limitations of citational analysis, Callon et al. (1983) introduced a new measure of relatedness among documents based on the co-occurrence of words, rather than references. The promise of this approach was to have a representation more anchored in the actual content of the documents, thus able to map the dynamics of the scientific process in more fine-grained and nuanced ways. Of course, this approach was also not exempt from criticism, pointing in particular to the shift in use and meaning of words over time and across contexts (Leydesdorff, 1997).

While initially meant to rely on a theoretically-based dictionary, text-based methods rapidly started to incorporate techniques that allow to build the word dictionary directly from the corpus of selected documents. This is typically done by first "cleaning" the data by means of either stemming or lemmatization – in order to, for example, remove the distinction between singular and plural forms or different tenses of the verbs –, the removing the so-called "stop words" – e.g. conjunctions, adverbs, etc –, and finally by applying a simple threshold function to remove words that occur less than a certain amount of times. In the end, a data matrix is built by counting the co-occurrence of words into the documents. A slightly more sophisticated technique – widely applied also in information retrieval tasks – is the term frequency–inverse document frequency weighting (TF-IDF) (Sparck Jones, 1972). This weighting scheme aims at giving more relevance to words that are both sufficiently frequent in the overall corpus *and* sufficiently sparse to be distinctive of the documents they occur in. So, for example, a word occurring exclusively in a single document – no matter how many

times – will receive a low weight, as well as a word occurring in almost all of them, while higher weights will be given to words that characterize a specific group of documents.

Text-based approaches gained momentum with the diffusion of a new technique called *topic modeling*, based on LDA (Latent Dirichlet Allocation), a statistical model explicitly targeted at dealing with collection of texts considered as sets of words (Blei, 2003). Beyond the observed dimensions of documents and words, LDA explicitly introduces the latent dimension of *topics*. Each document is then modelled as a mixture of topics, while topics themselves are modelled as mixtures of words. Given that each word has a different probability of occurring within a certain topic, word frequencies in documents provide information about the topical content of the documents. While the output of the model is technically, for each document, a vector of weights – that represent the strength of association between said document and each topic – in practice quite often the result is discretized by assigning each document to the topic with the strongest weight.

Following the success of LDA, several variations were proposed in order to address limitations given by the simplifying assumptions at the basis of the starting model. In particular, Correlated Topic Model (Blei & Lafferty, 2007) aimed at relaxing the assumption of independence among topics, while Structural Topic Model (Roberts et al., 2014) was introduced to allow the distribution of topics to vary according to document-level covariates in a regression-like fashion – thus enabling for example to account for the difference between types of documents, or between periods of time.

While these methods represented a great improvement over previous approaches, they also display two main limitations. First, the original definition of the model, as well as the subsequent ones, require the analyst to *a priori* – ideally based on domain expertise – provide the number of topics to be identified. Although some heuristics have been proposed to simplify this task, this is especially challenging since in practice text-modeling techniques are often used in an exploratory way, which by definition implies little prior information. The second main limitation of LDA and its variants is the reliance on the *bag-of-words* representation of the documents: by considering document uniquely as frequency distributions over a

vocabulary, all the contextual information encoded in the structure of the sentences and the order of words is lost.

2.2 New perspectives: text embeddings models

Another possible way to capture semantic similarity relationships between texts comes from Machine Learning, which has recently developed cutting-edge techniques in the field of Natural Language Processing (NLP).

In this context the first radical change compared to the approaches discussed in the previous Section is in how words are treated. While *bag-of-words* models consider each word as a unit in its own right – completely independent of any other – *text embedding* models represent each word as a vector – or, equivalently, a point – in a vector space. This fundamental difference allows embedding-based models to capture the semantic similarity between words, translating this relationship into the formal one of spatial proximity. Concretely, the first example of this kind of model – word2vec, introduced by Mikolov et al. (2013) – was implemented as shallow neural network and trained via backpropagation in word sequence prediction tasks.

The second key advancement was the introduction of the Transformer architecture. Vaswani et al. (2017) proposed a model based on the *attention* mechanism, which allows to selectively focus on specific parts of the input data when making predictions, thus allowing to give more relevance to the more salient parts of a text related to a given word even if those are not immediately adjacent to it. Furthermore, this approach also overcomes the inherent limitation of word2vec of having a *single* vector representing each word; in fact, in models based on the Transformer architecture the vector representing each word varies depending on the context. This advancement is the basis of models that have quickly become state-of-the-art in NLP, like BERT (Devlin et al., 2018).

All the models discussed so far have been developed to produce embeddings of individual words. In the specific setting of scientometrics, however, we typically aim to estimate the similarity – or relatedness – of texts, rather than simple words. Reimers & Gurevych (2019) proposed an adaptation of BERT – called Sentence-

BERT – specifically oriented to the embedding of short texts, providing an efficient way of computing semantic similarity among, for example, the abstract of publications.

As discussed in a previous Section, quite often estimating the relatedness among a collection of papers is just an intermediate step towards the end goal of a discrete categorisation into research topics. However, classical clustering and community detection techniques are not necessarily amenable to be used together with this new way of representing data as embedded in vector spaces. The typical workflow with text-embeddings thus involves two steps – i.e., dimensionality reduction and clustering – that again make use of recently introduced techniques.

Since embedding vectors have typically several hundred dimensions – and each of those does not singularly have any substantive meaning – the first step is to use non-linear dimensionality reduction techniques in order to obtain a lower dimensional representation of the relationship among documents. The most popular algorithms, UMAP (McInnes et al., 2018) and t-SNE (van der Maaten & Hinton, 2008), both work by attempting to preserve the nearest neighbours of each point. Finally, clustering is applied on the lower dimensional representation of the data. While in principle this step can be performed also with more traditional tools, a density-based clustering algorithm named HDBSCAN (Campello et al., 2013; McInnes et al., 2017) has grown in popularity due to the fact that i) it does not require to specify *ex ante* the number of clusters to be identified, and ii) it does not force each data point to belong to a cluster, thus being more robust to noise.

While specific implementations of this workflow have been used in impressive science mapping exercises (e.g., González-Márquez et al., 2023), the quantitative science studies community seems to not have started to explore the potential of these techniques yet. This is probably partly due to most of these developments being very recent, but this will likely change in the near future, also thanks to the increased availability of interfaces lowering the barriers to use for the end users (e.g., Grootendorst, 2022).

Despite the promising developments of these techniques, some limitations must be underlined both in general and in the specific context of their use for the analysis of the scientific literature. First, the development of NLP embedding techniques is

relatively recent and very fast. Second, the understanding of the emerging properties of these systems is an ongoing and challenging research effort. Third, given the amount of data and computing power required to train such systems – currently available only to very large firms and organisations – the end users typically employ out-of-the-box pretrained models. This can represent a challenge since transparency regarding the training data is lacking. In particular for scientometric applications, it is crucial to rely on models trained on data including scientific publications – like SentenceBERT all-mpnet-base-v2, trained on data including millions of articles from Semantic Scholar (Lo et al., 2020) – or even explicitly designed for the purpose – (e.g., Cohan et al., 2020). In addition, while general purpose models able to embed different languages in the same vector space are increasingly common, it might be challenging to find pretrained multi-language models able to accurately process scientific publications data.

Finally, it must be stressed the need for a systematic comparison of these new techniques with established ones. As previously discussed, benchmarking of different topics identification methods has been an open challenge for scientometric research even before the introduction of text embedding approaches. However, a wealth of option can also be seen as an opportunity to use different techniques to track the different aspects of research evolution.

3. Summary of the studies and discussion

The first Chapter has introduced the theoretical framework, discussing the role of norms and incentives in shaping the protected space of individual scholars. The second has instead discussed established and new methods for research topics identification, based on both citational and textual data. This Chapter will summarise the results of the three independent contributions reported in Part II, while highlighting their commonalities and how they contribute to addressing the overarching goal and question of this work. It will finally discuss the limitations and possible developments of the approach taken in the whole dissertation.

The overall objective of this work was investigating the dimension of the research topics and the factors affecting their choice by individual scholars. This was done by separately tackling macro, meso and micro-level scientific dynamics. Indeed, clarifying the level of analysis is fundamentally important to understand the dynamics at play. For instance, in the case of specialization, opposite processes are perfectly compatible at different levels of description: i.e. we might observe a decrease in specialization at the field-level – as the number of topics covered expands – and at the same time an increase at the individual level – as authors increasingly focus on their areas of expertise, or viceversa.

The first independent study, reported in Chapter I, takes a macro level perspective, by focusing on the role of macro-regional factors and challenges in determining the research focus at the country level. The study analyses a corpus of 193,471 scientific publications related to climate change research, covering a span of thirty years, from 1990 to 2020. It leverages the text embedding and clustering techniques discussed in the previous Chapter to identify 472 research topics in the corpus. It then maps topical trends to the country-level affiliation of the authors, in order to get a general overview of the topic specialisation of each country. The analysis reveals the existence of five major country blocks, with distinct research specializations. Furthermore, countries with a lower level of scientific publication output tend to be also the ones more dramatically impacted by the consequences of climate change. The study highlighted the fact that these countries display a unique focus on applied research, aimed at tackling impeding issues. Conversely, both basic research and works related to broad global regulatory policies are more

distinctive of the western countries. On a more general level, this study shows that the overall societal context and the level of available resources can have a substantial impact on the kind of research topic specialisation we can expect to observe.

The second contribution, instead, looks at the meso-level dynamics, and gives us a deeper understanding of the extent to which institutional "filters" modify the landscape of topics explored in practice. It does so by analysing proposals submitted to the European Research Council, the largest research funding programme in the EU. The explicit goal of the ERC is to fund bottom-up innovative research without any top-down prioritization, but the selection of funded proposals is performed by elite academics. The study identified 188 research topics across 91,273 proposals, by using text-embedding and clustering techniques on the corpus of proposals' summaries. Despite the bottom-up focus of the programme, research topics were found to have highly heterogeneous success rates. While this result might partially depend on the unobserved quality of individual proposals, these differences also point in the direction of a pattern that reflects elite scientists' priorities, exactly allowing to see one of the "filters" of the scientific community in action.

Finally, the third study investigates the micro-level dynamics of the evolution of the individual research agenda along the scientific career. The aim was not only to quantify the level of diversity of each individual's agenda, but also to track it longitudinally, in order to identify the factors affecting the point in time at which this diversification occurs. The study considers a sample of 4,785 scholars across four countries and four disciplines, covering a spectrum of different institutional contexts and disciplinary norms. We systematically collected data about the scientific production of the scholars in the sample – collectively comprised of 141,690 publications – and used text-embedding techniques to estimate each publication's deviation from the previous one from the same author. The results corroborate the importance of protected space for researchers. Indeed, scholars seem to diversify their research agenda with the increase in their seniority, which is typically connected to an increase in the protected space. Diversification though, can be also the consequence of a release from local hierarchy. This is likely the case, for instance, in laboratory-based fields, where mobility is shown to have a

dramatic impact on diversification, especially in the early career phase. A plausible explanation for such a pattern is the need for early career researchers to adapt their research agenda to that of their new laboratory. Additionally, mobility events in general – especially the early ones – are associated with increased diversification. This pattern can be in general point to the effect of changing environment and an increased variety of external stimuli. Finally, scientific collaboration is shown to be a great enabler of diversification, since by working together scholars can access approaches, methods, and areas of expertise that would otherwise be individually unavailable to them.

These three studies are an example of how quantitative and computational approaches can help to unveil the inner workings of the scientific community. In particular, the dimension of research topics is difficult to map and measure. This poses a serious challenge to our understanding of the directions towards which science is – or might be – evolving. By taking different angles, all the three studies provided an entry point to the factors affecting this evolutionary process.

They also testify to the potential of the new text-embedding techniques in providing a fine-grained treatment of the content of scientific publications and in giving us access to a latent dimension that would otherwise be invisible. Finally, the contribution by these three studies aspired to draw attention towards the role of institutional and social filters in the selection of the research topics that are actually investigated.

From a substantial point of view, the three studies contribute to our understanding of research topic dynamics from three different angles. The first highlights the link between contextual challenges and the differential focus on research questions, and underlines the importance of the boundaries in the mapping of research literature. The second shows how explicitly stated policies can be counteracted – or potentially even undermined – by the selection effect of institutional and social filters. The third – linking theories coming from the qualitative literature to quantitative evidence – uncovers a process of progressive expansion of the individuals' protected space along their careers, with discontinuities arising from their institutional mobility.

Furthermore, from a methodological standpoint, this contribution showcases three ways in which text-embedding techniques can be applied in the quantitative study of science. The first study provides a detailed description of the research on climate change, makes use of the research topics as a mapping and exploratory device. The second focuses on research proposals' success rate and uses research topics as a predictor. Finally, the third study has research topics as its outcome and goes beyond the usual definition of topic as a discrete-boundary classification by using instead a continuous measure of research agenda diversification.

This said, certain limitations must be underlined. As previously mentioned, text-embedding models are still an active area of development. While we might in the future have a clearer grasp on their emergent properties, this is for now an additional reason to invest effort into benchmarking exercises, and in comparing results obtained with several different methods. Aside from the technical challenges though, the introduction of these new tools might spur a renewed attention on the theoretical definition of the constructs used in scientometrics. Up to now, indeed, most of the work on research topics has relied on operational model-based definitions.

Another potential limitation of the current work is the reliance on the abstracts as representatives of the content of scientific products. While arguably being a step forward with respect to the exclusive use of citation-based measures, abstracts could – in some cases – not faithfully represent the actual content of the full-text. Aside from issues of data availability – severely limited in the case of full-texts – the comparison of these two sources of information has still not been thoroughly explored, and the existing literature is inconclusive in establishing a clear preference for either (Lin, 2009; Cohen et al., 2010).

A further challenge for more extensive research projects on this area is the limited availability of data regarding crucial processes that unfold in the scientific community. As this kind of data – for instance on peer-review (Squazzoni et al., 2020) – become more available, the tools to investigate the distribution and evolution of research topics will become increasingly relevant.

Additionally, a broader application of text-embedding opens up the necessity of developing new statistical able to deal with the high-dimensional data with

complex dependance structures that arise from embedding models. Other potential avenues for future research include, for example, a deeper dive into the topic composition of research agendas in specific fields – tracking not only the diversity of the agendas but also their focus – and study of the relation between semantic proximity and peer-review evaluations – along the lines of Boudreau et al. (2016).

In closing, it is worth mentioning the importance of these new instruments also in the context of science policy. Indeed, as of today, decision makers take crucial decisions on the future of scientific research without much information on the distribution of research topics, nor on the potential effect their decisions could have on the evolution of said distribution. Equipped with new tools we might be able to acquire insights on the current as well as on potential states of things.

References

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2024). Do research assessment systems have the potential to hinder scientists from diversifying their research pursuits? *Scientometrics*. https://doi.org/10.1007/s11192-024-04959-8

Aleta, A., Meloni, S., Perra, N., & Moreno, Y. (2019). Explore with caution: Mapping the evolution of scientific interest in physics. *EPJ Data Science*, *8*(1), 27. https://doi.org/10.1140/epjds/s13688-019-0205-9

Avin, S. (2019). Centralized Funding and Epistemic Exploration. *The British Journal for the Philosophy of Science*, *70*(3), 629–656. https://doi.org/10.1093/bjps/axx059

Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and creativity: Evidence from the academic life sciences. *The RAND Journal of Economics*, *42*(3), 527–554. https://doi.org/10.1111/j.1756-2171.2011.00140.x

Baccini, A., & Petrovich, E. (2023). A global exploratory comparison of country self-citations 1996-2019. *PLOS ONE*, *18*(12), e0294669. https://doi.org/10.1371/journal.pone.0294669

Baldwin, M. (2015). *Making "Nature": The History of a Scientific Journal*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226261591.001.0001

Baldwin, M. (2018). Scientific Autonomy, Public Accountability, and the Rise of "Peer Review" in the Cold War United States. *Isis*, *109*(3), 538–558. https://doi.org/10.1086/700070

Blei, D. M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1). https://doi.org/10.1214/07-AOAS114

Bloor, D. (1984). The Strengths of the Strong Programme. In *Scientific Rationality: The Sociological Turn* (pp. 75–94). Springer Netherlands. https://doi.org/10.1007/978-94-015-7688-8_3

Bloor, D. (1991). *Knowledge and social imagery* (2. ed., [Nachdr.]). Univ. of Chicago Press.

Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, *8*(1), 224. https://doi.org/10.1057/s41599-021-00903-w

Broucker, B., & De Wit, K. (2015). New Public Management in Higher Education. In J. Huisman, H. De Boer, D. D. Dill, & M. Souto-Otero (Eds.), *The Palgrave International Handbook of Higher Education Policy and Governance* (pp. 57–75). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-137-45617-5_4

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science. *Management Science*, *62*(10), 2765–2783. https://doi.org/10.1287/mnsc.2015.2285

Bush, V. (1945). *Science The Endless Frontier. A Report to the President by Vannevar Bush, Director of the Office of Scientific Research and Development*.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, *22*(2), 191–235. https://doi.org/10.1177/053901883022002003

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining* (Vol. 7819, pp. 160–172). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

Case, C. M. (1928). Scholarship in Sociology. *Sociology and Social Research*, *12*, 323–340.

Chubin, D. E. (1976). State of the Field The Conceptualization of Scientific Specialties. *The Sociological Quarterly*, *17*(4), 448–476. https://doi.org/10.1111/j.1533-8525.1976.tb01715.x

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). *SPECTER: Document-level Representation Learning using Citation-informed Transformers*. https://doi.org/10.48550/ARXIV.2004.07180

Cohen, K. B., Johnson, H. L., Verspoor, K., Roeder, C., & Hunter, L. E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. BMC Bioinformatics, 11(1), 492. https://doi.org/10.1186/1471-2105-11-492

Cronin, B. (1981). The Need for a Theory of Citing. *Journal of Documentation*, *37*(1), 16–24. https://doi.org/10.1108/eb026703

Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. T. Graham.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://doi.org/10.48550/ARXIV.1810.04805

Eck, N. J. V., & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology*, *60*(8), 1635–1651. https://doi.org/10.1002/asi.21075

Edwards, M. A., & Roy, S. (2017). Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. *Environmental Engineering Science*, *34*(1), 51–61. https://doi.org/10.1089/ees.2016.0223

Ferlie, E., Pettigrew, A., Lynn, A., & Louise, F. (1996). *The new public management in action*. Oxford University Press.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), eaao0185. https://doi.org/10.1126/science.aao0185

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and Innovation in Scientists' Research Strategies. *American Sociological Review*, *80*(5), 875–908. https://doi.org/10.1177/0003122415601618

Fyfe, A., Moxham, N., McDougall-Waters, J., & Mørk Røstvik, C. (2022). *A History of Scientific Journals. Publishing at the Royal Society, 1665–2015*. UCL Press. https://doi.org/10.14324/111.9781800082328

Fyfe, A., Squazzoni, F., Torny, D., & Dondio, P. (2020). Managing the Growth of Peer Review at the Royal Society Journals, 1865-1965. *Science, Technology, & Human Values*, *45*(3), 405–429. https://doi.org/10.1177/0162243919862868

Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science*, *122*(3159), 108–111. https://doi.org/10.1126/science.122.3159.108

Gilbert, N. (1977). Referencing as Persuasion. *Social Studies of Science*, *7*(1), 113–122. https://doi.org/10.1177/030631277700700112

Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, *111*(2), 981–998. https://doi.org/10.1007/s11192-017-2296-z

Gläser, J., & Laudel, G. (2015). A Bibliometric Reconstruction of Research Trails for Qualitative Investigations of Scientific Innovations. *Historical Social Research / Historische Sozialforschung Vol. 40, No. 3*. https://doi.org/10.12759/HSR.40.2015.3.299-330

González-Márquez, R., Schmidt, L., Schmidt, B. M., Berens, P., & Kobak, D. (2023). *The landscape of biomedical research*. https://doi.org/10.1101/2023.04.10.536208

Groen-Xu, M., Bös, G., Teixeira, P. A., Voigt, T., & Knapp, B. (2023). Short-term incentives of research evaluations: Evidence from the UK Research Excellence Framework. *Research Policy*, *52*(6), 104729. https://doi.org/10.1016/j.respol.2023.104729

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. https://doi.org/10.48550/ARXIV.2203.05794

Hanson, M. A., Barreiro, P. G., Crosetto, P., & Brockington, D. (2023). *The strain on scientific publishing*. https://doi.org/10.48550/ARXIV.2309.15884

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, *41*(2), 251–261. https://doi.org/10.1016/j.respol.2011.09.007

Horlings, E., & Gurney, T. (2013). Search strategies along the academic lifecycle. *Scientometrics*, *94*(3), 1137–1160. https://doi.org/10.1007/s11192-012-0789-3

Huang, S., Lu, W., Bu, Y., & Huang, Y. (2022). Revisiting the exploration-exploitation behavior of scholars' research topic selection: Evidence from a large-scale bibliographic database. *Information Processing & Management*, *59*(6), 103110. https://doi.org/10.1016/j.ipm.2022.103110

Jia, T., Wang, D., & Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. *Nature Human Behaviour*, *1*(4), 0078. https://doi.org/10.1038/s41562-017-0078

Kessler, M. M. (1963a). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25. https://doi.org/10.1002/asi.5090140103

Kessler, M. M. (1963b). Bibliographic coupling extended in time: Ten case histories. *Information Storage and Retrieval*, *1*(4), 169–187. https://doi.org/10.1016/0020-0271(63)90016-0

Kitcher, P. (1990). The Division of Cognitive Labor. *The Journal of Philosophy*, *87*(1), 5. https://doi.org/10.2307/2026796

Kuhn, T. S. (1959). The essential tension: Tradition and innovation in scientific research. In C. W. Taylor (Ed.), *The third (1959) university of utah research conference on the identification of scientific talent* (Reprint, pp. 162–174). University of Utah Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Univ. of Chicago Press.

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton University Press.

Laudel, G. (2017). How do National Career Systems Promote or Hinder the Emergence of New Research Lines? *Minerva*, *55*(3), 341–369. https://doi.org/10.1007/s11024-017-9314-4

Laudel, G., & Bielick, J. (2018). The Emergence of Individual Research Programs in the Early Career Phase of Academics. *Science, Technology, & Human Values*, *43*(6), 972–1010. https://doi.org/10.1177/0162243918763100

Laudel, G., & Gläser, J. (2008). From apprentice to colleague: The metamorphosis of Early Career Researchers. *Higher Education*, *55*(3), 387–406. https://doi.org/10.1007/s10734-007-9063-7

Leydesdorff, L. (1997). Why words and co-words cannot map the development of the sciences. *Journal of the American Society for Information Science*, *48*(5), 418–427. https://doi.org/10.1002/(SICI)1097-4571(199705)48:5<418::AID-ASI4>3.0.CO;2-Y

Lin, J. (2009). Is searching full text more effective than searching abstracts? BMC Bioinformatics, 10(1), 46. https://doi.org/10.1186/1471-2105-10-46

Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. (2020). S2ORC: The Semantic Scholar Open Research Corpus. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, *40*(5), 342–349. https://doi.org/10.1002/(SICI)1097-4571(198909)40:5<342::AID-ASI7>3.0.CO;2-U

Madsen, E. B., & Nielsen, M. W. (2024). Do thematic funding instruments lead researchers in new directions? Strategic funding priorities and topic switching among British grant recipients. *Research Evaluation*, rvae015. https://doi.org/10.1093/reseval/rvae015

March, J. G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, *2*(1), 71–87. https://doi.org/10.1287/orsc.2.1.71

Marshakova, I. (1973). System of Document Connections Based on References. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2-Informatsionnye Protsessy I Sistemy 2*, *6*, 3–8.

McCain, K. W. (1991). Mapping economics through the journal literature: An experiment in journal cocitation analysis. *Journal of the American Society for*

Information Science, *42*(4), 290–296. https://doi.org/10.1002/(SICI)1097-4571(199105)42:4<290::AID-ASI5>3.0.CO;2-9

McInnes, L., Healy, J., & Astels, S. (2017). Hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. https://doi.org/10.48550/ARXIV.1802.03426

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

Merton, R. K. (1979). *The Sociology of Science: Theoretical and Empirical Investigations*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. https://doi.org/10.48550/ARXIV.1301.3781

Mulkay, M. J. (1974). Methodology in the sociology of science: Some reflections on the study of radio astronomy. *Social Science Information*, *13*(2), 107–119. https://doi.org/10.1177/053901847401300206

Osterloh, M. (2010). Governance by Numbers. Does It Really Work in Research? *Analyse & Kritik*, *32*(2), 267–283. https://doi.org/10.1515/auk-2010-0205

Peirce, C. S. (1879). Note on the Theory of the Economy of Research. In *Report of the Superintendent of the United States Coast Survey Showing the Progress of the Work for the Fiscal Year Ending with June 1876* (pp. 197–201).

Peroni, S., Ciancarini, P., Gangemi, A., Nuzzolese, A. G., Poggi, F., & Presutti, V. (2020). The practice of self-citations: A longitudinal study. *Scientometrics*, *123*(1), 253–282. https://doi.org/10.1007/s11192-020-03397-6

Power, M. (1999). *The audit society: Rituals of verification* (Repr). Oxford University Press.

Price, D. J. D. S. (1963). *Little Science, Big Science*. Columbia University Press. https://doi.org/10.7312/pric91844

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. https://doi.org/10.48550/ARXIV.1908.10084

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Rudwick, M. J. S. (1985). *The great Devonian controversy: The shaping of scientific knowledge among gentlemanly specialists*. University of Chicago Press.

Seeber, M., Cattaneo, M., Meoli, M., & Malighetti, P. (2019). Self-citations as strategic response to the use of metrics for career decisions. *Research Policy*, *48*(2), 478–491. https://doi.org/10.1016/j.respol.2017.12.004

Shapin, S. (1992). Discipline and Bounding: The History and Sociology of Science as Seen through the Externalism-Internalism Debate. *History of Science*, *30*(4), 333–369. https://doi.org/10.1177/007327539203000401

Shapin, S. (2008). *The scientific life: A moral history of a late modern vocation*. University of Chicago press.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269. https://doi.org/10.1002/asi.4630240406

Snyder, T. D. (Ed.). (1993). *120 Years of American Education: A Statistical Portrait*. U.S. Department of Education. Center for Education Statistics.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Squazzoni, F., Ahrweiler, P., Barros, T., Bianchi, F., Birukou, A., Blom, H. J. J., Bravo, G., Cowley, S., Dignum, V., Dondio, P., Grimaldo, F., Haire, L., Hoyt, J.,

Hurst, P., Lammey, R., MacCallum, C., Marušić, A., Mehmani, B., Murray, H., … Willis, M. (2020). Unlock ways to share data on peer review. *Nature*, *578*(7796), 512–514. https://doi.org/10.1038/d41586-020-00500-y

Strevens, M. (2003). The Role of the Priority Rule in Science. *Journal of Philosophy*, *100*(2), 55–79. https://doi.org/10.5840/jphil2003100224

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. https://doi.org/10.48550/ARXIV.1706.03762

Wang, J., Lee, Y.-N., & Walsh, J. P. (2018). Funding model and creativity in science: Competitive versus block funding and status contingency effects. *Research Policy*, *47*(6), 1070–1083. https://doi.org/10.1016/j.respol.2018.03.014

Wei, T., Li, M., Wu, C., Yan, X.-Y., Fan, Y., Di, Z., & Wu, J. (2013). Do scientists trace hot topics? *Scientific Reports*, *3*(1), 2207. https://doi.org/10.1038/srep02207

Weisberg, M., & Muldoon, R. (2009). Epistemic Landscapes and the Division of Cognitive Labor. *Philosophy of Science*, *76*(2), 225–252. https://doi.org/10.1086/644786

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, *32*(3), 163–171. https://doi.org/10.1002/asi.4630320302

Whitley, R. (2000). *The intellectual and social organization of the sciences*. Oxford Univ. Press.

Whitley, R., & Gläser, J. (Eds.). (2014). *Organizational transformation and scientific change: The impact of institutional restructuring on universities and intellectual innovation* (1. ed). Emerald.

Whitley, R., Gläser, J., & Laudel, G. (2018). The Impact of Changing Funding and Authority Relationships on Scientific Innovations. *Minerva*, *56*(1), 109–134. https://doi.org/10.1007/s11024-018-9343-7

Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., Wang, Y., Stanley, H. E., & Havlin, S. (2019). Increasing trend of scientists to switch between topics. *Nature Communications*, *10*(1), 3439. https://doi.org/10.1038/s41467-019-11401-8

Zollman, K. J. S. (2010). The Epistemic Benefit of Transient Diversity. *Erkenntnis*, *72*(1), 17–35. https://doi.org/10.1007/s10670-009-9194-6

Part II

List of Papers

1. Debernardi, C., Seeber, M. & Cattaneo, M. (2024) Thirty Years of Climate Change Research: A Fine-Grained Analysis of Geographical Specialization. *Environmental Science & Policy* 152:103663. doi: 10.1016/j.envsci.2023.103663.
2. Debernardi, C., Seeber, M., & Piro, F.N. (under review) Investigator-driven or academic elite: preferential selection of research topics in the European Research Council's funding decisions.
3. Debernardi, C., Seeber, M., (to be submitted) Exploration and exploitation along the scientific career: the impact of seniority, mobility, and collaboration on individual research agendas diversification.

1. Thirty Years of Climate Change Research: A fine-grained analysis of geographical specialization

This chapter has been published as:

Debernardi, C., Seeber, M. & Cattaneo, M. (2024) Thirty Years of Climate Change Research: A Fine-Grained Analysis of Geographical Specialization. *Environmental Science & Policy* 152:103663. doi: 10.1016/j.envsci.2023.103663.

Abstract

Bibliometric overviews of climate change research typically focus on the main topical trends and few countries with the largest share of the scientific production. These are important limitations: most of the world's population live in countries that are heavily affected by climate change but have a relatively small scientific production, so that their topics of interest might be neglected. This contribution aims to address both limitations by investigating variations across countries in climate change research specialization. We employ a combination of state-of-the-art language modelling techniques to gain a fine-grained representation of the research topics on climate change, considering abstracts of 193,471 publications from 1990 to 2020. The analysis reveals the existence of five major country blocks , with distinct research specializations. Countries' research specialization is driven by the specific challenges posed by climate change, such as extreme precipitation and floods and food, as well as the level of resources at disposal, so that research into the phenomenon of climate change and its global solutions is more important in affluent western countries. Less affluent countries – which host several billion people – develop distinct research focuses on local problems' causes and mitigation strategies, but typically have limited resources to address these challenges. Hence, leading scientific countries should possibly contribute even more to addressing such issues.

1.1 Introduction

Climate change is one of the fundamental challenges of our times. Human activities have impacted – and are still impacting – local and global climate dynamics, causing a rise in temperatures, altering weather events patterns and disrupting biodiversity (IPCC 2023). This globally salient set of phenomena has been drawing the attention of the scientific community for decades, giving birth to a very active and interdisciplinary area of research, experiencing a strong growth (Haunschild, Bornmann, and Marx 2016). Making sense of a continuously growing scientific literature is extremely difficult. The problem is compounded by the sheer volume of publications on the subject and the challenges to organize, search, and analyse them. Literature reviews are very labour intensive, and they are not easily scalable to the level of broad scientific fields. As a consequence, bibliometric overviews and science mapping exercises often provide a general picture of big scientific areas at the expense of a coarser granularity. It has to be noticed, though, that between these two extremes is possible to find an integration in cases in which the desired level of description is still manageable by small groups of researchers (e.g. Fang et al. 2023).

Climate change research has been investigated several times with bibliometric techniques, ever since its outset, during the ’90s (e.g., Schwechheimer and Winterhager 1999). The most well documented finding is the rapid and substantial growth in the number of publications, together with a trend of increased co-authorship, international collaboration and interdisciplinarity (Stanhill 2001; Grieneisen and Zhang 2011; Haunschild, Bornmann, and Marx 2016). There are few general overviews of climate change research, while topic-specific analyses are commonly used to assess the state of the art of specific research areas. Some recent examples include overviews of research on the relation between climate change and infectious diseases (Li et al. 2020), and on carbon sinks (Huang, Chen, and Zhou 2020).

These studies – see section 1.2 – share a limitation common to much of the bibliometric overviews in general, namely, a focus on a macro level description of the research trends. Limiting the analysis to the identification of main topical trends hampers the level of detail of bibliometric studies and hides the variety of

research topics. In addition, many of these studies either ignore the geographical dimension or focus solely on the leading countries in terms of gross scientific production. Essentially, they focus only on major scientific players, and implicitly assume that these are representative of the global distribution of scientific topics. This approach is particularly problematic for climate change research because most of the world population is concentrated in developing countries, which experience the most severe consequences of climate change but have a very low scientific production per capita. For example, 1.4 billion people live in India and the Indian sub-continent is among the areas of the world most threatened by climate change (ND-GAIN vulnerability index 2020), yet its entire scientific production on climate change is one third of the United Kingdom's, despite a 20 times larger population. In turn, the combination of using macro level trends and focusing only on countries with high absolute scientific production underrepresents or hides research topics that are specific to countries highly impacted by climate change and home to a large share of the global population.

The aim of our contribution is therefore twofold. First, to provide a fine-grained overview of climate change research, highlighting the wide variety of research topics. Second, to explore how climate change research varies across countries, in terms of its relative importance and specific topics' importance, and what factors seemingly affect countries' focus on different topics. We do so by analysing a dataset of 193,471 scientific publications on climate change spanning 30 years (see Data and Methods). We address the first limitation by leveraging recently developed natural language processing techniques to gain a fine level of granularity. Regarding the second limitation, we start comparing the number of climate change related publications to the total scientific production by country to highlight efforts in this research area in a simple way; we then delve deeper into 472 research topics and reveal their unequal geographical distribution. A further limitation, that pertains to our study as to most of the studies that make use of textual data, is the exclusive focus on the literature written in English.

In the following section, we summarize the features and limitations of bibliometric overviews on climate change. We also discuss the factors that may affect the choice of research topics, and hence countries' specialization. We then proceed to describe the data and methods, where we also explain the added value of a new

approach combining different techniques for bibliometric studies in unveiling fine-grained research topics. In section four, we present the empirical analysis, which includes the identification and description of five country blocks based on their research specialization, as well as the discussion of their drivers of specialization. In the final section, we discuss the main findings and implications of the article.

1.2. Literature review and theoretical framework

1.2.1 Climate change research

Since 1970s, climate change has become a relevant topic that attracted considerable attention of scholars in the decades ahead. The field has grown at a fast pace, with climate change papers doubling every 5-6 years (Haunshild et al. 2016). This major interest and the huge societal and political implications have consequently led to the development of review papers aimed at assessing the state of climate change research. Bibliometric analyses have played a major role in this regard, providing both a description of the overall productivity of the field over time, and an exploration of more specific aspects related to climate change science.

The most comprehensive bibliometric analyses exploring climate change research have been developed in the last decade. Grieneisen and Zhang (2011) leveraged an in-depth selection of key words related to the climate change research and investigated a pool of 110,139 publications in the period 1997-2009. The authors stressed the fast-growing publication rate of the field in the first 2000s and presented the occurrence of climate change papers in relation to both the ten largest categories of Web of Science and the 253 subject categories. Also, they identified the most prolific institutions globally. Li et al. (2011) examined 41,457 climate change publications based on the online version of Science Citation Index Expanded between 1992 and 2009. The authors discussed the research patterns of major western countries and the most productive institutions. Further, they identified, through a keyword analysis, the most prominent areas of climate change research in the 21st century. Li and Zhao (2015) collected 113,468 publications (1993 to 2012) on environmental assessment (EA), i.e., a broader area of research including climate change, and described productivity patterns across several dimensions, including the evaluation of research performance of the most

productive countries, major subject categories, and most representative journals in the field. Relying on 15,000 publications (1999-2010), Pasgaard and Strange (2013) investigated the gap between rich and poor countries in the supply of climate change knowledge and showed a significant unbalance between resource-rich and poor contexts and the presence of divergent climate change concerns. Rich countries focused more on mitigation issues (CO_2 emission reduction), while in less developed countries, research was more concerned about climate change adaptation and impacts. Haunschild et al. (2016) considered 222,060 articles and reviews (1980-2014, Web of Science), and highlighted the strong growth across macro fields (e.g., natural sciences, medicine, engineering) and seven climate sub-systems (e.g., atmosphere, oceanic water); further, they provided descriptive statistics on aggregate publication outputs for 32 countries. Callaghan et al. (2020) identified 140 topics on climate change research, using a broad sample of over 400,000 climate change related publications (until 2018), and hence investigated the over and under representation of research topics in the IPCC reports' reference list. More recently, Fu and Waltman (2022) conducted a descriptive analysis on 120,000 articles (2011-2018) and implemented a term mapping and burst detection analysis (limited to a sub-set of 25,000 articles) to provide an overall picture of climate change research and insights about its evolution. The analysis was centred on five clusters of research topics (physical sciences, paleoclimatology, climate change ecology, climate policy) and focused on the eight most productive countries.

