Understanding scientific collaboration and mobility : a creativity perspective

Kévin Wirtz

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE AUGUSTIN COURNOT ED 221

BUREAU D'ÉCONOMIE THÉORIQUE ET APPLIQUÉE UMR 7522

THÈSE

pour l'obtention du titre de Docteur en Sciences Économiques

Présentée et soutenue le 24 Novembre 2023 par

Kevin WIRTZ

Understanding Scientific Collaboration and Mobility: A Creativity Perspective

---

Préparée sous la direction de Patrick LLERENA et Stefano BIANCHINI

Membres du jury :

Julien PENIN	Professeur des Universités, Université de Strasbourg	Président
Magda FONTANA	Associate Professor, Université de Turin	Rapportrice
Pablo D'ESTE	Senior Researcher, Spanish Council for Scientific Research (CSIC)	Rapporteur
Fabiana VISENTIN	Assistant Professor, Université de Maastricht	Examinatrice
Patrick LLERENA	Professeur des Universités, Université de Strasbourg	Directeur
Stefano BIANCHINI	Maître de conférences HDR, Université de Strasbourg	CoDirecteur

L'Université de Strasbourg n'entend donner aucune approbation, ni improbation aux opinions émises dans cette thèse ; elles doivent être considérées comme propres à leur auteur.

Remerciements

Je voudrais en tout premier lieu exprimer toute ma reconnaissance à mes deux superviseurs Patrick Llerena et Stefano Bianchini pour m'avoir encadré. Leur grande rigueur scientifique, leurs conseils, mais aussi leur confiance, ont grandement contribué à la réussite de cette thèse. Je voudrais également remercier Julien Pénin, Fabiana Visentin, Pablo d'Este et Magda Fontana pour m'avoir fait l'honneur de composer mon jury. Mes remerciements vont aussi à Nicolas Lachiche et Robin Cowan qui ont accepté de faire partie de mon Comité de Suivi dès ma seconde année de thèse. Je remercie également les coauteurs ayant collaboré aux travaux de cette thèse. Merci à Stefano Bianchini, Pierre Pelletier, Moritz Müller, Roman Jurowetski et Daniel Hain. Je remercie le Bureau d'Économie Théorique et Appliquée (BETA) et l'École doctorale Augustin Cournot pour avoir mis à ma disposition l'ensemble des moyens intellectuels pour la réalisation de ce travail.

Je tiens à remercier particulièrement certains doctorant.e.s et docteur.e.s qui m'ont accompagné durant ce long périple. En premier une personne sans qui je ne me serais jamais lancé dans une thèse et sans qui celle-ci ne serait probablement pas fini: Pierre. Merci pour tout, je suivrais ta carrière d'économiste italien avec intérêt ! Merci également à Eva, je sais que tu t'attends à un paragraphe entier sur ta personne, mais même en utilisant uniquement des mots-clés des délires qu'on a eu le paragraphe serait plus long que ma thèse (j'abuse à peine). Merci à la team Smash Nico et Heman, quand vous voulez pour me prendre un Bo3. Merci à la team de runner Théo, Louis et particulièrement Sarah qui a dû me carry sur le Ekiden alors que j'étais au bout de ma life. Merci à mon bureau beaucoup trop rempli; Agathe et les récits de sa vie qui pourrait probablement être une série Netflix; Guillaume avec sa bouture et ses grains de café; Emilien et son tabassage de bureau; Shengxi pour sa gentillesse; mais aussi Lucas, Diletta et Mathilde. Merci aux autres doctorant.e.s: Anne-Gaëlle pour les soirées Doctor Who; Jérome et son début de carrière de twittos; Romane et sa motivation sans faille pour le clubbing; Kenza et son intérêt pour les séries et twitch; Alexandre et ses passions jeux vidéo/Karaté/stata; Vincent et sa non-possession de carte pro.

Je n'oublie également pas mes ami.e.s hors des murs du BETA. Tanguy qui en l'espace de ma thèse à accompli beaucoup trop de choses et qui sait probablement mieux coder que moi maintenant (Et qui a un Yoshi de qualité). Léo qui est à l'origine de beaucoup d'heures perdu en m'introduisant à la peinture et aux jeux de sociétés. Merci Aurore pour tous les cinés, les moscow mule, l'eau courante et autres

inside joke que toi seule connaît. Merci Nico et toutes les heures passées en co-op pendant le Covid. Merci à Claire, c'est bien parce que c'est toi que j'accepte de partager mon anniversaire (et mention spéciale à Lucy). Merci Matt pour m'avoir supporté depuis tellement d'années (Ou alors c'est moi qui t'ai supporté ?). Merci à Phil et Ana qui eux aussi m'ont soutenu depuis bien des années et sur qui je sais que je pourrais encore compter. Merci à Arthur pour sa transparence, bienveillance et ses refs perchés. Merci Hugo pour ces discussions jusqu'à pas d'heure. Merci à Dresch, tu reviens quand tu veux pour parler de blockchain. Merci les frères grillet pour m'avoir soutenu dans le début chaotique de la thèse.

Je tiens enfin à exprimer ma reconnaissance éternelle à ma famille pour leur soutien indéfectible depuis le début de mes années d'études. Merci mes parents pour leur soutien malgré le flou de la thèse. Merci à mon frère sans qui j'aurais probablement pas fait autant d'années d'études. Merci à ma soeur et ses talents de pâtissière. Merci à mon petit chien qui m'a accompagné pendant bien des années.

Merci à vous tous, sans vous, je n'aurais sans doute jamais pu réaliser cette thèse.

Contents

General Introduction

"*A knowledge economy is one in which knowledge assets are deliberately accorded more importance than capital and labor assets, and where the quantity and sophistication of the knowledge pervading economic and societal activities reaches very high levels*" [World Bank, 2007].

The concept of a Knowledge-Based Economy (KBE) has evolved and been debated among economists and policymakers, with different definitions and interpretations of what constitutes a knowledge economy and how to measure its progress and performance [Powell and Snellman, 2004]. The fundamental principles of a KBE are the generation, dissemination, and use of knowledge [Milewska, 2018]. At the core of a Knowledge-Based Economy lies the process of creativity, which can be broken down into two essential components: *originality* and *impact* [Runco and Jaeger, 2012]. Originality can be defined as the quality of being inventive and novel, characterized by the introduction of innovative elements that distinguish an idea or concept from existing paradigms. It involves creating new knowledge, ideas, and insights that advance our understanding of the world. This process includes discovering new information, synthesizing existing knowledge, and developing innovative ways of approaching problems and phenomena. On the other hand, impact measures the significance and influence of the knowledge created. The impact of an idea is determined by its ability to bring about change, whether in scientific understanding, technological advancement, societal improvement, or other areas.

One theory of knowledge creation in an organizational setting was developed by Nonaka et al. [2006]. The generation of new knowledge occurs through interactions between explicit and tacit knowledge via a process known as the "socialization, externalization, combination, and internalization" (SECI) spiral, which means that to understand how knowledge is created and diffused we need to look at the actors, the interaction between them, and also the environment in which they co-evolve.

Science plays a central role in the KBE and drives technological advancements as well as contributes significantly to both economic development and growth [Nelson and Romer, 1996, Caliari and Chiarini, 2021]. Furthermore, science helps us understand the natural world, improve our quality of life, and address societal issues, but it can also help policymakers understand potential risks and hazards, such as natural disasters, pandemics, and emerging technologies [Mazzucato, 2018]. This enables them to create strategies and regulations to minimize risks or promote initiatives that benefit society.