Parallel to these comprehensive reviews, several bibliometric assessments have been performed to examine specific aspects of climate change research. These span from studies interested in reviewing the current state of climate science in relation to certain domains, like adaptation, vulnerability, resilience, and specific areas, like geoengineering and tourism (e.g., Janssen et al. 2006, Belter and Seidel 2013, Wang et al. 2014, Aleixandre-Benavent et al. 2017, Wang et al. 2018, Rana 2020), the analysis of collaborative research patterns in the field (Jappe 2007, Sangam and Savitha 2019, Engels and Ruschenburg, 2008), the understanding of climate change controversies (e.g., Jankó et al. 2014, 2017), the detailed investigation of single target journals' performance like *Climate Change* and *The Bulletin of the American Meteorological Society* (e.g., Stanhill, 2001, Hellsten and Leydesdorff,

2016), to the examination of the impact of IPCC reports on climate change research (e.g., Vasileiadou et al. 2011). These studies were based on specific focus and mostly descriptive approaches.

In general, few studies provided an extensive evaluation of the climate change research whereas many more contributions explored detailed subjects. In both cases, these bibliometric assessments were often limited in their geographical scope and focused on the performance of few top prolific countries (generally located in the western part of the world), and mostly related to a restricted set of topics or subject areas.

1.2.2 Factors affecting countries' specialization in climate change research.

Climate change research is a vast interdisciplinary area of study stretching from theoretical problems and modelling aiming to understand the reasons and possible evolution of the phenomenon, to the exploration of the ecological, economic, socio-political impact of climate change, as well as political and technological solutions. Therefore, several factors are likely to matter in affecting which research topics are more salient in each country.

In most cases, the decision about what to research is ultimately in the hands of individual scientists or research groups. Sociologists and economists of science have explored the drivers of researchers' behaviour - including decisions on what to research - and revealed the importance of several types of motivations, namely: *intrinsic* motivations, such as intellectual curiosity and pleasure of solving puzzles (Merton 1974), *extrinsic* motivations – like recognition from colleagues for being the first to make a scientific discovery (Merton 1974) or obtaining funds and progressing in career (Stephan 1996) – and *prosocial* motivations, i.e. to have a positive impact through the research work (Iorio, Labory, and Rentocchini 2017; Sormani and Uude 2022). Scientists are embedded into countries that differ along economic, social, political, cultural, and geographical traits, that may affect such motivations, the interest in specific questions and arguments, and hence the preference for certain research topics. Variations between countries may also depend on the existence of regional and national disciplinary communities. The discipline is the most important source of values, sense of belonging, and scientific goals for a scientist (Clark, 1983; Whitley 2000). Disciplines, however, vary

considerably in their intellectual and social organization, the degree of internal cohesion and hierarchical control, and therefore on the influence on what their members research and how (Whitley 2000). For example, *Theoretical Physics* displays a high level of cohesion that enables an international elite to influence the goals and procedures used by most researchers around the globe, whereas *Management* is an a-paradigmatic, fragmented discipline, where several schools coexist and address diverse problems and with diverse approaches (Whitley 2000). In turn, national disciplinary communities are partly independent from an international community, and can develop specific interests, traditions, or schools. The impact of each one of these potential sources of variation arguably depends on the specific research problem and topic.

Research problems may be largely disconnected from the surrounding context, while others are affected by the specific context. The specific traits of a country are of little importance for abstract conceptual problems that do not vary depending on the specific context, such as research in *Mathematics*. Variations between countries may still exist, in similar fields, due to the existence of specific schools or traditions, leading to partly different specializations. Research in *Astrophysics* or *Nuclear Physics* also investigates phenomena and natural laws that are not affected by the context in which the research is conducted. However, research in these fields requires extremely large investments in scientific instruments such as telescopes and particle accelerators, and it is concentrated in countries with large available resources. In a similar vein, we can expect that topics in climate change research that require large investments, for example on simulation and computational tools, will be comparatively more important in affluent countries.

Climate change is a global phenomenon, but its impact varies considerably from context to context, as well as the solutions to mitigate such impact. These variations can affect specialization in two regards. First, countries' specialization will likely reflect their respective challenges and search for targeted solutions. Environmental factors, namely how climate change manifests in a specific area, arguably directly and indirectly affects which topics will be more relevant and researched. Recent studies suggest, for example, that experiences of climate change disasters raised citizens' environmental concerns and affected their choices, such as increasing voting for green parties (Hoffmann et al. 2022).

Another example showed that extreme events and perceived risk affect the development of local adaptation measures (Braunschweiger and Ingold 2023). In a similar vein, experiencing the effects of climate change on themselves and their local community can directly affect scientists' choices. Influence can also be indirect, through governments earmarking research funds for the specific challenges that a country is facing. Second, research on abstract and theoretical problems, regarding political aspects of climate change, about its global causes and political solutions, is also expected to be comparatively more relevant in countries at the research frontier, with large financial resources and more influence on world politics. Scientists in less affluent countries, on their side, are likely pressed to focus on urgent practical challenges.

Finally, researchers are embedded in networks of collaborations and interactions that provide new ideas and puzzles, stimulate a researcher's curiosity and perceptions about which are the most important scientific questions and knowledge gaps. Hence, we can expect that countries that collaborate intensely may display a similar specialization, also in spite of different contextual conditions and resources. The creation of network ties – including scientific collaborations – are driven by several networking mechanisms. By affecting the probability of scientists to communicate and collaborate, these mechanisms may also affect whether two countries display a similar research specialization. For example, people tend to connect with people perceived to be like themselves (McPherson, Smith-Lovin, and Cook 2001), and such *homophily* mechanism favours collaboration between scholars holding similar traits, like their culture, language, or ethnicity (Freeman and Huang 2014). The probability of two nodes being connected is also greater if both are connected to one or more common nodes (Newman 2001); such *transitivity* mechanism is common in scholarly collaboration (Franceschet 2011; Newman 2001; Schilling and Phelps 2007) and creates a path-dependency effect (Zhang C. et al. 2018). *Homophily* and *transitivity* mechanisms lead to the expectation that scientists that share similar traits, like culture and language, will collaborate frequently, influence each other, and in turn display a similar research specialization. Accordingly, countries like Australia and United Kingdom, which are very far away and face very different climatic

conditions and challenges, may still collaborate intensely because of their common language and cultural background, and hence display a similar profile.

1.3 Data and methods

We retrieved scientific publications on climate change from Scopus database, using the search terms "climate change" and "global warming." The search was run in October 2021. Scientific production before 1990 is sparse (around 400 publications from 1911 to 1989) while Scopus coverage gets worse going back in time; we hence considered the period from 1990 until 2020. We kept articles, conference papers, reviews, book chapters and books,[4] and removed all the entries without an abstract written in English. The final sample includes 193,471 publications.

We restricted the choice of language for two main reasons: first of all, to ensure uniform data coverage – i.e., by using only one source we do not have to deal with over or under representation of local languages or specific disciplines due to the use of specialized databases –, second the Large Language Model we employ to embed the texts is trained on an English corpus. On the one hand, this limitation is conservative regarding the aim of the article to show a wider variety of topics across countries that are typically disregarded: i.e., by hypothetically including non-English literature our point could only be strengthened. On the other hand, regarding the country level specializations, our work is representative of specialization in the internationally oriented research in English. Further work would be needed to highlight the specialization of research in local languages (and/or not indexed in Scopus), e.g., research directly dealing with local decision makers in non-English speaking countries.

The empirical analysis is organized in fourth sections. In first and second sections, we analyse the evolution of climate change research production over time and its relative importance by country. In the third section we explore geographic specialization and in the fourth, the drivers of such specialization.

[4] We excluded, e.g., editorials, notes, letters.

To explore geographic specialization, we adopted the following text analysis procedure. First, we composed the corpus to be analysed by concatenating the textual data available, i.e., title and abstract. The next step was to embed the documents in a vector space by using a state-of-the-art pretrained language model, namely Sentence-BERT (Reimers and Gurevych 2019). This kind of model is particularly useful when the aim is to group similar documents, and significantly outperforms methods based on simple bag-of-words assumptions. Since the resulting space is still very high-dimensional, we used a nonlinear dimensionality reduction algorithm – UMAP (McInnes and Healy 2018) – in order to have a representation amenable to be managed by a density-based clustering algorithm – i.e., HDBSCAN (Campello et al. 2013, McInnes et al. 2017) – but still retaining the local relationships between nearest neighbours. The dimensionality reduction step involves stochasticity; hence, we repeated the reduction-clustering phase several times to ensure a higher stability of the results. As core of the final clusters, we used the sets of documents co-occurring in every iteration of the procedure. We then extended the classification to other documents through a diffusion process which included them in the closest cluster. After a manual inspection to ensure semantic coherence, we obtained the most distinguishing words by cluster – through a simple TF-IDF weighting of words and bigrams – to support the interpretation of the content. Overall, this pipeline is similar to the approach used by the BERTopic algorithm (Grootendorst 2022).

Through this process we obtained 870 clusters of documents representing research topics. In the subsequent analysis we then only kept the 472 clusters consisting of at least 100 publications – which covers 85% of the original sample. Then, we focused on the 54 countries with at least 500 publications on climate change. We built a matrix of the *relative* weight of each topic in each country – measured by the fraction of publications included in the target topic over the total from the target country. To assess the over/under-representation of topics across countries we then standardized the matrix topic-wise so that each entry represents the distance from the global mean in relative weight of the target topic in the target country, measured in standard deviations (i.e., z-score). Finally, we used Ward's clustering to identify five Country Blocks (CBs) with a similar research focus – i.e., groups of countries in which the same topics are over/under-represented. To describe the

Country Blocks' distinctive specialization, we manually analysed the 30 topics with the highest average z-score for each of them.

1.4 Results

1.4.1 Climate change research production over time

Figure 1 presents the yearly total number of scientific publications on climate change from 1996 until 2020 and shows that the number of scientific articles on climate change increased exponentially. Figure 2 displays the share of scientific publications on climate change on the total scientific production and reveals that the share of scientific publications on this theme has increased from 1.1 per thousand in the year 2000 to almost ten per thousand publications in 2020, or 1% of the global scientific production.[5]

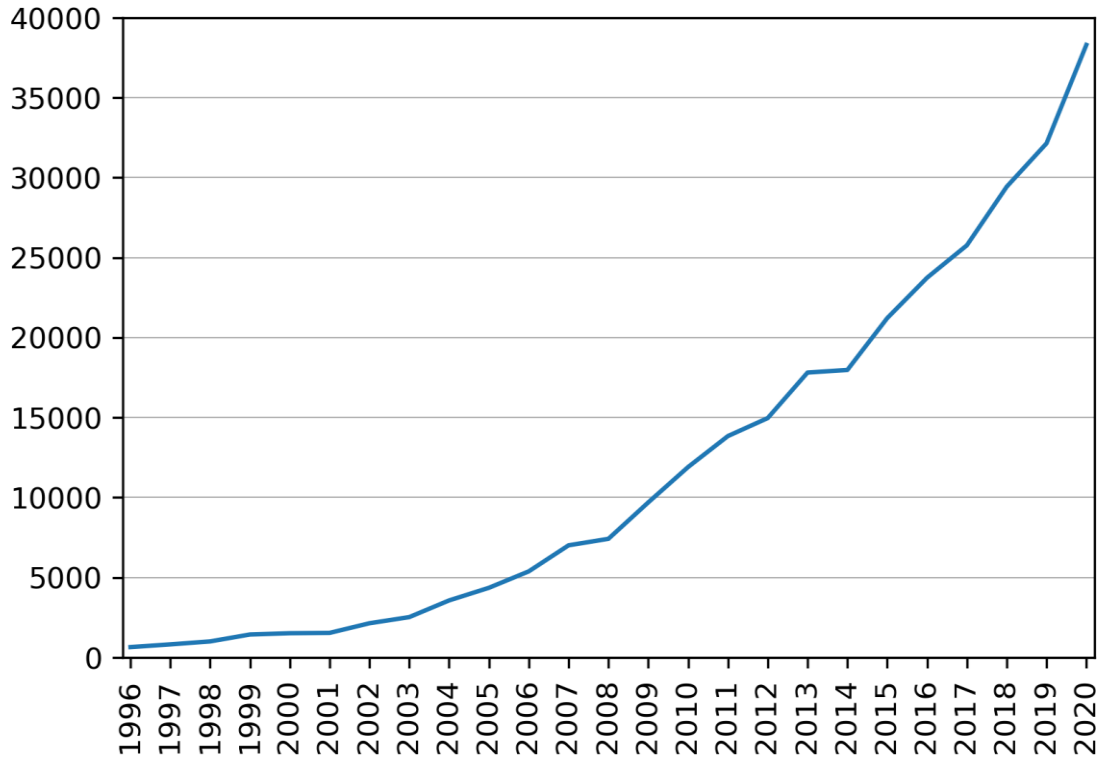

Figure 1: Absolute number of climate change related publications by year

[5] As an estimate of the global scientific production, we considered the sum of the Citable Documents by country (source: scimago-scopus). Those include articles, reviews, and

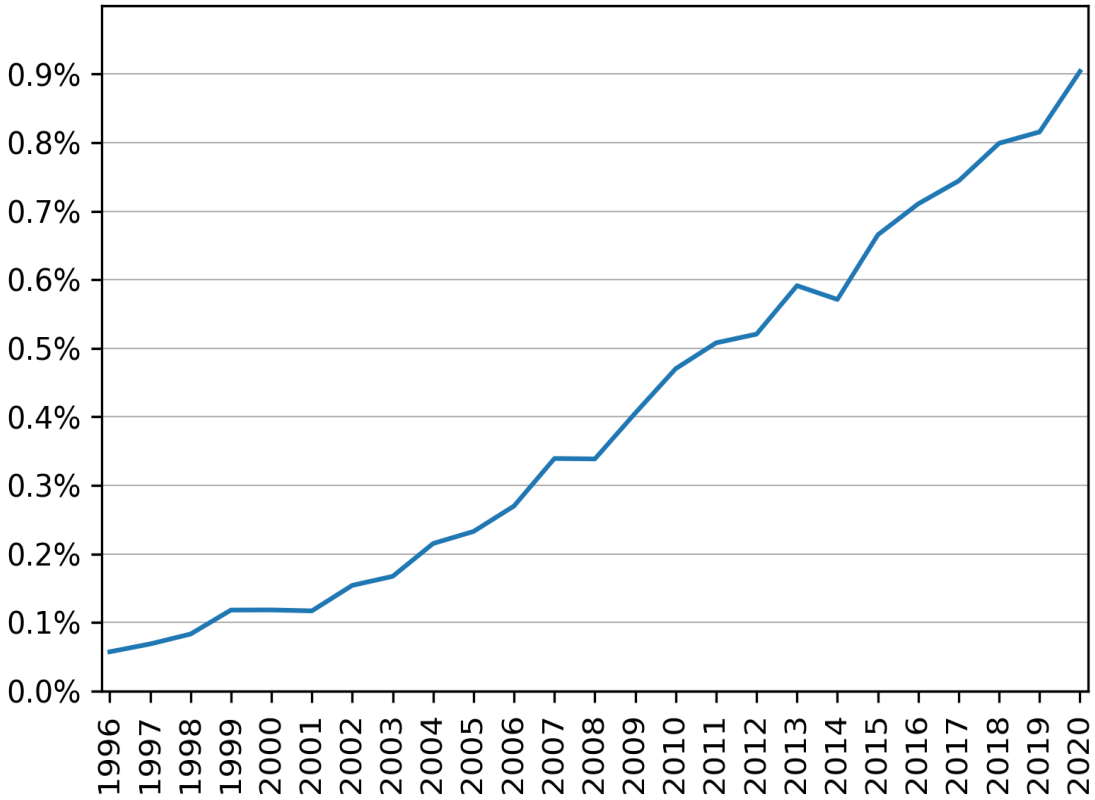

Figure 2: Estimated percentage of climate change related publications over the global scientific production

The distribution of this massive literature is not uniform across countries. As expected, at an absolute scale, the main countries are the major players of the global scientific production. The top ten countries – i.e., United States, China, United Kingdom, Germany, Australia, Canada, France, Spain, Italy, India – are involved in 62.5% of the publications. The scientific production has increasingly decentralized over time: the top ten countries went from producing over 70% of the new literature on climate change in the ’90s to less than 60% in 2020. The dominant role of western countries, and especially the US, has been rounded off (Figure 3).

conference papers. The time frame is limited since this information is available only from 1996.

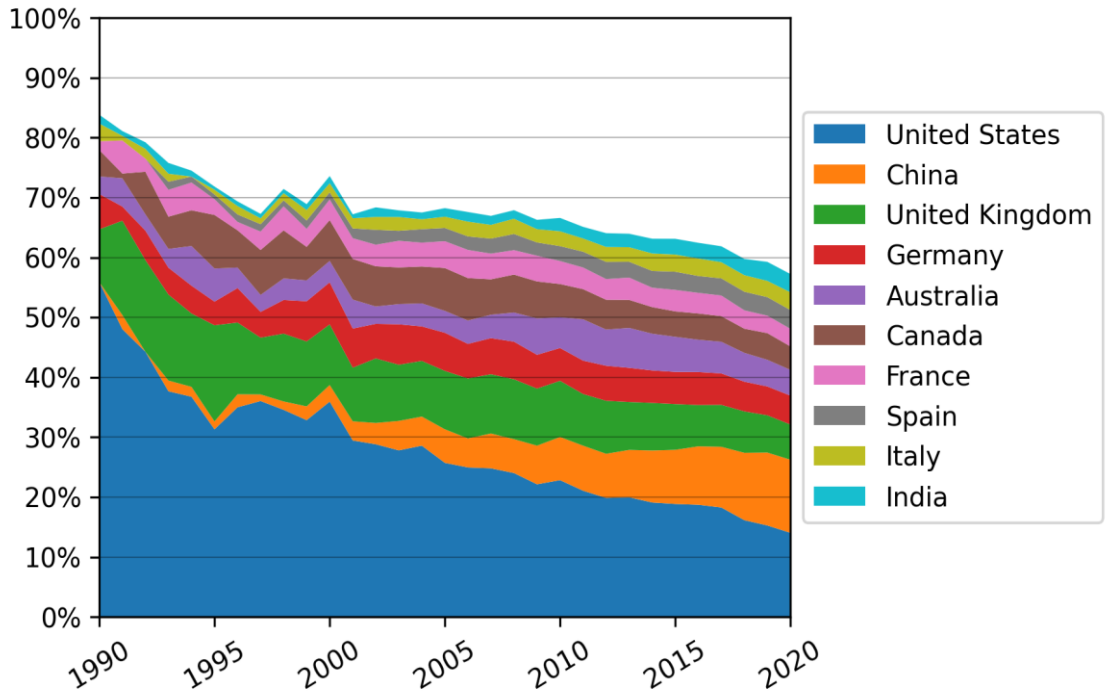

Figure 3: Percentage of the global scientific production by each of the top 10 countries for absolute production of climate change related publications by year.

1.4.2 Importance of climate change research

The percentage of publications focused on climate change on total publications (1996-2020) varies remarkably across the 54 countries considered for in-depth analysis, between 0.2% to almost 3% (see Figure A in the Appendix).

Figure 4 shows that there is not a clear geographical pattern. Countries strongly focused on climate change research, such as Nepal (2.95%), Kenia (2.45%), Ethiopia (1.88%), Norway (1.39%), Bangladesh (1.35%), Australia (1.07%), belong to different geographical and political areas, as much as those very little focused, like Russia (0.22%), Japan (0.24%), Israel (0.24%).

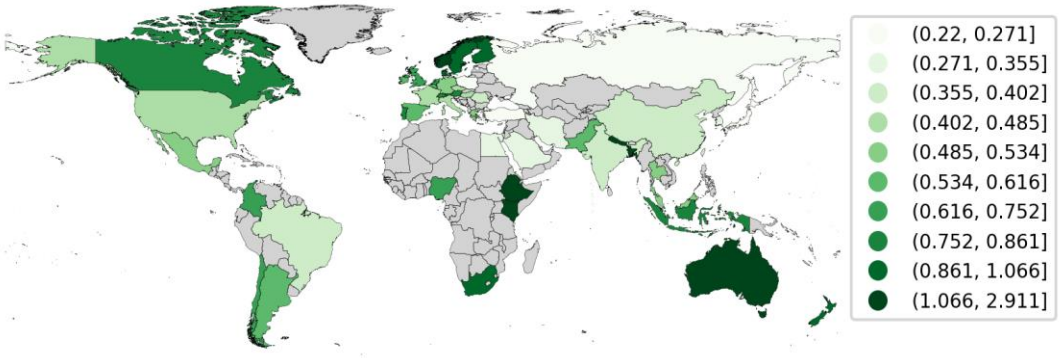

Figure 4: Fraction of climate change related publication per country (1996-2020). Coloured by decile.

Figure 5 presents the ND-GAIN vulnerability index for the selected countries, namely the propensity to be negatively impacted by climate hazards (Chen et al. 2015)[6] – from low vulnerability (dark blue) to high vulnerability (dark red) – and juxtaposes the relative importance of climate change research (Y-axis), with the absolute (top) and per million inhabitants (bottom) number of scientific publications on climate change (X-axis).

The figures show that there is not a clear relationship between country's vulnerability and the relative importance of climate change research: some of the most vulnerable countries are strongly focused on climate change research, while others are not.

The figures also shows that countries that are more vulnerable to climate change hazards (red dots) tend to be less productive in absolute term (left picture) and even more in proportion to their population (right picture) compared to countries less

[6] The ND-GAIN vulnerability index assesses the countries propensity to be negatively impacted by climate hazards by considering six life-supporting sectors: food, water, health, ecosystem services, human habitat, and infrastructure. Each sector is represented by six indicators that represent three components: the exposure of the sector to climate-related or climate-exacerbated hazards; the sensitivity of that sector to the impacts of the hazard and the adaptive capacity of the sector to cope or adapt to these impacts (Chen et al. 2015)

vulnerable (blue dots). Such negative relationship is not due to a lower interest on climate change (Y-axis) but on the fact that the most vulnerable countries tend to be less wealthy, invest less in research, and therefore produce less publications in general. This corroborates the idea that focusing only on research from countries with high absolute scientific production hides topics that are most important for countries highly impacted by climate change.

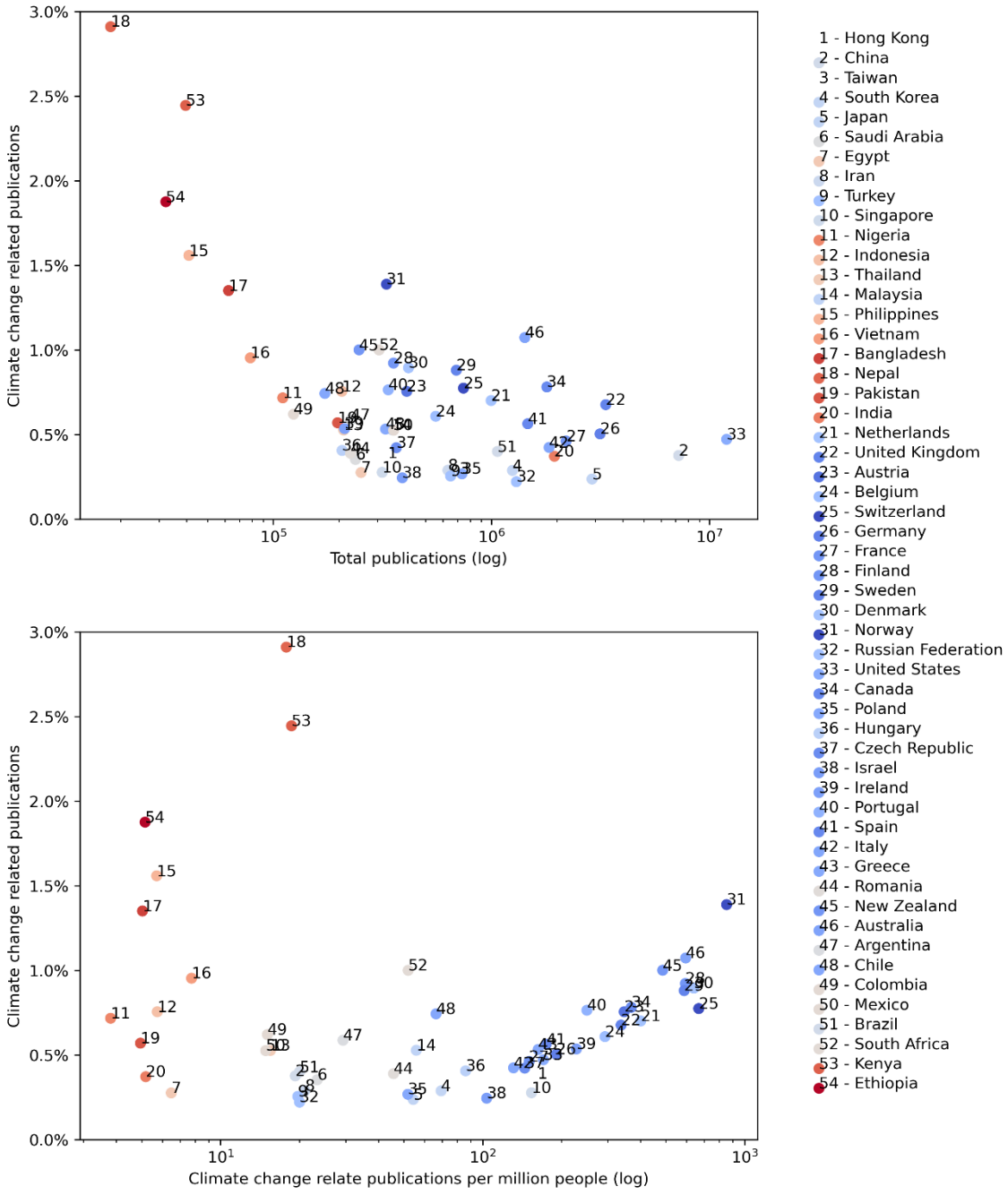

Figure 5: Total (top side) and per capita (bottom side) publications on climate change (X-axis) vis-a-vis percentage of climate change related publication (Y-axis) and climate change vulnerability: from low (dark blue) to high (dark red)

1.4.3 Geographical specialization

Figure 6 reports the full matrix of standardized relative weight of the 472 research topics across the 54 considered countries, together with the dendrograms depicting the hierarchical relations of similarity among countries' specialization. The red colour stands for over-represented topics, while the blue one for the under-represented ones; white spaces represent topics that are absent from a given country. For more details on the definition of the matrix see the Section "Data and Methods".

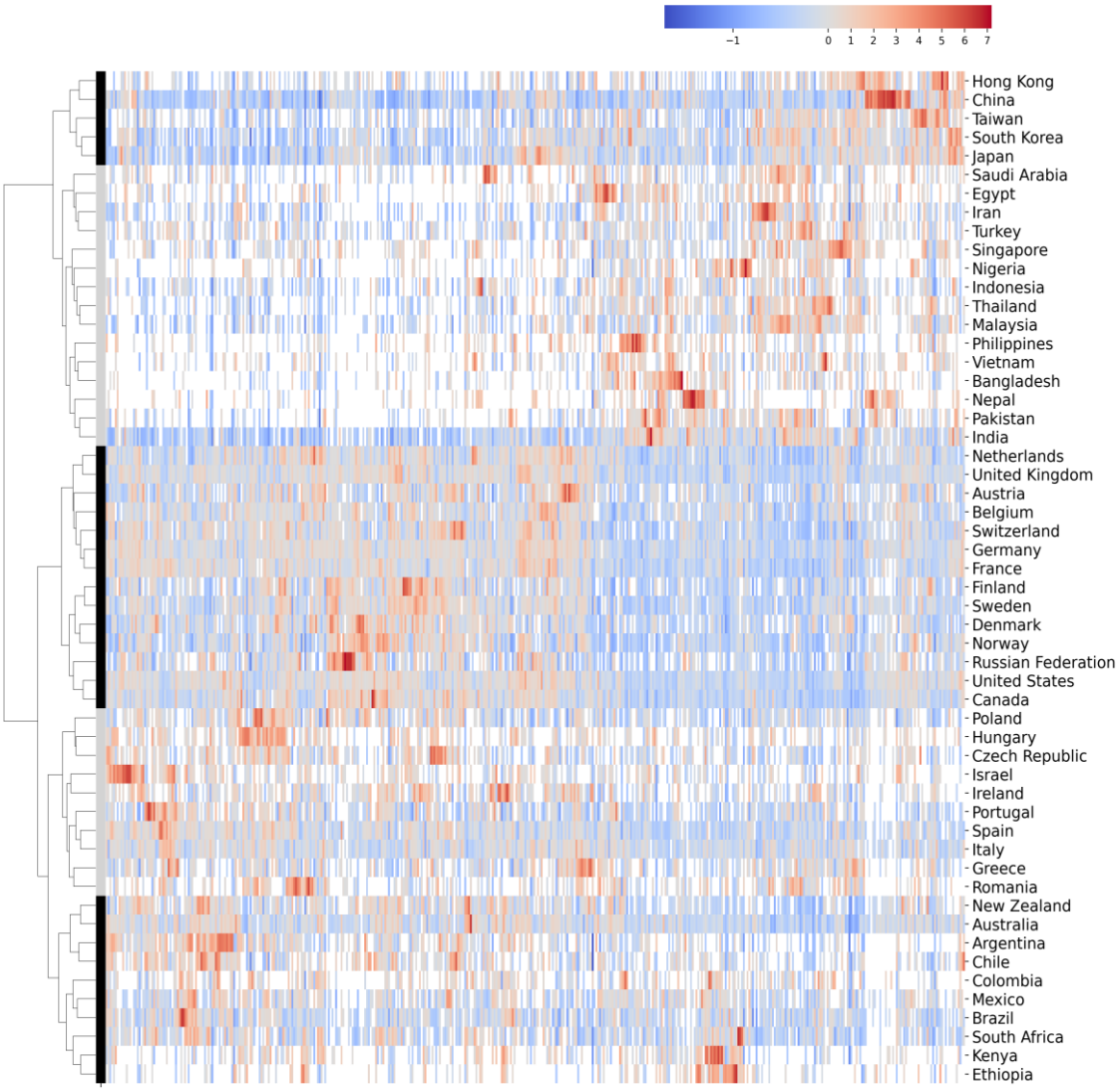

Figure 6: Standardized weight of topic by country matrix. White spaces indicate topics that are fully absent from a given country.

To make a manageable description of the research focus by geographic area, we aggregated the countries into five Country Blocks (CBs) according to their similarity in terms of over/under-representation of the research topics (see Data and Methods). This grouping relies on a fine-grained level of description, and – more importantly – allows us to delve deeper into each block.

Figure 7 presents the countries coloured by their Country Block.[7]

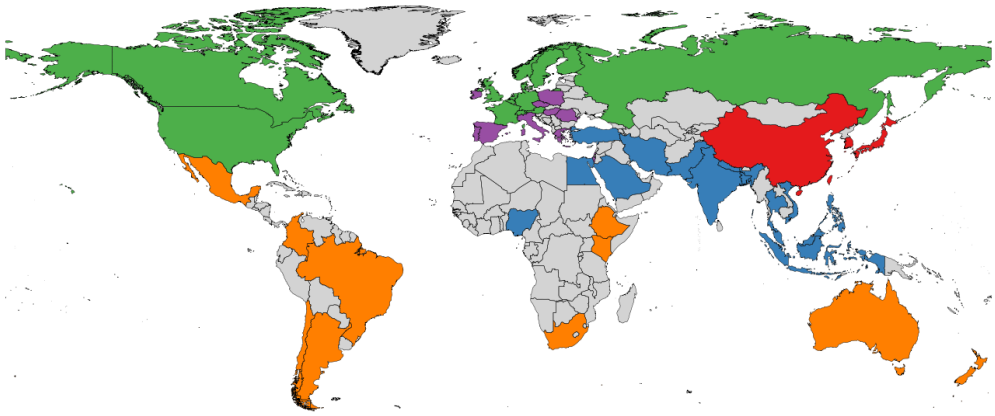

Figure 7: Countries coloured by block. Countries belonging to the same block have similar specialization profiles over research topics.

In the following, we describe each Country Block's distinctive specialization in climate change research.

Country Block 1 includes five east-Asian countries, with a total population of 1.6 billion people,[8] and represents 16% of the global publications on climate change (on average, 19 per million inhabitant).[9] Two main focuses characterize research on climate change in this block. The first includes topics related to: i) weather

[7] The full list of countries included in each block is reported in the appendix.

[8] All the population data are from the World Bank, year of reference 2020. Available at: https://data.worldbank.org/indicator/SP.POP.TOTL

[9] The percentage of global publications on climate change in which each Country Block is involved is computed considering our sample of publications to be representative of the global scientific production on climate change.

events, namely extreme precipitations, cyclones, monsoon, hurricanes, and extreme heat, also with a focus on specific areas like Yangtze and Yellow rivers, and Tibetan plateau, as well as ii) the causes of extreme events, like Sea Surface Temperature (SST) anomalies, and iii) their specific consequences, like landslides, floods, and flash droughts. The second area of focus is research about carbon emissions in the perspective of reducing them by increasing energy efficiency, through low carbon technologies, renewable sources, energy taxes, and reducing the carbon impact of the supply chain.

This specialization pattern reflects distinctive challenges posed by climate change on this geographical area. East Asia, together with India and Southeast Asia, is a region that combines extreme precipitations – related to monsoons and typhoons – with high density population (Zhang W. et al. 2018), hence resulting in the extremely high flood hazard and flood risk, with the displacement of millions of people (Carozza and Boudreault 2021).

Increasing energy efficiency is also a recognized objective and challenge for the largest country in this group: China, which is not only the largest emitter of CO_2 globally but has consistently displayed one of the highest ratios of CO_2 emissions per wealth produced. While China is by far the largest producer of renewable energy in the world,[10] and reduced the CO_2 per wealth ratio from 1.95 Kg of CO_2 per US Dollar of GDP in PPP (Purchasing Power Parity) in 1990 to 0.46 in 2019 (-73%), nevertheless, it still displays the 8th highest ratio in the world out of 228 countries.[11]

Country Block 2 includes countries in Southeast Asia, Indian Subcontinent, the Middle East, Egypt, and Nigeria – for a total population of 2.7 billion inhabitants, and 8% of the global publications on climate change (on average, 6 per million inhabitant). The most prominent area of research in CB2 revolves around food,

[10] Source: Statista. Available at: https://www.statista.com/statistics/267233/renewable-energy-capacity-worldwide-by-country/

[11] CO_2 emissions (kg per PPP $ of GDP). Wolrd Bank data. Available at: https://data.worldbank.org/indicator/EN.ATM.CO2E.PP.GD

with a central role played by topics related to the farming of rice, maize, and wheat, but also including fishing, aquaculture, and water management. It is also possible to observe a wide range of different approaches to the issue, from the study of rural households and small farmers adaptation strategies to the selection of resilient and productive species of crops. Renewable energies – mainly photovoltaic, solar, heat pumps and wind turbines – also appear frequently, followed by topics related to biogas and biofuel. Two additional areas of interest are disaster management – with research about floods and monsoons – and public health – with a focus on the diffusion of malaria and heat stress related diseases.

Also in this case, specialization patterns reflect distinctive challenges – but also opportunities – of countries in this block. Countries in this block are particularly vulnerable to hazards that climate change poses on their food production and supply. For example, the ND-global adaptation index measures *food vulnerability* with a composite indicator that considers, among others: projected change of cereal yields, food import dependency, agriculture capacity – and countries in this block have the highest average score (0.51) compared to the other blocks (1: 0.30, 3: 0.24; 4: 0.30; 5: 0.42). Interest about floods, monsoons, malaria and heat stress related diseases are also highly relevant for countries in this area. Research focus on solar energy and photovoltaics is also clearly driven by opportunities since this area of the planet displays very high levels of irradiation.

Country Block 3 includes countries in the Northern part of the Northern hemisphere – namely most of the western countries and Russia – for a total population of 805 million inhabitants, and 53% of the global publications on climate change (on average, 128 per million inhabitant). Two research areas play a prominent role. The first one is related to climate policy – featuring research on e.g., the Paris agreement, carbon tax implementation, climate governance and climate change narratives. The second area comprehends basic research on the phenomenon of climate change itself and topics like the albedo effect, the level of oceanic heat uptake, the carbon cycle and the AMOC (Atlantic meridional overturning circulation). Other focus areas include regional biology – especially of arctic species – and studies on the state of ice and snow coverage.

The research focus of this area is related to the role played in shaping policies and responses to climate change, both through the influence on global governance and through a leading role in research.

Country Block 4 includes South and East European countries, with a total population of 219 million inhabitants, and 11% of the global publications on climate change (on average, 97 per million inhabitant). The topics of specialization are rather heterogenous. It's single largest and most relevant topic of specialization is related to urban car traffic, electric vehicles, and sustainable transportation in general. Other prominent topics of specialization include disaster management – with studies on heat waves, extreme precipitations, and floods – soil degradation and erosion of the coastline. Among applied topics we also find research on renewable energy policies, water management and viticulture. Basic research, on the other hand, includes many topics from biology, with a focus on the sea fauna.

In this case the focus on vehicles and traffic might be related to the prevalence of road transportation both in private mobility and in freight shipping. Vehicles – especially commercial and industrial ones – also tend to be older, on average, in south and east Europe[12]. Another factor, partially related to the first one, is the low air quality – in particular with respect of PM10 pollutants – of several areas of eastern and southern Europe, especially Poland and northern Italy[13]. The focus on the erosion of the coastline, on the other hand, is probably explained by the advancement of the process in the Mediterranean area that dramatically affects countries with extensive coastlines like Italy and Greece[14].

Country Block 5 includes countries in the southern hemisphere, with a total population of 713 million inhabitants, and 13% of the global publications on

[12] ACEA Vehicles in Use report (2022), see also: https://www.acea.auto/figure/average-age-of-eu-vehicle-fleet-by-country/

[13] European Environment Agency data: https://www.eea.europa.eu/data-and-maps/dashboards/air-quality-statistics

[14] European Environment Agency data: https://www.eea.europa.eu/data-and-maps/figures/coastal-erosion-patterns-in-europe-1

climate change (on average, 35 publications per million inhabitant). Like for CB2, a major specialization is in food-related research topics: from studies on food security to research on livestock or maize and coffee cultivation. Another important area of research is the impact of climate change on public health, particularly on the diffusion of malaria and the emergence of zoonotic diseases. A third area regards local adaptation policies, the enactment of protected areas to preserve biodiversity and the study of forestry management. More region-specific topics revolve around the Amazon Forest and the presence of indigenous communities.

The focus on food is arguably due, like in the case of CB2, to the high-level food vulnerability (see above). The rest of the areas of specialization are instead clearly related to the geography of the countries in this block. The presence of some of the areas with the highest biodiversity in the world[15] – like the Amazon Forest – increases the efforts in preserving the wildlife. At the same time, tropical regions with a high wildlife biodiversity display a high risk for the emergence of new zoonotic diseases (Allen et al. 2017).

1.4.4 Forces affecting countries' specialization.

In section 2.1 we discussed and anticipated some factors that may affect countries' specialization. In a first place, it is important to note that there is indeed a high degree of differentiation and specialization, leading to five main country blocks. The extent of variation between countries' specialization is evident from Figure 8, which reports the cosine similarity among the countries' specialization profile (i.e., the cosine similarity among the rows of the matrix represented in Figure 6).[16]

We expected that the specific environmental challenges faced by countries, directly and indirectly drive an interest into specific topics exploring the nature,

[15] National Biodiversity Index, available at: https://www.cbd.int/gbo1/annex.shtml

[16] The cosine similarity is a measure of similarity among vectors. Vectors pointing in the same direction (i.e. countries with the same specialization profile) have similarity = 1, orthogonal vectors have 0, while vectors pointing in opposite direction have -1. In Figure 8 the scale of color is bounded at 0 to ease the visualization.

causes and possible solution for those challenges, and that countries facing similar challenges will display a similar research specialization. The results corroborate this perspective: country blocks' common drivers of specialization are often their shared challenges, such as extreme precipitation and floods (CB1), food (CB2 and CB5), sustainable transportation (CB4).

At the same time, we also observe, as expected, that the research focus is affected by the level of available resources. Less affluent CBs focus on applied topic, local problems' causes and mitigation strategies, whereas basic research into the phenomenon of climate change and its global solutions, such as policies and international negotiations, is comparatively more important in affluent western countries (CB3).

Finally, since collaboration affects topics' selection and a common interest drive collaboration, then countries with a strong level of collaboration should display a similar specialization. This is not confirmed. Figure 9 presents the co-publication matrix between the countries in our sample and shows that for most countries' the most common co-publication partners are the same, namely the United States (12.3% of all collaborative publications), the United Kingdom (8.8%) and Germany (7.2%). However, this does not lead to a similar specialization to those countries. Rather, it emerges a quite different picture from the topic specialization pattern (Figure 6) and the topic cosine similarity (Figure 8). Hence, collaboration does not have a decisive impact on topic specialization, nor vice versa.

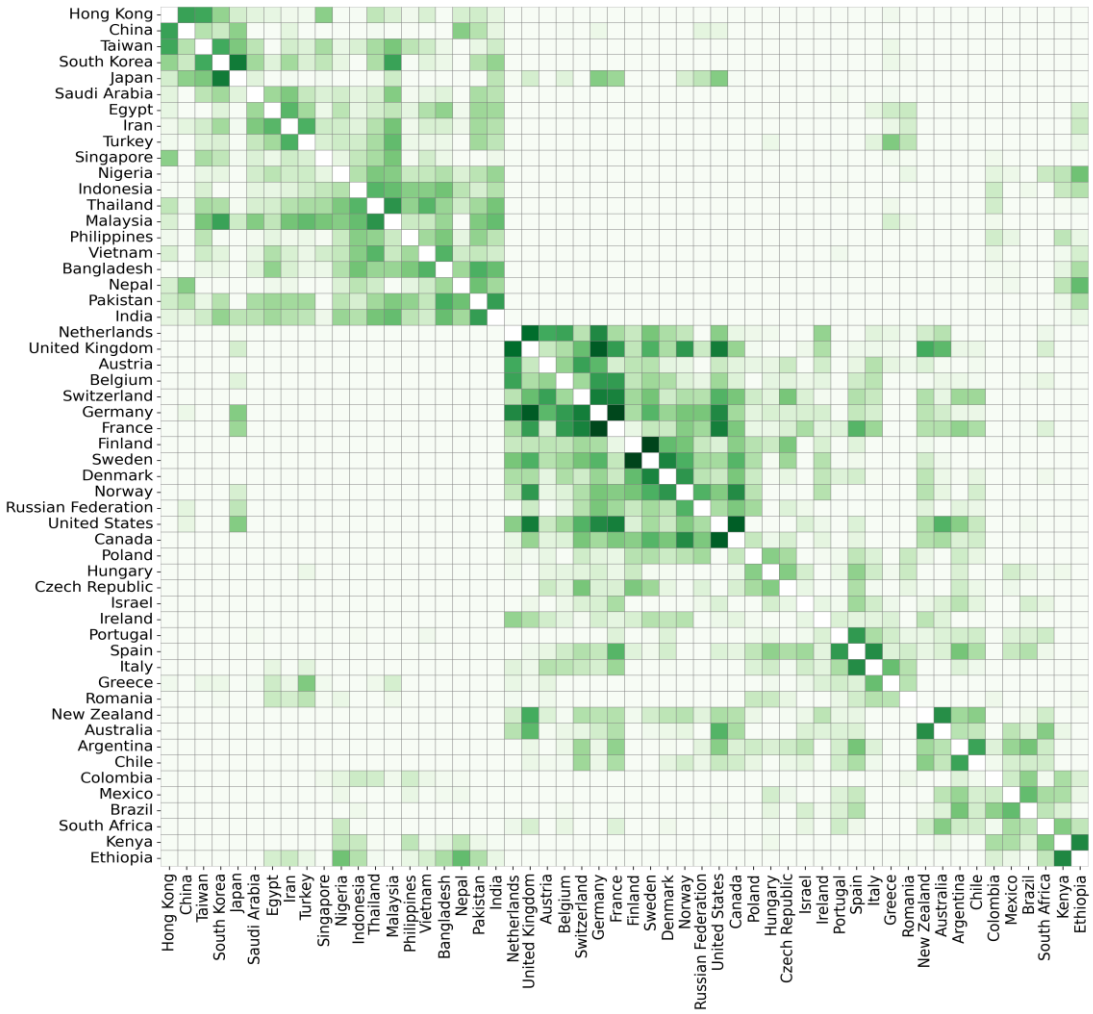

Figure 8: Cosine similarity of the countries' research profile

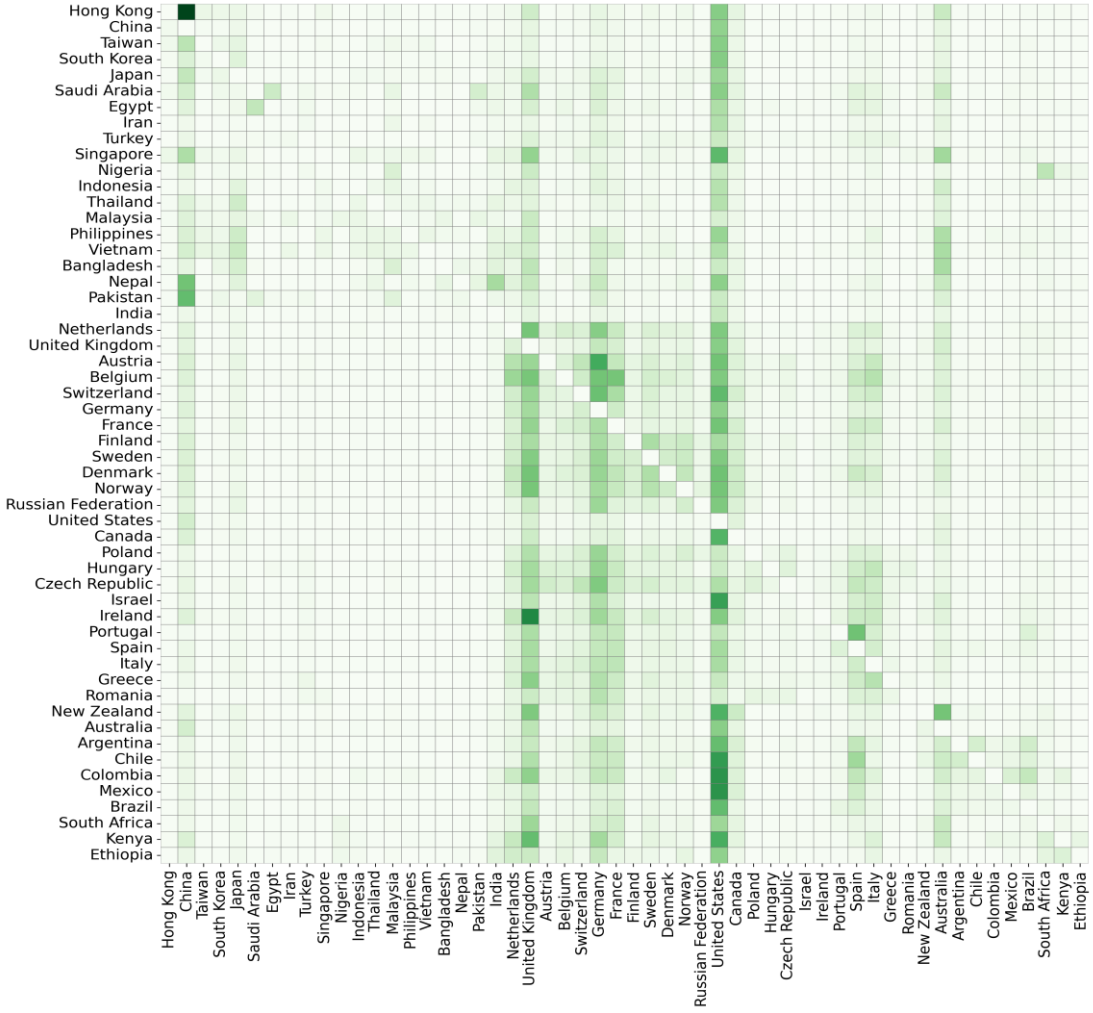

Figure 9: Fraction of scientific collaboration among countries (row normalized)

1.5 Discussion

This article developed a bibliometric overview of climate change research to address two main limitations of existing analyses of this kind in an integrated fashion. First, existing overviews are limited to the identification of main topical trends, hiding the large variety of research topics. In addition, they either ignore the geographical dimension or focus only on major scientific players, implicitly assuming that these are representative of the global distribution of scientific topics. These are important limitations of bibliometric overviews, especially for climate change research, since most of the world's population live in countries that are heavily affected by climate change but have a relatively small scientific production, so that their topics of interest might be neglected.

Therefore, the article aimed to provide a fine-grained overview of climate change research, as well as to explore how climate change research specialization varies across countries and what factors seemingly drives specialization. It did so, by employing a combination of state-of-the-art language modelling techniques to analyse the abstracts of 193,471 publications produced from 1990 to 2020.

Before discussing the empirical results, some choices and limitations should be addressed. First, the restriction of the sample to 54 countries and 472 research topics is based on two thresholds that – despite being adopted to ensure the inclusion of cases with a reasonable number of observations – are somehow arbitrary. Another limit to be considered is the restriction of the sample to publications in English indexed on Scopus; despite this choice, we were able to observe a wide variety of research topics, and our main claims about the variety of topics and country variations would not be weakened in the event of an even higher variety resulting from the inclusion of publications from other sources or in other languages. Future works could expand on our results by exploring country-level specialization in local languages.

The empirical analysis revealed that scientific production on climate change has increased drastically over time – in the last two decades from one per thousand to almost one percent of the total scientific production – and became less concentrated in the most productive countries. Countries that are more vulnerable to climate change hazards tend to be less productive of climate change research –

both in absolute and per capita terms – not due to a lower interest on climate change, but because they typically produce less publications in general. The relative importance of climate change research varies remarkably across countries, from 0.2% to over 3% of their research output, but there is not a clear geographical pattern nor a clear relationship with their vulnerability to climate change.

The analysis identified 472 research topics and found remarkable differences in the countries' research foci. Five blocks of countries with a similar specialization emerged. Country blocks' common drivers of specialization are often their shared challenges, such as extreme precipitation and floods (CB1- East Asia), food (CB2 - Southeast Asia, Indian Subcontinent, the Middle East, Egypt, and Nigeria - and CB5 - Southern hemisphere), sustainable transportation (CB4 - South and East European countries). The research focus is also affected by the level of available resources: less affluent CBs focus on applied topic, local problems' causes and mitigation strategies, whereas basic research into the phenomenon of climate change and its global solutions, such as policies and international negotiations, is comparatively more important in affluent western countries (CB3 – Northern part of the Northern hemisphere). This finding is in line with the evidence provided by Pasgaard and Strange (2013) on a narrower corpus of publications.

It is also interesting to note that, while most countries in our sample collaborate strongly with leading scientific countries like the US, UK, and Germany, nevertheless this is not associated to a similar specialization. This is an important finding for studies of science and for science policy. Namely, strong scientific collaborations between central and peripheral countries do not necessarily drive what topics are studied in the latter, and peripheral countries can develop a distinct research focus. At the same time, these countries typically have limited resources to address these challenges and host several billion people. Hence, a further research policy implication is that leading scientific countries should possibly contribute even more to addressing such issues.

A related aspect highlighted by this study is that the uneven geographical spread of the topics represents a risk for bibliometric overviews that only consider the countries with the greatest scientific output and/or the main topical trends. These might systematically underestimate the importance of research areas concentrated

in less productive countries and overlook the diverse and original research focus of many developing countries (see also Pasgaard and Strange 2013). This issue is particularly problematic in the case of climate change research, given the necessity to develop global strategies that consider also the challenges faced by countries with relatively low scientific output. For example, previous research has shown that developing countries are underrepresented among IPCC reviewers (Palutikof et al. 2023). A broader and more detailed approach to bibliometric overviews can help in highlighting these issues and restoring the balance.

Finally, this study represents an example of the benefits that recently developed techniques in natural language processing – still not widely adopted in scientometric research – might bring to the mapping of the scientific literature, especially in the case of big and fast-growing fields like climate change research.

1.6 Conclusions

In this article, we provide evidence of a progressive decentralization of climate change research from leading countries to developing ones. This process takes place along a diversification of the research specialization at the country level, with developing countries focusing more on applied topics and prominent ones on basic research and global policies. Awareness of these developments is crucial for the public discourse on climate change research in order not to underplay the role of developing countries and the topics they prioritize.

Future developments of this work might provide information on the temporal evolution of the specializations identified in this study, unveil the growth dynamic of research topics in relation to climatic challenges, investigate the role of specific institutions in developing country-level research foci and analyse the representation of countries and research topics in the narrative of policy documents.

References

Aleixandre-Benavent, R., Aleixandre-Tudó, J. L., Castelló-Cogollos, L., & Aleixandre, J. L. (2017). Trends in scientific research on climate change in agriculture and forestry subject areas (2005). *Journal of Cleaner Production*, *147*, 406–418. https://doi.org/10.1016/j.jclepro.2017.01.112

Allen, T., Murray, K. A., Zambrana-Torrelio, C., Morse, S. S., Rondinini, C., Di Marco, M., Breit, N., Olival, K. J., & Daszak, P. (2017). Global hotspots and correlates of emerging zoonotic diseases. *Nature Communications*, *8*(1), 1124. https://doi.org/10.1038/s41467-017-00923-8

Belter, C. W., & Seidel, D. J. (2013). A bibliometric analysis of climate engineering research: Bibliometric analysis of climate engineering research. *Wiley Interdisciplinary Reviews: Climate Change*, *4*(5), 417–427. https://doi.org/10.1002/wcc.229

Braunschweiger, D., & Ingold, K. (2023). What drives local climate change adaptation? A qualitative comparative analysis. *Environmental Science & Policy*, 145, 40–49. https://doi.org/10.1016/j.envsci.2023.03.013

Callaghan, M. W., Minx, J. C., & Forster, P. M. (2020). A topography of climate change research. *Nature Climate Change*, *10*(2), 118–123. https://doi.org/10.1038/s41558-019-0684-5

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining* (Vol. 7819, pp. 160–172). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

Carozza, D. A., & Boudreault, M. (2021). A Global Flood Risk Modeling Framework Built With Climate Models and Machine Learning. *Journal of Advances in Modeling Earth Systems*, *13*(4). https://doi.org/10.1029/2020MS002221

Chen, C., Noble, I., Hellmann, J., Coffee, J., Murillo, M., & Chawla, N. (2015). *University of Notre Dame Global Adaptation Index - Country Index Technical Report*.

Clark, B. R. (1983). *The higher education system: Academic organization in cross-national perspective*. Univ. of California Press.

Engels, A., & Ruschenburg, T. (2008). The uneven spread of global science: Patterns of international collaboration in global environmental change research. *Science and Public Policy*, *35*(5), 347–360. https://doi.org/10.3152/030234208X317160

Fang, X., Jingwei, L. & Qun, M. (2023) Integrating green infrastructure, ecosystem services and nature-based solutions for urban sustainability: A comprehensive literature review. *Sustainable Cities and Society*, (Vol. 98, p. 104843). Elsevier BV. https://doi.org/10.1016/j.scs.2023.104843

Franceschet, M. (2011). Collaboration in computer science: A network science approach. *Journal of the American Society for Information Science and Technology*, *62*(10), 1992–2012. https://doi.org/10.1002/asi.21614

Freeman, R. B., & Huang, W. (2014). Collaboration: Strength in diversity. *Nature*, *513*(7518), 305–305. https://doi.org/10.1038/513305a

Fu, H.-Z., & Waltman, L. (2022). A large-scale bibliometric analysis of global climate change research between 2001 and 2018. *Climatic Change*, *170*(3-4), 36. https://doi.org/10.1007/s10584-022-03324-z

Grieneisen, M. L., & Zhang, M. (2011). The current status of climate change research. *Nature Climate Change*, *1*(2), 72–73. https://doi.org/10.1038/nclimate1093

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. https://doi.org/10.48550/ARXIV.2203.05794

Haunschild, R., Bornmann, L., & Marx, W. (2016). Climate Change Research in View of Bibliometrics. *PLOS ONE*, *11*(7), e0160393. https://doi.org/10.1371/journal.pone.0160393

Hellsten, I., & Leydesdorff, L. (2016). The construction of interdisciplinarity: The development of the knowledge base and programmatic focus of the journal

Climatic Change , 1977-2013. *Journal of the Association for Information Science and Technology*, *67*(9), 2181–2193. https://doi.org/10.1002/asi.23528

Hoffmann, R., Muttarak, R., Peisker, J., & Stanig, P. (2022). Climate change experiences raise environmental concerns and promote Green voting. *Nature Climate Change*, *12*(2), 148–155. https://doi.org/10.1038/s41558-021-01263-8

Huang, L., Chen, K., & Zhou, M. (2020). Climate change and carbon sink: A bibliometric analysis. *Environmental Science and Pollution Research*, *27*(8), 8740–8758. https://doi.org/10.1007/s11356-019-07489-6

Iorio, R., Labory, S., & Rentocchini, F. (2017). The importance of pro-social behaviour for the breadth and depth of knowledge transfer activities: An analysis of Italian academic scientists. *Research Policy*, *46*(2), 497–509. https://doi.org/10.1016/j.respol.2016.12.003

IPCC (2023). *Climate Change 2023: Synthesis Report.* A Report of the Intergovernmental Panel on Climate Change. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland, (in press)

Jankó, F., Móricz, N., & Papp Vancsó, J. (2014). Reviewing the climate change reviewers: Exploring controversy through report references and citations. *Geoforum*, *56*, 17–34. https://doi.org/10.1016/j.geoforum.2014.06.004

Jankó, F., Papp Vancsó, J., & Móricz, N. (2017). Is climate change controversy good for science? IPCC and contrarian reports in the light of bibliometrics. *Scientometrics*, *112*(3), 1745–1759. https://doi.org/10.1007/s11192-017-2440-9

Janssen, M. A., Schoon, M. L., Ke, W., & Börner, K. (2006). Scholarly networks on resilience, vulnerability and adaptation within the human dimensions of global environmental change. *Global Environmental Change*, *16*(3), 240–252. https://doi.org/10.1016/j.gloenvcha.2006.04.001

Jappe, A. (2007). Explaining international collaboration in global environmental change research. *Scientometrics*, *71*(3), 367–390. https://doi.org/10.1007/s11192-007-1676-1

Li, F., Zhou, H., Huang, D.-S., & Guan, P. (2020). Global Research Output and Theme Trends on Climate Change and Infectious Diseases: A Restrospective Bibliometric and Co-Word Biclustering Investigation of Papers Indexed in PubMed (1999). *International Journal of Environmental Research and Public Health*, *17*(14), 5228. https://doi.org/10.3390/ijerph17145228

Li, W., & Zhao, Y. (2015). Bibliometric analysis of global environmental assessment research in a 20-year period. *Environmental Impact Assessment Review*, *50*, 158–166. https://doi.org/10.1016/j.eiar.2014.09.012

McInnes, L., Healy, J., & Astels, S. (2017). Hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. https://doi.org/10.48550/ARXIV.1802.03426

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Merton, R. K. (1974). *The sociology of science: Theoretical and empirical incvestigations* (4. Dr.). Univ. of Chicago Pr.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404–409. https://doi.org/10.1073/pnas.98.2.404

Palutikof, J. P., Boulter, S. L., Field, C. B., Mach, K. J., Manning, M. R., Mastrandrea, M. D., Meyer, L., Minx, J. C., Pereira, J. J., Plattner, G.-K., Ribeiro, S. K., Sokona, Y., Stadler, F., & Swart, R. (2023). Enhancing the review process in global environmental assessments: The case of the IPCC. *Environmental Science & Policy*, 139, 118–129. https://doi.org/10.1016/j.envsci.2022.10.012

Pasgaard, M., & Strange, N. (2013). A quantitative analysis of the causes of the global climate change research distribution. *Global Environmental Change*, *23*(6), 1684–1693. https://doi.org/10.1016/j.gloenvcha.2013.08.013

Rana, I. A. (2020). Disaster and climate change resilience: A bibliometric analysis. *International Journal of Disaster Risk Reduction*, *50*, 101839. https://doi.org/10.1016/j.ijdrr.2020.101839

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. https://doi.org/10.48550/ARXIV.1908.10084

Sangam, S. L., & Savitha, K. S. (2019). Climate change and global warming : A scientometric study. *COLLNET Journal of Scientometrics and Information Management*, *13*(1), 199–212. https://doi.org/10.1080/09737766.2019.1598001

Schilling, M. A., & Phelps, C. C. (2007). Interfirm Collaboration Networks: The Impact of Large-Scale Network Structure on Firm Innovation. *Management Science*, *53*(7), 1113–1126. https://doi.org/10.1287/mnsc.1060.0624

Schwechheimer, H., & Winterhager, M. (1999). Highly dynamic specialities in climate research. *Scientometrics*, *44*(3), 547–560. https://doi.org/10.1007/BF02458495

Sormani, E., & Uude, K. (2022). Academics' prosocial motivation for engagement with society: The case of German academics in health science. *Science and Public Policy*, *49*(6), 962–971. https://doi.org/10.1093/scipol/scac042

Stanhill, G. (2001). *The Growth of Climate Change Science: A Scientometric Study*. 10.

Stephan, P. E. (1996). The Economics of Science. *Journal of Economic Literature*, *34*(3), 1199–1235. https://www.jstor.org/stable/2729500

Vasileiadou, E., Heimeriks, G., & Petersen, A. C. (2011). Exploring the impact of the IPCC Assessment Reports on science. *Environmental Science & Policy*, *14*(8), 1052–1061. https://doi.org/10.1016/j.envsci.2011.07.002

Wang, B., Pan, S.-Y., Ke, R.-Y., Wang, K., & Wei, Y.-M. (2014). An overview of climate change vulnerability: A bibliometric analysis based on Web of Science database. *Natural Hazards*, *74*(3), 1649–1666. https://doi.org/10.1007/s11069-014-1260-y

Wang, Z., Zhao, Y., & Wang, B. (2018). A bibliometric analysis of climate change adaptation based on massive research literature data. *Journal of Cleaner Production*, *199*, 1072–1082. https://doi.org/10.1016/j.jclepro.2018.06.183

Whitley, R. (2000). *The intellectual and social organization of the sciences*. Oxford Univ. Press.

Zhang, C., Bu, Y., Ding, Y., & Xu, J. (2018). Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *Journal of the Association for Information Science and Technology*, *69*(1), 72–86. https://doi.org/10.1002/asi.23916

Zhang, W., Zhou, T., Zou, L., Zhang, L., & Chen, X. (2018). Reduced exposure to extreme precipitation from 0.5 less warming in global land monsoon regions. *Nature Communications*, *9*(1), 3153. https://doi.org/10.1038/s41467-018-05633-3

Appendix

Sources:

ND-GAIN Vulnarability (Chen et al. 2015), available at: https://gain.nd.edu/our-work/country-index/

Population data from the World Bank. Available at: https://data.worldbank.org/indicator/SP.POP.TOTL

Taiwan population retrieved from: https://worldpopulationreview.com/countries/taiwan-population

Country	Country Block	ND-GAIN Vulnerability 2020	Population in 2020	Percentage of country publications on climate change	Climate change publications	Climate change publications per million people
Hong Kong	1		7,481,000	0.36	1,215	162.14
China	1	0.40	1,411,100,000	0.38	27,225	19.28
Taiwan	1		23,821,464	0.27	1,924	80.77
South Korea	1	0.38	51,836,239	0.29	3,604	69.45
Japan	1	0.38	126,261,000	0.24	6,905	54.36
Saudi Arabia	2	0.41	35,997,107	0.35	841	23.36
Egypt	2	0.44	107,465,134	0.28	699	6.49
Iran	2	0.39	87,290,193	0.29	1,830	20.96
Turkey	2	0.35	84,135,428	0.26	1,661	19.72
Singapore	2	0.39	5,685,807	0.28	874	153.72
Nigeria	2	0.50	208,327,405	0.72	794	3.81
Indonesia	2	0.45	271,857,970	0.76	1,561	5.73
Thailand	2	0.44	71,475,664	0.53	1,112	15.50
Malaysia	2	0.38	33,199,993	0.53	1,855	55.84
Philippines	2	0.46	112,190,977	1.56	641	5.70
Vietnam	2	0.48	96,648,685	0.95	749	7.75
Bangladesh	2	0.54	167,420,951	1.35	843	5.04
Nepal	2	0.52	29,348,627	2.91	523	17.82
Pakistan	2	0.53	227,196,741	0.57	1,124	4.95
India	2	0.51	1,396,387,127	0.37	7,250	5.18
Netherlands	3	0.35	17,441,500	0.70	7,051	401.11
United Kingdom	3	0.30	67,081,000	0.68	22,838	337.71
Austria	3	0.28	8,916,864	0.76	3,103	346.87
Belgium	3	0.35	11,538,604	0.61	3,392	293.10
Switzerland	3	0.26	8,638,167	0.77	5,794	667.97
Germany	3	0.29	83,160,871	0.50	15,938	190.99
France	3	0.31	67,571,107	0.46	10,223	150.61
Finland	3	0.31	5,529,543	0.92	3,301	594.08
Sweden	3	0.30	10,353,442	0.88	6,104	587.82
Denmark	3	0.35	5,831,404	0.89	3,742	639.81
Norway	3	0.26	5,379,475	1.39	4,606	853.43
Russian Federation	3	0.35	144,073,139	0.22	2,898	20.02
United States	3	0.33	331,501,080	0.47	57,054	170.62

Canada	3	0.30	38,037,204	0.78	14,121	369.11
Poland	4	0.33	37,899,070	0.27	1,972	51.85
Hungary	4	0.37	9,750,149	0.41	843	86.15
Czech Republic	4	0.30	10,697,858	0.42	1,556	144.80
Israel	4	0.32	9,215,100	0.24	960	103.63
Ireland	4	0.32	4,985,382	0.54	1,138	228.07
Portugal	4	0.33	10,297,081	0.76	2,577	250.07
Spain	4	0.30	47,365,655	0.56	8,290	174.83
Italy	4	0.32	59,438,851	0.42	7,813	131.24
Greece	4	0.33	10,698,599	0.53	1,747	162.92
Romania	4	0.41	19,265,250	0.39	881	45.73
New Zealand	5	0.31	5,090,200	1.00	2,492	486.62
Australia	5	0.32	25,655,289	1.07	15,343	595.67
Argentina	5	0.41	45,376,763	0.59	1,335	29.33
Chile	5	0.32	19,300,315	0.74	1,284	66.42
Colombia	5	0.42	50,930,662	0.62	769	15.10
Mexico	5	0.42	125,998,302	0.53	1,874	14.87
Brazil	5	0.40	213,196,304	0.40	4,276	20.03
South Africa	5	0.42	58,801,927	1.00	3,063	51.99
Kenya	5	0.53	51,985,780	2.45	972	18.66
Ethiopia	5	0.56	117,190,911	1.88	604	5.15

Table A: Detailed information on the countries in the sample

2. Investigator-driven or academic elite: preferential selection of research topics in the European Research Council's funding decisions

This chapter is the result of a collaborative effort among Carlo Debernardi, Marco Seeber and Fredrik Niclas Piro. It has been submitted as a research article at *Research Policy*, receiving a "revise and resubmit" judgement.

Abstract

The European Research Council (ERC) is one of the most important research funding schemes of the European Union supporting investigator-driven, excellent, frontier research. The ERC governing bodies have fiercely defended the "bottom-up" nature of the scheme and opposed any attempt to promote specific research topics "top-down". However, the selection process is not necessarily topic neutral. Proposals are evaluated by excellent scientists, namely an academic elite who practically decide what research will be funded. This article investigates what the research topics of ERC proposals are, how the proposals' research topics vary across framework programs, countries, type of organization and grant, and whether a proposal's research topic affects the chances to be funded. We use state-of-the-art natural language processing techniques to examine the textual content of the abstract of 91,273 proposals for Starting, Consolidator, and Advanced Grants in the Seventh Framework Program (FP7) and Horizon 2020 (H2020) – i.e., 95.3% of all the proposals, identifying 188 unique research topics. The proportion of proposals in each topic changed substantially from FP7 to H2020 programs, with one in four increasing or decreasing by more than 20% in relative terms. The research topic of a proposal is an important predictor of success, with 37.4% of the proposals having significantly higher or lower chances of being funded due to their topic. At the same time, we estimate that the research topics of the funded proposals were similar to a scenario in which the topics did not matter (82%

overlap). Overall, the results hint to a notable impact of the selection process on what is researched, which nevertheless does not overturn its bottom-up orientation.

2.1 Introduction

Decisions about what to research shape the direction of science and the production of knowledge. These decisions are mostly up to individual scientists or research groups, affected by individual motivations, like curiosity and pleasure of solving puzzles (Merton, 1974), the wish to obtain recognition from colleagues (Merton, 1974), funds and career progression (Stephan, 1996) as well as having a positive impact through one's work (Iorio, Labory, and Rentocchini, 2017; Sormani and Uude, 2022). Scientists' choices are also affected by a multiplicity of contexts in which they are embedded. In the first place, the disciplinary community to which a scientist belongs influences perceptions about the most important questions and goals (Clark, 1983; Whitley, 2000). Scientists are also embedded into different national systems, that differ in terms of economic, social, political, cultural, and geographical characteristics, that can impact on motivations, level of resources, and the salience of specific problems (Debernardi et al., 2024).

In recent decades, higher education reforms have also attempted to increase university leaders' and government steering power through new policy and managerial instruments. While a common perception is that these reforms have weakened the academic profession, Musselin (2013) argued instead that they reconfigured the academic profession because they created evaluative processes which rely on peer review by academics themselves. Notably, an increasing share of resources for research is managed by funding agencies and foundations, through competitive schemes relying on peer review processes. Funding agencies were expected to be more accountable but turned into a "double-edged sword for policymakers" (Braun, 1993), because they are often controlled by the academic community and their existence make direct political influence on academic issues less legitimate and opposed by academics in the agencies, who demand respect of their autonomy (Musselin, 2013).

In turn, competitive funding has increased the influence and power of an academic elite in charge of deciding who and what research will be funded (Whitley, 2007;

Musselin, 2013; Hoening, 2017). However, there is scant empirical evidence on the extent to which grant selection affects what research is supported. In fact, several studies have explored different factors that predict a grant proposal's success, but one important factor that has been overlooked is the research topic of the proposal itself.

The main objective of this article is therefore to fill this gap by exploring to what extent the evaluation process of project proposals affects what research topics are supported. Or, in other words, is the research topic of a proposal an important predictor of its success?

We consider a funding instrument that aims to support topics proposed by scientists themselves, namely the European Research Council (ERC) grants. This is the most important funding scheme of the European Union supporting investigator-driven, excellent, frontier research, across all fields. We examine the content of 91,237 proposals submitted during the FP7 and Horizon 2020 programs (2007-2020) and investigate what the research topics of ERC proposals were, how the research topics varied across framework programs, countries, type of organization and grant, and whether a proposal's research topic affects its chances to be funded.

The following section discusses the literature on the factors that affect research proposal success, as well as how the mission and evaluation process of the ERC can affect the type of research supported. We then present the data and the methods of analysis, followed by the results section. The final section discusses the article's main findings and their implications for scholars and policy makers interested in research policy and research grant evaluation procedures.

2.2 Theoretical framework

2.2.1 Factors affecting success in funding program evaluation.

Research exploring the factors predicting success in funding programs has examined the characteristics of the consortia of applicants or their organizations, the traits of the proposer and of the proposal itself. We review this literature and identify a gap, as no studies have explored the research topic of the proposal.

When considering the factors that predict success, it is also important to bear in mind that the characteristics of the grant affect in the first place *who is going to apply*, and indirectly, the chances of success (Seeber et al., 2022a). ERC applicants tend to have an above-average scientific output and impact (Neufeld et al., 2013), because the highly selective nature of ERC grants discourage scientists with a weak scientific track record. The sum of the grant can also affect the propensity to apply. While the greater the sum, the greater the attractiveness, it is also true that *ceteris paribus*, a grant will be more attractive for researchers in countries lacking alternative source of funding (Piro et al., 2024), and where the cost of the research personnel is lower (Seeber et al., 2022a). That said, the endowment and reputational payoff of an ERC grant is very large and then likely to be highly and similarly attractive in all countries.

Several studies have explored which *organizational characteristics* are associated with participation in EU Framework programs and found that they are dominated by institutions with strong academic reputation and scientific productivity (Geuna, 1998; Henriques et al., 2009; Enger and Castellacci, 2016; Lepori et al., 2015), as well as low teaching load and high proposal intensity (Piro et al. 2020), and that proposals from consortia and organizations with high levels of experience and reputation have greater chances of success (Enger and Castellacci, 2016; Wanzenböck et al., 2020).

The assessment of grant proposals is influenced by the *applicants' traits*, such as past performance, either in terms of scientific production (Van den Besselaar and Leydesdorff, 2009; Van den Besselaar et al., 2018) or previous grant awards (Tamblyn et al., 2018; Bol et al., 2018). As to ERC Starting Grants, Neufeld et al. (2013) found that in most panels there was not a significant difference in productivity between funded and non-funded proposals. Veugelers et al. (2022) examined whether the ERC selected researchers with a track record of conducting risky research. They found that applicants with a history of risky research were less likely to be selected for funding, especially for early career applicants.