To understand creativity in the science system, it becomes essential to examine the interplay among the various individuals and entities engaged in research activities across multiple levels of analysis. In recent decades, there have been significant changes in the way research is conducted and how the scientific network is structured. These changes have been characterized first and foremost by increased mobility and collaborations among researchers and a decrease in solo authorship [Geuna, 2015]. Collaboration and mobility are often intertwined. Researcher mobility refers to the movement of researchers between different research environments, either within the same country (domestic mobility) or across international borders (international mobility). This mobility can occur at various stages of a researcher's career and is influenced by multiple factors. Researchers can move across locations, sectors, and career stages [Fernandez-Zubieta et al., 2015]. Mobility plays a positive role in working with new co-authors while keeping existing ties intact [Liu and Hu, 2022]. While considerable attention has been dedicated to the globalization of science, studies have concentrated their efforts on the link between collaboration, mobility, and impact and very little on originality. Yet originality plays a pivotal role in addressing emerging challenges and thinking of innovative solutions [Witt, 2016, Fortunato et al., 2018].

Understanding the structure and dynamics of the scientific system, how resources are orchestrated, and its consequences on creativity is crucial for both fostering efficient growth of the pool of knowledge and the transfer of it, allowing us to better support scholars when facing challenges of our time, including climate change, poverty, and, as seen recently, global pandemics.

This thesis contributes to this endeavor and is organized into four chapters. **Chapter 1** focuses on the structure and resilience of the scientific collaboration network, the usage of existing resources, and the adaptability of countries follow-

ing an exogenous shock, namely the Covid-19 pandemic. **Chapter 2** presents a methodological paper that introduces *Novelpy*, an open-source tool developed in Python. The primary objective of this package is to facilitate the computation of novelty and disruptiveness indicators, thereby enhancing transparency in evaluating research through the study, comparison, creation, and application of these indicators. **Chapter 3** breaks down the creativity aspect of scientific publications and analyses how collaboration of researchers affects novelty and impact, the pillars of creativity. Finally, **Chapter 4** investigates the intersectoral mobility of researchers in a specific domain, namely Artificial Intelligence (AI), and the potential consequences on their research outcomes. More specifically, we addressed the following overarching questions:

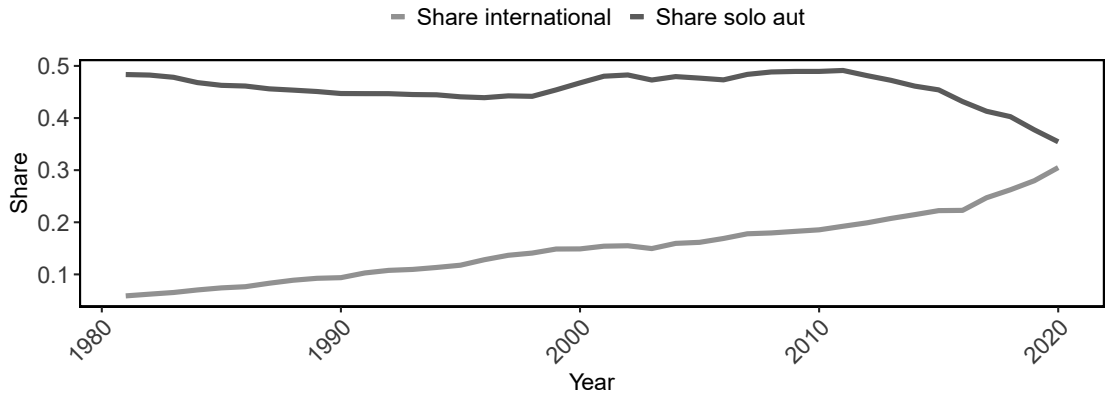
- How does a country leverage its existing resources, knowledge, and past partnerships to address urgent challenges stemming from an exogenous shock?
- How do team composition and the researchers' abilities to explore the knowledge space influence the originality and impact of their research ?
- What transformations occur in individuals' research following a transition between public and private sectors?

Collaboration dynamics

The rise of international collaborations has been documented in multiple studies, indicating an increasing trend in the number of authors per paper and cross-country collaboration [Adams, 2012]. This result is corroborated in Wagner et al. [2017] observing a growth of 120% of scientific international collaborations between 1990 and 2013. This is partly explained by the increase of countries participating in research (i.e., more actors, more possibility of collaboration). This increase is accompanied by a decrease in the proportion of papers solo-authored, which appears to have a lesser impact on younger researchers [Kuld and O'Hagan, 2018]. One possible interpretation for this collaboration surge is the knowledge space's evolution. While the accessibility to scientific knowledge is growing, team size is growing, and agents increasingly specialize their competencies to navigate this ever-increasing knowledge landscape as it is too vast and complex for any individual agent to master it [Boudreau et al., 2016]. Using OpenAlex, a fully open-source library designed to

represent the global research system, containing over 240 million research documents [Priem et al., 2022], Figure 2 shows that these two trends continue to be observed recently.

Figure 1: Evolution of research collaboration in OpenAlex



Notes: Evolution of the share of papers solo-authored (blue) and the share of papers with at least two authors with distinct countries (orange) in OpenAlex. Source: own elaboration

In this context, it is crucial to understand how the scientific collaboration structure emerges and how it evolves. Multiple dimensions increase the likelihood of a collaboration between two authors. Boschma [2005] for instance, distinguishes five proximity dimensions that influence the possibility of co-authorship: cognitive, geographical, institutional, organizational, and social proximity. Cognitive proximity refers to the similarity or compatibility of researchers' knowledge, expertise, and research interests. When researchers have similar cognitive backgrounds or work in related fields, it becomes easier to communicate and collaborate effectively. Geographical proximity relates to the physical distance between researchers. When researchers are located in close proximity, such as working in the same institution or region, it facilitates face-to-face interactions, frequent meetings, and easier coordination for collaborative projects. Institutional proximity could be better understood as the national context in which scientist evolves (i.e., the "societal macro level"). Each country has its agenda, law, and budget, which then influences grants and the creation of collaboration. Organizational proximity focuses on the affiliation context of researchers. Collaboration becomes more likely when researchers belong to the same or closely aligned organization, as they share resources, facilities, and research networks. Finally, social proximity relates to the social relationships and networks among researchers. Collaboration is facilitated when researchers have established so-

cial connections, such as shared professional networks, common collaborators, friendships, or mentor-mentee relationships, as trust and familiarity are already present.

Network theory is frequently employed to comprehend the dynamics and structure of scientific collaborations. Researchers have used this approach to identify several noteworthy characteristics that define the system. One such characteristic is the emergence of a community structure within the network, often attributed to the triadic closure phenomenon.[1] This phenomenon, characterized by the tendency for individuals to form connections with others who share mutual connections, suggests that if person A collaborates with person B and person B collaborates with person C, there is a higher likelihood that person A will also collaborate with person C. Additionally, this community structure can manifest as a core-periphery structure, as demonstrated in previous research [Wedell et al., 2022]. The preferential attachment mechanism further reinforces the consolidation of this core-periphery structure. This mechanism posits that newcomers to the network are more inclined to connect with the most reputable individuals initially. As a result, the scientific system ends up as a robust and resilient structure [Wagner et al., 2017].