Several scholars have explored whether the *gender* of the applicant matters. The gender of the applicants was not a significant predictor of success in studies on the Austrian Science Fund, the Dutch NOW, and ERC Proof of Concept (Mutz et al.,

2012; Albers, 2015; Volker and Steenbeek, 2015; Seeber et al., 2022a), whereas Van den Besselaar and Leydesdorff (2009) found that female proponents to the Netherlands Research Council for the Economic and Social Sciences are more likely to be funded *vis-à-vis* their past scientific performance and reviewers' evaluations and Sandström and Hällsten (2008) that – controlling for other co-variates – proposals to the Swedish Research Council from female proponents receive a "bonus" of around nine percentage points.

Proposals received higher scores when the proponents had the *same affiliation* as their reviewers or panel members (in Sweden: Sandström and Hällsten, 2008; Korea: Jang et al., 2017; Canada: Tamblyn et al., 2018) and, in the ERC StG 2014, when a panelist was from the institution where the applicant had agreed to use the grant (Mom & van den Besselaar, 2021).

Some studies examined how the *content of the proposal* could affect the chances of success. Considering 369 proposals to the National Institutes of Health (NIH) in the US, Boyack et al. (2018) found that proposals were more likely to be funded when being clearly articulated and when the topic of the references overlapped with the applicant's prior publications. Other studies have found both that evaluators tend to give lower scores to research proposals that are closer to their own areas of expertise (Boudreau et al., 2016), and that they provide higher scores (Li, 2017). It is debated whether the interdisciplinarity of a proposal affects the chances of success: in the Australian Research Council's Discovery Program interdisciplinary research proposals had lower success rate (Bromham et al., 2016), while the degree of interdisciplinarity did not affect the success of proposals to the Collaboration in Science and Technology (COST) European program (Seeber et al., 2022b). Van den Besselaar et al. (2018) examined the evaluation of the ERC Starting Grants' panels and found that they focused on the applications' weak points rather than looking for ground-breaking ideas.

Thus far, however, no research has explored the impact of the proposal's *research topic* on its chances of being funded.

2.3 What research is supported by the ERC: mission and evaluation process

2.3.1 ERC goals and principles

The ERC grants aim to support *investigator-driven* excellent *frontier research*, across all fields.

Frontier research encompasses four main aspects (European Commission, 2005; Hörlesberger et al., 2013). First, *novelty*, namely research at the forefront of creating new knowledge and developing new understanding, able to lead to fundamental discoveries and advances in theoretical and empirical understanding, and possibly revolutionary breakthroughs. Second, *risk*, because the path leading to greatest advances in knowledge often implies a high risk of failure (i.e., high risk-high reward). Third, it may generate not only new theoretical knowledge but also *useful* knowledge, entailing a closer connection between science and technology. Fourth, it may transcend established disciplinary boundaries and involve *multi-, inter- or trans-disciplinary* research.

The investigator-driven concept means that the ERC is a bottom-up funding scheme, namely the agency should support researchers with the most exciting ideas, rather than trying to identify top-down priorities. The importance of this principles emerged in a dispute between the ERC's President and the Scientific Council – the ERC's most important governing body, composed of 21 highly reputed scientists from all scientific fields. In January 2020, a new President was appointed: Professor Mauro Ferrari, a reputed scientist in the field of nanomedicine and cancer research, who had spent most of his career at the University of California Berkeley. Carlos Moedas, EU Commissioner for Research, Science, and Innovation, declared:

"*The ERC has a global reputation for funding excellent, curiosity-driven research. (….) With his understanding of the societal value of science, his strong leadership and exceptional communication skills, Prof. Ferrari is the right person to take the ERC and European science to new heights*" (European Commission, 2019).

After a few months, however, Ferrari resigned. In a letter to the Financial Times, he explained that:

"*As it became evident that the pandemic would be a tragedy of possibly unprecedented proportions, I moved that the ERC should establish a special program directed at combating Covid-19. (…) The proposal was rejected unanimously by the governing body of the ERC, (…) based on the notion that the ERC funds "Bottom-Up" research: it does not specify focus areas or its funding objectives, nor does it consider beneficial impact on society as a funding criterion.*" (Financial Times, 2020).

A few days later, the Scientific Council replied that they unanimously asked for Ferrari's resignation, among others, because Ferrari displayed a complete lack of appreciation for the *raison-d'être* of the ERC to support excellent frontier science (European Research Council, 2020).

Ferrari's initiatives challenged more or less explicit founding principles of the ERC grant scheme, regarding *what to research*, namely i) the priority of knowledge over usefulness: research may – but not necessarily must – have immediate practical relevance, ii) autonomy and expertise: the topics are proposed bottom up by the researchers and selected by peer scientists (experts); iii) egalitarianism: disciplines are worth the same, no discipline can have the priority over the others.

2.3.2 The evaluation process

What type of research is supported also depends on how the evaluation process is organized. The ERC selection is conducted by 25 disciplinary panels in three major fields: Physical Sciences and Engineering (PE) – ten subpanels; Life Sciences (LS) – nine subpanels; Social Sciences and Humanities (SSH) – six subpanels. Between 2008-2011 also one Interdisciplinary (ID) panel was in place. Each panel is composed of 11 to 18 internationally recognized scientists and there are two groups of panel members, in charge of the review process in alternating years.

The evaluation consists of two steps. In *step 1* only panel members are involved and evaluate 5-page short proposals and 5-page CVs; the proposals are divided among panel members, who evaluate them remotely. Each panel reviewer is assigned about 30–50 proposals and evaluates them based on their generalist scientific expertise. Each proposal is reviewed by three panel reviewers who have

about a month to complete their reviews and then meet in Brussels for 3 days for panel discussions, ranking of proposals, and assignment of lead reviewers and the selection of external reviewers. Help from other panels is solicited for some of the interdisciplinary proposals. *Step 2* considers longer proposals that passed the first round (15 pages). The evaluation is conducted by panel members with the support of topic-specific external reviewers, invited by each member. After external reviews have been collected, interviews with ranked candidates are conducted in Brussels by all panel members and two officers, followed by panel discussions and the final ranking.

According to Hoenig (2017) peer review is crucial for the cultural legitimacy of the ERC acting as an intermediary organization 'by scientists for scientists', representing the collective self-regulation and intellectual autonomy of science *vis-a-vis* political actors, such as the European Commission. ERC's approach to peer review is based on the idea that 'excellent researchers will recognize excellent researchers and research' (Müller, 2021). According to van den Besselaar, Sandström & Schiffbaenker (2018), however, the criterion of excellence is intentionally vague "*as the ERC wants to give panel members the freedom to do the selection in their own way. The selected panel members are seen as the leading scholars in their fields, and therefore know the best how and what to select*". Their study used e.g., project descriptions, CVs, reviewer scores and review reports of 3030 applicants to the ERC, and from content analyses, concluded that panelists did not differ between past performance and promising new research ideas, and to a little degree focused on high-risk/high-gain groundbreaking ideas.

According to a former member, proposals should be written keeping both the generalists and the specialists in mind, and "*proposal need at least 2-3 strong supporters to pass. Sometimes a simple remark by a candidate tips the balance in the panel discussions*."[17]

Brunet and Müller (2022) examined how members of ERC review panels assess applications in the first, highly competitive step of evaluations and found that

[17] https://tto.ku.edu.tr/news/an-interview-with-ku-researcher-hakan-urey-on-erc-panel-evaluations/

panelists employ four evaluation devices: i) base judgments on applicants' prior achievements; ii) adjust the evaluation of individual applications to the quality of a given set of applications; iii) combine multiple elements to assess the feasibility of proposals and iv) consider the impact of the proposed research on science and society. They argue that this pragmatic way of evaluation is shaped by the need to review many high-quality proposals in a short time with possibly limited expert knowledge, and that it might influence what kinds of research and researchers are selected for funding. Namely, it favors proposals that are easily understood across various fields and knowledge levels. CVs with accepted markers of excellence, such as a continuous and high publication output and a strong record in third-party funding, are favored. Due to high selectivity and unanimity rule, panel reviewers "*set personal preferences aside because they can usually only champion those applications that will be backed by others and advocating for outliers too often could negatively impact their credibility*." (Brunet and Müller, 2022).

In turn, whereas ERC aims to promote ground-breaking research, several elements of the evaluation process engender a certain level of conservatism in the selection process. The evaluation mechanisms also seemingly prevent proposals from being selected due to the specific interests of one or a few panelists. Moreover, ERC disciplinary panels display a similar success rate, which abide to the principles of parity between disciplines. However, what is not known is whether the success rate is also similar across different research topics. Namely, to what extent and how does the selection process affect what research topics are supported, altering the bottom-up research topics mix of the proposals?

2.4 Data

2.4.1 The ERC

The ERC was established in 2007 by the European Commission under the Seventh Framework Program for Research and currently allocates 2 billion euros every year to investigator-driven frontier research projects to be conducted in an organization in the EU or associated country for up to 5 years. ERC grants typically range between 1.5 million € for Starting Grants (2–7 years after PhD), 2.5 million

€ for Consolidator Grants (7–12 years after PhD) and 3.5 million € for Advanced Grants (candidate with significant research achievements in the last 10 years).

The European Commissioner for Innovation, Research, Culture, Education and Youth selects (invites) six high-level scientists as members of an *independent committee,* which has the task to identify the members of the *Scientific Council*, then appointed by the European Commission.[18] The Scientific Council establishes the work program for the implementation of the ERC activities[19] and selects the *Chairs* of the thematic panels. Each Panel's Chair select members of the panel about one year before the panels start their work; to become a member, one must be included in the European Commission's expert evaluator database, often they are part of the ERC database – either as previous recipients or evaluators.

2.4.2 Data

The main data source for this study was proposal data from the European Commission's database eCORDA, used under the *Nordic Institute for Studies in Innovation, Research and Education* (NIFU) license according to the confidentiality rules for Framework Programme data stored in eCORDA. At NIFU the institutional affiliations of all applicants have been standardized, so that the unique number of applicants was reduced from the original 11,349 unstandardized institutional names in eCORDA to 4,979 unique institutional standardized names. For our study we have only included proposals for Starting Grants, Consolidator Grants and Advanced Grants, and considered information only for the principal investigators of the proposals. The final dataset included information from eCORDA, regarding 91,273 proposals submitted to ERC during the FP7 and Horizon 2020 programs (2007-2020), along the variables described in section 3.3. The 91,273 proposals represent 95.3% of all proposals submitted for the three grant types. The remaining 4.7% were excluded from our study because their textual

[18]https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/top-scientists-identify-future-members-european-research-councils-scientific-council-2020-05-19_en

[19] https://erc.europa.eu/about-erc/erc-president-scientific-council

summaries were too short – i.e. have less than 200 characters – making them ineligible for the research topic classification phase (see Section 2.5.1).

2.4.3 Variables

Dependent variable

Status. Whether the proposal was rejected or funded.

Independent variable

Research topic. The research topic of the proposal: 188 topics were identified (see Section 2.5.1). About one in three proposals was not linked to any of the 188 topics. Hence, they were assigned to a separate "mixed" topic, which also represents the benchmark category in the regression analysis to assess the impact of each research topic.

Control variables

The control variables include characteristics of the proposal, and of the country and organization of the applicant. In particular, variables related to the call to which the proposals were submitted – i.e. year of the call, type of project, program, disciplinary panel – account for the variation in the success rate that is not due to the proposals themselves, but rather to factors like the numerosity of applications or other organizational constraints (e.g. avoiding an excessive disparity in success rate between panels). The amount of funding requested can impact the success rate of the proposals in at least two ways: first, by contributing to the organizational constraint of a fixed amount of money to distribute; second, by potentially providing hints to the panel members – i.e. a funding request considered too low might be perceived as not worth or appropriate, while a request higher than the maximum could make the proposal ineligible to be funded. Finally, the type of the host organization, as well as the country and institution, can potentially signal the quality of the proposal and/or the prestige of the applicants.

Call Year. The year of application, from 2007 to 2020.

Project type. Three types of projects: Starting Grant (StG), Consolidator Grand (CoG), and Advanced Grants (AdG).

Framework program. Whether the proposal was submitted to the Seventh Framework Program (FP7) or the Horizon 2020 (H2020) program.

Disciplinary Panel. The disciplinary panel to which the proposal was submitted. The panel information is available for 86% of the proposals; proposals with missing information on the panel were attributed to a "mixed panels" group.

Funding requested. The amount of funding requested, ranging – with few exceptions – from 0.25 million € to 3.5 million €. We normalized the sum requested by dividing it for the maximum allowable amount by funding scheme and then created a dummy variable with ten categories by decile.

Organization. The type of organization to which the scientist was affiliated, either Higher Education institution (HES), Research Center (REC), Private Companies (PRC), Public Organizations (PUB), or Other organizations (OTH).

Country of affiliation: dummy variables.

Institution of affiliation: dummy variables.

2.5 Methods

2.5.1 Identification of the proposals' research topics

The identification of topical trends in textual corpora has experienced dramatic advancements in recent years. Previous methods (e.g. Blei et al., 2003) – based on simple word co-occurrence – represented documents as unordered sets of words and thus struggled to represent contextual information and capture semantic content. Text embedding models proposed to represent individual words as vectors, enabling the representation of semantic similarity (Mikolov et al., 2013). Further technical advancements, with the introduction of the attention mechanism (Vaswani et al., 2017), led to the development of the Transformer architecture, at the basis of contemporary Large Language Models. Models based on this architecture can provide a much finer granularity then previous approaches, and have rapidly become the state-of-the-art in measuring semantic proximity among textual data.

To identify research topics in the proposals we adopted the following procedure, with a pipeline that is similar to the approach used by the BERTopic algorithm (Grootendorst, 2022). First, we composed a textual corpus by combining title and abstract of the proposals. We then embedded the documents in a vector space by using Sentence-BERT (Reimers and Gurevych, 2019), a state-of-the-art pretrained language model. Since the result is very high-dimensional we used UMAP (McInnes and Healy, 2018), a nonlinear dimensionality reduction algorithm, to get a lower dimensional representation of the corpus. In this reduced space we identified dense areas of similar documents via HDBSCAN (Campello et al., 2013, McInnes et al., 2017), a density-based clustering algorithm. Among the advantages of this approach is the fact that the number of topics does not have to be set in advance as a parameter. Rather, it is inferred by the clustering algorithm by identifying dense areas of the embedding space and constrained by a parameter regulating the minimum size of clusters (that we set to 90 proposals). Additionally, unlike in most traditional clustering approaches, observations are not necessarily bound to be grouped into a cluster (thus incorporating noise) but can be left unlabelled. Through this process we obtained 188 clusters of documents-proposals representing research topics, that include 61.5% of the proposals (56,078) that could be clearly allocated to a given research topic.

To support the interpretation of the content of the topics – and in line with Grootendorst (2022) – we identified the five most distinguishing words by cluster by employing a simple TF-IDF (term frequency–inverse document frequency) weighting of words and bigrams (see Appendix – Table B). This scheme, widespread in information retrieval, assigns higher scores to words that are more distinguishing of a certain document (or cluster, in our case) while filtering out those that have a frequency too low to be informative. It must be noticed that this operation was performed as an analytical step and did not impact the classification of proposals into the topics, which was entirely driven by the text embedding and clustering procedures.

2.5.2 Descriptive and inferential analysis

The empirical analysis is organized into two main sections.

In the first section we identify the proposals' research topics, and examine variations over framework programs, across countries, type of organization, type of grant, and the success rate of proposals in different topics.

In the second section we explore to what extent the proposals' research topics affect the chances of success and identify what research topics have the greatest positive or negative impact. The purpose of the inferential analysis is to disentangle a composition effect. Namely, proposals in a given topic may be more successful because, for example, they were submitted to a call with a higher success rate, or by scientists in countries or organizations with a higher success rate. Therefore, we investigate what factors predict a proposal's success, to estimate the importance of the research topics as a predicting factor and what research topics predict a success rate significantly higher or lower than the reference category. We do this by means of a hierarchical Bayesian logistic regression, implemented via the *CmdStanPy*, a Python interface to *Stan* platform for statistical modelling (Stan Dev Team, 2024). This enables us to take advantage of partial pooling in handling the estimation of coefficients with relatively low coverage in terms of observations. This is particularly helpful in this setting since we expect topics to vary substantially in size, as well as countries and institutions to have a skewed distribution in terms of number of proposals or even display complete separation (e.g., countries or institutions with no or all funded proposals).

The model is specified according to the following formulation, where the beta coefficients are fixed to 0 for the reference category, namely, the most frequent category of each covariate.[20]

$$y_n \sim \text{bernoulli}(\text{logit}^{-1}(\mu))$$

$$\begin{aligned}\mu = \; & \alpha + \beta_{\text{topic}} + \beta_{\text{year}} + \beta_{\text{panel}} + \beta_{\text{country}} + \beta_{\text{funding-req}} \\ & + \beta_{\text{action-type}} + \beta_{\text{inst-type}} + \beta_{\text{institution}}\end{aligned}$$

[20] Priors and hyperpriors are reported in the appendix, Figure F.

2.6 Empirical analysis

The 91,273 proposals included in the dataset consisted of: i) 39,511 FP7 proposals and 51,762 H2020 proposals, ii) 45,681 ERC Starting Grant, 19,550 Consolidator Grant, and 26,042 Advanced Grant; iii) 69,626 proposals from scientists affiliated to Higher Education institutions (HES), 20,765 to Research Centers (REC), 464 to Private Companies (PRC), 217 to Public Organizations (PUB), 199 other organizations (OTH), 2 unknown; iv) by macro-disciplinary panel: 37,179 in Physical Sciences and Engineering (PE), 25,768 Life Sciences (LS), 15,369 Social Sciences and Humanities (SSH), 12,859 unknown, 98 Interdisciplinary (ID).

Figure 1 presents the percentage of proposals and grants by country, as well as their success rate, for the 15 countries with the greatest number of proposals and grants: together, they account for 90% of the proposals and 96% of the grants. The country with most proposals (16.7%) was the United Kingdom, which obtained 19.3% of the grants (success rate 13.2%); followed by Germany: 12.3% of proposals, 16.1% of the grants, 14.9% success rate; and Italy: 10.5% of proposals, 4.8% of the grants, 5.2% success rate (full data available in Table A in the appendix). The countries with the highest success rates were Switzerland (21%), Israel (18.2%) and the Netherlands (15.4%).

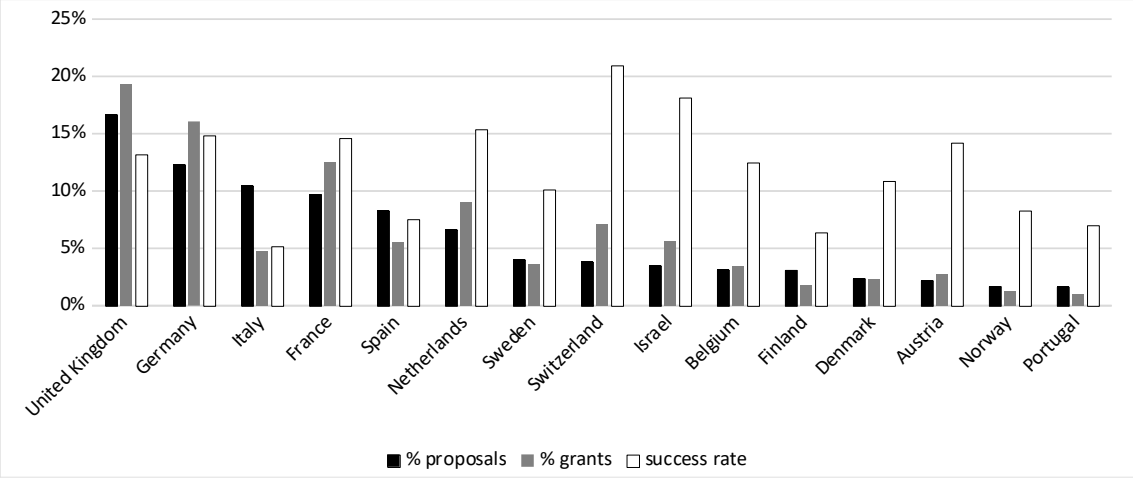

Figure 1: Percentage of proposals and grants and success rate for the largest 15 countries

2.6.1 Descriptive analysis

In terms of size, 176 research topics include between 90 and 707 proposals, eleven topics include between 756 and 1,716 proposals, and one topic includes 4,312 proposals. For 151 out of 188 research topics, more than 85% of proposals were submitted to subpanels within the same macro area.

The proportion of proposals in each research topic varied considerably from FP7 to H2020, with one in four topics increasing or decreasing by more than 20% (see Table B in the appendix). Among the topics with the strongest growth, several are related to renewable energy, such as new materials for photovoltaics (topic 132: from 1 to 115 proposals) and batteries, lithium, electrolyte (topic 133 from 56 to 171). There was also a strong growth of proposals regarding governance and legal copyrights in Artificial Intelligence (topic 172: 20 to 84), music and opera (topic 180: 39 to 113); news, conspiracy, and democracy (topic: 179: 35 to 97).

Proposal topics by country

We explored the similarity between proposals' research topics across countries via the cosine similarity index. Cosine similarity is a measure of similarity among vectors; in this case, the vectors express the proportion of proposals from a country in any given research topic. Vectors pointing in the same direction (i.e. countries with the same proposals' topics profile) have similarity = 1, orthogonal vectors have 0, while vectors pointing in opposite direction have -1.

Due to the substantial differences in size between topics, the proportion of proposals in each topic tend to be similar across countries (see Figure A in the appendix). Hence, we also explored the over/under-representation of topics across countries compared to their average frequency, by standardizing the matrix topic-wise so that each entry represents the weight of the topic in a country, measured in standard deviations (i.e., z-score). The standardized importance of topics varies considerably across countries (see Figure B in the Appendix). Hence, cosine similarity scores are generally low, and only one small group of countries consistently show positive scores, namely Austria, Germany, Israel, France, Switzerland, and United Kingdom (see Figure C in the Appendix).

Proposals' topics by type of organization

Focusing on the proposals from scientists working in the Higher Education Sector (HES) and Research Centres (REC), the share of proposals in each topic is similar (correlation of 0.86) yet the relative frequency of each topic[21] varies considerably. REC scientists submit less proposals on topics in Social Sciences and Humanities, such as topic 164 (keywords: markets, firms, finance, credit, monetary), with 0.43% of the proposals compared to 1.28% for the HES, or topic 163 (keywords: welfare, inequality, labour, political, solidarity): 0.09% compared to 0.21%, whereas several topics in Physics and Astrophysics research are more common in REC proposals, such as neutrino detection 0.56% vs. 0.21% (see Table B appendix).

Proposals' topics by type of grant

The frequency of each topic by type of grant is similar (see Table B appendix). There are a few exceptions. Notably, several topics in Astrophysics and Matter Physics are relatively more common (20-30% more than the average) for Consolidator Grants (COG) proposals, as well as proposals regarding photovoltaics and batteries (topics 131 to 134) (+25-40%) and much less common in Advanced Grants proposals (ADG) (-26% and -37%) (see Table B appendix).

Success rate of proposals in different topics

The average success rate for the proposals allocated to any topic is 12%. Proposals not linked to any specific topic have a success rate of 10.4%. The success rate for proposals within the same topic ranges from 1.2% for research on teacher/s - teaching, education, learners (topic 144) to 25.2% for research on chromatin, chromosome, cohesion, chromosomes, meiotic (topic 90), with a standard deviation of 4.8%. Some topics display a very low or high success rate, namely 16 topics display a success rate 50% lower (<6.0%) than the average, and 19 topics 50% higher (>18.0%) than the average.

[21] Excluding proposals not assigned to any topic.

2.6.2 Inferential analysis

Proposals in different topics display substantial variation in their success rate. As explained in the methodology, this variation may be affected by a compositional effect. Through an inferential analysis we control for several covariates to obtain a more reliable estimate of the impact of the research topic on a proposal's chances of success and to identify what topics have the greatest impact. One limitation of the analysis is that we could not control for some individual researcher' traits, nor – of course – for the intrinsic quality of the proposal.

First, we aim to assess to what extent the research topic affects the chances of a proposal to be funded. The results of the logistic regression predicting whether a proposal was funded reveals that a considerable number of research topics significantly affect the chance of success of a proposal (Table C in the appendix). More specifically, 49 out of 188 topics determine a significantly greater or lower success rate compared to the group of proposals not assigned to any topic, which is taken as a reference category. Numerically speaking, these topics represent 20,958 proposals, or 37.4% of the proposals assigned to any given topic, excluding the reference category of mixed proposals.

Figure 2 presents the regression coefficients on the odds ratio scale[22] for the 49 research topics for which 95% of the posterior distribution is above or below zero. The reference category (with coefficient fixed to 0) is the largest topic-cluster, i.e. the one including proposals not assigned to a specific topic. In the figure, the thinner whiskers represent the 5th and 95th percentile of the posterior distribution of the parameter, while the thicker ones correspond to the 25th and 75th (i.e. quartiles). The dots represent the posterior mean. In turn, there are quite large differences in success rate between topics, also when a compositional effect is considered. Such differences and the order of the topics are similar to the one observed descriptively (see Table B in the appendix).

[22]Thus, for example, looking at the point estimate a proposal in topic 26 has roughly double the chances of being funded than one in the cluster of unlabeled proposals (other things being equal).

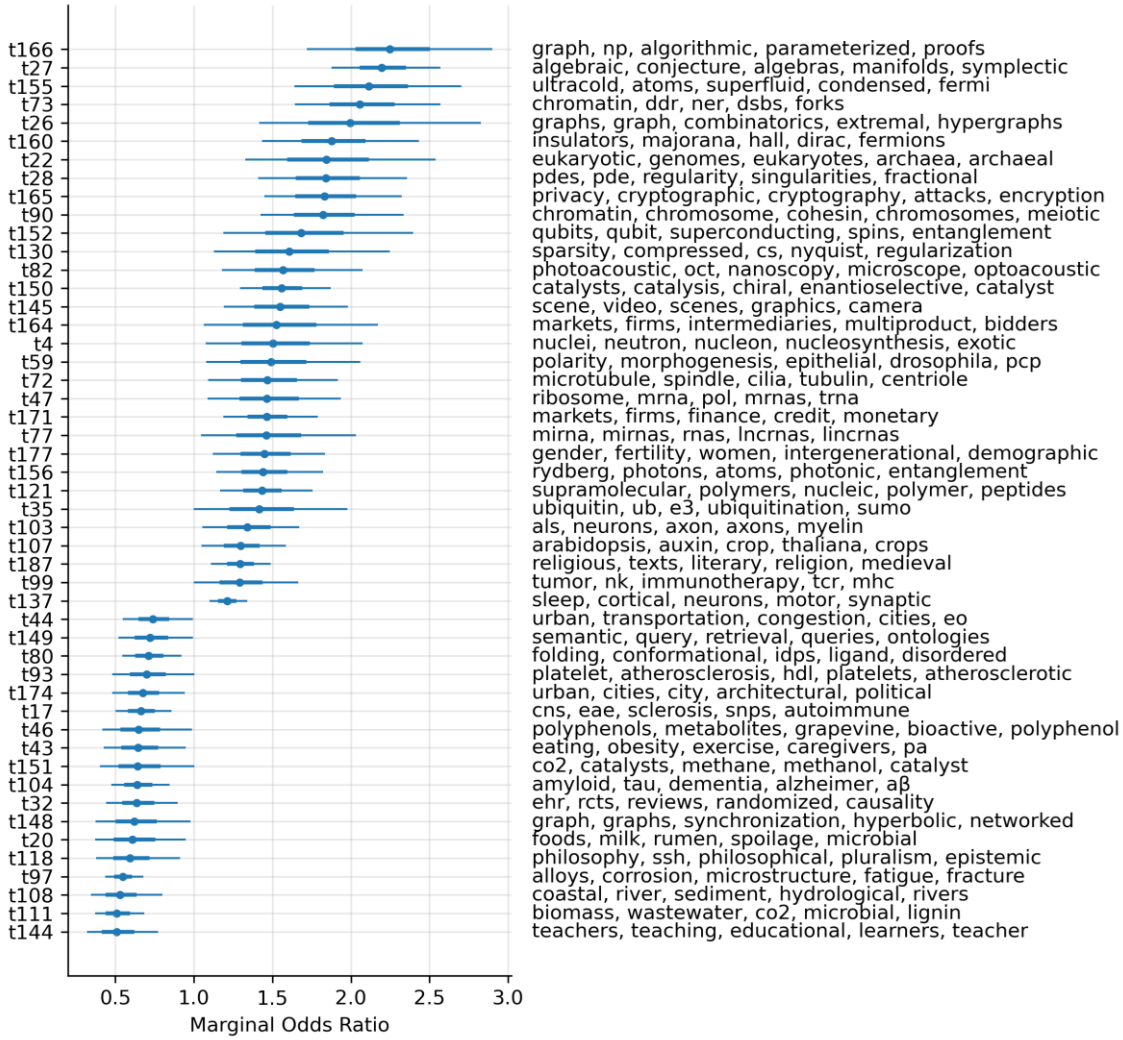

Figure 2: Significant estimated odds ratio of success by topic. Note that the keywords are reported to help the interpretation and are not directly part of the topic identification in this form.

To further explore the importance of the research topic for a proposal's success, we compared the observed distribution of topics among funded proposals with the predicted distribution of topics among funded proposals assuming no effect of the research topic (i.e. imputing a zero coefficient on the logit scale). The overlap of these two distributions is estimated at 82.6%. By comparison, applying the same procedure, the overlap for the country distribution is 73.3%, 43.9% for the organization, and 91.7% for the disciplinary panel variables. We report visual representations of the impact of different countries and panels in the Appendix (Figures D and F).

Finally, we explore changes in research topics' impact on success rate in FP7 and H2020 by running regression analyses separately for the two periods. Figure 3 presents the estimated OR of proposals in different topics in the two frameworks, where the size of the dots expresses the total number of proposals in the entire period. While revealing that the estimated OR changed remarkably for some topics, it is important to notice that this happened for topics with few proposals. Hence, in no case did the impact of the topic turned from significantly negative to positive (or vice versa), and only for 24 topics – that in total representing 7,358 proposals, or 13.1% of the proposals assigned to a specific topic-, the impact turned from non-significant to significant (or vice-versa).

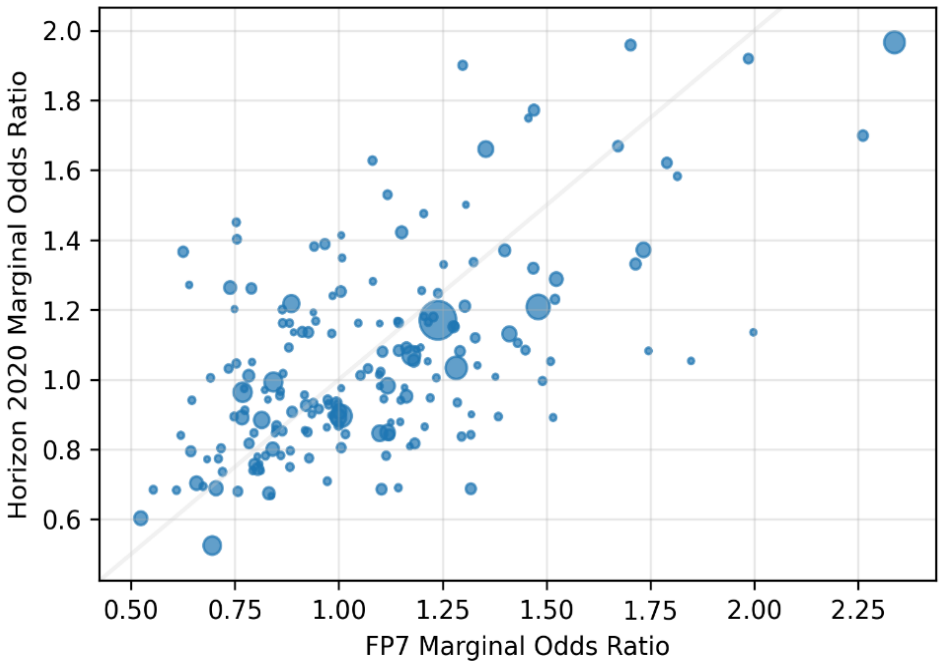

Figure 3: Estimated Odd Ratios of proposals by topics in FP7 and H2020. The size of the dot represents the total number of proposals

2.7 Conclusions

Decisions about what to research shape the direction of science and the production of knowledge. These decisions are mostly up to individual scientists or research groups, but they are affected by the availability of resources. An increasing share of resources for research is managed by funding agencies and foundations, through

competitive schemes relying on peer review processes. Scholars have argued that competitive funding has increased the influence and power of an academic elite in charge of deciding who and what kind of research will be funded (Whitley, 2007; Musselin, 2013; Hoening, 2017). However, there is scant empirical evidence on the extent to which grant selection affects what research is supported.

This article aimed to address this gap by exploring to what extent the evaluation process of the ERC grants modifies the bottom-up research topics mix of the proposals. Hence, we investigated whether a proposal's research topic is an important predictor of its success. We used state-of-the-art natural language processing techniques to examine the textual content of the abstract of 91,273 proposals for Starting, Consolidator, and Advanced Grants in the Seventh Framework Program (FP7) and Horizon 2020 (H2020), i.e., 95.3% of all the proposals, and identified 188 unique research topics. We found that the research topic of a proposal is an important predictor of success, with 37.4% of the proposals having significantly higher or lower chances of being funded due to their topic compared to the chosen baseline group of proposals not assigned to any topic. At the same time, we estimated that the research topics mix of the funded proposals overlap substantially, i.e., 82.5%, with the research topic mix of the proposals funded in a scenario in which the topics did not matter. Overall, the results hint to a notable impact of the selection process on what is researched, which nevertheless does not overturn its bottom-up orientation.

It is important to highlight that our analysis was not based on simple success rates but controlled for several other factors to disentangle a potential compositional effect, namely that some topics were successful because, for example, frequently submitted by researchers from excellent research systems. We show that differences in success rates across topics remain similar also when considering other covariates. This also implies that differences in success rate across countries, which are quite large, are not explained by the research topic mix of their proposals either.

This article is also the first to explore in detail the research topics and geographical unevenness of ERC research funds. It reveals the important contextual and thematic conditions that affect proposals' success rate, even for a program which

targets individuals, like the ERC. An important implication for university managers and policy makers in many European institutions and countries is to consider the importance of such factors, and to ascertain that the right conditions for successful applications are in place, because, if these conditions are lacking, merely incentivizing grant applications will not be an effective nor efficient way for attracting resources.

Some limitations should be discussed. First, the characteristics of the individual applicant like its scientific productivity, age, academic rank, gender, could not be considered in the analysis, due to privacy rules restrictions. Arguably, the variables about the country and institution of affiliation incorporate part of the effect of differences in scientific production and impact. Second, we explored the extent to which the selection process affects what research is supported only in one funding scheme, with specific traits like a bottom-up orientation. Future research should explore whether the impact is similar, or probably greater, in programs that do not have such a mission. In general, research can explore what traits of the funding program and evaluation procedure might affect the degree to which the evaluation procedure can impact what research topics are supported and modify the topic mix of the proposals.

Finally, we would like to hint some promising research ideas to extend this line of research to understand why some topics are more successful than others. The ERC governing bodies have steadily opposed a direct – and external – identification of priorities. Nevertheless, an in-depth analysis may shed light on whether differences and variation over time in the success rate of the research topics reflect shifting societal, political, and economic priorities or preferences (e.g., Clark et al., 2023). Success may also depend on intrinsic traits of the topics, such as a different propensity for fundamental-groundbreaking research. Another possibility is that the success rates of different topics can be accounted in part by the composition of the panels, such as due to experts being more lenient or critical to research in their own areas of expertise (e.g., Li, 2017).

References

Albers, C. J., 2015. Dutch research funding, gender bias, and Simpson's paradox. *Proceedings of the National Academy of Sciences*, 112(50), E6828-E6829.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Bol, T., de Vaan, M., van de Rijt, A., 2018. The Matthew effect in science funding. *Proceedings of the National Academy of Sciences*, 115(19), 4887–4890.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., Riedl, C., 2016. Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science*, 62(10), 2765–2783.

Boyack, K. W., Smith, C., Klavans, R., 2018. Toward predicting research proposal success. *Scientometrics*, 114, 449-461.

Braun, D., 1993. Who governs intermediary agencies? Principal-agent relations in research policy-making. *Journal of Public Policy*, 13(2), 135-162.

Bromham, L., Dinnage, R., Hua, X., 2016. Interdisciplinary research has consistently lower funding success. *Nature*, 534(7609), 684–687.

Brunet, L., Müller, R., 2022. Making the cut: How panel reviewers use evaluation devices to select applications at the European Research Council. *Research Evaluation*, 31(4), 486-497.

Campello, R.J.G.B., Moulavi, D., Sander, J., 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7819. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

Clark, B. R., 1983. *The higher education system: Academic organization in cross-national perspective*. Univ. of California Press.

Clark, C.J., Jussim, L., Freu, K., Stevens, S.T., et al., 2023. Prosocial motives underlie scientific censorship by scientists: A perspective and research agenda. *Proceedings of the National Academy of Sciences*, 120(48), e2301642120.