Scientific collaboration between countries, including developing ones, is influenced by various factors, and analyzing it from a macro-level perspective encompasses all the aspects mentioned above. According to Dosso et al. [2023], region-specific factors significantly shape scientific collaboration. The study found that shared ethnic language, membership in the African and Malagasy Council for Higher Education (CAMES), and the presence of a common European partner as a third partner in co-publication were essential factors in scientific collaboration. In addition, other determinants of scientific collaboration between countries include geographical distance, colonial heritage, and common language. However, these factors are shown to be decreasing in importance. This suggests that as scientific research becomes more globalized, other factors such as shared research interests and complementary expertise become more important. Another important factor in scientific collaboration between countries is the presence of international networks and organizations facilitating collaboration. For example, scientists from eight full member countries are working together within the SESAME community, as described by UNESCO.[2]

[1] Recent studies have challenged this explanation [Kim and Diesner, 2017]

[2] `https://www.unesco.org/en/scientific%2Dresearch%2Dcooperation%2Dwhy%2Dcollaborate%2Dscience%2Dbenefits%2Dand%2Dexamples`

This organization provides a platform for scientists to collaborate and share expertise, despite political tensions between some of the member countries. In addition to regional organizations, global organizations such as the International Science Council and the World Academy of Sciences promote scientific collaboration. These organizations provide funding and support for international research projects and facilitate collaboration between scientists from different countries. National policies and funding priorities can also influence international collaboration [Ben-David, 1971, Pavitt, 1991, Stephan, 2010]

The literature is still scarce on the dynamics of collaborations after exogenous shocks. The COVID-19 pandemic has raised questions about the adaptability of the scientific system in response to urgent global shocks. The goal of **Chapter 1** is to gain insights on this phenomenon and answer two problematics. How did the structure of the coronavirus research evolve after Covid-19 and how does the structure of global health sciences relate to the scientific response to the pandemic? We find a close coupling of national and international positioning in Coronavirus Related Research (CRR). CRR capacity accumulated before the pandemic has been influential for national CRR output at the beginning of the pandemic. Broader health science capacity became the dominant factor after. In the context of international collaboration, the primary driver of CRR collaboration is the presence of established collaborations unrelated to CRR before the pandemic. Global CRR during the pandemic rapidly converges towards the global order of broader health science capacity, but the mirroring is imperfect. Following the shock, existing grants intended for global research were redirected towards CRR. The emergence of dedicated CRR projects has led to a diminishing emphasis over time on the significance of existing funding in response capabilities. In the initial stages, nations reliant on international funding for their research experienced minimal disruptions in their CRR. However, the subsequent years following the pandemic saw a detrimental effect of dependence on their output.

These findings illustrate that when faced with a global shock, the scientific community tends to reorganize itself by imitating, though not flawlessly, the pre-existing structure and by leveraging existing resources and capabilities. However, it's essential to note that in this chapter, we did not assess the quality of the research papers. Is the replicated network effective? How can we evaluate the quality of the scientific system? If the CRR converges toward an established structure, it becomes imper-

ative to assess the efficiency of that structure. How can we establish an "efficient" network configuration under such circumstances? **Chapters 2 and 3** aim to shed light on these critical questions.

Collaboration and creativity

The origins of what we now refer to as the "Science of Science" and the methods for measuring scientific progress can be traced back to the early 1900s, Paul Otlet initially introduced the concept of bibliometrics in 1934 [Otlet, 1990]. Later on, in 1969, Pritchard reintroduced bibliometrics in a paper titled "Statistical Bibliography or Bibliometrics?" [Pritchard et al., 1969]. The main objective of bibliometrics was to address the challenge of information overload and assist librarians in efficiently selecting relevant materials for their collections [Sugimoto and Larivière, 2018]. In 1955, chemist and documentalist Garfield proposed the idea of a citation index, which led to the establishment of the Institute for Scientific Information in 1960. This institute went on to develop the Science Citation Index (SCI), launched in 1963, which simplified the process of locating and referencing specific publications. Additionally, the SCI introduced citation metrics, such as counting the number of times an article is cited by others, enabling the evaluation of the impact and significance of scientific research. This initiative laid the foundation for the Web of Science. This comprehensive platform encompasses scientific metadata and is extensively utilized in the field of Science of Science research.

Since then, there has been an exponential growth in the volume of scientific documents and the emergence of various channels (e.g., preprints, reports, working papers) and databases containing metadata associated with papers (e.g., Knowledge Graphs, Scite, Crossref). Citation metrics became more prominent, offering a diversity of evaluation tools [Waltman, 2016]. Moreover, new metrics have been devised to capture different aspects of research. For instance, novelty indicators have been developed to assess the originality of a document. In contrast, disruptiveness indicators aim to measure the extent to which a research contribution transforms existing paradigms and has the potential to significantly impact current practices. Alternative metrics, known as Altmetrics, have also been created exploiting web-based data such as Twitter likes and shares, with the purpose of measuring the visibility of a document. These advancements highlight the importance of not only managing the vast pool of available knowledge but also comprehending the tools necessary for its

effective management.

Novelty and disruptiveness indicators are relatively new, and there is a lack of studies that compare these indicators. At the same time, the absence of available codes makes it challenging for those unfamiliar with the subject to utilize them and replicate research findings. The objective of the **Chapter 2** is to standardize the mathematical notation of some existing indicators using graph theory and introduce a Python package called *Novelpy*. This package aims to simplify the application of these emerging indicators and enhance their accessibility for researchers and practitioners.

Novelty, often referred to as originality and atypicality, is crucial in science. First, originality in science helps to promote a diverse range of ideas and perspectives. This diversity is essential for scientific progress, as it allows researchers to explore new avenues and challenge established paradigms. If originality is lacking, the field may stagnate, and the body of knowledge will not grow [Flexner, 2017, Arnott et al., 2020]. Second, a highly novel paper is more likely to be a breakthrough paper than a more conventional [Wang et al., 2017, Shibayama et al., 2021]. Research with high levels of novelty tends to receive a greater number of citations on average, although it is accompanied by heightened uncertainty as well [Wang et al., 2017]. In parallel, the incentive for originality is low. For instance, a recent study shows that if a combination of topics in a grant proposal is too far from the grant Principal Investigator's (PI) knowledge domain, then it is less likely to win the grant [Franzoni et al., 2022]. This implies that to participate in the funding process, one must adhere to the existing norms or conventions within their field. Indeed, scholars who experience rejection for a grant subsequently propose less innovative research in the future [Franzoni et al., 2022].

The factors contributing to the success of some highly novel papers in terms of impact (i.e., creative papers) remain relatively unexplored. Understanding how one can create an environment that fosters creativity and successful science is of crucial interest to tackle societal issues [Fortunato et al., 2018].

The **third Chapter** of this thesis is dedicated to exploring the creativity of the Health Science System. We explore the relationship between cognitive diversity in scientific teams and their capacity to both foster innovative ideas and attain scientific recognition. We propose an author-level metric based on the semantic representation of researchers' past publications to measure cognitive diversity at individual and team levels. We can think of our indicator as a measure of *potential novelty*: we connect the likelihood of novel combinations of knowledge with the diversity of

academic backgrounds within the team and individuals' ability to serve as effective intermediaries, bridging the gap between their fellow team members. However, when we assess the novelty, there is a lack of clear output produced by this team, emphasizing the *potential* aspect of novelty. In comparison, existing novelty indicators can be considered as *realized novelty*, i.e., they measure the final output of the research conducted by this team. Finally, Faculty Opinion labeling and other external validation methods can describe the *perceived novelty*, i.e., the peers' perception of this study. Seen from this perspective, we investigate whether *potential* novelty contributes to *realized* and *perceived* novelty and its scientific recognition, measured with metrics of disruptiveness [Wu et al., 2019, Bornmann et al., 2019a, Bu et al., 2019]. Using 1.8M articles from the period 2000-2005[3] in PubMed Knowledge Graph (PKG), we analyze the impact of cognitive diversity on novelty, as measured by combinatorial novelty indicators but also peer labeling using Faculty Opinion. The findings reveal an inverse U-shaped relationship between cognitive diversity and average exploratory profiles within teams with combinatorial novelty and citation impact. It is demonstrated that the presence of highly exploratory individuals is beneficial for generating distant knowledge combinations, but only when balanced by a significant proportion of highly exploitative individuals. Moreover, teams with a high share of exploitative profiles consolidate science, while those with a high share of exploratory profiles disrupt it, particularly when associated with exploitative researchers. These results emphasize the significance of team composition in scientific creativity, indicating that a combination of exploratory and exploitative individuals leads to the most disruptive and distant knowledge combinations. Our findings also emphasize the critical role of the cognitive dimensions in creativity, as they significantly influence originality and success. We show that cognitive diversity always seems beneficial to combine more distant knowledge. In contrast, the within-team average exploratory profile follows an inverse U-shaped relation with combinatorial novelty (i.e., there is a turning point where it is no longer beneficial). The same relation can be found when examining the impact in terms of citations.

Mobility and Creativity

In previous chapters, the focus was on the collaborative aspect of the scientific

[3]Restricting our sample to early 2000 allows us to avoid the potential bias of long-term citation caused by "sleeping beauties" [Lin et al., 2021]

system with no distinction on the affiliation of the actors. However, as mentioned earlier, the past few decades have witnessed a rise in mobility [Geuna, 2015], particularly in terms of intersectoral mobility between academic and non-academic organizations. The increasing participation of industry in fundamental research as well as the growing intersectoral mobility of researchers, raises the question of the dynamics of creativity for a scientist that underwent a transition.

Universities play a critical role in knowledge creation, and their missions reflect this importance. The first two missions of universities are teaching and research. Universities are responsible for educating the next generation of scholars, professionals, and leaders. This involves providing high-quality instruction in a wide range of fields. On the other hand, universities are also responsible for conducting cutting-edge research that advances our understanding of the world and addresses some of the most pressing issues facing society [Secundo et al., 2017]. This involves pursuing new ideas and discoveries, testing hypotheses, and developing new technologies and innovations. In recent decades, changes in the national and international context have transformed how universities and economic actors, such as companies, collaborate and the roles they play in society. The purpose of this transformation is to establish a connection between universities and society, which is commonly referred to as universities' "third mission" (TM) [Zomer and Benneworth, 2011, Secundo et al., 2017, Compagnucci and Spigarelli, 2020]. The TM is generally concerned with knowledge transfer or exchange, which involves diffusing the knowledge generated by universities to society and contributing to the surrounding community's economic, social, and cultural development. One of the channels of this TM is the link between academia and industry which can be decomposed into different collaboration types [D'Este and Patel, 2007, Ankrah and Omar, 2015].

Since then, numerous studies have focused on exploring university-industry collaboration (UIC) outcomes. Academic researchers engaged in partnerships with industry professionals tend to exhibit a higher publication output and a reduced patent output compared to their counterparts without such industry affiliations. This trend is ascribed to the availability of corporate resources, contingent upon the specific research domain being pursued [Bikard et al., 2019, Garcia et al., 2020]. Di Maria et al. [2019] found a positive effect on firms' performance following a UIC in the environmental sustainability field but with no effect on the productivity of researchers. While prior research largely emphasizes the benefits of UIC, concerns about potential disadvantages for universities have also emerged. Behrens and Gray [2001]

expressed worries about academic research independence when collaborating with industry. Commercial interests hinder knowledge advancement due to the presence of intellectual property rights on academic research outcomes [Foray and Lissoni, 2010, Larsen, 2011].

While the consequences of university-industry collaboration continue to be explored, there remains limited understanding of the transition of researchers between universities and companies and how it connects with their creativity.
Fernandez-Zubieta et al. [2015] stress that this mobility should be viewed as a long-term process rather than focused solely on short-term outcomes, also discussing the varied incentives and evaluation systems that might impact transitioning researchers' work. Notably, there is a positive correlation between researcher recruitment and the number of patents developed by companies [Herstad et al., 2015]. D'este et al. [2019] results reveal that the relation between Interdisciplinary Research (IDR) and University-Industry (U-I) interaction is influenced by the type of interaction mode. Scientists with an IDR focus show a stronger link to transactional and low goal specificity modes, such as academic entrepreneurship and technology transfer, highlighting their ability to recognize and capitalize on commercial opportunities in their research. Buenstorf and Heinisch [2020] found that PhDs whose dissertation topics diverge from the firm's existing knowledge contribute to more exploratory patents. Those who file such patents early in their careers tend to draw more on their dissertation work.

One of the research field where the industry became prominent in research in recent years is Artificial Intelligence (AI). The recent events with ChatGPT are a testimony to the importance that the industry has on new models that have an impact on society at large. In the last decade, we have seen the dominance of the Deep Learning paradigm supported by private companies. Currently, we face a narrowing of AI research, raising the question of whether the dominance of this paradigm is not sub-optimal and requires the implication of policymakers [Klinger et al., 2020]. The pervasiveness of AI technology and its potential as a General Purpose Technology [Bianchini et al., 2022] raises concerns about an inferior paradigm promoted by companies where the private sector absorbs public resources. Such a scenario could pose risks to society.

Chapter 4 is focused on establishing a deeper understanding of the connection between intersectoral mobility among AI researchers and creativity. The study aims

to contribute to discussions on science policy related to AI development, emphasizing the importance of maintaining a public research space for AI that prioritizes long-term research over short-term commercial interests.

Using OpenAlex, we create an affiliation history of researchers in order to analyze their transition. We identify 1.7M AI papers written by 2.3M researchers. The involvement of the private sector, particularly large tech companies like Google, Microsoft, and Facebook, in AI research, has led to a significant migration of researchers from academia to industry. This trend raises concerns about a potential brain drain from the public sector. Using a survival analysis, we found that researchers specializing in deep learning techniques, which have driven recent AI advancements, are more likely to transition to industry. Researchers with highly cited research and prestigious affiliations are more likely to move to industry. Those with strong connections to companies through co-publication networks are also more prone to transition. Using a difference-in-difference analysis, we found that scholars transitioning to industry experience a decline in academic creativity and novelty, potentially due to the focus on exploiting existing technology. The chapter reveals a significant migration of AI researchers from academia to industry, particularly deep learning specialists and exploitative profiles. Industry involvement raises concerns about the impact on public interest in AI research and academic creativity. The study underscores the need for careful science policy discussions to balance the growth of AI technology with the preservation of a vibrant, explorative public research space.

Introduction générale

"*Une économie de la connaissance est une économie dans laquelle le capital et la main-d'œuvre revêtent une importance accrue, et où la quantité et la complexité des connaissances qui imprègnent les activités économiques et sociétales atteignent des niveaux très élevés.*" [World Bank, 2007].

Le concept de l'économie de la connaissance a évolué et fait l'objet de débats parmi les économistes et les décideurs politiques, avec différentes définitions et interprétations de ce qui constitue une économie de la connaissance et comment mesurer sa progression et ses performances [Powell and Snellman, 2004]. Les principes fondamentaux d'une économie de la connaissance sont la production, la diffusion et l'utilisation de la connaissance [Milewska, 2018]. Au cœur d'une économie du savoir se trouve le processus de créativité, qui peut être décomposé en deux composantes essentielles : l'*originalité* et l'*impact* [Runco and Jaeger, 2012]. L'originalité peut être définie comme la capacité d'être inventif et novateur, ce qui est caractérisée par l'introduction d'éléments innovants qui distinguent une idée ou un concept des paradigmes existants. Elle implique la création de nouvelles connaissances, idées et perspectives qui font progresser notre compréhension du monde. Ce processus comprend la découverte de nouvelles informations, la synthèse des connaissances existantes et le développement de moyens innovants d'aborder les problèmes et les phénomènes. De son côté, l'impact mesure l'influence de la connaissance créée. L'impact d'une idée est déterminé par sa capacité à provoquer des changements, que ce soit dans la compréhension scientifique, l'avancement technologique, l'amélioration de la société ou d'autres domaines.

Une théorie de la création de connaissances dans un contexte organisationnel a été développée par Nonaka et al. [2006]. Selon cette théorie, la production de nouvelles connaissances se produit par des interactions entre la connaissance explicite et la connaissance tacite, via un processus connu sous le nom de « spirale de sociali-

sation, d'extériorisation, de combinaison et d'intériorisation » (SECI). Cela signifie que pour comprendre comment la connaissance est créée et diffusée, il faut examiner les acteurs, les interactions entre ces-derniers, ainsi que l'environnement dans lequel ils co-évoluent.