Debernardi, C., Seeber, M., Cattaneo, M., 2024. Thirty years of climate change research: A fine-grained analysis of geographical specialization. *Environmental Science & Policy*, 152, 103663.

European Commission, 2005. Frontier research: The European challenge high level expert group report. European Commission.

European Commission, 2019. Commission appoints Mauro Ferrari as next President of the European Research Council. European Commission press release. Brussels, 14 May 2019. https://ec.europa.eu/commission/presscorner/detail/en/IP_19_2471

Enger, S.G., Castellacci F., 2016. Who gets Horizon 2020 research grants? Propensity to apply and probability to succeed in a two-step analysis. *Scientometrics* 109: 1611-1638.

European Research Council, 2020. Resignation of Mauro Ferrari – Statement by the Scientific Council. https://erc.europa.eu/news/resignation-mauro-ferrari-%E2%80%93-statement-scientific-council

Financial Times, 2020. EU science chief resigns with blast at coronavirus response. *Financial Times*. https://www.ft.com/content/f94725c8-e038-4841-a5f6-2e046ae78e95

Geuna, A., 1998. Determinants of university participation in EU-funded R&D cooperative projects. *Research Policy*, 26(6), 677-687.

Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv*:2203.05794.

Henriques, L., Schoen, A., Pontikakis, D., 2009. Europe's top research universities in FP6: scope and drivers of participation. *JRC Technical Notes*, 53681.

Hoenig, B., 2017. Europe’s new scientific elite: social mechanisms of science in the European research area. Routledge, London & New York.

Hörlesberger, M., Roche, I., Besagni, D., Scherngell, T., François, C., Cuxac, P., ... Holste, D., 2013. A concept for inferring ‘frontier research’in grant proposals. *Scientometrics*, 97, 129-148.

Iorio, R., Labory, S., Rentocchini, F., 2017. The importance of pro-social behaviour for the breadth and depth of knowledge transfer activities: An analysis of Italian academic scientists. *Research Policy*, *46*(2), 497–509. https://doi.org/10.1016/j.respol.2016.12.003

Jang, D., Doh, S., Kang, G.M., Han, D.S., 2017. Impact of Alumni Connections on Peer Review Ratings and Selection Success Rate in National Research. *Science Technology & Human Values* 42 (1): 116-143

Lepori B., Veglio V., Heller-Schuh B., Scherngell T. and Barber M., 2015. Participations to European Framework Programs of higher education institutions and their association with organizational characteristics. *Scientometrics* 105: 2149–2178.

Li, D., 2017. Expertise versus Bias in Evaluation: Evidence from the NIH. American Economic Journal: Applied Economics, 9(2): 60–92.

McInnes, L., Healy, J., & Astels, S., 2017. hdbscan: Hierarchical density based clustering, *Journal of Open Source Software*, 2(11), 205, doi:10.21105/joss.00205

McInnes, L., Healy, J., & Melville, J., 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29), 861, https://doi.org/10.21105/joss.00861

Merton, R. K., 1974. *The sociology of science: Theoretical and empirical investigations* (4. Dr.). Univ. of Chicago Pr.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*. arXiv:1310.4546

Mom, C.S. van den Besselaar, P., 2021. Do interests affect grant application success? The role of organizational proximity. *ArXiv*:2206.03255 – V3

Mutz, R., Bornmann, L., Daniel, H.D., 2012. Does gender matter in grant peer review? An empirical investigation using the example of the Austrian science fund. *Zeitschrift für Psychologie*, 220(2), 121–129.

Müller, R., 2021. Time as a Judgment Device: How Time Matters When Reviewers Assess Applicants for ERC Starting and Consolidator Grants. In *Inquiring into Academic Timescapes* (pp. 195-209). Emerald Publishing Limited.

Musselin, C., 2013. How peer review empowers the academic profession and university managers: Changes in relationships between the state, universities and the professoriate. Research Policy, 42(5), 1165-1173.

Neufeld, J., Huber, N., Wegner, A., 2013. Peer review-based selection decisions in individual research funding, applicants' publication strategies and performance: The case of the ERC starting grants. *Research Evaluation*, 22, 237–247.

Piro, F.N., Børing, P., Scordato, L. Aksnes, D.W., 2020. University characteristics and probabilities for funding of proposals in the European Framework Programs. *Science and Public Policy*, 47(4), 581-593.

Piro, F. N., Seeber, M., Wang, L., 2024. Regional and sectoral variations in the ability to attract funding from the European Union's Seventh Framework Program and Horizon 2020. *Scientometrics*, 129, 1493-1521.

Reimers N., Gurevych. I., 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *ArXiv*:1908.10084

Sandström, U., Hällsten, M., 2008. Persistent nepotism in peer-review. *Scientometrics*, 74(2), 175-189.

Schiebel, E., Roche, I., Besagni, D., Hörlesberger, M. Francois, C., 2014. A textual approach to measure the interdisciplinary character of research proposals for ERC starting grants. Pp. 534-542 in *The 19th edition of the STI*, Leiden, the Netherlands, 3–5 September 2014 (sti2014.cwts.nl), book of abstracts.

Seeber, M., Alon, I., Pina, D., Piro, F.N. Seeber, M., 2022a. Predictors of European Research Council (ERC) Valorization Grant Application and Winning: A Machine Learning Approach, *Technological Forecasting & Social Change*, 184, 122009.

Seeber, M., Vlegels, J., Cattaneo, M., 2022b. Conditions that do or do not disadvantage interdisciplinary research proposals in project evaluation. *Journal of the Association for Information Science and Technology*, 73(8), 1106-1126.

Sormani, E., Uude, K., 2022. Academics' prosocial motivation for engagement with society: The case of German academics in health science. *Science and Public Policy*, *49*(6), 962–971. https://doi.org/10.1093/scipol/scac042

Stan Development Team, 2024. Stan Modeling Language Users Guide and Reference Manual. https://mc-stan.org

Stephan, P. E., 1996. The Economics of Science. *Journal of Economic Literature*, *34*(3), 1199–1235. https://www.jstor.org/stable/2729500

Tamblyn, R., Girard, N., Qian, C. J., Hanley, J., 2018. Assessment of potential bias in research grant peer review in Canada. *CMAJ*, 190(16), E489–E499.

Van den Besselaar, P., Leydesdorff, L., 2009. Past performance, peer review and project selection: a case study in the social and behavioral sciences. *Research Evaluation*, 18(4), 273-288.

Van den Besselaar, P., Sandström, U. Schiffbaenker, H., 2018. Studying grant decision-making: a linguistic analysis of review reports. *Scientometrics*, 117, 313–329.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I., 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30).

Veugelers, R., Wang. J, Stephan, P., 2022. Do Funding Agencies Select and Enable Risky Research: Evidence from ERC Using Novelty as a Proxy of Risk Taking. US: The National Bureau of Economic Research (NBER), Working paper 30320, October 2022. DOI 10.3386/w30320.

Volker, B., Steenbeek, W., 2015. No evidence that gender contributes to personal research funding success in The Netherlands: A reaction to van der Lee and Ellemers. *Proceedings of the National Academy of Sciences*, 112(51), E7036–E7037.

Wanzenböck, I., Lata, R., Ince, D., 2020. Proposal success in Horizon 2020: A study of the influence of consortium characteristics. *Quantitative Science Studies*, 1(3), 1136-1158.

Whitley, R., 2000. *The intellectual and social organization of the sciences*. Oxford Univ. Press.

Whitley, R., 2007. Changing governance of the public sciences: the consequences of establishing diverse research evaluation systems. In: Whitley, R., Glaeser, J. (Eds.), *The Changing Governance of the Sciences: The Advent of Research Evaluation Systems*. Springer, Dordrecht, Netherlands, pp. 3–27.

Appendix

	proposals		grants		success
Country	**N**	**%**	**N**	**%**	**rate**
United Kingdom	15243	16.7%	2014	19.3%	13.2%
Germany	11255	12.3%	1673	16.1%	14.9%
Italy	9569	10.5%	497	4.8%	5.2%
France	8899	9.7%	1302	12.5%	14.6%
Spain	7581	8.3%	573	5.5%	7.6%
Netherlands (the)	6079	6.7%	936	9.0%	15.4%
Sweden	3686	4.0%	374	3.6%	10.1%
Switzerland	3509	3.8%	736	7.1%	21.0%
Israel	3208	3.5%	583	5.6%	18.2%
Belgium	2871	3.1%	359	3.4%	12.5%
Finland	2858	3.1%	183	1.8%	6.4%
Denmark	2185	2.4%	238	2.3%	10.9%
Austria	2031	2.2%	289	2.8%	14.2%
Norway	1565	1.7%	130	1.2%	8.3%
Portugal	1522	1.7%	107	1.0%	7.0%
Greece	1481	1.6%	57	0.5%	3.8%
Ireland	1393	1.5%	111	1.1%	8.0%
Poland	1208	1.3%	30	0.3%	2.5%
Hungary	829	0.9%	69	0.7%	8.3%
Turkey	773	0.8%	28	0.3%	3.6%
Czechia	764	0.8%	44	0.4%	5.8%
Romania	667	0.7%	9	0.1%	1.3%
Slovenia	443	0.5%	14	0.1%	3.2%
Cyprus	270	0.3%	14	0.1%	5.2%

Estonia	217	0.2%	10	0.1%	4.6%
Bulgaria	196	0.2%	2	0.0%	1.0%
Croatia	176	0.2%	6	0.1%	3.4%
Slovakia	152	0.2%	1	0.0%	0.7%
Serbia	109	0.1%	2	0.0%	1.8%
Lithuania	105	0.1%	1	0.0%	1.0%
Luxembourg	96	0.1%	11	0.1%	11.5%
Iceland	84	0.1%	5	0.0%	6.0%
Ukraine	64	0.1%	1	0.0%	1.6%
Latvia	56	0.1%	1	0.0%	1.8%
Malta	30	0.0%	0	0.0%	0.0%
Georgia	26	0.0%	0	0.0%	0.0%
Republic of North Macedonia	21	0.0%	0	0.0%	0.0%
Bosnia and Herzegovina	11	0.0%	0	0.0%	0.0%
Moldova	10	0.0%	0	0.0%	0.0%
Montenegro	8	0.0%	0	0.0%	0.0%
Albania	5	0.0%	0	0.0%	0.0%
Other countries	18	0.0%	0	0.0%	0.0%
Grand Total	91,273	100.0%	10410	100.0%	11.4%

Table A: Proposals and grants by country; sample used for the analysis (95.3% of all proposals). Due to the 95.4 coverage, there are minor differences with the full sample. E.g., Malta obtained one project.

		proposals		grants		proposals		proposals		
ID	keywords	FP7	H2020	FP7	H2020	HES	REC	STG	COG	ADG
-1		15244	19951	1360	2308	27003	7803	17660	7370	10165
0	circadian, clock, rhythms, clocks, sleep	56	84	7	11	109	30	63	37	40
1	gravity, string, gauge, gravitational, supersymmetric	317	390	38	39	558	147	306	187	214
2	neutrino, neutrinos, detector, detectors, neutrinoless	64	96	10	9	88	72	72	42	46
3	detectors, detector, silicon, pixel, cmos	46	51	2	2	52	45	44	11	42
4	nuclei, neutron, nucleon, nucleosynthesis, exotic	107	103	11	20	115	95	109	45	56
5	qcd, quark, qgp, quarks, gluon	116	109	6	9	163	61	127	42	56
6	stars, gravitational, neutron, gw, star	110	211	14	21	227	91	136	85	100
7	lhc, higgs, boson, sm, collider	205	331	16	37	344	192	246	150	140
8	dm, wimp, searches, universe, axion	55	136	3	17	135	56	95	50	46
9	cosmic, rays, neutrinos, neutrino, gamma	47	67	3	6	67	47	57	31	26
10	coronal, plasmas, corona, stellar, sun	176	206	11	24	232	149	165	84	133
11	galaxies, galaxy, stars, universe, star	598	850	74	110	955	492	655	333	460
12	planets, exoplanets, exoplanet, stars, atmospheres	57	100	10	12	111	46	74	43	40
13	planets, planetary, planet, protoplanetary, disks	25	65	3	11	54	36	37	26	27
14	retinal, retina, amd, corneal, myopia	100	141	14	19	186	50	113	55	73
15	pain, migraine, endocannabinoid, nociceptive, neurons	106	109	10	13	170	43	110	39	66
16	kin, cognition, sociality, chimpanzees, birds	90	100	15	17	150	40	105	47	38
17	cns, eae, sclerosis, snps, autoimmune	300	256	21	21	430	113	287	92	177
18	grid, grids, electricity, renewable, hvdc	59	116	3	6	164	11	99	36	40
19	senescent, senescence, lifespan, longevity, elegans	64	114	15	16	122	56	94	39	45
20	foods, milk, rumen, spoilage, microbial	82	53		2	117	17	80	20	35

21	raman, sers, vibrational, ters, hyperspectral	65	86	5	10	111	40	70	45	36
22	eukaryotic, genomes, eukaryotes, archaea, archaeal	62	71	12	21	89	44	52	32	49
23	autophagy, er, hsp90, lysosomal, chaperone	60	101	6	16	98	60	77	34	50
24	mesh, meshes, pdes, discretization, iga	107	111	12	11	178	39	99	48	71
25	glycosylation, glycan, glycans, carbohydrate, carbohydrates	86	103	7	15	140	47	101	38	50
26	graphs, graph, combinatorics, extremal, hypergraphs	77	80	14	14	141	16	65	23	69
27	algebraic, conjecture, algebras, manifolds, symplectic	597	758	106	137	1164	191	663	276	416
28	pdes, pde, regularity, singularities, fractional	159	183	22	31	264	78	155	57	130
29	lipid, lds, er, ld, lipids	33	59	7	6	55	35	38	19	35
30	philosophy, metaphysics, philosophical, reasoning, epistemology	79	119	6	10	181	17	88	52	58
31	tb, mtb, tuberculosis, mycobacterium, mycobacterial	64	54	7	5	79	38	65	28	25
32	ehr, rcts, reviews, randomized, causality	76	165	3	10	214	25	120	49	72
33	bcl, tumor, apoptosis, mitochondrial, caspase	96	110	9	14	143	63	99	55	52
34	mantle, magma, seismic, volcanic, subduction	433	600	50	78	720	301	450	220	363
35	ubiquitin, ub, e3, ubiquitination, sumo	76	53	14	13	82	44	61	29	39
36	liver, hcc, hepatocytes, cirrhosis, fibrosis	35	55	3	5	67	22	39	21	30
37	angiogenesis, vascular, endothelial, tumor, angiogenic	88	55	12	11	99	44	73	25	45
38	combustion, flame, flames, fuels, ignition	81	101	5	15	132	48	84	38	60
39	kidney, renal, fibrosis, lung, glomerular	83	100	9	13	133	47	91	39	53
40	asthma, allergen, allergic, allergy, allergens	72	90	6	7	127	30	78	27	57
41	auditory, hearing, cochlear, speech, tinnitus	87	89	7	11	155	20	87	37	52
42	afm, tip, cantilever, microscope, stm	57	40	9	6	67	27	55	19	23

43	eating, obesity, exercise, caregivers, pa	75	101	2	5	161	13	97	42	37
44	urban, transportation, congestion, cities, eo	157	226	8	17	324	51	223	61	99
45	telomere, telomeres, g4, telomerase, telomeric	56	48	9	7	53	51	48	17	39
46	polyphenols, metabolites, grapevine, bioactive, polyphenol	81	77	2	4	109	46	93	25	40
47	ribosome, mrna, pol, mrnas, trna	109	121	21	24	142	88	99	48	83
48	pluripotent, reprogramming, pluripotency, lineage, somatic	59	89	13	15	109	38	72	34	42
49	dft, functionals, initio, ab, atoms	76	143	9	21	162	56	106	62	51
50	hsc, hscs, hematopoietic, hspc, bone	43	50	3	9	64	29	53	11	29
51	aml, leukemia, cll, myeloid, leukemic	75	123	5	24	132	59	107	46	45
52	infertility, sperm, pcos, ovarian, endometrial	82	80	3	9	125	34	80	21	61
53	pregnancy, maternal, placental, fetal, pe	79	104	5	12	134	41	96	48	39
54	gbm, tumor, glioma, glioblastoma, gliomas	38	60	2	7	64	34	47	30	21
55	mitochondrial, mtdna, mitochondria, oxphos, mitophagy	141	169	11	27	237	67	171	54	85
56	bacterial, antibiotic, antibiotics, biofilm, infections	347	474	28	75	611	202	440	195	186
57	microbiota, gut, intestinal, ibd, microbiome	150	265	21	36	289	123	212	96	107
58	contaminants, pollutants, aquatic, edcs, ecotoxicological	52	42	1	1	64	29	63	8	23
59	polarity, morphogenesis, epithelial, drosophila, pcp	68	84	14	22	59	93	68	33	51
60	bone, cartilage, oa, biomaterials, scaffolds	310	446	25	49	607	141	414	166	176
61	actin, actomyosin, cytoskeleton, myosin, filaments	68	73	14	11	71	70	61	26	54
62	parasite, malaria, parasites, mosquito, plasmodium	182	259	22	33	274	163	216	104	121

63	cyanobacteria, microalgae, cyanobacterial, rubisco, fixation	56	61	7	6	86	31	58	22	37
64	nmr, dnp, nmos, epr, ssnmr	97	119	15	19	151	65	92	47	77
65	cardiac, cardiovascular, arterial, coronary, plaque	63	68	8	8	94	36	62	32	37
66	microbial, marine, microbes, isoprene, biogeochemical	92	100	14	10	119	73	111	24	57
67	mars, prebiotic, arsenic, microbial, meteorites	34	57	4	10	63	28	32	28	31
68	dose, pet, breast, detector, ct	72	85	2	9	96	61	84	37	36
69	tumor, breast, metastatic, tumour, metastasis	278	398	35	53	441	221	334	162	180
70	ultrasound, hifu, lung, mri, ct	37	69	4	6	81	25	56	23	27
71	lipid, ionization, breath, peptides, bilayer	159	155	17	12	236	75	157	60	97
72	microtubule, spindle, cilia, tubulin, centriole	98	90	17	18	98	87	97	43	48
73	chromatin, ddr, ner, dsbs, forks	167	158	33	43	187	138	151	84	90
74	virus, vaccines, infectious, viruses, vaccination	54	61	6	5	79	36	53	25	37
75	glasses, glass, amorphous, metallic, glassy	51	49	5	6	69	31	43	24	33
76	ionic, liquids, ils, solvents, polymer	77	68	5	7	124	21	76	31	38
77	mirna, mirnas, rnas, lncrnas, lincrnas	72	50	18	10	68	52	70	26	26
78	ca2, clc, calcium, voltage, gated	106	57	7	4	113	50	83	31	49
79	obesity, hypothalamic, leptin, neurons, hypothalamus	45	68		14	83	29	55	27	31
80	folding, conformational, idps, ligand, disordered	250	167	22	13	298	115	226	68	123
81	influenza, virus, viruses, iav, antiviral	38	56	5	5	71	22	47	19	28
82	photoacoustic, oct, nanoscopy, microscope, optoacoustic	58	138	9	34	145	51	95	45	56

83	nanopore, fret, nanopores, nucleic, biomolecules	63	67	8	12	107	23	67	37	26
84	biosensors, electrochemical, analytes, biosensor, analyte	100	89	9	5	137	51	103	48	38
85	muscle, skeletal, exercise, sarcopenia, dmd	80	81	7	4	119	39	88	30	43
86	insulin, diabetes, adipose, t2d, obesity	169	170	21	24	271	63	153	62	124
87	multiphase, boltzmann, reactors, turbulent, bubble	38	57	5	6	81	12	40	15	40
88	turbulence, wind, turbulent, turbines, turbine	157	178	14	19	284	43	169	54	112
89	methylation, histone, chromatin, lsd1, tet	75	71	10	13	91	47	82	32	32
90	chromatin, chromosome, cohesin, chromosomes, meiotic	120	174	29	45	162	132	133	75	86
91	pet, mri, tumor, tumour, mr	46	74	7	7	92	24	68	22	30
92	tumor, hyperthermia, tumour, nanocarriers, pdt	113	185	8	17	201	93	167	69	62
93	platelet, atherosclerosis, hdl, platelets, atherosclerotic	94	87	4	7	140	35	100	34	47
94	hiv, virus, infected, cd4, env	91	63	8	6	99	54	76	27	51
95	hcv, hcmv, hbv, virus, cmv	73	58	4	8	86	42	58	29	44
96	textile, printing, mems, elastomers, stretchable	41	69	8	11	89	21	52	27	31
97	alloys, corrosion, microstructure, fatigue, fracture	446	523	26	31	745	215	477	197	295
98	mycorrhizal, fungi, soil, symbiosis, amf	65	108	9	13	119	53	93	39	41
99	tumor, nk, immunotherapy, tcr, mhc	109	170	11	34	189	79	144	56	79
100	treg, ra, autoimmune, il, arthritis	137	158	16	25	188	103	136	56	103
101	cardiac, hf, hypertrophy, myocardial, mirnas	60	80	7	7	108	29	71	31	38
102	cardiac, af, cardiomyocytes, arrhythmias, scd	80	142	8	17	163	53	111	57	54
103	als, neurons, axon, axons, myelin	214	188	28	38	260	140	210	81	111

104	amyloid, tau, dementia, alzheimer, aβ	245	271	12	20	378	132	266	108	142
105	pd, neurons, lrrk2, hd, htt	77	74	2	8	116	35	62	34	55
106	pathogen, pathogens, fungal, crop, immunity	83	84	10	13	112	54	92	30	45
107	arabidopsis, auxin, crop, thaliana, crops	284	348	38	55	388	241	292	163	177
108	coastal, river, sediment, hydrological, rivers	75	81	2	3	127	25	85	34	37
109	ice, ocean, arctic, weather, precipitation	237	363	24	41	416	179	270	145	185
110	gender, feminist, political, rights, sexual	41	94	4	8	117	18	67	35	33
111	biomass, wastewater, co2, microbial, lignin	224	312	7	16	417	113	310	107	119
112	aerosol, clouds, cloud, aerosols, ice	56	69	6	13	80	44	64	22	39
113	aerosol, soa, aerosols, emissions, anthropogenic	67	88	5	13	99	55	80	40	35
114	soil, microbial, soils, som, ecosystem	67	105	3	10	130	40	88	38	46
115	moral, preferences, economics, psychology, rationality	126	176	10	29	280	20	145	68	89
116	ocean, marine, co2, biogeochemical, phytoplankton	150	212	16	26	235	126	167	99	96
117	forest, forests, vegetation, ecosystem, tropical	76	129	7	10	154	50	110	48	47
118	philosophy, ssh, philosophical, pluralism, epistemic	68	85	1	3	137	16	73	33	47
119	cavitation, suspensions, colloids, colloidal, wetting	128	171	19	24	231	67	148	56	95
120	coral, reef, reefs, marine, corals	49	66	3	9	73	42	64	30	21
121	supramolecular, polymers, nucleic, polymer, peptides	231	292	36	49	432	89	259	92	172
122	teeth, fossil, skull, fossils, tooth	40	67	2	8	81	26	46	33	28
123	gan, qds, semiconductor, nws, leds	92	76	7	13	113	55	96	16	56
124	thermoelectric, te, conductivity, pyroelectric, seebeck	41	57	6	3	67	30	48	28	22
125	sex, sexual, sperm, male, chromosomes	60	74	9	13	105	29	65	33	36

126	biodiversity, ecosystem, ecosystems, dispersal, habitat	117	160	13	15	196	80	144	61	72
127	speciation, birds, biodiversity, fitness, migratory	141	180	20	14	241	79	156	64	101
128	mofs, porous, zeolites, mof, zeolite	41	109	3	16	111	39	71	39	40
129	attosecond, pulses, ultrafast, femtosecond, lasers	329	333	41	37	423	236	294	151	217
130	sparsity, compressed, cs, nyquist, regularization	70	43	16	6	96	17	59	20	34
131	pv, photovoltaic, silicon, photovoltaics, cigs	39	75	4	12	74	38	54	36	24
132	perovskite, perovskites, halide, pv, optoelectronic	1	115		20	88	27	61	31	24
133	batteries, battery, electrochemical, lithium, electrolyte	56	171	6	17	160	65	119	63	45
134	photocatalytic, photocatalysts, co2, splitting, pec	94	139	7	13	175	58	120	71	42
135	oleds, semiconductors, exciton, emitting, photovoltaic	130	135	10	12	214	51	158	60	47
136	olfactory, neurons, sensory, taste, neuronal	89	118	18	19	125	79	105	38	64
137	sleep, cortical, neurons, motor, synaptic	1883	2429	223	361	3403	876	2291	894	1127
138	speech, linguistic, languages, syntactic, linguistics	595	814	50	81	1204	194	723	270	416
139	electrochemical, electrolyte, electrode, electrocatalytic, electrochemistry	28	64	4	8	65	27	37	25	30
140	thz, terahertz, ghz, antennas, wireless	45	56	5	7	78	22	48	28	25
141	robots, robot, robotics, robotic, rl	93	183	9	26	228	44	140	50	86
142	haptic, robots, robot, tactile, robotics	73	130	6	14	172	30	106	42	55
143	sofc, electrochemical, h2, sofcs, catalysts	60	89	3	6	92	55	76	34	39
144	teachers, teaching, educational, learners, teacher	89	72		2	156	4	94	33	34
145	scene, video, scenes, graphics, camera	144	199	27	36	276	65	195	59	89

146	photonic, metamaterials, plasmonic, photonics, metasurfaces	171	150	20	20	232	86	168	73	80
147	lasers, comb, combs, fiber, fibre	51	50	2	9	67	34	45	18	38
148	graph, graphs, synchronization, hyperbolic, networked	51	50		2	90	10	48	30	23
149	semantic, query, retrieval, queries, ontologies	174	149	10	11	278	40	180	71	72
150	catalysts, catalysis, chiral, enantioselective, catalyst	326	370	43	75	590	105	334	158	204
151	co2, catalysts, methane, methanol, catalyst	34	58	2	1	78	14	48	18	26
152	qubits, qubit, superconducting, spins, entanglement	55	57	16	10	87	24	67	25	20
153	graphene, heterostructures, fullerenes, cnts, dichalcogenides	132	205	19	30	248	85	160	89	88
154	entanglement, cryptography, cryptographic, multipartite, nonlocality	128	166	12	21	236	56	171	53	70
155	ultracold, atoms, superfluid, condensed, fermi	126	176	29	37	225	77	154	63	85
156	rydberg, photons, atoms, photonic, entanglement	151	178	29	35	245	84	175	76	78
157	oxide, ferroelectric, oxides, multiferroic, magnetoelectric	77	64	5	3	102	39	77	27	37
158	superconductivity, superconductors, superconducting, tc, cuprates	115	90	8	20	142	62	99	39	67
159	smms, magnets, spintronics, spins, qubits	75	55	13	6	104	26	69	28	33
160	insulators, majorana, hall, dirac, fermions	71	162	11	42	153	80	107	52	74
161	spintronics, spintronic, magnon, magnetization, magnons	54	97	8	12	99	52	76	40	35
162	wireless, cloud, radio, iot, antennas	233	267	12	21	416	78	270	123	107

163	welfare, inequality, labour, political, solidarity	38	64	3	12	89	12	53	20	29
164	markets, firms, intermediaries, multiproduct, bidders	33	64	6	15	77	20	39	29	29
165	privacy, cryptographic, cryptography, attacks, encryption	106	194	20	40	235	61	155	67	78
166	graph, np, algorithmic, parameterized, proofs	127	132	24	31	215	44	135	50	74
167	processors, processor, cores, reconfigurable, exascale	79	79	5	7	116	41	84	33	41
168	cps, concurrency, correctness, cyber, specification	45	60	6	11	83	21	59	19	27
169	languages, reasoning, developers, bugs, agile	104	98	13	17	173	28	108	36	58
170	firms, entrepreneurship, business, entrepreneurial, fdi	165	128	5	19	259	32	154	50	89
171	markets, firms, finance, credit, monetary	302	303	47	51	544	56	267	132	206
172	legal, governance, copyright, rights, ai	20	84	3	10	98	1	43	32	29
173	governance, political, biodiversity, land, ecosystem	372	662	25	70	888	135	578	212	244
174	urban, cities, city, architectural, political	113	164	4	12	252	20	145	59	73
175	hiv, grief, africa, governance, aids	58	84	7	4	126	15	75	24	43
176	migrants, refugee, refugees, citizenship, immigrants	151	301	8	44	403	46	266	94	92
177	gender, fertility, women, intergenerational, demographic	154	238	21	37	336	54	213	87	92
178	retirement, inequalities, pension, inequality, workers	67	116	6	11	138	43	102	37	44
179	political, news, conspiracy, democracy, citizens	35	97	2	17	124	7	81	28	23
180	music, musical, musicology, opera, musicians	39	113	1	10	121	25	62	36	54
181	music, musical, listening, musicians, ai	71	93	4	17	140	23	92	26	46
182	political, democracy, democratic, electoral, parties	129	222	16	31	316	33	183	66	102

183	artistic, artists, arts, aesthetics, photography	95	147	8	12	218	17	102	67	73
184	legal, rights, criminal, political, justice	322	463	25	43	701	72	414	180	191
185	war, political, soviet, holocaust, violence	92	116	7	10	168	38	93	52	63
186	archaeological, roman, archaeology, neolithic, prehistoric	377	689	24	75	807	236	461	226	379
187	religious, texts, literary, religion, medieval	639	1077	70	134	1527	167	731	406	579

Table B: Proposals' topics by call, type of organization and grant

variable	mean	5%	95%	*	variable	mean	5%	95%	*
alpha	-2,12	-2,29	-1,97	1	panel[LS1]	-0,24	-0,35	-0,14	1
sigma_panel	0,16	0,12	0,21	1	panel[LS2]	-0,12	-0,22	-0,02	1
sigma_topic	0,35	0,31	0,40	1	panel[LS3]	-0,31	-0,43	-0,20	1
sigma_country	0,88	0,64	1,18	1	panel[LS4]	0,06	-0,05	0,15	0
sigma_id_inst	0,60	0,55	0,66	1	panel[LS5]	-0,09	-0,20	0,01	0
year[2007]	-1,11	-1,23	-0,98	1	panel[LS6]	-0,06	-0,17	0,04	0
year[2008]	0,28	0,15	0,41	1	panel[LS7]	0,21	0,11	0,30	1
year[2009]	-0,06	-0,17	0,05	0	panel[LS8]	0,04	-0,08	0,14	0
year[2010]	0,21	0,12	0,30	1	panel[LS9]	0,20	0,08	0,32	1
year[2011]	0,03	-0,06	0,11	0	panel[na]	0,00	0,00	0,00	0
year[2012]	0,03	-0,05	0,11	0	panel[PE1]	0,07	-0,05	0,18	0
year[2013]	-0,39	-0,47	-0,31	1	panel[PE2]	0,03	-0,07	0,13	0
year[2014]	-0,12	-0,20	-0,05	1	panel[PE3]	-0,04	-0,14	0,05	0
year[2015]	0,00	0,00	0,00	0	panel[PE4]	0,08	-0,01	0,18	0
year[2016]	-0,19	-0,27	-0,11	1	panel[PE5]	0,01	-0,08	0,10	0
year[2017]	-0,13	-0,20	-0,05	1	panel[PE6]	0,03	-0,07	0,12	0
year[2018]	-0,21	-0,29	-0,13	1	panel[PE7]	0,14	0,03	0,25	1
year[2019]	-0,22	-0,30	-0,14	1	panel[PE8]	0,33	0,23	0,43	1
year[2020]	-0,44	-0,54	-0,34	1	panel[PE9]	-0,05	-0,18	0,07	0
grant[STG]	0,00	0,00	0,00	0	panel[PE10]	-0,03	-0,15	0,09	0
grant[ADG]	-0,22	-0,28	-0,16	1	panel[SH1]	0,17	0,04	0,30	1
grant[COG]	0,43	0,35	0,50	1	panel[SH2]	-0,05	-0,14	0,05	0
type[HES]	0,00	0,00	0,00	0	panel[SH3]	-0,02	-0,14	0,11	0
type[REC]	0,42	0,32	0,51	1	panel[SH4]	-0,10	-0,20	0,01	0
type[PRC]	-0,71	-1,14	-0,31	1	panel[SH5]	-0,01	-0,13	0,10	0
type[PUB]	-0,43	-1,00	0,14	0	panel[SH6]	0,04	-0,08	0,15	0
type[OTH]	-0,46	-1,07	0,11	0	country[UK]	0,00	0,00	0,00	0
funding[1]	-1,22	-1,35	-1,10	1	country[DE]	0,19	0,03	0,35	1
funding[2]	-0,72	-0,82	-0,62	1	country[IT]	-0,60	-0,79	-0,41	1
funding[3]	-0,35	-0,45	-0,24	1	country[FR]	0,31	0,16	0,46	1
funding[4]	-0,52	-0,62	-0,42	1	country[ES]	-0,46	-0,65	-0,28	1
funding[5]	-0,39	-0,48	-0,30	1	country[NL]	0,42	0,20	0,65	1
funding[6]	-0,24	-0,33	-0,15	1	country[SE]	-0,22	-0,52	0,08	0
funding[7]	-0,05	-0,14	0,05	0	country[CH]	0,62	0,38	0,86	1
funding[8]	-0,14	-0,22	-0,07	1	country[IL]	0,30	-0,03	0,62	0
funding[9]	0,00	0,00	0,00	0	country[BE]	0,24	-0,07	0,56	0
funding[10]	0,01	-0,07	0,09	0	country[FI]	-0,83	-1,18	-0,50	1

Table C: Logistic Regression predicting proposals' probability of success.
**no overlap with zero*

variable	mean	5%	95%	*	variable	mean	5%	95%	*
country[DK]	-0,04	-0,39	0,34	0	country[SK]	-1,43	-2,51	-0,51	1
country[AT]	0,37	0,12	0,63	1	country[RS]	-1,02	-1,86	-0,24	1
country[NO]	-0,53	-0,88	-0,18	1	country[LT]	-1,22	-2,22	-0,32	1
country[PT]	-0,56	-0,86	-0,28	1	country[LU]	-0,12	-0,90	0,58	0
country[EL]	-1,01	-1,35	-0,69	1	country[IS]	-0,27	-1,04	0,50	0
country[IE]	-0,21	-0,57	0,15	0	country[UA]	-0,88	-1,92	0,07	0
country[PL]	-1,33	-1,74	-0,93	1	country[LV]	-0,90	-1,92	0,03	0
country[HU]	-0,11	-0,54	0,29	0	country[MT]	-0,81	-2,18	0,42	0
country[TR]	-0,99	-1,45	-0,58	1	country[GE]	-0,76	-2,09	0,39	0
country[CZ]	-0,61	-1,07	-0,14	1	country[MK]	-0,70	-1,99	0,41	0
country[RO]	-1,60	-2,16	-1,08	1	country[BA]	-0,47	-1,75	0,67	0
country[SI]	-1,16	-1,72	-0,61	1	country[MD]	-0,55	-1,87	0,59	0
country[CY]	-0,53	-1,17	0,07	0	country[ME]	-0,55	-1,85	0,67	0
country[EE]	-0,52	-1,15	0,08	0	country[MY]	-0,15	-1,60	1,23	0
country[BG]	-1,26	-2,17	-0,43	1	country[AL]	-0,26	-1,56	1,01	0
country[HR]	-0,73	-1,47	-0,04	1	country[AM]	-0,17	-1,62	1,24	0
country[TZ]	-0,10	-1,56	1,30	0	country[TW]	-0,14	-1,50	1,16	0
country[IN]	-0,04	-1,37	1,35	0	country[TN]	-0,11	-1,54	1,40	0
country[CN]	-0,09	-1,66	1,35	0	country[PF]	-0,07	-1,55	1,41	0
country[BD]	-0,05	-1,48	1,37	0	country[AZ]	-0,07	-1,41	1,22	0
					country[UZ]	-0,07	-1,51	1,38	0

Table C: Logistic Regression predicting proposals' probability of success [continues from previous page].