La science joue un rôle central dans l'économie basée sur la connaissance (KBE) et stimule les avancées technologiques, tout en contribuant de manière significative au développement économique et à la croissance [Nelson and Romer, 1996, Caliari and Chiarini, 2021]. De plus, la science nous aide à comprendre le monde naturel, à améliorer notre qualité de vie et à résoudre des problèmes sociétaux. Elle permet également aux décideurs politiques de comprendre les risques potentiels et les dangers, tels que les catastrophes naturelles, les pandémies et les technologies émergentes [Mazzucato, 2018]. Cela leur permet de créer des stratégies et des réglementations pour minimiser les risques ou promouvoir des initiatives bénéfiques pour la société.

Afin de comprendre la créativité dans le système scientifique, il devient essentiel d'examiner l'interaction entre les individus et entités engagés dans des activités de recherche à différents niveaux d'analyse. Au cours des dernières décennies, il y a eu des changements significatifs dans la manière dont la recherche est menée et la structure du réseau scientifique. Ces changements ont été caractérisés principalement par une mobilité et des collaborations entre chercheurs accrues, ainsi qu'une diminution de la recherche individuelle [Geuna, 2015]. La collaboration et la mobilité sont souvent étroitement liées. La mobilité des chercheurs fait référence au déplacement des chercheurs entre différents environnements de recherche, soit à l'intérieur du même pays (mobilité nationale), soit à travers les frontières internationales (mobilité internationale). Cette mobilité peut survenir à différentes étapes de la carrière d'un chercheur et est influencée par divers facteurs. Les chercheurs peuvent se déplacer entre des lieux, des secteurs et à des étapes de carrière différents [Fernandez-Zubieta et al., 2015]. La mobilité semble jouer un rôle positif dans la collaboration avec de nouveaux coauteurs tout en maintenant des liens déjà existants intacts [Liu and Hu, 2022]. Alors que beaucoup d'attention a été consacrée à la mondialisation de la science, les études se sont concentrées sur le lien entre la collaboration, la mobilité et l'impact, et très peu sur l'originalité. Pourtant, l'originalité joue un rôle central dans la résolution des défis émergents et la recherche de solutions innovantes [Witt, 2016, Fortunato et al., 2018].

Comprendre la structure et la dynamique du système scientifique, la manière

dont les ressources sont orchestrées, et ses conséquences sur la créativité, est crucial à la fois pour favoriser une croissance efficace de l'accumulation de connaissances et pour faciliter le transfert de ces connaissances. Cela nous permet de mieux soutenir les chercheurs confrontés aux défis de notre époque, comme ceux du changement climatique, la pauvreté, et, comme nous l'avons récemment constaté, les pandémies.

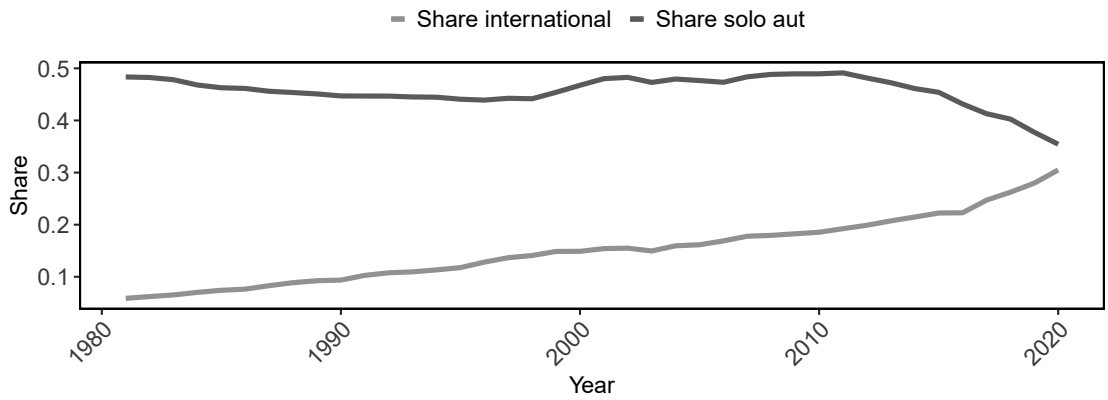
Cette thèse contribue à cette compréhension et est organisée en 4 chapitres. Le **Chapitre 1** se concentre sur la structure et la résilience du réseau de collaboration scientifique mais aussi l'utilisation des ressources existantes et l'adaptabilité suite à un choc exogène, à savoir la pandémie de Covid-19. Le **Chapitre 2** présente un article méthodologique qui introduit *Novelpy*, un outil open source développé sous Python. L'objectif principal de ce package est de faciliter le calcul d'indicateurs de nouveauté et de disruption, améliorant ainsi la transparence dans l'évaluation de la recherche grâce à l'étude, la comparaison, la création et l'application de ces indicateurs. Le **Chapitre 3** décompose l'aspect créatif des publications scientifiques et examine comment la collaboration entre chercheurs affecte la nouveauté et l'impact qui sont les piliers de la créativité. Enfin, le **Chapitre 4** examine la mobilité intersectorielle des chercheurs dans un domaine spécifique, à savoir l'intelligence artificielle (IA), et les conséquences potentielles sur les caractéristiques de leur recherche. Plus précisément, dans cette thèse nous abordons les questions globales suivantes :

- Comment un pays exploite-t-il ses ressources existantes, ses connaissances et ses collaborations passés pour faire face aux défis urgents découlant d'un choc exogène ?
- Comment la composition de l'équipe et la capacité des chercheurs à explorer l'espace de la connaissance influencent-elles l'originalité et l'impact de leur recherche ?
- Quelles transformations surviennent dans la recherche des individus suite à une transition entre les secteurs public et privé ?

Dynamique de la collaboration

L'augmentation des collaborations internationales a été documentée dans de nombreuses études, indiquant une tendance croissante du nombre d'auteurs par article et de collaborations entre pays [Adams, 2012]. Ce résultat est corroboré par Wagner et al. [2017], qui observent une croissance de 120% des collaborations internationales scientifiques entre 1990 et 2013. Cela s'explique en partie par l'augmentation du nombre de pays participant à la recherche (c'est-à-dire plus d'acteurs, plus de possibilités de collaboration). Cette augmentation s'accompagne d'une diminution de la proportion d'articles rédigés par des individus seuls, ce qui semble avoir moins d'impact sur les jeunes chercheurs [Kuld and O'Hagan, 2018]. Une interprétation possible de cette augmentation des collaborations est l'évolution de l'espace de la connaissance. Alors que l'accessibilité à la connaissance scientifique augmente, la taille des équipes augmente également, et les agents se spécialisent de plus en plus. En effet, il est trop complexe pour un agent individuel de naviguer seul dans ce paysage de connaissances en constante augmentation [Boudreau et al., 2016]. La Figure 2, basée sur OpenAlex, une bibliothèque open source complète destinée à représenter le système de recherche mondial avec plus de 240 millions de documents de recherche [Priem et al., 2022], met en évidence la persistance de ces deux tendances récemment observées.

Figure 2: Évolution de la Collaboration en Recherche dans OpenAlex



Notes : Evolution de la part des articles rédigés par un auteur unique (en bleu) et de la part des articles avec au moins deux auteurs provenant de pays distincts (en orange) dans OpenAlex

Dans ce contexte, il est crucial de comprendre comment la structure de la collaboration scientifique émerge et évolue. Plusieurs facteurs contribuent à augmenter

la probabilité qu'une collaboration se produise entre deux auteurs. Par exemple, Boschma [2005] distingue cinq dimensions de proximité qui influencent la possibilité de co-écriture : la proximité cognitive, géographique, institutionnelle, organisationnelle et sociale. La proximité cognitive fait référence à la similitude ou à la compatibilité des connaissances, de l'expertise et des intérêts de recherche des chercheurs. Lorsque les chercheurs ont des antécédents cognitifs similaires ou travaillent dans des domaines connexes, il devient plus facile de communiquer et de collaborer efficacement. La proximité géographique concerne la distance physique entre les chercheurs. Lorsque les chercheurs sont situés à proximité les uns des autres, comme travailler dans la même institution ou la même région, cela facilite les interactions en face à face, les réunions fréquentes et la coordination des projets de collaboration. La proximité institutionnelle peut être mieux comprise comme le contexte national dans lequel évolue le scientifique (c'est-à-dire le "niveau macro-sociétal"). Chaque pays a ses propres objectifs, lois et budgets, ce qui influence les subventions et la création de collaborations. La proximité organisationnelle se concentre sur le contexte d'affiliation des chercheurs. La collaboration devient plus probable lorsque les chercheurs appartiennent à la même organisation ou à des organisations étroitement alignées, car ils partagent des ressources, des installations et des réseaux de recherche. Enfin, la proximité sociale concerne les relations sociales et les réseaux entre les chercheurs. La collaboration est facilitée lorsque les chercheurs ont établi des liens sociaux, tels que des réseaux professionnels communs, des collaborateurs communs, des amitiés ou des relations mentor-mentoré, car la confiance et la familiarité sont déjà présentes.

Pour comprendre la dynamique et la structure des collaborations scientifiques, la théorie des réseaux est fréquemment utilisée. Les chercheurs ont utilisé cette approche pour identifier plusieurs caractéristiques importantes qui définissent le système. L'une de ces caractéristiques est l'émergence d'une structure communautaire au sein du réseau, souvent attribuée au phénomène de la fermeture triadique.[4] Ce phénomène, caractérisé par la tendance des individus à établir des connexions avec d'autres qui partagent des connexions mutuelles, suggère que si la personne A collabore avec la personne B et que la personne B collabore avec la personne C, il est plus probable que la personne A collabore également avec la personne C. De plus, cette structure communautaire peut se manifester sous la forme d'une structure noyau-périphérie, comme cela a été démontré dans des recherches antérieures [Wedell

[4]Des études récentes ont remis en question cette explication [Kim and Diesner, 2017]

et al., 2022]. La consolidation de cette structure noyau-périphérie est renforcée par le mécanisme d'attachement préférentiel. Ce mécanisme suppose que les nouveaux venus dans le réseau sont plus enclins à se connecter initialement avec les individus les plus réputés. En conséquence, le système scientifique devient une structure robuste et résiliente [Wagner et al., 2017].

La collaboration scientifique entre les pays, y compris les pays en développement, est influencée par divers facteurs, et son analyse d'un point de vue macro englobe tous les aspects mentionnés ci-dessus. Selon Dosso et al. [2023], les facteurs spécifiques à la région jouent un rôle significatif dans la formation de la collaboration scientifique entre les pays africains. L'étude a constaté que la langue ethnique partagée, l'appartenance au Conseil africain et malgache de l'enseignement supérieur (CAMES) et la présence d'un partenaire européen commun en tant que troisième partenaire dans les co-publications étaient des facteurs importants de la collaboration scientifique. En plus des autres déterminants de la collaboration scientifique entre les pays, on trouve la distance géographique, l'héritage colonial similaire et la langue commune. Cependant, il a été démontré que ces facteurs perdent en importance. Cela suggère que, à mesure que la recherche scientifique devient de plus en plus mondialisée, d'autres facteurs tels que les intérêts de recherche partagés et l'expertise complémentaire deviennent plus importants. Un autre facteur important dans la collaboration scientifique entre les pays est la présence de réseaux et d'organisations internationaux qui facilitent la collaboration. Par exemple, des scientifiques issus de huit pays membres travaillent ensemble au sein de la communauté SESAME, comme décrit par l'UNESCO.[5] Cette organisation offre une plateforme aux scientifiques pour collaborer et partager leur expertise, malgré les tensions politiques entre certains des pays membres. En plus des organisations régionales, il existe également des organisations mondiales qui favorisent la collaboration scientifique, telles que le Conseil international de la science et l'Académie mondiale des sciences. Ces organisations fournissent un financement et un soutien pour les projets de recherche internationaux et facilitent la collaboration entre les scientifiques de différents pays. La collaboration internationale peut également être influencée par les politiques nationales et les priorités en matière de financement [Ben-David, 1971, Pavitt, 1991, Stephan, 2010].

[5] https://www.unesco.org/en/scientific%2Dresearch%2Dcooperation%2Dwhy%2Dcollaborate%2Dscience%2Dbenefits%2Dand%2Dexamples

La littérature sur la dynamique des collaborations après des chocs exogènes demeure encore limitée. La pandémie de COVID-19 a soulevé des questions sur l'adaptabilité du système scientifique en réponse à des chocs mondiaux urgents. L'objectif du **Chapitre I** est d'obtenir des informations sur ce phénomène et de répondre à deux problématiques. Comment la structure de la recherche sur le coronavirus a-t-elle évolué après la COVID-19 et comment la structure des sciences de la santé au niveau mondial est-elle liée à la réponse scientifique face à la pandémie ? Nous constatons une étroite relation entre la position nationale et internationale dans la recherche liée au coronavirus (CRR). La capacité de CRR accumulée avant la pandémie a été détérminante pour la production nationale de CRR au début de la pandémie. Cependant, la capacité plus large dans le domaine des sciences de la santé est devenue le facteur dominant par la suite. Dans le contexte de la collaboration internationale, le principal moteur de la collaboration en matière de CRR est la présence de collaborations pré-établies qui n'étaient pas liées à la CRR avant la pandémie. La CRR mondiale au cours de la pandémie converge rapidement vers l'ordre mondial de la capacité en sciences de la santé, mais la correspondance n'est pas parfaite. Après le choc, les subventions existantes destinées à la recherche mondiale ont été réorientées vers la CRR. L'émergence de projets de CRR dédiés a conduit à une diminution progressive de l'importance du financement existant dans les capacités de réponse. Dans les premières étapes, les nations dépendantes de financements internationaux pour leur recherche ont connu des perturbations minimales dans leur CRR. Cependant, les années suivant la pandémie a eu un effet préjudiciable sur leur production de papier en raison de cette dépendance.

Ces résultats montrent que lorsque la communauté scientifique est confrontée à un choc mondial, elle a tendance à se réorganiser en imitant, bien que de manière imparfaite, la structure préexistante et en tirant parti des ressources et des capacités existantes. Cependant, il est essentiel de noter que dans ce chapitre, nous n'avons pas évalué la qualité des articles de recherche. Le réseau reproduit est-il efficace ? Comment pouvons-nous évaluer la qualité du système scientifique ? Si la CRR converge vers une structure établie, il devient impératif d'évaluer l'efficacité de cette structure. Comment pouvons-nous établir une configuration de réseau "efficace" dans de telles circonstances ? Les **Chapitres 2 et 3** visent à apporter des éclaircissements sur ces questions cruciales.

Collaboration et creativité

Les origines de ce que nous appelons maintenant la “Science de la Science” et les méthodes de mesure du progrès scientifique remontent au début des années 1900. Paul Otlet a introduit initialement le concept de bibliométrie en 1934 [Otlet, 1990]. Plus tard, en 1969, la bibliométrie a été réintroduite par Pritchard dans un article intitulé “Statistical Bibliography or Bibliometrics?” [Pritchard et al., 1969]. L'objectif principal de la bibliométrie était de relever le défi de la surcharge d'informations et d'aider les bibliothécaires à sélectionner efficacement des documents pertinents pour leurs collections [Sugimoto and Larivière, 2018]. En 1955, le chimiste et documentaliste Garfield a proposé l'idée d'un index des citations, ce qui a conduit à la création de l'Institut de l'Information Scientifique en 1960. Cet institut a ensuite développé le Science Citation Index (SCI), lancé en 1963, qui a simplifié le processus de localisation et de référencement de publications spécifiques. De plus, le SCI a introduit des métriques de citation, telles que le comptage du nombre de fois qu'un article est cité par d'autres, ce qui permet d'évaluer l'impact de la recherche scientifique. Au sein de la communauté scientifique, cette initiative a posé les bases du Web of Science, une plateforme complète qui englobe des métadonnées scientifiques et qui est largement utilisée dans le domaine de la recherche en Science de la Science.