variable	mean	5%	95%	*	variable	mean	5%	95%	*
topic[t-1]	0,00	0,00	0,00	0	topic[t36]	-0,08	-0,50	0,36	0
topic[t0]	0,02	-0,33	0,36	0	topic[t37]	0,26	-0,08	0,59	0
topic[t1]	-0,06	-0,27	0,15	0	topic[t38]	0,05	-0,33	0,40	0
topic[t2]	0,02	-0,33	0,37	0	topic[t39]	0,06	-0,23	0,36	0
topic[t3]	-0,39	-0,87	0,05	0	topic[t40]	-0,24	-0,63	0,12	0
topic[t4]	0,41	0,08	0,73	1	topic[t41]	-0,08	-0,41	0,26	0
topic[t5]	-0,20	-0,55	0,12	0	topic[t42]	0,22	-0,16	0,60	0
topic[t6]	-0,08	-0,36	0,21	0	topic[t43]	-0,44	-0,84	-0,06	1
topic[t7]	-0,19	-0,42	0,03	0	topic[t44]	-0,30	-0,59	-0,01	1
topic[t8]	-0,10	-0,45	0,25	0	topic[t45]	0,22	-0,17	0,60	0
topic[t9]	-0,20	-0,59	0,17	0	topic[t46]	-0,43	-0,86	-0,02	1
topic[t10]	-0,08	-0,36	0,19	0	topic[t47]	0,38	0,09	0,66	1
topic[t11]	0,13	-0,04	0,28	0	topic[t48]	0,25	-0,07	0,55	0
topic[t12]	0,08	-0,26	0,41	0	topic[t49]	0,20	-0,09	0,47	0
topic[t13]	0,19	-0,19	0,56	0	topic[t50]	0,06	-0,31	0,42	0
topic[t14]	0,21	-0,11	0,49	0	topic[t51]	0,25	-0,08	0,56	0
topic[t15]	0,08	-0,22	0,38	0	topic[t52]	-0,20	-0,56	0,15	0
topic[t16]	0,26	-0,05	0,55	0	topic[t53]	0,01	-0,38	0,38	0
topic[t17]	-0,41	-0,67	-0,16	1	topic[t54]	-0,15	-0,57	0,28	0
topic[t18]	-0,29	-0,69	0,11	0	topic[t55]	0,19	-0,08	0,44	0
topic[t19]	0,21	-0,10	0,52	0	topic[t56]	0,12	-0,06	0,30	0
topic[t20]	-0,50	-0,98	-0,06	1	topic[t57]	0,14	-0,10	0,38	0
topic[t21]	-0,08	-0,46	0,27	0	topic[t58]	-0,38	-0,88	0,08	0
topic[t22]	0,61	0,29	0,93	1	topic[t59]	0,40	0,08	0,72	1
topic[t23]	0,11	-0,23	0,45	0	topic[t60]	-0,19	-0,39	0,02	0
topic[t24]	0,07	-0,24	0,37	0	topic[t61]	0,17	-0,15	0,48	0
topic[t25]	-0,08	-0,39	0,23	0	topic[t62]	0,04	-0,19	0,26	0
topic[t26]	0,69	0,35	1,04	1	topic[t63]	-0,02	-0,40	0,37	0
topic[t27]	0,79	0,63	0,94	1	topic[t64]	0,24	-0,03	0,54	0
topic[t28]	0,61	0,35	0,86	1	topic[t65]	0,07	-0,30	0,43	0
topic[t29]	0,08	-0,33	0,46	0	topic[t66]	0,03	-0,34	0,34	0
topic[t30]	-0,05	-0,39	0,29	0	topic[t67]	0,14	-0,24	0,50	0
topic[t31]	-0,18	-0,55	0,22	0	topic[t68]	-0,21	-0,59	0,17	0
topic[t32]	-0,45	-0,80	-0,12	1	topic[t69]	0,02	-0,18	0,21	0
topic[t33]	-0,06	-0,41	0,25	0	topic[t70]	-0,32	-0,70	0,07	0
topic[t34]	0,14	-0,06	0,32	0	topic[t71]	-0,18	-0,49	0,13	0
topic[t35]	0,35	0,01	0,68	1	topic[t72]	0,38	0,09	0,65	1

Table C: Logistic Regression predicting proposals' probability of success [continues from previous page].

variable	mean	5%	95%	*	variable	mean	5%	95%	*
topic[t73]	0,72	0,50	0,94	1	topic[t111]	-0,67	-0,97	-0,39	1
topic[t74]	-0,10	-0,52	0,31	0	topic[t112]	0,27	-0,12	0,65	0
topic[t75]	0,00	-0,42	0,40	0	topic[t113]	0,10	-0,26	0,42	0
topic[t76]	-0,08	-0,46	0,30	0	topic[t114]	-0,23	-0,56	0,09	0
topic[t77]	0,38	0,05	0,71	1	topic[t115]	0,07	-0,21	0,33	0
topic[t78]	-0,25	-0,62	0,12	0	topic[t116]	-0,01	-0,29	0,26	0
topic[t79]	0,02	-0,35	0,41	0	topic[t117]	-0,22	-0,59	0,15	0
topic[t80]	-0,34	-0,60	-0,09	1	topic[t118]	-0,52	-0,95	-0,10	1
topic[t81]	-0,03	-0,45	0,37	0	topic[t119]	0,17	-0,09	0,42	0
topic[t82]	0,45	0,17	0,73	1	topic[t120]	-0,04	-0,43	0,35	0
topic[t83]	0,13	-0,24	0,47	0	topic[t121]	0,36	0,16	0,56	1
topic[t84]	-0,11	-0,47	0,24	0	topic[t122]	-0,11	-0,51	0,28	0
topic[t85]	-0,28	-0,69	0,11	0	topic[t123]	0,05	-0,31	0,41	0
topic[t86]	0,13	-0,13	0,37	0	topic[t124]	-0,05	-0,48	0,37	0
topic[t87]	-0,14	-0,52	0,24	0	topic[t125]	0,21	-0,15	0,54	0
topic[t88]	-0,13	-0,40	0,15	0	topic[t126]	-0,06	-0,37	0,23	0
topic[t89]	0,21	-0,13	0,53	0	topic[t127]	-0,11	-0,39	0,16	0
topic[t90]	0,60	0,36	0,85	1	topic[t128]	0,13	-0,24	0,47	0
topic[t91]	-0,04	-0,43	0,32	0	topic[t129]	-0,04	-0,23	0,15	0
topic[t92]	-0,12	-0,44	0,18	0	topic[t130]	0,47	0,13	0,81	1
topic[t93]	-0,36	-0,72	0,00	1	topic[t131]	0,21	-0,17	0,57	0
topic[t94]	-0,18	-0,54	0,17	0	topic[t132]	0,22	-0,15	0,57	0
topic[t95]	-0,14	-0,52	0,23	0	topic[t133]	-0,10	-0,44	0,23	0
topic[t96]	0,15	-0,24	0,53	0	topic[t134]	-0,19	-0,51	0,11	0
topic[t97]	-0,60	-0,81	-0,40	1	topic[t135]	-0,25	-0,58	0,07	0
topic[t98]	0,09	-0,24	0,40	0	topic[t136]	0,24	-0,06	0,52	0
topic[t99]	0,26	0,01	0,51	1	topic[t137]	0,19	0,10	0,29	1
topic[t100]	0,05	-0,21	0,31	0	topic[t138]	-0,05	-0,21	0,11	0
topic[t101]	-0,12	-0,51	0,26	0	topic[t139]	0,06	-0,32	0,44	0
topic[t102]	-0,05	-0,36	0,24	0	topic[t140]	0,05	-0,35	0,48	0
topic[t103]	0,29	0,06	0,51	1	topic[t141]	0,05	-0,24	0,33	0
topic[t104]	-0,45	-0,73	-0,18	1	topic[t142]	-0,10	-0,44	0,23	0
topic[t105]	-0,29	-0,68	0,10	0	topic[t143]	-0,29	-0,67	0,07	0
topic[t106]	0,27	-0,07	0,59	0	topic[t144]	-0,68	-1,12	-0,27	1
topic[t107]	0,26	0,05	0,46	1	topic[t145]	0,44	0,18	0,68	1
topic[t108]	-0,63	-1,05	-0,23	1	topic[t146]	0,12	-0,15	0,37	0
topic[t109]	-0,06	-0,28	0,17	0	topic[t147]	-0,01	-0,43	0,42	0
topic[t110]	0,04	-0,33	0,40	0	topic[t148]	-0,48	-0,96	-0,03	1

Table C: Logistic Regression predicting proposals' probability of success [continues from previous page].

variable	mean	5%	95%	*	variable	mean	5%	95%	*
topic[t149]	-0,33	-0,64	-0,01	1	topic[t168]	0,16	-0,20	0,53	0
topic[t150]	0,44	0,26	0,62	1	topic[t169]	0,29	-0,01	0,58	0
topic[t151]	-0,44	-0,89	0,00	1	topic[t170]	-0,03	-0,34	0,27	0
topic[t152]	0,52	0,18	0,87	1	topic[t171]	0,38	0,18	0,58	1
topic[t153]	0,23	-0,02	0,46	0	topic[t172]	0,14	-0,26	0,53	0
topic[t154]	0,13	-0,16	0,42	0	topic[t174]	-0,39	-0,72	-0,07	1
topic[t155]	0,75	0,50	0,99	1	topic[t175]	-0,23	-0,59	0,14	0
topic[t156]	0,37	0,14	0,60	1	topic[t176]	0,10	-0,15	0,34	0
topic[t157]	-0,36	-0,76	0,03	0	topic[t177]	0,37	0,12	0,60	1
topic[t158]	0,11	-0,20	0,42	0	topic[t178]	-0,12	-0,45	0,21	0
topic[t159]	0,19	-0,14	0,54	0	topic[t179]	0,29	-0,07	0,63	0
topic[t160]	0,63	0,36	0,89	1	topic[t180]	-0,10	-0,51	0,29	0
topic[t161]	0,07	-0,27	0,43	0	topic[t181]	0,17	-0,18	0,50	0
topic[t162]	-0,26	-0,56	0,03	0	topic[t182]	0,23	-0,02	0,49	0
topic[t163]	0,33	-0,06	0,72	0	topic[t183]	-0,19	-0,52	0,13	0
topic[t164]	0,42	0,07	0,77	1	topic[t184]	-0,09	-0,30	0,13	0
topic[t165]	0,61	0,38	0,84	1	topic[t185]	-0,18	-0,53	0,17	0
topic[t166]	0,81	0,55	1,06	1	topic[t186]	-0,09	-0,28	0,09	0
topic[t167]	-0,25	-0,63	0,12	0	topic[t187]	0,26	0,11	0,39	1

Table C: Logistic Regression predicting proposals' probability of success [continues from previous page].

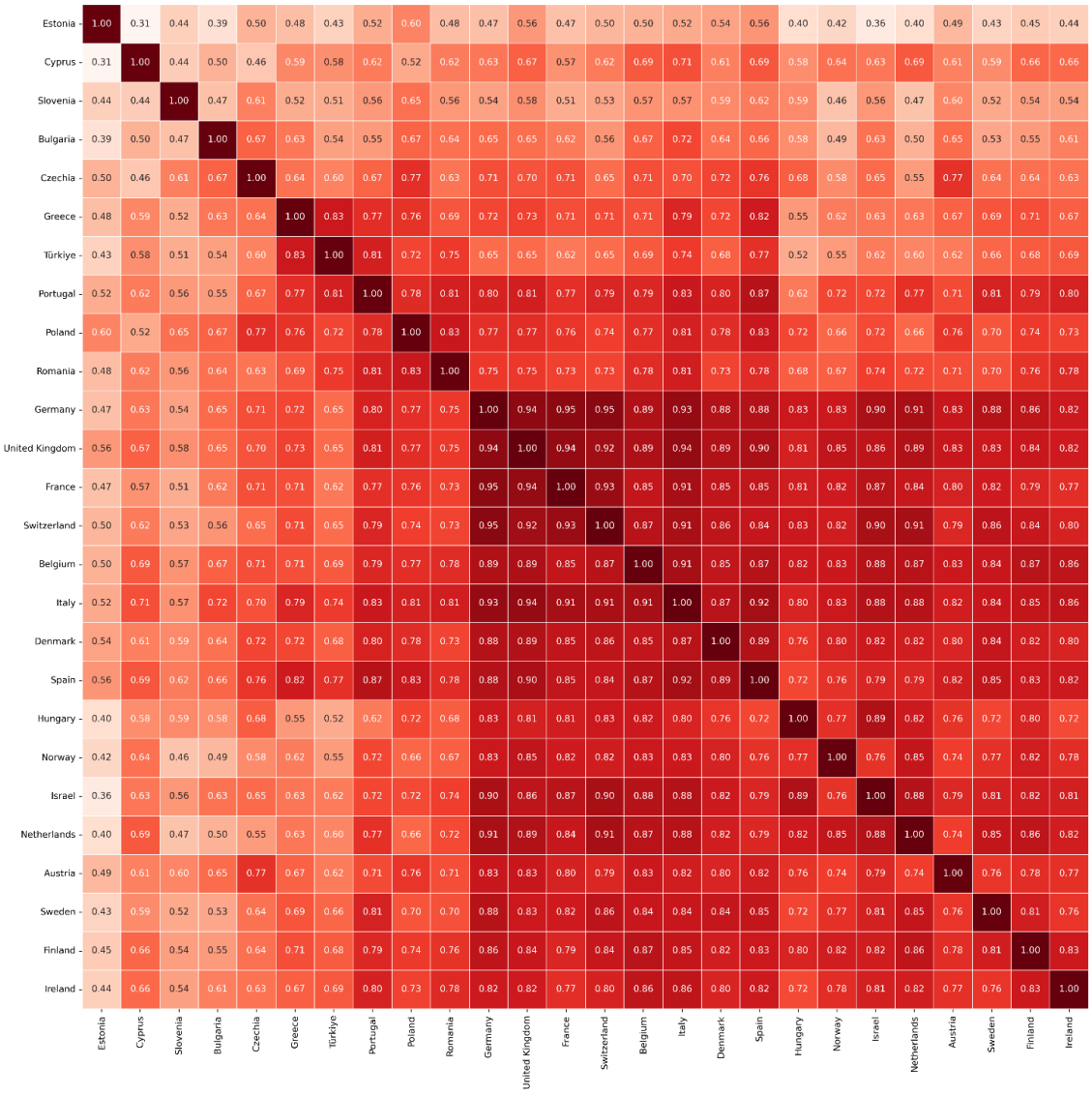

Figure A: Cosine similarity of the countries' research profile by weight of each research topic

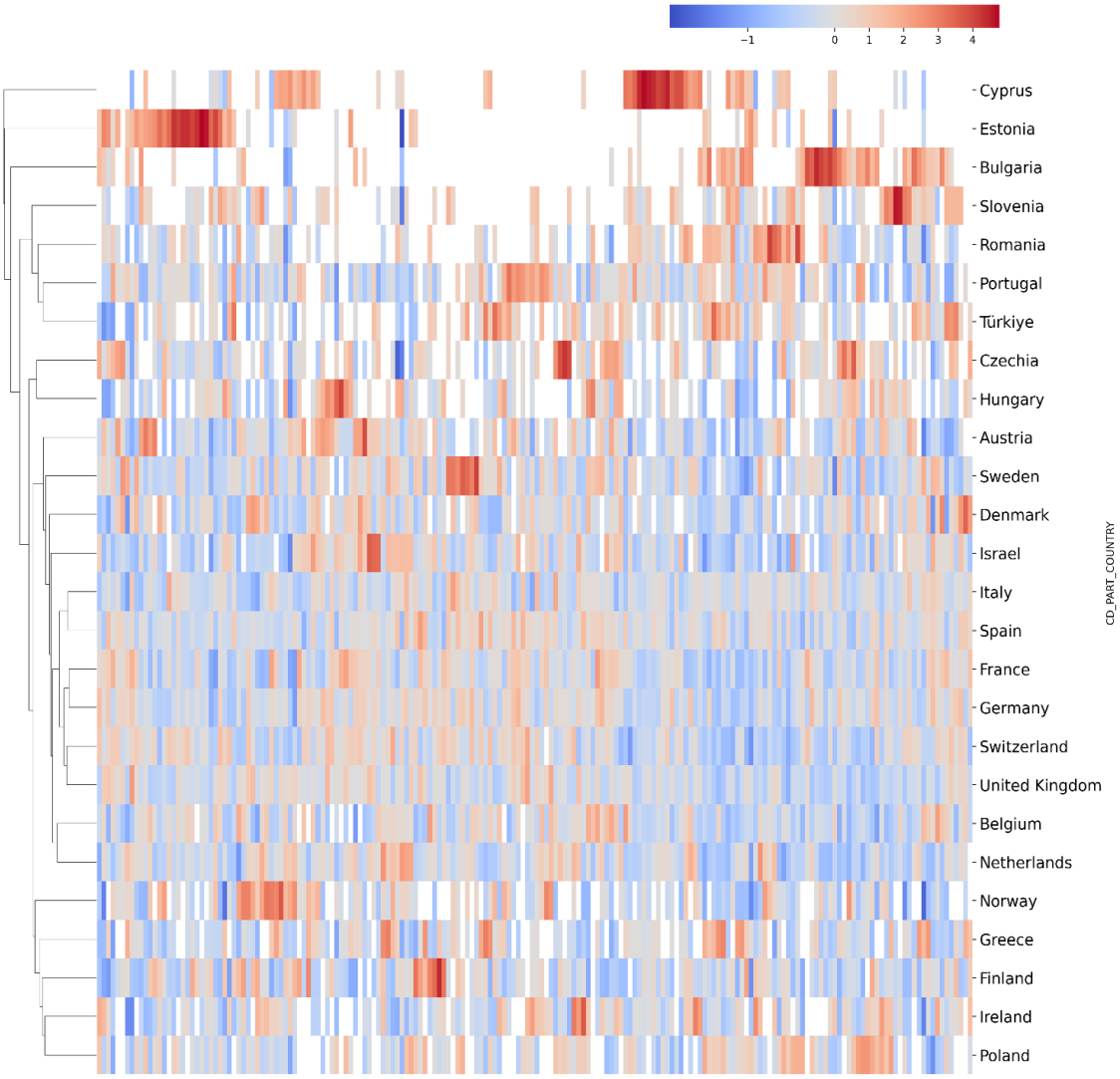

Figure B: Standardized weight of the research topics by country matrix Note: White spaces indicate topics that are fully absent from a given country

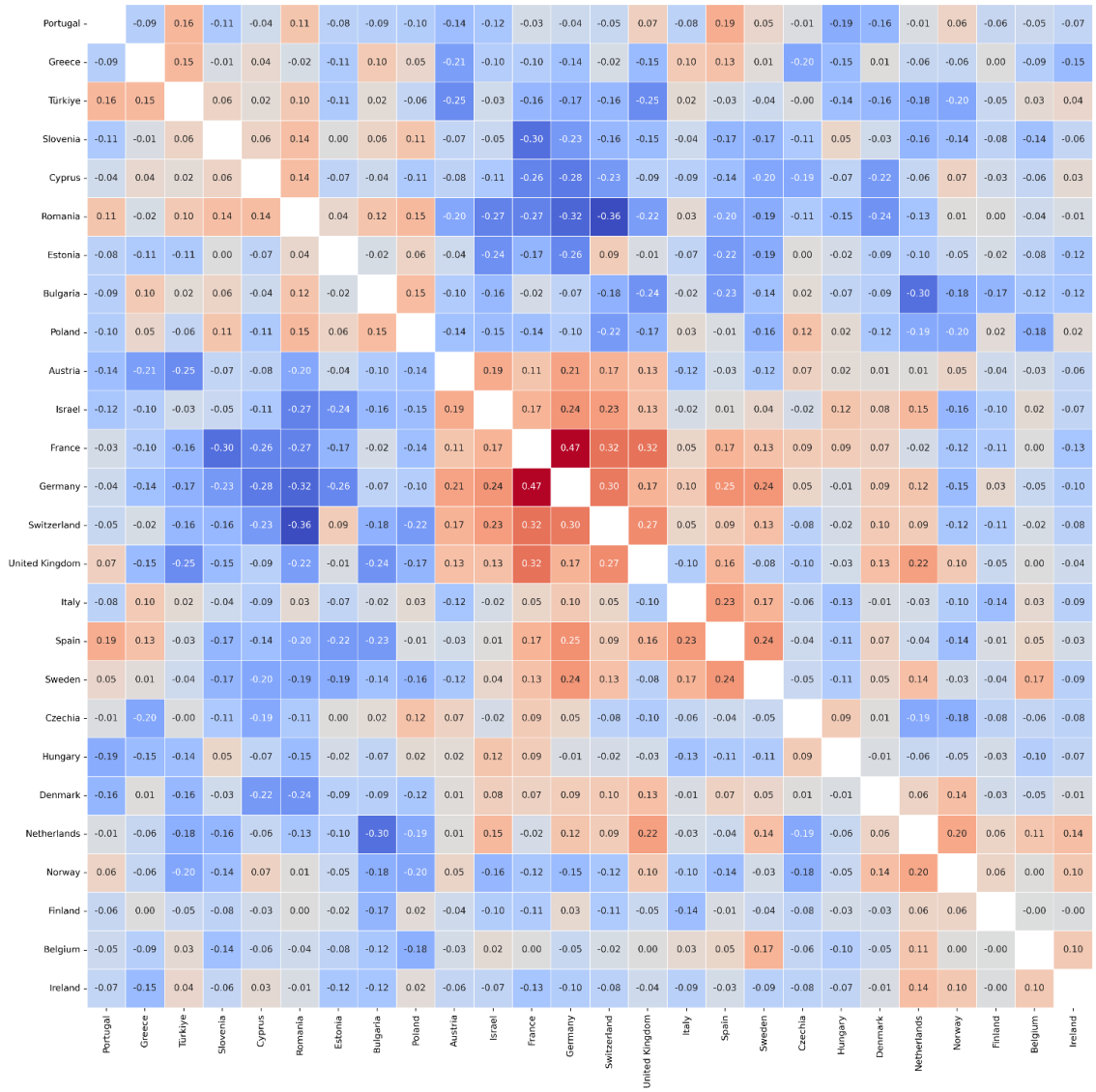

Figure C: Cosine similarity of research topics specializations by country

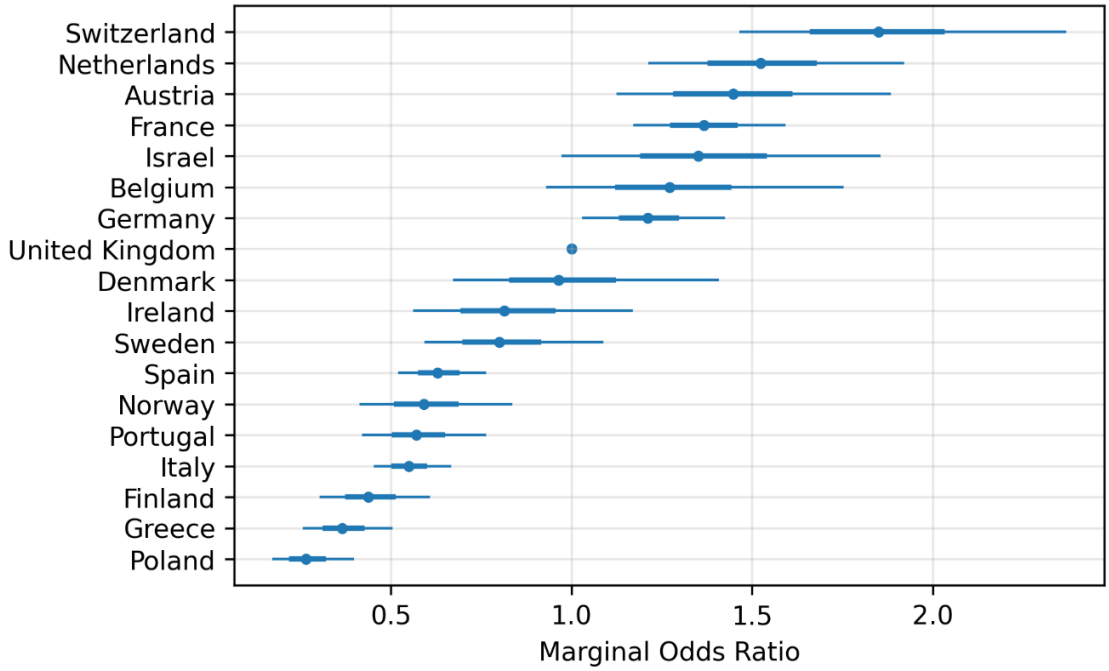

Figure D: Marginal Odds Ratios by Country.
Countries with at least 1,000 proposals

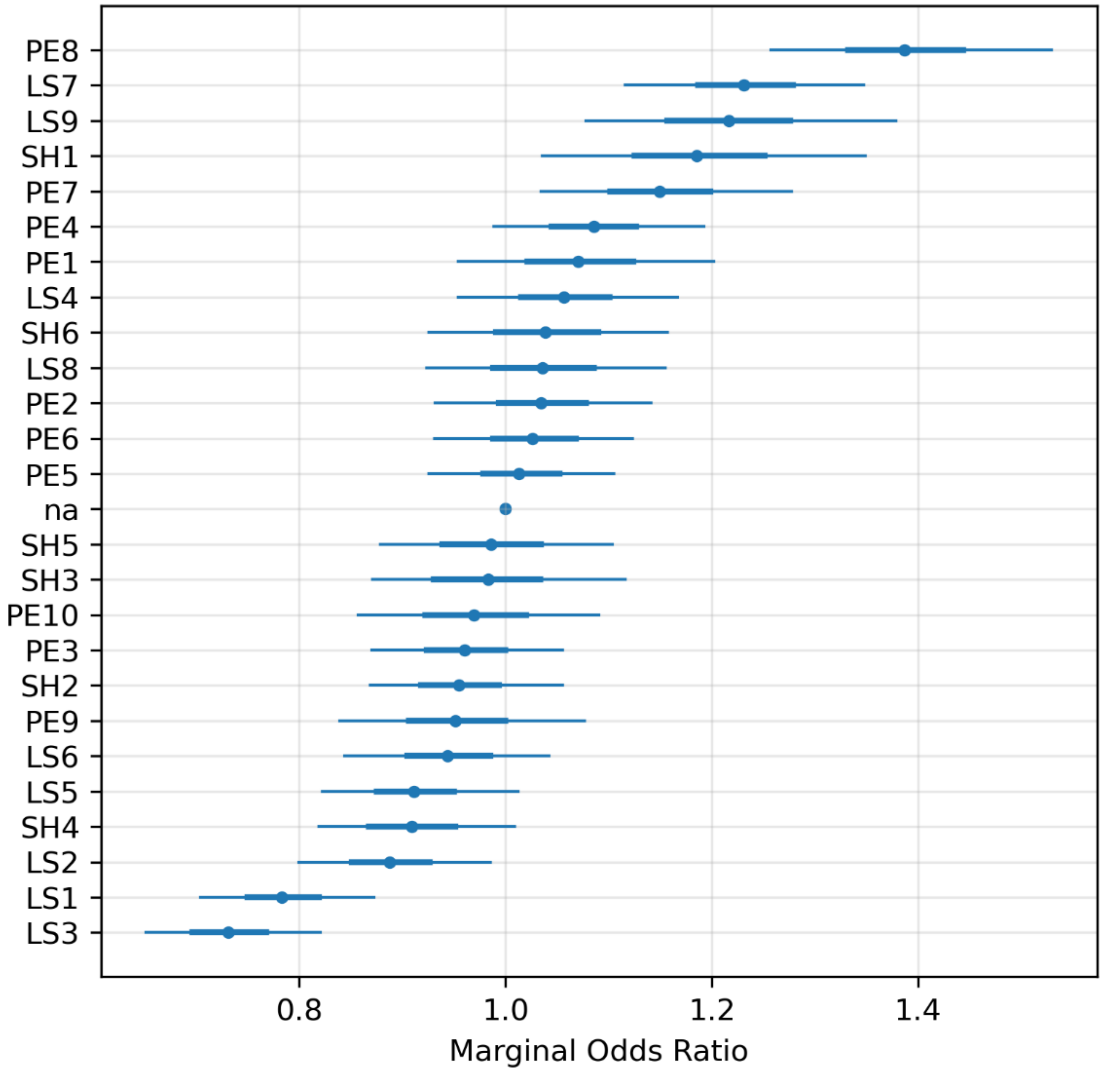

Figure E: Marginal Odds Ratios by Panel

$$
\begin{aligned}
\alpha &\sim \mathcal{N}(0, 2) \\
\beta_{\text{topic}} &\sim \mathcal{N}(0, \sigma_{\text{topic}}) \\
\beta_{\text{panel}} &\sim \mathcal{N}(0, \sigma_{\text{panel}}) \\
\beta_{\text{country}} &\sim \mathcal{N}(0, \sigma_{\text{country}}) \\
\beta_{\text{institution}} &\sim \mathcal{N}(0, \sigma_{\text{institution}}) \\
\beta_{\text{year}} &\sim \mathcal{N}(0, 1) \\
\beta_{\text{funding-req}} &\sim \mathcal{N}(0, 1) \\
\beta_{\text{action-type}} &\sim \mathcal{N}(0, 1) \\
\beta_{\text{inst-type}} &\sim \mathcal{N}(0, 1)
\end{aligned}
$$

$$
\begin{aligned}
\sigma_{\text{topic}} &\sim \mathcal{N}(0, 1) \\
\sigma_{\text{panel}} &\sim \mathcal{N}(0, 1) \\
\sigma_{\text{country}} &\sim \mathcal{N}(0, 1) \\
\sigma_{\text{institution}} &\sim \mathcal{N}(0, 1)
\end{aligned}
$$

Figure F: Weakly informative priors (left), and hyperpriors (right) needed to implement partial pooling across the parameters

3. Exploration and exploitation along the scientific career: the impact of seniority, mobility, and collaboration on individual research agendas diversification

This chapter is the result of a collaborative effort among Carlo Debernardi and Marco Seeber. It will soon be submitted as an independent research article to a scientific journal.

Abstract

Being productive and innovative is a requirement for a career in academia. This leads to an essential tension between exploitation and exploration. Qualitative research has highlighted some micro-level factors affecting scientists' development of their research agenda, but there is scant knowledge on the role played by the institutional context and practices, and little quantitative evidence. This article provides a longitudinal comparative study of 4,785 scientists in four research areas (i.e., Sociology and Political Science, Economics and Econometrics, Immunology, Statistical and Nonlinear Physics) in four European university systems. The existence of a "protected space" has been identified as a crucial condition for exploration. We argue that interactions between a researcher's seniority, scientific mobility, and – crucially – the discipline in which they work, affect the protected space and hence the capability to diversify their individual research agenda. We examined the textual content of the abstracts of 141,690 publications collected from Scopus and explored which factors affected research diversification over the career of individuals, namely a publication's diversity compared to the most similar past publication of the same author. We reconstructed the spread of each scholar's publications within an embedding space obtained via a Large Language Model. The results show that seniority positively affect diversification. Mobility is also positively associated with diversification, especially in disciplinary areas where individuals have more limited protected

space. Finally, new scientific collaborations are the stronger predictor of an increase in diversification.

3.1 Introduction

Understanding what drives scientists to choose a particular problem or shift their research focus has been a central interest of science studies and the sociology of science (Foster et al. 2015). When developing their research agenda, scientists face two main options: they can either specialize and exploit the accumulated knowledge – by incrementally pursuing marginal novelties – or venture into new research problems and topics. As a matter of fact, the competitive academic context requires scientists to be simultaneously productive and innovative. This implies an "essential tension" between productive *tradition* and risky *innovation* (Kuhn 2000),[23] because productivity is more easily achieved through specialization within a focused and consolidated research agenda, whereas innovative contributions often emerge from the exploration and exposure to new and diverse problems, topics, and knowledge domains (Rodan and Galunic, 2004; Colquitt and George, 2011).

Tradition and innovation both entail advantages as well as risks. Specialization increases productivity but reduces the probability of radical innovation. Innovative endeavors fail more frequently and penalize productivity but can stimulate greater innovation and acclaim (Kuhn 1962). This tension leads scientists to develop research agendas that balance tradition and innovation, thus determining various levels of diversity and diversification over time.

Examining what factors affect scientists' diversification process is key to understanding the evolution of scientific knowledge and can help to design better training and career systems (Foster et al. 2015). While research has tried to identify

[23] Note that this dichotomy has been identified and discussed by authors adopting different terminologies, such as reliable "succession" versus risky "subversion" (Bourdieu 1975), "discipline" versus "rebellion" (Polanyi 1969), and "exploitation" versus "exploration" in the study of innovation (March 1991).

factors that affect research agenda and diversification (see Section 2.1), there are still gaps to fill. On the one hand, research incorporating information regarding the overall conditions of the disciplinary and institutional context has been mainly conducted with qualitative methods (e.g. Gläser and Laudel 2015; Laudel 2017; Whitley et al. 2018), thus providing fruitful insights on potential mechanisms but having the limit of being case-specific and hardly generalisable. On the other hand, quantitative studies focused on very general field-level dynamics rather than at the individual level (e.g. Huang et al. 2022) or relied on classification schemes that are discipline specific and top down – like the Physics and Astronomy Classification Scheme (PACS) codes (e.g., Zeng et al 2019; Jia et al. 2017; Wei et al. 2013; Abramo 2024). Hence these studies did not exploit the full potential of textual information, the availability of digital publication data, and of language-based topic models. Finally, so far only a few studies have explored the diversification of scientists' research agenda over their career, which is key to understanding the exploration/exploitation tension. For instance, Aleta et al. (2019) relied on PACS codes to compare the research topics of the first publication year with the last one and observe flows among subfields. Zeng et al. (2019) inferred research topics from the citational network and devised a model inspired by a random walk process. Both these studies capture interesting discipline level dynamics but do not investigate factors affecting the process at the individual level. On the contrary, Abramo et al. (2024) considered individual characteristics, but relied on the Scopus journal classification as a proxy for research topics and had data only for a single country. In summary, while available studies provide useful insights on the process of research agenda diversification, there are clearly open paths to be explored. Finally, Madsen and Nielsen (2024) used statistical matching on individual level characteristics to investigate the effect of thematic or bottom-up grants in the UK on individual research agendas, and operationalised research topics through community detection on the citational network.

This article aims to contribute to the understanding of the factors affecting research agenda diversification. In this regard, a central role is played by the so-called *protected space* (Whitley and Glaser 2014) of a researcher, namely a space that enables the autonomy and the room for taking risks, and which previous studies identified as a major factor affecting diversification. Hence, we argue that the

interactions between a researcher's seniority, scientific mobility, and – crucially – the discipline in which they work, affect the protected space of a scientists and hence diversification dynamics, namely, the capability to diversify its research. Therefore, we examine the scientific production of 4,785 scientists in four disciplines – Sociology and Political Science, Economics and Econometrics, Immunology, Statistical and Nonlinear Physics – thus covering a spectrum of protected spaces' typologies, ranging from (semi) autonomous work, frequent scientific collaboration, and local hierarchical dependence from a group or a laboratory. We do so to explore the impact of seniority, mobility and collaboration on the diversification of individual research agendas.

In addition, we also include in the analysis different country systems, to control for an important context that might affect diversification dynamics. Hence, we consider four systems that all belong to the European Research Area (ERA) but display remarkable differences as to norms, customs, and rules regarding career progression. More specifically, we selected Italy and Norway, where career progression occurs predominantly within the same institution and with relatively low mobility between different locations, and Netherlands and Germany, where mobility is quite common or even required for career progression (Seeber and Mampaey 2022; Macháček et al. 2022).

Empirically, we use a quantitative approach based on the embedding of articles' abstract text via a Large Language Model to explore diversification of the research agendas over the course of individual careers, by measuring the diversity of a scientific output, i.e. an article – compared to the previous individual scientific production.

The remainder of the article is organized as follows. In Section 2, we review studies on the factors affecting scientists' choice of what research problems and topics to explore, the decision to diversify and shift to new research problems and topics, and the diversity of their research production. We then formulate hypotheses about the factors shaping the process of diversification of the research agendas. In Section 3, we describe the data and methods. In Section 4, we present the empirical results while in the last section we discuss the main findings, limitations, and prospects for future research.

3.2 Theoretical framework and hypotheses development

3.2.1 Factors affecting the diversification of the research production

Only a few quantitative studies have examined scientists' diversification of their research over their career. Jia et al. (2017) found that the research interest of individual physicists could shift significantly from the beginning to the end of their career. Zeng et al. (2019) found that scientists cover a narrow array of topics but in recent decades switched more frequently between topics. They showed that frequent switching at the beginning of a career is associated with an initial lower productivity but greater productivity in later career stages. At the same time, they observed that a few switches of topics were associated with high average citation per paper in all career periods and argued that switching topic frequently might reduce the leadership in a specific field, yielding fewer citations. Hence, diversification is initially a riskier choice than sticking to a specific area of specialization. Beyond risk considerations, other factors at individual and institutional level influence scientific interests, whether they will change over time, and how a scientist will combine exploration vs. exploitation, tradition vs. innovation.

Individual level

The decision about what to research is affected by the need to satisfying one's curiosity (Merton 1974), recognition from colleagues (Merton 1974), attract research funds and achieve a career progression (Stephan 1996), but also the desire to produce a positive impact (Iorio, Labory, and Rentocchini 2017; Sormani and Uude 2022). Similar drivers can push a scientist to specialize or diversify their research, at different points in their career. For instance, scientific seniority or age (Jones and Weinberg 2011) can influence the level of curiosity and the stake associated with an innovation, along with one's academic rank. Scientists' gender affects their preference for focusing on a single topic and avoiding shifts (Santos et al. 2020; Abramo et al. 2024). Academic research interests are also driven by personal values and interests, and a misalignment can be common among early-career academics and lead to lower satisfaction (Zhang and Horta 2023). Scientists

can also – consciously or unconsciously – mimic their mentors' career choices (Malmgren et al. 2010).