Depuis lors, il y a eu une croissance exponentielle du volume de documents scientifiques, ainsi que l'émergence de divers canaux (par exemple, prépublications, rapports, working papers) et de bases de données contenant des métadonnées associées aux articles (par exemple, Knowledge Graphs, Scite, Crossref). Les métriques de citation sont devenues plus importantes, offrant une diversité d'outils d'évaluation [Waltman, 2016]. De plus, de nouvelles métriques ont été conçues pour capturer différents aspects de la recherche. Par exemple, des indicateurs de nouveauté ont été développés pour évaluer l'originalité d'un document, tandis que des indicateurs de disruption visent à mesurer dans quelle mesure une contribution de recherche transforme les paradigmes existants et a le potentiel d'avoir un impact significatif sur les pratiques actuelles. Des métriques alternatives, connues sous le nom d'Altmetrics, ont également été créées en exploitant des données basées sur le web telles que les “j'aime” et les partages sur Twitter, dans le but de mesurer la visibilité d'un document. Ces évolutions mettent en évidence l'importance non seulement de gérer la vaste quantité de connaissances disponible, mais aussi de comprendre les outils nécessaires à sa gestion efficace.

Les indicateurs de nouveauté et de disruption sont relativement nouveaux et

il existe un manque d'études visant à les comparer. De plus, l'absence de codes disponibles rend difficile pour ceux qui ne sont pas familiers avec le sujet de les utiliser et de reproduire les résultats de la recherche. L'objectif du **deuxième Chapitre** est de normaliser la notation mathématique de certains indicateurs existants en utilisant la théorie des graphes et d'introduire une bibliothèque Python appelée *Novelpy*. Cette bibliothèque vise à simplifier l'application de ces indicateurs émergents et à améliorer leur accessibilité pour les chercheurs et les praticiens.

La nouveauté, souvent désignée comme l'originalité et l'atypicité, est cruciale en science. Tout d'abord, l'originalité en science contribue à promouvoir une diversité d'idées et de perspectives. Cette diversité est essentielle pour le progrès scientifique, car elle permet aux chercheurs d'explorer de nouvelles voies et de remettre en question les paradigmes établis. Si l'originalité fait défaut, le domaine peut stagner, et le corpus de connaissances ne progressera pas [Flexner, 2017, Arnott et al., 2020]. Deuxièmement, un article très novateur a plus de chances d'être un article révolutionnaire qu'un article plus conventionnel [Wang et al., 2017, Shibayama et al., 2021]. Les recherches très novatrices ont tendance à recevoir en moyenne un plus grand nombre de citations, bien qu'elles soient également accompagnées d'une plus grande incertitude [Wang et al., 2017]. Parallèlement, l'incitation à l'originalité est faible. Par exemple, une étude récente montre que si la combinaison de sujets dans une proposition de subvention est trop éloignée du domaine de connaissance du chercheur principal (PI) de la subvention, il est moins probable que la subvention soit attribuée [Franzoni et al., 2022]. Cela implique que pour participer au processus de financement, il faut adhérer aux normes ou conventions existantes dans son domaine. En effet, les chercheurs qui essuient un refus pour une subvention proposent par la suite des recherches moins novatrices [Franzoni et al., 2022].

Les facteurs contribuant au succès de certains articles très novateurs en termes d'impact (c'est-à-dire les articles créatifs) restent relativement peu explorés. Comprendre comment créer un environnement qui favorise la créativité et la réussite en science est d'un intérêt crucial pour aborder les problèmes de société [Fortunato et al., 2018].

Le **troisième Chapitre** de cette thèse est consacré à l'exploration de la créativité du système de sciences de la santé. Nous examinons la relation entre la diversité cognitive au sein des équipes scientifiques et leur capacité à favoriser à la fois des idées innovantes et à obtenir une reconnaissance scientifique. Nous proposons une métrique au niveau de l'auteur basée sur la représentation sémantique des publications passées

des chercheurs pour mesurer la diversité cognitive aux niveaux individuel et de l'équipe. On peut considérer notre indicateur comme une mesure de la *nouveauté potentielle* : nous établissons un lien entre la probabilité de combinaisons nouvelles de connaissances avec la diversité des parcours académiques au sein de l'équipe et la capacité des individus à servir d'intermédiaires efficaces, comblant le fossé entre leurs collègues. Cependant, lors de l'évaluation de la nouveauté, le papier de cette équipe n'existe pas, mettant en avant l'aspect *potentiel* de la nouveauté. En comparaison, les indicateurs de nouveauté existants peuvent être considérés comme une *nouveauté réalisée*, c'est-à-dire qu'ils mesurent la production finale de la recherche menée par cette équipe. Enfin, l'opinion des professeurs et d'autres méthodes de validation externe peuvent décrire la *nouveauté perçue*, c'est-à-dire la perception de cette étude par les pairs. Vu sous cet angle, nous examinons si la nouveauté *potentielle* contribue à la nouveauté *réalisée* et à la nouveauté *perçue*, ainsi qu'à sa reconnaissance scientifique, mesurée avec des indicateurs de disruption [Wu et al., 2019, Bornmann et al., 2019a, Bu et al., 2019]. En utilisant 1,8 million d'articles de la période 2000-2005[6] dans le PubMed Knowledge Graph (PKG), nous analysons l'impact de la diversité cognitive sur la nouveauté, telle que mesurée par des indicateurs de nouveauté combinatoire, mais aussi par la labélisation par les pairs à l'aide de l'opinion d'experts. Les résultats révèlent une relation en forme de U inversé entre la diversité cognitive et les profils exploratoires moyens au sein des équipes avec la nouveauté combinatoire et l'impact des citations. Il est démontré que la présence d'individus hautement exploratoires est bénéfique pour la création de combinaisons de connaissances éloignées, mais seulement lorsque cela est équilibré par une proportion significative d'individus hautement exploitatifs. De plus, les équipes avec une forte proportion de profils exploitatifs consolident la science, tandis que celles avec une forte proportion de profils exploratoires la perturbent, notamment lorsqu'elles sont associées à des chercheurs exploitatifs. Ces résultats soulignent l'importance de la composition des équipes dans la créativité scientifique, indiquant qu'une combinaison d'individus exploratoires et exploitatifs conduit aux combinaisons de connaissances les plus disruptives et éloignées. Nos résultats mettent également en évidence le rôle essentiel des dimensions cognitives dans la créativité, car elles influencent significativement l'originalité et le succès. Nous montrons que la diversité cognitive semble toujours bénéfique pour combiner des connaissances plus éloignées. En revanche, le profil

[6]La restriction de notre échantillon au début des années 2000 nous permet d'éviter le biais potentiel de citation à long terme causé par les "sleeping beauties" [Lin et al., 2021]

exploratoire moyen au sein de l'équipe suit une relation en forme de U inversé avec la nouveauté combinatoire (c'est-à-dire qu'il existe un point de retournement où il n'est plus bénéfique). Cette même relation peut être observée lorsque l'on examine l'impact en termes de citations.