Institutional context

Scientists are embedded in a complex institutional context that affects their choices through values, norms, and resources (Bazeley 2010). Pressures can come from academic organizations (e.g., departments and universities), national and supranational systems, which regulate and provide financial support to academics, as well as their own disciplinary community. In particular, pressure towards high levels of productivity – i.e. "publish or perish" – can incentivize a selection of topics functional to a rapid influx of publications, to the detriment of research driven by curiosity or societal impact (Ramassa et al., 2024).

A useful framework to understand the role of the institutional context in shaping individual's ability to diversify their agenda is the one of *protected space*, that Whitley and Glaser (2014) define as "the discretion over the use of needed resources, including their own efforts, to pursue particular problems and approaches before having to produce publishable and collectively valued results. It incorporates authority over the choice and formulation of which topics to study, how to do so, and how to obtain and manage resources". Through a qualitative study, Laudel and Bielick (2018) found that the protected space increased with career advancement in terms of academic rank, prestige, and recognition from peers. Furthermore, disciplinary context, country-level hiring systems, funding models and funding availability have been identified as factors that likely play a role in shaping the protected space of individuals in the different phases of their careers (Laudel 2017; Whitley et al. 2018).

A major institutional context affecting the protected space and other dimensions of a scientists' work is the scientific discipline. The scientific discipline is in fact the most important source of values, sense of belonging, and scientific goals for scientists (Abbott 2001; Clark 1983; Knorr-Cetina 1999; Whitley 2000). Disciplines vary considerably in their own intellectual and social organization, the degree of internal cohesion and hierarchical control, and therefore on the influence on topics, methods, and way of work of their members (Whitley 2000). Research problems in some disciplines may be largely disconnected from the surrounding

context, such as research in theoretical or abstract fields like Mathematics, others may depend on the level of available resource, such as Physics or Astrophysics, which require large investments in equipment and infrastructure, while certain research areas can reflect specific challenges, such as research on Climate Change (Debernardi et al. 2024). Furthermore, emerging fields often lack consolidated assumptions and paradigms, and this can grant scientists more freedom (Foster et al. 2015).

The norms, goals, and level of resources of the department and the university of affiliation can also affect scientists' research agenda and their attitudes towards risk (Ebadi and Schiffauerova, 2015; Hillson and Murray-Webster, 2007). For instance, in some institutions, researchers rely primarily on external grants, which can push them to conform to the research goals of the funding organization (Leisyte and Dee, 2012; Seeber 2013). A survey by Horta and Santos (2020) showed that the scientists' level of autonomy and support granted by their organization is crucial for social scientists to develop ambitious, multidisciplinary, collaborative, and risk-taking research agendas. Furthermore, an increased reliance on competitive grant funding, as opposed to block funding, has been linked to a decrease in the protected space, especially for early career researchers (Whitley et al. 2018). The size of a department can also matter, as large departments are less likely to be dominated by a single topic and paradigm, giving scientists more leeway (Foster et al. 2015).

The external context can affect the choice and shift of research topics, through commercial opportunities (Evans 2010), scientific funding opportunities (Hoonlor et al. 2013; Seeber 2013), pressures (Vallas and Kleinman 2007), and commercially related policies (Berman 2012) that increase interest in the supported areas. The characteristics of a funding program also impact on the rate of scientific exploration of the recipients: programs tolerating early failure, rewarding long-term success, and giving appointees great freedom to experiment, favor the exploration of novel lines of inquiry and the production of high-impact articles (Azoulay et all. 2011). Finally, awards can also play a role and stimulate scientists to take risks (Wright 1983).

3.2.2 Exploring factors affecting diversification

In this section we first discuss three key conditions to diversify one's research production: the possibility, opportunity, and capability to do so. Accordingly, we identify three factors that affect these conditions – seniority, mobility, and co-authorship – and develop several hypotheses about their impact on the diversification of a scientist's research production, namely on the level of diversity of a publication compared to the most similar past publication of the same author.

The possibility for diversification: a protected space.

In the early stages of their career, researchers experience a limited protected space. Laudel and Glaser (2008) propose a framework in which individual careers can be tracked according to three dimensions: an organizational career, a cognitive career, and a community one. The first relates to formal positions (e.g. PhD student, postdoc, etc.), the second relates to the research agenda content, while the last one represents the social standing in the community as captured by prestige and recognition. At the start of their career researchers are typically, from the standpoint of the community career, in the *apprentice* phase. They must undergo a process of learning in order to understand and master a specific research domain under the supervision of senior scholars. This means that a junior scientist will initially focus on specific themes, problems, and methods. This master-apprentice relationship also frequently implies a strong embeddedness in a laboratory, group, or dependency on a chair – according to the disciplinary and institutional context – and therefore, a relatively low freedom to embark into new and independent research directions. From this starting point, career progression is then dependent on scientific productivity, which can also be more efficiently and effectively achieved through specialization. In sum, several factors lead to the expectation that only over time a scientist can gradually achieve scientific maturity and a stable position, enabling autonomy and freedom to risk, i.e., a greater protected space.

Therefore, we formulate the hypothesis that:

> *Hypothesis 1.1) Scientific seniority positively affect the diversification of a scientist's research production*

Arguably, the effect of scientific seniority will be particularly strong in scientific disciplines characterized by intense laboratory work, tight research group collaborations, greater dependence on the research group and embeddedness in a common agenda, and hence more limited protected space. In fact, there are disciplinary variations in how strong local hierarchy are. Comparing molecular biology and history, Laudel (2017) underlined how crucial the choice of applying for specific positions is in the former case, given the lack of control that Early Career Researchers (ECRs) have on the development of the research to be carried out: "they could not control the range of topics of their protected space. Once a researcher had chosen a position, research had to be undertaken within the research area of the group leader. In the following case, the researcher had selected a position which turned out not to fit his interests. Since the topic was determined by the group leader, the only option he had was to leave this position before the end of its term" (Laudel 2017).

Another classical example of local hierarchy in lab-based disciplinary areas is provided by the field work of Latour and Woolgar (1986). In this analysis they described with a wealth of details the hierarchical structure of a laboratory, where technicians, ECRs, and sometimes even less prestigious professors, are highly specialized in a well-defined set of tasks, have a very low degree of autonomy, and are explicitly considered as easily replaceable. We thus hypothesize that:

> *Hypothesis 1.2) The effect of seniority on the diversification of a scientist's research production will be greater in disciplines characterized by stronger local hierarchy (e.g. laboratory-based fields).*

The opportunity and capacity for diversification: scientific mobility and co-authorship

The chances of venturing into a new research problem or topic are favoured not only by the power to do so, in terms of autonomy and freedom to risk, but also by the opportunity and capability to do so. Namely, a researcher is more likely to diversify their agenda, when they are exposed to new problems or topics, as well as when they have the skills, knowledge, understanding of the problem in which she desires to work on.

In this regard, a major factor leading to diversification is arguably scientific mobility – especially when it involves a stable change of institutional affiliation or even moving to a new location. Two mechanisms are likely to matter. By moving, a scientist becomes durably embedded in a new social and working environment, which exposes her/him to new ideas, methods, problems, as well as provide opportunities for new co-authorships.

This leads to the following hypothesis:

> *Hypothesis 2.1) Publications following a durable mobility episode will display greater diversity.*

This will be particularly the case for fields characterized by laboratory work and tight group collaboration and when mobility occurs in the early career stages, since a junior scientist must adapt its research agenda to the new group in which they will be working.

> *Hypothesis 2.2) The effect of mobility on publication diversity will be greater in disciplines characterized by stronger local hierarchy (e.g. laboratory-based fields).*

> *Hypothesis 2.3) The effect of mobility on publication diversity will be greater for early episodes.*

A further pivotal mechanism that can realize both the opportunity and capability to diversify is collaboration with new colleagues. Working with new scientists, a researcher may learn about or get involved in new problems; at the same time, a scientists may look for new collaborators to fill some gaps in skills or knowledge that are necessary to conduct research in an area that she does not fully master. In other words, a new collaboration may be the cause and/or the instrument to explore new problems or topics.

> *Hypothesis 3) Publications with a new co-author will display greater diversity.*

3.3 Data and Methods

3.3.1 Selection of the cases

We selected four disciplinary areas and four country systems aiming to cover contexts with variations in the characteristics that are explored in our hypotheses. Hence, we selected disciplines spanning across the social, life and natural sciences and that are likely to differ in terms of the *protected space* of their members. Since the selection of cases had to be coordinated with data collection, we relied on the Scopus Subfields classification to cover the disciplinary areas of interest.

In the social sciences we opted for two disciplinary areas that are representative of different structures. On the one hand, the Subfield of *Sociology and Political Science* represents a disciplinary area lacking a unifying foundational theory, with parallel strands of literature dealing with different topics and/or moving from different assumptions. On the other hand, the Subfield of *Economics and Econometrics* represents a disciplinary area with a more centralized structure, both in terms of theoretical foundations and of prestige hierarchy – e.g., top journals having a key role in the discipline (Heckman and Moktan 2020) – even though the diversity of topics and approaches has been increasing, stretching outwards the boundary of the disciplinary mainstream (Cedrini and Fontana 2018).

Life sciences are represented with the Subfield of *Immunology.* This represents a disciplinary area with lab-based research. As previously discussed, this implies a strong and quite explicit hierarchy at the local level of research groups.

From the natural sciences we selected *Statistical and Nonlinear Physics.* This Subfield is characterised by a prevalence of theoretical and/or computational research, and the absence of extremely large-scale collaborations – with hundreds of authors – that are typical of some areas of Physics. Indeed, such large collaborations would put in question the relevance of any single publication for the development of individual research agendas, since it would be very difficult to disentangle individual contributions.

As for the level of the countries, we selected the cases similar enough to be comparable, but covering different career systems and levels of mobility. In particular, we chose to include four Western European countries part of the

European Research Area (ERA): i.e., Germany, Italy, the Netherlands and Norway.

Notably, academic career progression in Germany is based on a chair system, while the other countries implement a tenure system (Laudel 2017). Furthermore, Germany and the Netherlands display high levels of national and international mobility – as a result of a combination of norms and rules – while Italy and Norway have comparatively lower levels of mobility (Macháček et al. 2022; Seeber and Mampaey 2022).

3.3.2 Data collection and selection

We aimed to select a sample including scholars that i) are unambiguously embedded in the disciplines we selected; ii) have been exposed to the institutional incentives of the countries we are interested in for a large enough portion of their career; iii) have a publication track record long enough to enable us to examine their research agenda over a sufficient time.

Hence, we collected data from Scopus with an iterative approach to identify a sample of authors that could be reliably classified as members of the respective disciplinary areas. In general, scholarly databases include disciplinary classifications for publications rather than authors. We thus chose to proceed by first identifying relevant publications in the selected areas (according to the All Science Journal Classification – ASJC classification[24]). As a second step, we identified the authors of said publications and recovered information about their full scientific production. We finally filtered the authors, keeping only those for which the publications in the selected areas represented most of their total production. This general procedure was – in more detail – implemented as follows.

First, for each Subfield we gathered the metadata of all the publications (restricted to articles with an English abstract) that included at least one author affiliated in one of the selected countries (GE, IT, NL, NO).

[24] Available at: https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/.

Second, we selected only the authors with at least two publications in this first sample and collected data about all their publications – including those outside the AJSC Field and Subfield[25].

Third, we aimed to identify authors whose main subfield of research is the one under study. Hence, we first kept only the authors whose most frequent field of publication is the one under consideration, and fourth, in this sample, we kept the authors whose most frequent Subfield of publication is the one under consideration.

At the end of this procedure, we ended up with a sample of scholars that are embedded in the scientific disciplines we selected. To enforce our additional requirements, we implemented some additional filters and retained in our sample only authors i) with a publication track record of at least six years – measured as the distance in time of their last Scopus indexed publication from their first one, and in line with Macháček et al. (2022) – ii) have published at least a paper every two years, and iii) have reported to be affiliated to an institution in the considered countries in at least 30% of their publications.

Table 1 reports the considered Subfield, the ASJC Field and Subject Area they belong to, and the number of authors by field.

[25] In the case of the Immunology Subfield, we restricted the sample to 20,000 randomly selected authors among those with two publications in the initial sample (we did this to limit the otherwise excessive amount of data to be collected).

Subject area	Field (and codes)	Subfield (and codes)	Label	N authors
Social Sciences & Humanities	General Social Sciences (3300-3322)	Sociology and Political Science (3312)	SPS	878
Social Sciences & Humanities	General Economics, Econometrics and Finance (2000-2003)	Economics and Econometrics (2002)	ECO	1,648
Life Sciences	General Immunology and Microbiology (2400-2406)	Immunology (2403)	IMM	1,599
Physical Sciences	General Physics and Astronomy (3100-3110)	Statistical and Nonlinear Physics (3109)	SNP	660

Table 1: Fields and Subfields.

3.3.3 Method for diversification measurement

By concatenating title and abstract of the publications we obtained a textual corpus that we used as input for a Large Language Model – namely Sentence-BERT (Reimers and Gurevych 2019) – whose output is a high-dimensional embedding space in which texts that are semantically similar are represented as points close to each other.[26]

We can thus analyze each author's research production as a collection of publications – i.e. vectors/points in the space – gradually increasing in items over time. To examine and quantify the diversification of the research agenda of an individual, we consider the distance of a newly added publication from the most similar past publication of the same author. Namely, we measure the *cosine distance from the closest past publication from the same author*. From an intuitive point of view this measure can be better understood by considering it as the length of a metaphorical "step" out of the epistemic area the author is already familiar with. It measures the marginal contribution of an additional publication to the diversity of an author's scientific production.

Given this choice of outcome measure we consider as observations each author-publication pair. This means that the same publication can appear multiple times, each with a different focal author and thus with a different value of the outcome, given that the previous track-record of publications is different. For obvious reasons, this measure is defined only starting from the second publication by the focal author.

Figure 1 provides a graphical illustration of the rationale of the measure. In this example, each node represents a publication by the focal author – with numbers defining their chronological order – placed in a two-dimensional space where similar publications located closer to each other. Each arrow thus corresponds to

[26] While this is a pre-trained general model, not fine-tuned to scientific publications exclusively, it is worth noting that Semantic Scholar is one of the sources of training data and most likely a large part of the publications we are considering were included in the training sample.

one observation in our dataset. It can be noticed that arrow *d* encodes the diversification related to P5 (publication number 5); its length does not depend on the distance by P4, because P1 is closer. Vice versa, arrow *a* encodes the diversification brought in by P2, it starts from P1 despite P6 being closer to P2; this is because at the publication time of P2, P6 did not exist yet.

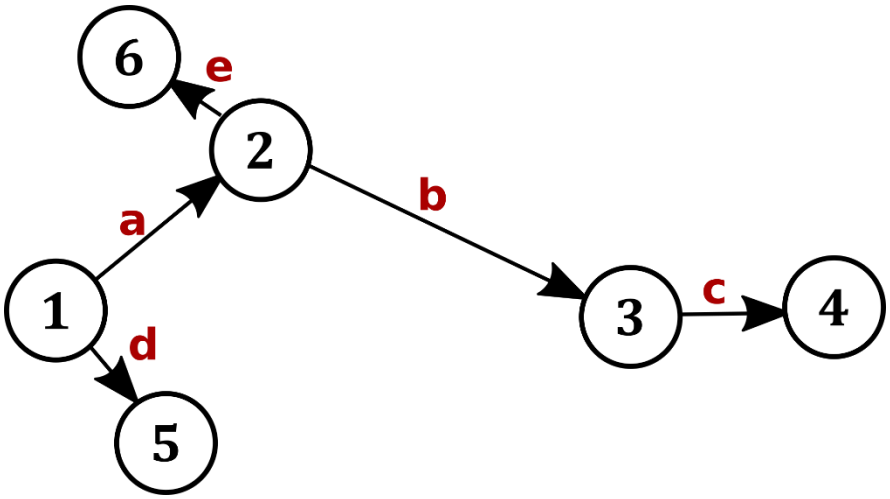

Figure 1: Graphical illustration of the diversification process. Nodes are publications by a focal author, numbered by their chronological order. Arrows correspond to observations in our dataset, and their length measure the amount of diversification from previous works by the author.

3.3.4 Inferential method

We estimated several Bayesian hierarchical models separately for each of the four disciplinary areas using *Stan*, with the *cmdstanpy* interface (Stan Dev Team 2024).

We used hierarchical linear models and included a varying intercept for each focal author. The intercepts are regularised through partial pooling. We selected weakly informative but broad priors given the scale of the outcome measure (see below). The model can thus be described as follows, with *X* being the predictor matrix and *j* an index running on the focal authors:

$$y \sim \mathcal{N}(\alpha + \beta X + \gamma_i, \sigma)$$

Equation 1: Main model

The overall intercept of the model and the effects are given a normal prior as follows, with a standard deviation allowing for a wide range of behaviors given the scale of the outcome variable.

$$\alpha, \beta \sim \mathcal{N}(0,0.5)$$

Equation 2: Priors for the intercept and the covariate effects

Finally, the varying intercepts gamma are assumed to be drawn from a normal distribution with unknown standard deviation. This allows to account for the individual variability among focal authors.

$$\gamma \sim \mathcal{N}(0, \zeta), \;\; \zeta \sim exp(1)$$

Equation 3: Prior and hyperprior for the varying intercepts and their standard deviation

3.3.5 Variables

As predictors, we consider the following variables – computed for each observation, which consists of a pair "focal author-publication":

Focal author: the author under consideration. By including it, we explicitly account for unmeasured individual level variability given for example by personal propensity, curiosity etc. Technically, we do this by introducing author-level varying intercepts, regularized through partial pooling.

Dependent variable

Diversification: measured as the cosine distance in the embedding space from the closest past publication from the same author, as defined in Section 3.3.

Independent variables

Seniority: measured using the number of years passed from the first publication in Scopus as a proxy. Log transformed. This measure is necessary a compromise – broadly used in the scientometric literature – due to the unavailability of demographic data about the individuals in the sample.

Mobility: a dummy variable encoding a durable change in affiliation. This variable was computed from the affiliation city reported in the publications, to which we applied a standardization to fix a single value in each year. This was done to overcome challenges linked with self-reported affiliations: e.g., multiple affiliation, inconsistent reporting, hierarchy of institutions or name changes, etc. A more detailed explanation of the standardization procedure is available in the Appendix. In a modified version of our baseline model, we further subdivide these variables into separate dummies for the *nth* occurrence of mobility events along individuals' careers.

New coauthor: a dummy variable that has value 1 if at least one of the authors of the current publication collaborates with the focal author for the first time.

Control variables

Besides the main variables of interest, we account for the possible effect of other factors, namely: gender, since other studies mentioned a possible impact (e.g. Abramo et al., 2024); country of affiliation, since a different institutional environment exposes to different incentives; number of coauthors, given that we want to differentiate between new and lasting collaborations; and past publications, since their number is mechanistically related to the outcome measure (i.e. the more past publications the less avenues for agenda diversification).

Gender: estimated with a simple matching procedure of the given names using the *World Gender Name Dictionary v2* (Lax-Martinez et al. 2021). We were able to classify around 90% of the authors (overall 64% M and 26% F) and labeled the remaining ones in a third category.

Country: the specific country (possibly more than one) in which the focal author is reported to be affiliated in the current publication. This information is included in the model as the fractional counting of affiliations reported by the focal author in each of the considered countries. This definition of the variable implies that the reference term for the regression coefficients are publications with reported affiliation of the focal author in a country other than the selected ones.

Number of authors: number of authors of the current publication. Log transformed.

Number of past publications: cumulative number of publications from the focal author, up to the year before the current publication. Log transformed.

3.4 Empirical analysis

3.4.1 Descriptives

In this Section, we summarise descriptive information in order to provide a clearer picture of our measures and samples.

Figure 2 reports boxplots for the outcome measure by disciplinary area. The distribution is approximately normal, with a slightly stronger right tail. While theoretically the measure is defined in the [0,2] interval (being 1 minus a cosine), the observed range is much narrower. Indeed, a value close to 2 would imply a step from a starting point in the embedding space to the orthogonally opposite one.

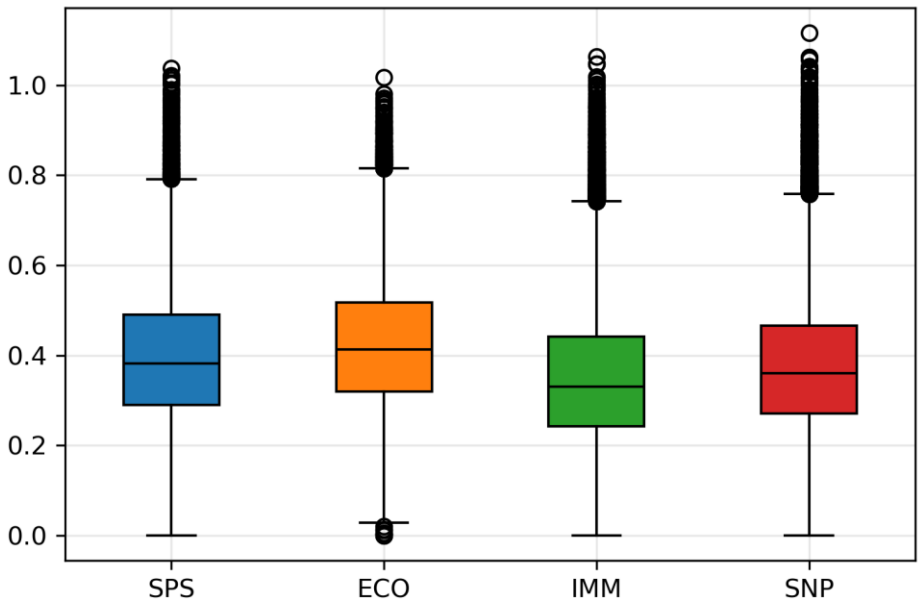

Figure 2: Distribution of the outcome variable (i.e., diversification) across Subfields.

Another key variable is seniority. Table 2 summarises the distribution of the maximum observed seniority per focal author. We can observe a substantial

heterogeneity across Subfields, with *Immunology* and *Statistics and Nonlinear Physics* displaying longer career trajectories.

	First quartile	Median	Mean	Third quartile
SPS	11	14	16.1	19
ECO	13	18	19.7	25
IMM	12	20	22.2	31
SNP	15	25	26.4	36

Table 2: Descriptive statistics of the maximum seniority per focal author by Subfield.

Table 3 reports summary information about the distribution of the cumulated number of publications per focal author. Also in this case we observe high levels of heterogeneity, with the two above mentioned Subfields having substantially higher publication outputs.

	First quartile	Median	Mean	Third quartile
SPS	11	15	19.2	22
ECO	10	16	20.4	25
IMM	14	26	43.6	52
SNP	23	42	62.7	76

Table 3: Descriptive statistics of the cumulated number of publications per focal author by Subfield.

Finally, we overall observe 13,124 publications (around 8% of the total observations) linked to 7,271 mobility events distributed across 3,225 focal

authors. These observations are divided across Subfields and countries as reported in Table 4.

	Germany	**Italy**	**Netherlands**	**Norway**	**Others**
SPS	684	245	443	112	409
ECO	1,484	654	494	66	536
IMM	1,655	785	629	102	1,000
SNP	1,337	1,166	145	33	1145

Table 4: Number of publications linked to mobility events across Subfields and countries.

3.4.2 Inferential analysis

In this Section we discuss the results of the inferential analysis. We present the results in graphical form, with dots representing the posterior mean, thin whiskers for 95% credible intervals, and thicker ones covering the first to third quartile of the posterior. The results are available in tabular version in the Appendix.

Figure 3 displays the regression coefficients of the first model, estimated separately on each of the four disciplinary areas. We use this model – i.e., Model A – as a baseline: the other models, as well as all the robustness checks, are implemented as slight variations of Model A.

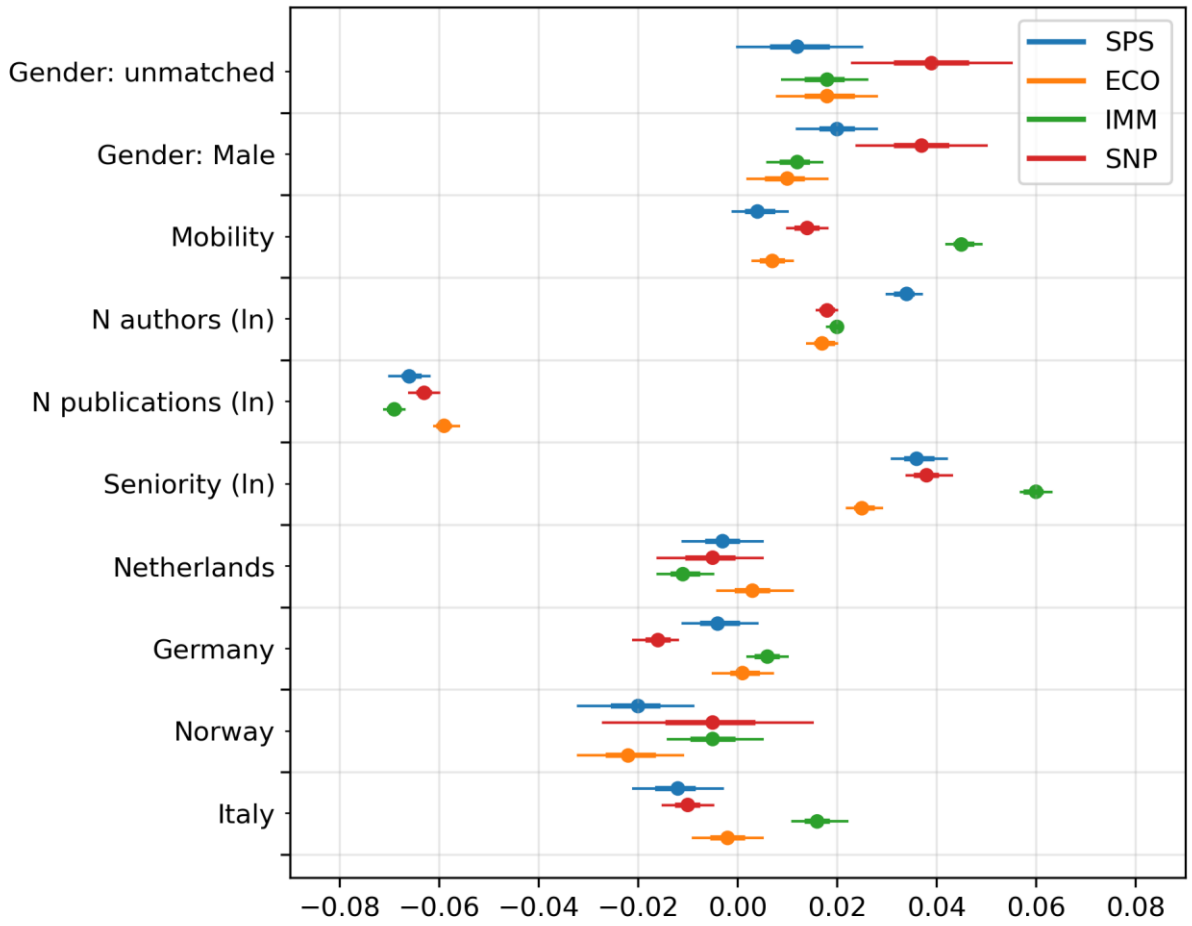

Figure 3: Model A results. Dots represent the posterior mean, thin whiskers are 95% credible intervals, thicker ones cover the first to third quartile of the posterior.

The results reveal that the marginal effect of the academic seniority is positive across all fields. This is in line with our hypothesis 1.1, and compatible with the idea that, while growing in seniority, scholars gain a more stable and legitimate position both in formal terms – e.g., by gaining tenure – and informal ones – e.g., by transitioning from being apprentices to recognised colleagues.

In line with our hypothesis 1.2, this association is stronger for the Subfield of Immunology. We argued that this is due to the presence of local hierarchies in the laboratory context, which limit the protected space. Seniority is thus more salient for the ability to independently set at least part of an individual's research agenda, since it gradually reduces the strength of the dependence relation at the local level, or even, eventually, allows researchers to have their own apprentices.

It should be stressed that the negative effect of the cumulative number of publications is due to the very definition of our measure of diversification as the marginal contribution to diversity of a new publication. This implies that the greater the number of past publications, the higher the chances that a new publication will have similar past publication and hence smaller diversification contribution. Instead, this does not imply that an additional publication may have a negative diversification score, because an additional publication will always have a positive cosine distance – diversification value and hence contribute in some way to the overall diversity of the scientific production (unless completely identical to a past one, in which case it will have a zero score). When interpreting these marginal contributions, the reference term is thus the average diversification expected at the start of the career – i.e., the diversification given by the second publication (since for the first one the measure is not defined).

The negative effect of the cumulative number of publications will often counterbalance the effect of seniority: as career progresses, the number of publications will increase – curbing diversification – whereas the protected space will increase – enabling diversification. Understanding the combined contribution of these two effects is not straightforward: first, their balance depends on the rate of publication, and second, both variables have been log transformed. To support this interpretation, Figure 4 reports the marginal diversification under the conjoint effect of these two factors back-transformed in their natural scale. This is computed by considering only the point estimate of the effects of seniority and number of publications for three ideal-profiles of scholars: i.e., a scholar at the first quartile of the number of publications, one at the median, and one at the third quartile. These quantiles are in turn computed from the observed cumulative distribution of the number of publications for each seniority level by field (reported in the Appendix). By looking at Figure 4, it is clear that – according to our model – despite the positive effect of seniority, the overall trend along the scientific career indicates decreasing diversification: i.e., the marginal contribution of diversity of the *nth* publication decreases.

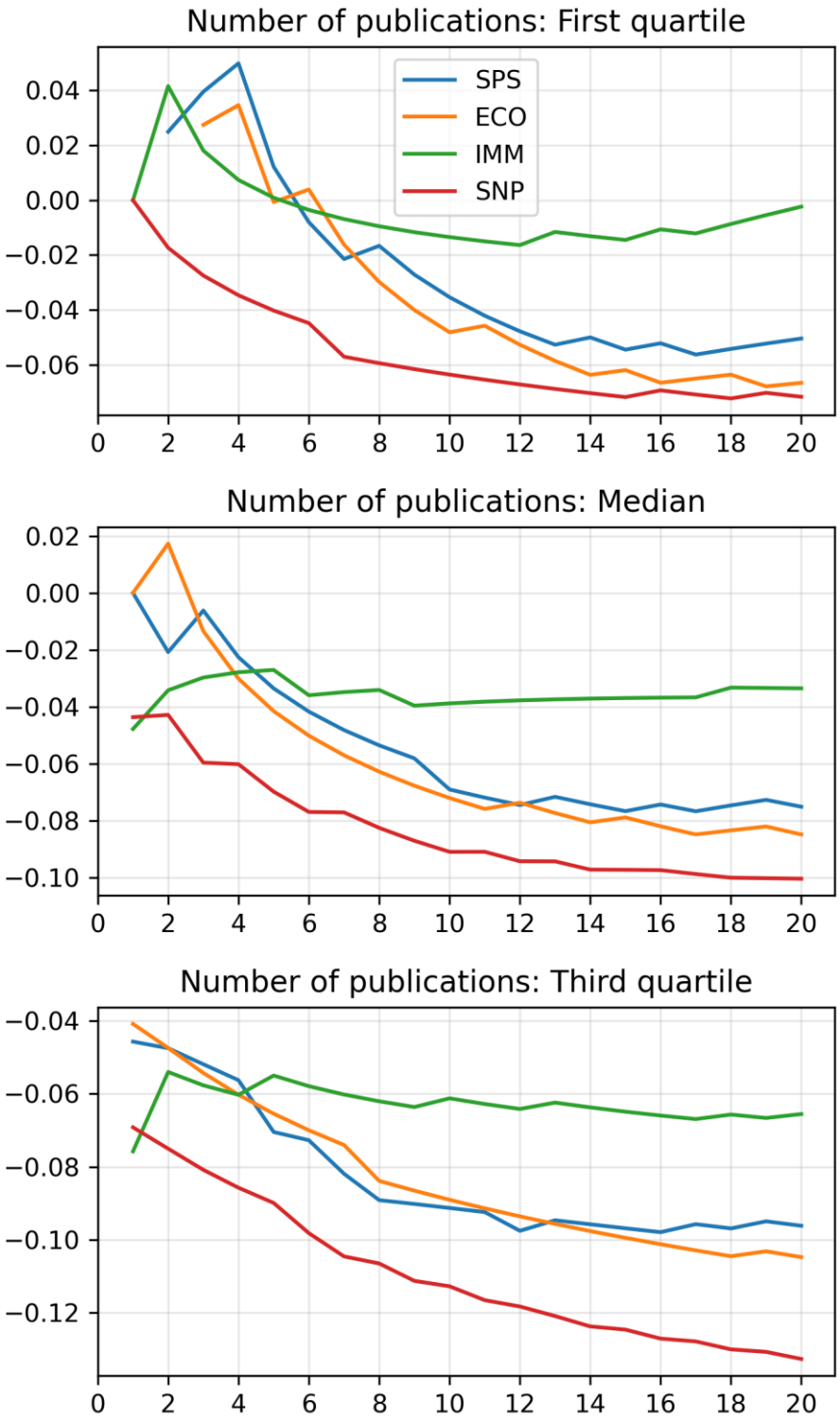

Figure 4: Trends of marginal diversification (y-axes) over the seniority levels (x-axis) and number of publications, computed following the observed publication rates for scholars in the first quartile, median and third quartile of the distribution.

Moving on to hypothesis 2, which regards the effect of mobility, we have enough evidence to reject a traditional null hypothesis $H_0 = 0$ in three out of the four disciplinary areas, with the exception being *Sociology and Political Science*. The magnitude of these effects though is severely limited, with the only exception of the Subfield of *Immunology*. Notably, the stark contrast of this latter case with the other ones is compatible with our hypothesis 2.2 of a stronger effect in laboratory-based disciplines.

To address hypothesis 2.3 – regarding the role of early mobility events as opposed to later ones – we need to consider Model B, where the mobility events have been included with different dummy variables related to their cumulative count – i.e., we distinguish the first mobility event in a career from the second, and so on. The results of this model, reported in Figure 5, provide support to the idea that the early mobility episodes are those that matter the most for the research agenda diversification process. Furthermore, we can see that in this version of the model, the magnitude of the effect of the first mobility event in the Subfield of *Statistical and Nonlinear Physics* is substantially higher. This further disaggregation thus allowed us to observe an effect otherwise hidden by attenuation.

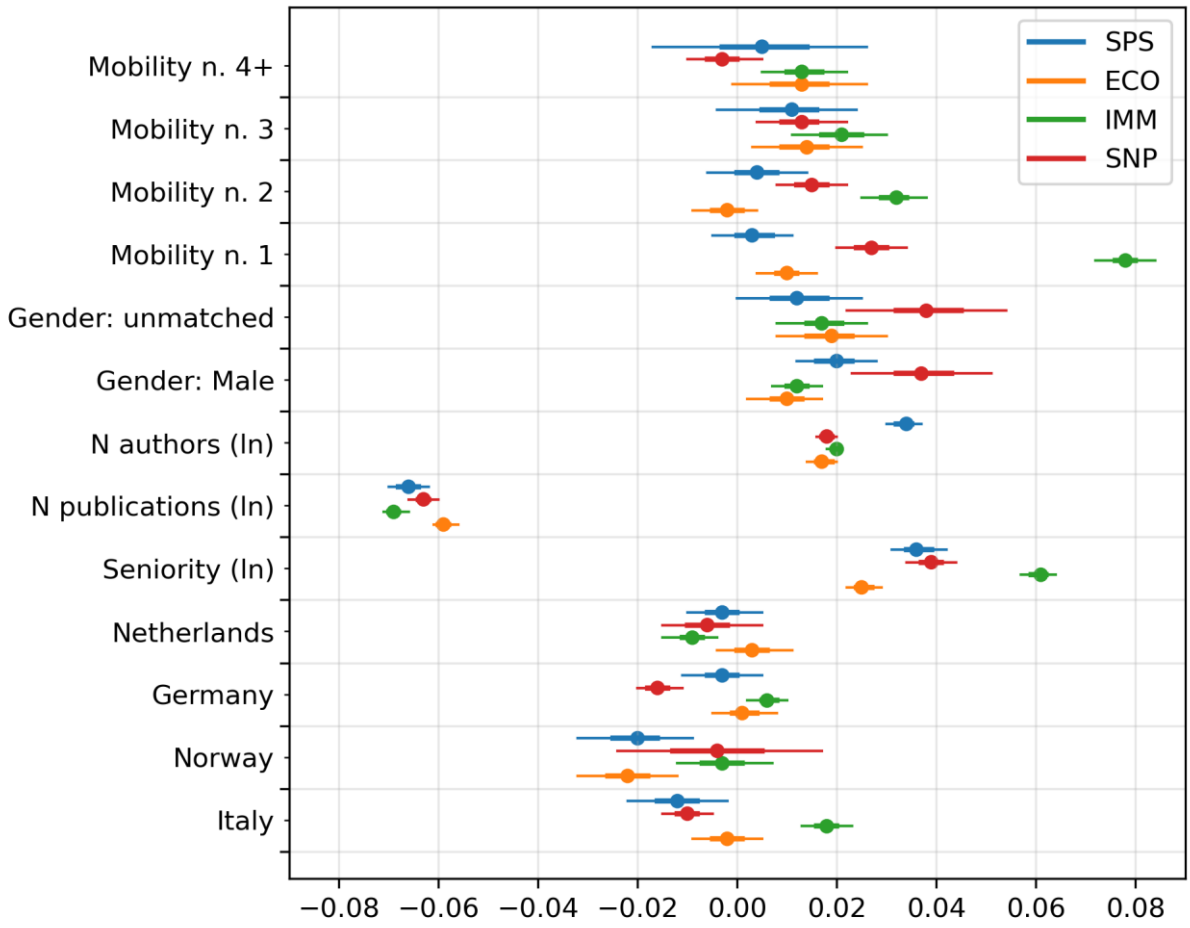

Figure 5: Model B results. Dots represent the posterior mean, thin whiskers are 95% credible intervals, thicker ones cover the first to third quartile of the posterior.

We finally consider our hypothesis 3, related to the role of new collaborations. To do so we can inspect the results of Model C, reported in Figure 6, which specifically includes a dummy variable for the occurrence of collaborations of the focal author with scholars they never worked with before. The effect of a new collaboration is positive, and has similar values, across all fields. It also notably has a very large magnitude, making this the single largest contributor to research agenda diversification events. By comparing the results of Model A and C, we can also see that the coefficient for the number of authors flips from positive to negative, indicating that – as we would expect – the diversification contribution mostly comes from *new* collaborations.

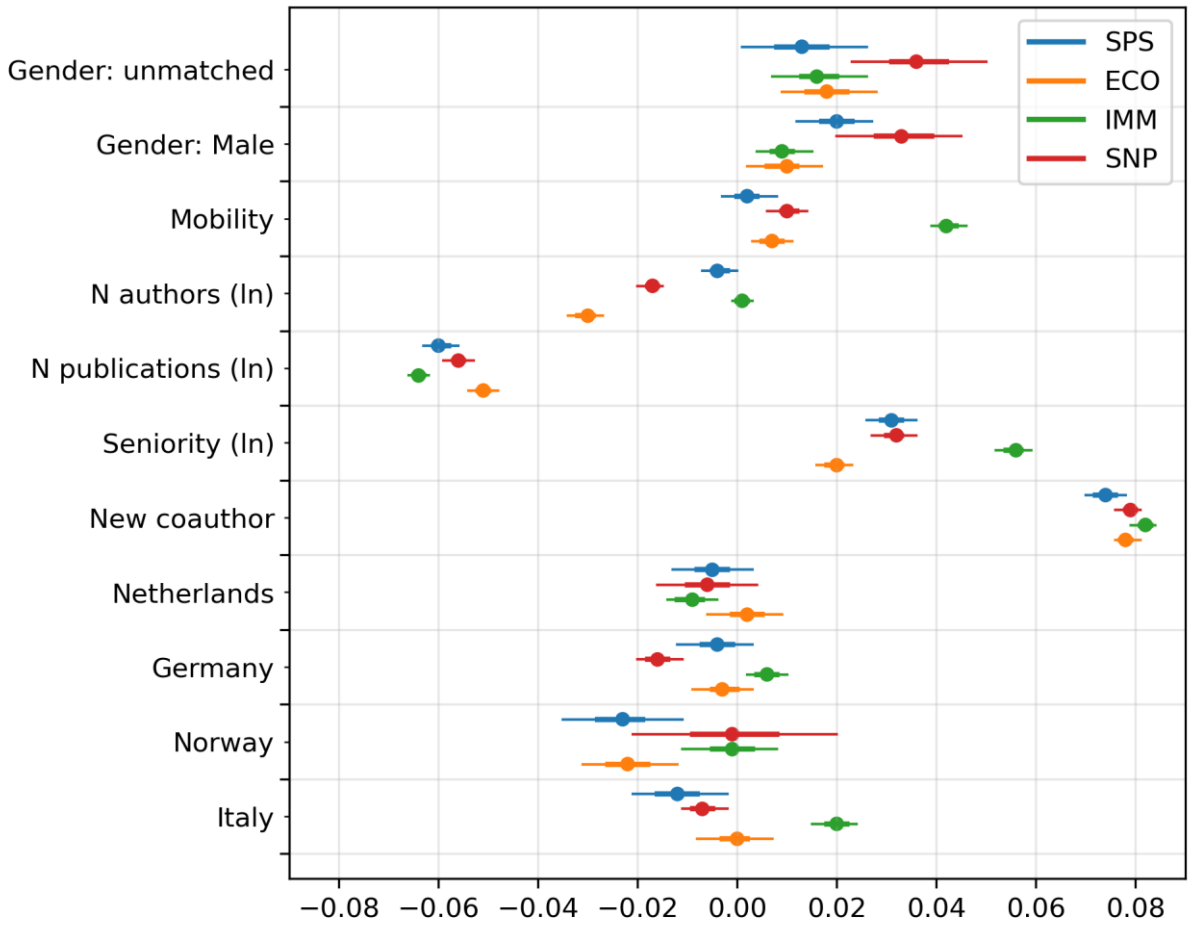

Figure 6: Model C results. Dots represent the posterior mean, thin whiskers are 95% credible intervals, thicker ones cover the first to third quartile of the posterior.

Apart from our starting hypotheses, it is worth mentioning some other interesting results. First, we observe an association between different levels of diversification and gender, with in particular women displaying on average a more specialised research agenda. This would confirm previous findings by Abramo et al. (2024) in a study on Italian scholars across all disciplinary areas. In the framework of the current contribution, this might be related to a more limited protected space for female scholars, but more in depth investigation is needed to formulate any hypothesis about the possible mechanisms generating this pattern, such as differences in risk-taking behaviour (e.g., Byrnes et al. 1999).

Furthermore, we do not observe any systematic pattern of variation across countries. Again, this is in line with findings from Laudel (2017), underlining that different institutional structures can produce functionally equivalent outcomes.

Overall, we observe that the effects of the main variables of interest across the four disciplinary areas are consistently either positive or negative, but also have important variations in the magnitude of the effects. Arguably, this reflects the capability of the model to capture the general underlying features of the phenomenon of research agenda diversification.

To sum up, we can say that, on average, the pattern of development of individual research agendas is decreasing diversification. This is largely due to the fact that the marginal contribution to diversification of subsequent publications is decreasing, i.e., the more past publications an individual has, the more likely the new ones are to be close to the older ones. The protected space gained through career progression and peer recognition – measured through academic seniority – is partially able to limit this overall tendency, especially for scholars with relatively lower publication rates. Mobility events – mostly the early ones – can represent points of discontinuity, allowing for a temporary surge in exploration, especially in disciplinary areas characterized by tight local hierarchy. Furthermore, co-authorship with new people is the single most important predictor of a temporary increase in research agenda diversification.

3.5 Conclusions

Scientist's individual decisions about what questions and topics to research collectively shape science and the knowledge that will be produced. It is hence of pivotal importance to understand what affect similar decisions. In this regard, scholars of science highlighted that when developing a research agenda, scientists must aim to be simultaneously productive and innovative, as both are key conditions to pursue an academic career. This leads to an "essential tension" between productive tradition and risky innovation (Kuhn 2000) and the need for scientists to balance specialization and innovation in their research agendas, leading to varying levels of diversity and diversification over time.

Research exploring the factors that affect research agenda diversification has generated important insights, through empirical studies almost exclusively qualitative and/or cross-sectional. This article aims to contribute to the literature through a longitudinal and quantitative approach based on the embedding of articles' abstract text via a Large Language Model to explore diversification of the research agendas over the course of individual careers. Conceptually, we focus on the so-called *protected space* (Whitley and Gläser 2014) of a researcher, namely a space that enables the autonomy and the room for taking risks, and which previous studies identified as a major factor affecting diversification. We explored how the interactions between a researcher's seniority, scientific mobility, and – crucially – the discipline in which they work, affect the protected space and hence diversification dynamics. Therefore, we examined the scientific production of 4,785 scientists in four disciplinary areas – Sociology and Political Science, Economics and Econometrics, Immunology, Statistical and Nonlinear Physics – characterized by different protected spaces, due to various rooms for (semi) autonomous work, frequent scientific collaboration, and local hierarchical dependence from a group or a laboratory. In addition, we also considered four country systems – Germany, Italy, the Netherlands, and Norway – to account for different norms and rules regarding career progression, and particularly to the extent to which mobility is common or even required (Seeber and Mampaey 2022; Macháček et al. 2022).

We found that academic seniority has a positive marginal effect on the expected diversification of individual research agendas. This is arguably linked to an increase in protected space due to both formal career progression and informal peer recognition. In line with the idea that protected space is more constrained for ECRs in laboratory disciplines – due to stronger local hierarchy – we observe a stronger effect in this disciplinary area. Furthermore, mobility is a possible way to obtain more independence. Indeed, we observe a positive effect of mobility on research agenda diversification, which is greater in disciplinary areas with strong local hierarchy. Finally, we show how new scientific collaborations are the single most impactful predictor of an increase in research agenda diversification. By collaborating with others, scholars are exposed to new ideas and topics, and/or can

leverage theories, methods and other forms of knowledge to address new ideas and topics and that would otherwise be unavailable to them.

This article is also the first to employ the new tools provided by Large Language Models to the study the evolution of individual research agendas and shows the potential of these methods to provide fine grained measures based on the textual content of scientific publications.

To sum up, it is important to underline that being able to understand and quantitatively measure the evolution of individual research agendas is crucial in the attempt to formulate effective science policies. Indeed, as of today, decision makers take crucial decisions on the future of scientific research neither with information on the distribution of research topics, nor considering the potential effect their decisions could have on the evolution of said distribution. This work provides both a deeper understanding of the factors at play and an example of a possible method to systematically gather relevant information about the phenomenon.

Of course, some limitations must be underlined. First, from the point of view of the methods, more benchmarking of classical techniques – like citational analysis – vis a vis the new ones, is required in order to have a better understanding of the measures practically employed. Second, further work is required in trying to apply these methods in more controlled contexts, in which it would be possible to go beyond observational data and towards the identification of causal effects.

References

Abbott, A. (2001). *Chaos of disciplines*. Univ. of Chicago Press.

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2024). Do research assessment systems have the potential to hinder scientists from diversifying their research pursuits? *Scientometrics*. https://doi.org/10.1007/s11192-024-04959-8

Aleta, A., Meloni, S., Perra, N., & Moreno, Y. (2019). Explore with caution: Mapping the evolution of scientific interest in physics. *EPJ Data Science*, *8*(1), 27. https://doi.org/10.1140/epjds/s13688-019-0205-9

Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and creativity: Evidence from the academic life sciences. *The RAND Journal of Economics, 42*(3), 527–554. https://doi.org/10.1111/j.1756-2171.2011.00140.x

Bazeley, P. (2010). Conceptualising research performance. *Studies in Higher Education, 35*(8), 889–903. https://doi.org/10.1080/03075070903348404

Berman, E. P. (2012). Explaining the move toward the market in US academic science: How institutional logics can change without institutional entrepreneurs. *Theory and Society*, *41*(3), 261–299. https://doi.org/10.1007/s11186-012-9167-7

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*(3), 367–383. https://doi.org/10.1037/0033-2909.125.3.367

Cedrini, M., & Fontana, M. (2018). Just another niche in the wall? How specialization is changing the face of mainstream economics. *Cambridge Journal of Economics*, *42*(2), 427–451. https://doi.org/10.1093/cje/bex003

Clark, B. R. (1983). *The higher education system: Academic organization in cross-national perspective*. Univ. of California Pr.

Colquitt, J. A., & George, G. (2011). Publishing in *AMJ* —Part 1: Topic Choice. *Academy of Management Journal*, *54*(3), 432–435. https://doi.org/10.5465/amj.2011.61965960

Debernardi, C., Seeber, M., & Cattaneo, M. (2024). Thirty years of climate change research: A fine-grained analysis of geographical specialization. *Environmental Science & Policy*, *152*, 103663. https://doi.org/10.1016/j.envsci.2023.103663

Ebadi, A., & Schiffauerova, A. (2015). How to Receive More Funding for Your Research? Get Connected to the Right People! *PLOS ONE*, *10*(7), e0133061. https://doi.org/10.1371/journal.pone.0133061

Evans, J. A. (2010). Industry collaboration, scientific sharing, and the dissemination of knowledge. *Social Studies of Science*, *40*(5), 757–791. https://doi.org/10.1177/0306312710379931

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and Innovation in Scientists' Research Strategies. *American Sociological Review*, *80*(5), 875–908. https://doi.org/10.1177/0003122415601618

Gläser, J., & Laudel, G. (2015). A Bibliometric Reconstruction of Research Trails for Qualitative Investigations of Scientific Innovations. *Historical Social Research / Historische Sozialforschung Vol. 40, No. 3*. https://doi.org/10.12759/HSR.40.2015.3.299-330

Heckman, J. J., & Moktan, S. (2020). Publishing and Promotion in Economics: The Tyranny of the Top Five. *Journal of Economic Literature*, *58*(2), 419–470. https://doi.org/10.1257/jel.20191574

Hillson, D., & Murray-Webster, R. (2007). *Understanding and managing risk attitude* (2nd ed). Gower.

Hoonlor, A., Szymanski, B. K., & Zaki, M. J. (2013). Trends in computer science research. *Communications of the ACM*, *56*(10), 74–83. https://doi.org/10.1145/2500892

Horta, H., & Santos, J. M. (2020). The Multidimensional Research Agendas Inventory—Revised (MDRAI-R): Factors shaping researchers' research agendas in all fields of knowledge. *Quantitative Science Studies*, *1*(1), 60–93. https://doi.org/10.1162/qss_a_00017

Huang, S., Lu, W., Bu, Y., & Huang, Y. (2022). Revisiting the exploration-exploitation behavior of scholars' research topic selection: Evidence from a large-scale bibliographic database. *Information Processing & Management*, *59*(6), 103110. https://doi.org/10.1016/j.ipm.2022.103110

Iorio, R., Labory, S., & Rentocchini, F. (2017). The importance of pro-social behaviour for the breadth and depth of knowledge transfer activities: An analysis of Italian academic scientists. *Research Policy*, *46*(2), 497–509. https://doi.org/10.1016/j.respol.2016.12.003

Jia, T., Wang, D., & Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. *Nature Human Behaviour*, *1*(4), 0078. https://doi.org/10.1038/s41562-017-0078

Jones, B. F., & Weinberg, B. A. (2011). Age dynamics in scientific creativity. *Proceedings of the National Academy of Sciences*, *108*(47), 18910–18914. https://doi.org/10.1073/pnas.1102895108

Knorr-Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Harvard University Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Univ. of Chicago Press.

Kuhn, T. S. (2000). *The essential tension: Selected studies in scientific tradition and change* (11. print.). Univ. of Chicago Press. (Original work published 1979)

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton University Press.

Laudel, G. (2017). How do National Career Systems Promote or Hinder the Emergence of New Research Lines? *Minerva*, *55*(3), 341–369. https://doi.org/10.1007/s11024-017-9314-4

Laudel, G., & Bielick, J. (2018). The Emergence of Individual Research Programs in the Early Career Phase of Academics. *Science, Technology, & Human Values*, *43*(6), 972–1010. https://doi.org/10.1177/0162243918763100

Laudel, G., & Gläser, J. (2008). From apprentice to colleague: The metamorphosis of Early Career Researchers. *Higher Education*, *55*(3), 387–406. https://doi.org/10.1007/s10734-007-9063-7

Lax-Martinez, G., Saenz De Juano-i-Ribes, H., Yin, D., Le Feuvre, B., Hamdan-Livramento, I., Saito, K., & Raffo, J. D. (2023). Expanding the World Gender-Name Dictionary: WGND 2.0. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4422333

Leisyte, L., & Dee, J. R. (2012). Understanding Academic Work in a Changing Institutional Environment: Faculty Autonomy, Productivity, and Identity in Europe and the United States. In J. C. Smart & M. B. Paulsen (Eds.), *Higher Education: Handbook of Theory and Research* (Vol. 27, pp. 123–206). Springer Netherlands. https://doi.org/10.1007/978-94-007-2950-6_3

Macháček, V., Srholec, M., Ferreira, M. R., Robinson-Garcia, N., & Costas, R. (2022). Researchers' institutional mobility: Bibliometric evidence on academic inbreeding and internationalization. *Science and Public Policy*, *49*(1), 85–97. https://doi.org/10.1093/scipol/scab064

Madsen, E. B., & Nielsen, M. W. (2024). Do thematic funding instruments lead researchers in new directions? Strategic funding priorities and topic switching

among British grant recipients. *Research Evaluation*, rvae015. https://doi.org/10.1093/reseval/rvae015

Malmgren, R. D., Ottino, J. M., & Nunes Amaral, L. A. (2010). The role of mentorship in protégé performance. *Nature*, *465*(7298), 622–626. https://doi.org/10.1038/nature09040

Merton, R. K. (1974). *The sociology of science: Theoretical and empirical incvestigations* (4. Dr.). Univ. of Chicago Pr.

Ramassa, P., Avallone, F., & Quagli, A. (2024). Can "publishing game" pressures affect the research topic choice? A survey of European accounting researchers. *Journal of Management and Governance*, *28*(2), 507–542. https://doi.org/10.1007/s10997-023-09667-8

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. https://doi.org/10.48550/ARXIV.1908.10084

Rodan, S., & Galunic, C. (2004). More than network structure: How knowledge heterogeneity influences managerial performance and innovativeness. *Strategic Management Journal*, *25*(6), 541–562. https://doi.org/10.1002/smj.398

Santos, J. M., Horta, H., & Amâncio, L. (2021). Research agendas of female and male academics: A new perspective on gender disparities in academia. *Gender and Education*, *33*(5), 625–643. https://doi.org/10.1080/09540253.2020.1792844

Seeber, M. (2013). Efficacy and limitations of research steering in different disciplines. *Studies in Higher Education*, *38*(1), 20–38. https://doi.org/10.1080/03075079.2011.561308

Seeber, M., & Mampaey, J. (2022). How do university systems' features affect academic inbreeding? Career rules and language requirements in France, Germany, Italy and Spain. *Higher Education Quarterly*, *76*(1), 20–35. https://doi.org/10.1111/hequ.12302

Sormani, E., & Uude, K. (2022). Academics' prosocial motivation for engagement with society: The case of German academics in health science. *Science and Public Policy*, *49*(6), 962–971. https://doi.org/10.1093/scipol/scac042

Stan Development Team. (2024) *Stan Modeling Language Users Guide and Reference Manual*. https://mc-stan.org

Stephan, P. E. (1996). The Economics of Science. *Journal of Economic Literature*, *34*(3), 1199–1235. JSTOR.

Vallas, S. P., & Kleinman, D. L. (2007). Contradiction, convergence and the knowledge economy: The confluence of academic and commercial biotechnology. *Socio-Economic Review*, *6*(2), 283–311. https://doi.org/10.1093/ser/mwl035

Wei, T., Li, M., Wu, C., Yan, X.-Y., Fan, Y., Di, Z., & Wu, J. (2013). Do scientists trace hot topics? *Scientific Reports*, *3*(1), 2207. https://doi.org/10.1038/srep02207

Whitley, R. (2000). *The intellectual and social organization of the sciences*. Oxford Univ. Press.

Whitley, R., & Gläser, J. (Eds.). (2014). *Organizational transformation and scientific change: The impact of institutional restructuring on universities and intellectual innovation* (1. ed). Emerald.

Whitley, R., Gläser, J., & Laudel, G. (2018). The Impact of Changing Funding and Authority Relationships on Scientific Innovations. *Minerva*, *56*(1), 109–134. https://doi.org/10.1007/s11024-018-9343-7

Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., Wang, Y., Stanley, H. E., & Havlin, S. (2019). Increasing trend of scientists to switch between topics. *Nature Communications*, *10*(1), 3439. https://doi.org/10.1038/s41467-019-11401-8

Zhang, L., & Horta, H. (2023). Research agendas and job dissatisfaction among Hong Kong academics. *Higher Education*, *85*(1), 103–122. https://doi.org/10.1007/s10734-022-00824-w

Appendix

Measuring mobility events

Institutional affiliations reported on published papers can be highly volatile and have a large degree of overlap. It is indeed quite common for authors to report multiple affiliations at the same time due to varying degrees of collaboration or involvement with several institution. These multiple affiliations can be distributed across several physical locations, often in different countries. However, our aim was to capture durable changes in affiliation that are associated with the physical relocation of the scholars. Thus, relying on the full spectrum of reported affiliations would largely overestimate the amount of mobility events and likely dilute their estimated effect on the level of research agenda diversification. To address such issue, we employed a standardization procedure that conservatively assigns a single geographical location to each scholar in each year. Additionally, since institutional affiliations present several challenges given by hierarchy, inconsistent reporting and disambiguation, we rely on the reported affiliation cities.

More in detail, for each individual scholar, the procedure runs as follows:

- Gather all the reported affiliation cities in each year (this information is included in Scopus metadata of each publication)
- Starting from the first year, if more than one city is reported, remove all the cities that do not appear neither in the previous nor in the subsequent year;
- If at this point more than one city is still present, remove all cities that do not appear neither in the previous two years nor in the subsequent two.

At the end of this procedure, there are still cases in which multiple cities appear in a single year. Since these cases represent affiliations which are chronologically isolated from both the preceding and following ones, we simply treat these as unique tokens. This is justified since our aim is not to determine the exact physical location of the scholars at each time, but rather to ascertain if their location is changed.

This first measure of mobility events runs into the risk of artificially postponing the time of relocation. To address such concern, we defined a second measure

which simply proceeds in reverse: starting from the last year, it reinstates removed cities, if the in original list they appeared earlier for the first time. A fictional example of the whole process is available in Table 5.

We used this second measure of mobility events as a robustness check in Model E, which yielded results in line with our main analysis.

Year	Reported cities	Standardised	Standardised (backwards)	Mobility	Mobility (backwards)
2008	Amsterdam	Amsterdam	Amsterdam	0	0
2009	Amsterdam-Oslo	Amsterdam	Amsterdam	0	0
2010	Amsterdam	Amsterdam	Amsterdam	0	0
2011	Utrecht	Amsterdam	Utrecht	0	1
2012	Utrecht	Utrecht	Utrecht	1	0
2013	Utrecht-Florence	Utrecht	Utrecht	0	0
2014	Utrecht	Utrecht	Utrecht	0	0
2015	Rome	Rome	Rome	1	1
2016	Rome	Rome	Rome	0	0
2017	Rome	Rome	Rome	0	0

Table 5: Fictional example of reported cities pattern and standardisation.

Robustness checks

In order to test the adequateness of several aspects of our model, we estimated three other variations of it. Overall, the results are consistent with our main analysis.

Model D includes mobility with four dummies periods: first year (i.e., the one included in the main model); 2-5 year; after 5 years. Most of the impact is visible in the first year after the mobility event.

Model E is the same as Model A, with the backward variant of the mobility measure defined according to the description above. The results are in line with Model A.

Model F includes an interaction between the mobility measure and a dummy variable for seniority <10. The results are in line with the ones of Model B: the most impactful mobility events are the early ones.

	SPS			ECO			IMM			SNP		
Parameter	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**
Intercept	0,422	0,411	0,433	0,473	0,462	0,483	0,358	0,351	0,364	0,446	0,432	0,459
Observation sd	0,132	0,131	0,133	0,128	0,128	0,129	0,133	0,132	0,134	0,136	0,135	0,137
Italy	-0,012	-0,021	-0,003	-0,002	-0,009	0,005	0,016	0,011	0,022	-0,01	-0,015	-0,005
Norway	-0,02	-0,032	-0,009	-0,022	-0,032	-0,011	-0,005	-0,014	0,005	-0,005	-0,027	0,015
Germany	-0,004	-0,011	0,004	0,001	-0,005	0,007	0,006	0,002	0,01	-0,016	-0,021	-0,012
Netherlands	-0,003	-0,011	0,005	0,003	-0,004	0,011	-0,011	-0,016	-0,005	-0,005	-0,016	0,005
Seniority (ln)	0,036	0,031	0,042	0,025	0,022	0,029	0,06	0,057	0,063	0,038	0,034	0,043
N publications (ln)	-0,066	-0,07	-0,062	-0,059	-0,061	-0,056	-0,069	-0,071	-0,067	-0,063	-0,066	-0,06
N authors (ln)	0,034	0,03	0,037	0,017	0,014	0,02	0,02	0,018	0,021	0,018	0,016	0,02
Mobility	0,004	-0,001	0,01	0,007	0,003	0,011	0,045	0,042	0,049	0,014	0,01	0,018
Author sd	0,054	0,051	0,057	0,05	0,048	0,052	0,054	0,052	0,056	0,049	0,046	0,051
Gender: Male	0,02	0,012	0,028	0,01	0,002	0,018	0,012	0,006	0,017	0,037	0,024	0,05
Gender: unmatched	0,012	0	0,025	0,018	0,008	0,028	0,018	0,009	0,026	0,039	0,023	0,055

Table 6: Regression coefficients for Model A

	SPS			ECO			IMM			SNP		
Parameter	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**
Intercept	0,423	0,412	0,434	0,473	0,462	0,483	0,354	0,347	0,361	0,443	0,429	0,459
Observation sd	0,132	0,131	0,133	0,128	0,128	0,129	0,133	0,132	0,134	0,136	0,135	0,136
Italy	-0,012	-0,022	-0,002	-0,002	-0,009	0,005	0,018	0,013	0,023	-0,01	-0,015	-0,005
Norway	-0,02	-0,032	-0,009	-0,022	-0,032	-0,012	-0,003	-0,012	0,007	-0,004	-0,024	0,017
Germany	-0,003	-0,011	0,005	0,001	-0,005	0,008	0,006	0,002	0,01	-0,016	-0,02	-0,011
Netherlands	-0,003	-0,01	0,005	0,003	-0,004	0,011	-0,009	-0,015	-0,004	-0,006	-0,015	0,005
Seniority (ln)	0,036	0,031	0,042	0,025	0,022	0,029	0,061	0,057	0,064	0,039	0,034	0,044
N publications (ln)	-0,066	-0,07	-0,062	-0,059	-0,061	-0,056	-0,069	-0,071	-0,066	-0,063	-0,066	-0,06
N authors (ln)	0,034	0,03	0,037	0,017	0,014	0,02	0,02	0,018	0,021	0,018	0,016	0,02
Author sd	0,054	0,051	0,057	0,05	0,048	0,052	0,054	0,052	0,056	0,049	0,046	0,051
Gender: Male	0,02	0,012	0,028	0,01	0,002	0,017	0,012	0,007	0,017	0,037	0,023	0,051
Gender: unmatched	0,012	0	0,025	0,019	0,008	0,03	0,017	0,008	0,026	0,038	0,022	0,054
Mobility n. 1	0,003	-0,005	0,011	0,01	0,004	0,016	0,078	0,072	0,084	0,027	0,02	0,034
Mobility n. 2	0,004	-0,006	0,014	-0,002	-0,009	0,004	0,032	0,025	0,038	0,015	0,008	0,022
Mobility n. 3	0,011	-0,004	0,024	0,014	0,003	0,025	0,021	0,011	0,03	0,013	0,004	0,022
Mobility n. 4+	0,005	-0,017	0,026	0,013	-0,001	0,026	0,013	0,005	0,022	-0,003	-0,01	0,005

Table 7: Regression coefficients for Model B

	SPS			ECO			IMM			SNP		
Parameter	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**
Intercept	0,411	0,399	0,421	0,464	0,454	0,475	0,319	0,313	0,326	0,438	0,425	0,451
Observation sd	0,128	0,127	0,13	0,124	0,124	0,125	0,131	0,13	0,131	0,131	0,131	0,132
Italy	-0,012	-0,021	-0,002	0	-0,008	0,007	0,02	0,015	0,024	-0,007	-0,011	-0,002
Norway	-0,023	-0,035	-0,011	-0,022	-0,031	-0,012	-0,001	-0,011	0,008	-0,001	-0,021	0,02
Germany	-0,004	-0,012	0,003	-0,003	-0,009	0,003	0,006	0,002	0,01	-0,016	-0,02	-0,011
Netherlands	-0,005	-0,013	0,003	0,002	-0,006	0,009	-0,009	-0,014	-0,004	-0,006	-0,016	0,004
New coauthor	0,074	0,07	0,078	0,078	0,076	0,081	0,082	0,079	0,084	0,079	0,076	0,081
Seniority (ln)	0,031	0,026	0,036	0,02	0,016	0,023	0,056	0,052	0,059	0,032	0,027	0,036
N publications (ln)	-0,06	-0,063	-0,056	-0,051	-0,054	-0,048	-0,064	-0,066	-0,062	-0,056	-0,059	-0,053
N authors (ln)	-0,004	-0,007	0	-0,03	-0,034	-0,027	0,001	-0,001	0,003	-0,017	-0,02	-0,015
Mobility	0,002	-0,003	0,008	0,007	0,003	0,011	0,042	0,039	0,046	0,01	0,006	0,014
Author sd	0,053	0,05	0,056	0,048	0,046	0,05	0,053	0,051	0,055	0,046	0,044	0,049
Gender: Male	0,02	0,012	0,027	0,01	0,002	0,017	0,009	0,004	0,015	0,033	0,02	0,045
Gender: unmatched	0,013	0,001	0,026	0,018	0,009	0,028	0,016	0,007	0,026	0,036	0,023	0,05

Table 8: Regression coefficients for Model C

	SPS			ECO			IMM			SNP		
Parameter	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**
Intercept	0,412	0,401	0,423	0,466	0,456	0,477	0,322	0,316	0,33	0,438	0,424	0,451
Observation sd	0,128	0,127	0,129	0,124	0,124	0,125	0,131	0,13	0,131	0,131	0,131	0,132
New coauthor	0,074	0,07	0,078	0,078	0,076	0,081	0,082	0,079	0,084	0,079	0,076	0,081
Italy	-0,012	-0,021	-0,002	-0,001	-0,008	0,006	0,019	0,014	0,024	-0,007	-0,011	-0,002
Norway	-0,022	-0,034	-0,011	-0,023	-0,033	-0,012	-0,002	-0,011	0,008	0	-0,021	0,021
Germany	-0,005	-0,013	0,003	-0,003	-0,009	0,003	0,006	0,001	0,01	-0,016	-0,02	-0,011
Netherlands	-0,005	-0,013	0,002	0,001	-0,007	0,008	-0,01	-0,015	-0,004	-0,007	-0,016	0,003
Seniority (ln)	0,03	0,025	0,036	0,019	0,016	0,023	0,055	0,052	0,059	0,032	0,027	0,037
N publications (ln)	-0,061	-0,065	-0,057	-0,051	-0,054	-0,049	-0,065	-0,067	-0,063	-0,056	-0,059	-0,053
N authors (ln)	-0,004	-0,008	0	-0,03	-0,034	-0,027	0,001	-0,001	0,003	-0,017	-0,02	-0,014
Author sd	0,053	0,05	0,057	0,048	0,046	0,05	0,053	0,051	0,055	0,046	0,044	0,049
Gender: Male	0,02	0,012	0,028	0,01	0,003	0,017	0,009	0,004	0,014	0,034	0,021	0,047
Gender: unmatched	0,014	0,002	0,027	0,018	0,008	0,029	0,016	0,007	0,026	0,037	0,021	0,053
Mobility year	0,005	-0,002	0,012	0,006	0,002	0,011	0,045	0,041	0,049	0,009	0,003	0,014
2-5 after Mobility	0,002	-0,003	0,008	-0,003	-0,007	0,001	0,001	-0,003	0,004	-0,002	-0,007	0,003
5+ after Mobility	0,013	0,005	0,021	0,004	-0,001	0,009	0,009	0,005	0,013	-0,001	-0,007	0,004

Table 9: Regression coefficients for Model D

	SPS			ECO			IMM			SNP		
Parameter	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**
Intercept	0,421	0,411	0,433	0,472	0,462	0,483	0,358	0,352	0,365	0,443	0,429	0,456
Observation sd	0,132	0,131	0,133	0,128	0,128	0,129	0,133	0,133	0,134	0,136	0,135	0,136
Italy	-0,012	-0,022	-0,002	-0,002	-0,009	0,005	0,016	0,011	0,022	-0,01	-0,015	-0,004
Norway	-0,02	-0,032	-0,009	-0,022	-0,032	-0,011	-0,005	-0,014	0,005	-0,004	-0,025	0,017
Germany	-0,003	-0,011	0,004	0,001	-0,005	0,007	0,006	0,001	0,01	-0,016	-0,02	-0,011
Netherlands	-0,003	-0,01	0,005	0,003	-0,004	0,01	-0,01	-0,016	-0,005	-0,006	-0,017	0,004
Seniority (ln)	0,036	0,031	0,042	0,025	0,021	0,029	0,06	0,056	0,063	0,039	0,034	0,043
N publications (ln)	-0,066	-0,07	-0,062	-0,059	-0,061	-0,056	-0,069	-0,071	-0,067	-0,063	-0,066	-0,059
N authors (ln)	0,034	0,03	0,037	0,017	0,014	0,02	0,02	0,018	0,021	0,018	0,016	0,02
Mobility	0,007	0,001	0,012	0,006	0,002	0,01	0,035	0,032	0,039	0,018	0,014	0,023
Author sd	0,054	0,051	0,057	0,05	0,047	0,052	0,054	0,052	0,056	0,049	0,046	0,052
Gender: Male	0,02	0,012	0,028	0,01	0,003	0,018	0,011	0,005	0,017	0,037	0,025	0,05
Gender: unmatched	0,013	0	0,025	0,019	0,009	0,029	0,017	0,009	0,026	0,039	0,023	0,056
Intercept	0,421	0,411	0,433	0,472	0,462	0,483	0,358	0,352	0,365	0,443	0,429	0,456
Observation sd	0,132	0,131	0,133	0,128	0,128	0,129	0,133	0,133	0,134	0,136	0,135	0,136
Italy	-0,012	-0,022	-0,002	-0,002	-0,009	0,005	0,016	0,011	0,022	-0,01	-0,015	-0,004

Table 10: Regression coefficients for Model E

	SPS			ECO			IMM			SNP		
Parameter	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**	**mean**	**5%**	**95%**
Intercept	0,423	0,412	0,435	0,473	0,462	0,484	0,354	0,348	0,361	0,442	0,429	0,456
Observation sd	0,132	0,131	0,133	0,128	0,128	0,129	0,133	0,132	0,134	0,136	0,135	0,136
Italy	-0,012	-0,023	-0,002	-0,002	-0,009	0,005	0,017	0,012	0,022	-0,01	-0,015	-0,005
Norway	-0,02	-0,032	-0,009	-0,022	-0,031	-0,011	-0,005	-0,016	0,005	-0,004	-0,025	0,016
Germany	-0,004	-0,011	0,004	0,001	-0,005	0,008	0,007	0,002	0,011	-0,016	-0,02	-0,011
Netherlands	-0,003	-0,011	0,005	0,003	-0,004	0,011	-0,01	-0,015	-0,004	-0,006	-0,016	0,004
Seniority (ln)	0,036	0,031	0,042	0,025	0,022	0,029	0,061	0,058	0,064	0,039	0,035	0,044
N publications (ln)	-0,066	-0,07	-0,062	-0,059	-0,062	-0,056	-0,069	-0,071	-0,067	-0,063	-0,066	-0,06
N authors (ln)	0,034	0,03	0,037	0,017	0,014	0,02	0,02	0,018	0,021	0,018	0,016	0,02
Mobility	0,009	-0,001	0,019	0,007	0,001	0,013	0,023	0,018	0,028	0,001	-0,005	0,007
Mobility x Sen<10	-0,006	-0,017	0,005	0	-0,008	0,008	0,043	0,035	0,05	0,023	0,016	0,031
Author sd	0,054	0,051	0,057	0,05	0,048	0,052	0,054	0,052	0,056	0,049	0,046	0,052
Gender: Male	0,02	0,012	0,028	0,01	0,002	0,017	0,011	0,006	0,017	0,038	0,024	0,051
Gender: unmatched	0,012	-0,001	0,025	0,019	0,009	0,029	0,017	0,008	0,027	0,04	0,024	0,056
Intercept	0,423	0,412	0,435	0,473	0,462	0,484	0,354	0,348	0,361	0,442	0,429	0,456
Observation sd	0,132	0,131	0,133	0,128	0,128	0,129	0,133	0,132	0,134	0,136	0,135	0,136

Table 11: Regression coefficients for Model F