Mobilité et creativité

Dans les chapitres précédents, l'accent était mis sur l'aspect collaboratif du système scientifique sans faire de distinction sur l'affiliation des acteurs. Cependant, comme mentionné précédemment, les dernières décennies ont été marquées par une augmentation de la mobilité [Geuna, 2015], en particulier en ce qui concerne la mobilité intersectorielle entre les organisations académiques et non académiques. La participation croissante de l'industrie à la recherche fondamentale ainsi que la mobilité intersectorielle croissante des chercheurs soulèvent la question de la dynamique de la créativité pour un scientifique qui a fait une transition.

Les universités jouent un rôle crucial dans la création de connaissances, et leurs missions reflètent cette importance. Les deux premières missions des universités sont l'enseignement et la recherche. Les universités sont responsables de l'éducation de la prochaine génération de chercheurs, de professionnels et de leaders. Cela implique de fournir un enseignement de haute qualité dans un large éventail de domaines. D'autre part, les universités sont également responsables de la réalisation de recherches de pointe qui font progresser notre compréhension du monde et qui abordent certains des problèmes les plus pressants de la société [Secundo et al., 2017]. Cela implique la poursuite de nouvelles idées et découvertes, la vérification d'hypothèses et le développement de nouvelles technologies et innovations. Au cours des dernières décennies, les changements dans le contexte national et international ont entraîné une transformation dans la manière dont les universités et les acteurs économiques tels que les entreprises collaborent et les rôles qu'ils jouent dans la société. L'objectif de cette transformation est d'établir un lien entre les universités et la société, communément appelé "troisième mission" des universités (TM) [Zomer and Benneworth, 2011, Secundo et al., 2017, Compagnucci and Spigarelli, 2020]. La TM concerne généralement le transfert de connaissances ou l'échange de connaissances, ce qui implique de diffuser les connaissances générées par les universités dans la société et de contribuer au développement économique, social et culturel de la communauté environnante. L'un des canaux de cette TM est le lien entre le monde

académique et l'industrie, qui peut se décomposer en différents types de collaborations [D'Este and Patel, 2007, Ankrah and Omar, 2015].

Depuis lors, de nombreuses études se sont concentrées sur l'exploration des résultats de la collaboration entre universités et l'industrie (UIC). Les chercheurs universitaires engagés dans des partenariats avec des professionnels de l'industrie ont tendance à produire davantage de publications mais à déposer moins de brevets que leurs homologues sans de telles affiliations industrielles. Cette tendance est attribuée à la disponibilité des ressources de l'entreprise, en fonction du domaine de recherche spécifique poursuivi [Bikard et al., 2019, Garcia et al., 2020]. Di Maria et al. [2019] ont trouvé un effet positif sur la performance des entreprises suite à une UIC dans le domaine de la durabilité environnementale, mais sans effet sur la productivité des chercheurs. Alors que la recherche précédente met largement l'accent sur les avantages de l'UIC, des préoccupations concernant les inconvénients potentiels pour les universités ont également émergé. Behrens and Gray [2001] ont exprimé des inquiétudes concernant l'indépendance de la recherche académique lors de la collaboration avec l'industrie. Les intérêts commerciaux entravent l'avancement des connaissances en raison de la présence de droits de propriété intellectuelle sur les résultats de la recherche académique [Foray and Lissoni, 2010, Larsen, 2011].

Bien que les conséquences de la collaboration entre universités et l'industrie continuent d'être explorées, on en sait encore peu sur la transition des chercheurs entre les universités et les entreprises et sur la manière dont elle est liée à leur créativité. Fernandez-Zubieta et al. [2015] soulignent que cette mobilité devrait être considérée comme un processus à long terme plutôt que de se concentrer uniquement sur les résultats à court terme, et discutent également des incitations et des systèmes d'évaluation variés qui pourraient avoir un impact sur le travail des chercheurs en transition. Notamment, il existe une corrélation positive entre le recrutement de chercheurs et le nombre de brevets développés par les entreprises [Herstad et al., 2015]. Les résultats de D'este et al. [2019] révèlent que la relation entre la recherche interdisciplinaire (IDR) et l'interaction université-industrie (U-I) est influencée par le type de mode d'interaction. Les scientifiques axés sur l'IDR montrent une plus grande relation avec les modes transactionnels et à faible spécificité des objectifs, tels que l'entrepreneuriat académique et le transfert de technologie, mettant en évidence leur capacité à reconnaître et à tirer parti des opportunités commerciales dans leur recherche. Buenstorf and Heinisch [2020] ont constaté que les doctorants dont les sujets de thèse divergent des connaissances existantes de l'entreprise contribuent à

des brevets plus exploratoires, et ceux qui déposent de tels brevets tôt dans leur carrière ont tendance à s'appuyer davantage sur leur travail de thèse.

L'un des domaines de recherche où l'industrie est devenue prépondérante ces dernières années est l'intelligence artificielle (IA). Les récents événements liés à ChatGPT témoignent de l'importance que l'industrie accorde aux nouveaux modèles ayant un impact sur l'ensemble de la société. Ces dix dernières années, nous avons assisté à la domination du paradigme de l'apprentissage profond soutenu par des entreprises privées. Actuellement, nous sommes confrontés à un rétrécissement de la recherche en IA, ce qui pose la question de savoir si la domination de ce paradigme n'est pas sous-optimale et nécessite l'implication des décideurs politiques [Klinger et al., 2020]. La pervasivité de la technologie IA et son potentiel en tant que technologie à usage général [Bianchini et al., 2022] suscite des inquiétudes concernant un paradigme inférieur promu par les entreprises, où les ressources publiques sont absorbées par le secteur privé. Un tel scénario pourrait présenter des risques pour la société.

Le **Chapitre 4** est axé sur l'établissement d'une compréhension plus approfondie du lien entre la mobilité intersectorielle des chercheurs en intelligence artificielle (IA) et la créativité. L'étude vise à contribuer aux discussions sur la politique scientifique liée au développement de l'IA, en mettant l'accent sur l'importance de maintenir un espace de recherche public pour l'IA qui privilégie la recherche à long terme par rapport aux intérêts commerciaux à court terme.

En utilisant OpenAlex, nous avons créé un historique des affiliations des chercheurs afin d'analyser leur transition. Nous avons identifié 1,7 million d'articles sur l'IA rédigés par 2,3 millions de chercheurs. L'implication du secteur privé, en particulier de grandes entreprises technologiques telles que Google, Microsoft et Facebook, dans la recherche en IA a entraîné une migration significative de chercheurs de l'académie vers l'industrie. Cette tendance suscite des inquiétudes quant à un possible exode des cerveaux du secteur public. En utilisant une analyse de survie, nous avons constaté que les chercheurs spécialisés dans les techniques d'apprentissage profond, qui ont impulsé les récents progrès en IA, sont plus susceptibles de passer à l'industrie. Les chercheurs dont la recherche est très citée et qui ont des affiliations prestigieuses sont également plus susceptibles de passer à l'industrie, et ceux qui ont des liens solides avec les entreprises grâce à des réseaux de co-publication sont également plus enclins à effectuer cette transition. En utilisant une analyse de doubles différences, nous avons constaté que les chercheurs qui passent à l'industrie connaissent une diminution de leur créativité et de leur originalité académique, potentiellement en

raison de leur focalisation sur l'exploitation de la technologie existante. En résumé, ce chapitre révèle une migration significative des chercheurs en IA de l'académie vers l'industrie, en particulier les spécialistes de l'apprentissage profond et les profils axés sur l'exploitation. L'implication de l'industrie suscite des inquiétudes quant à l'impact sur la recherche en IA d'intérêt public et la créativité académique. L'étude souligne la nécessite de débattre des politiques scientifique et de recherche à concilier la croissance de la technologie IA avec la préservation d'un espace de recherche publique exploratoire et dynamique.

Chapter 1

On The Global Health Science Response To COVID-19

This chapter was co-authored with

Moritz Müller, Pierre Pelletier and Stefano Bianchini

Summary of the chapter

How has the global health science system reacted to the COVID-19 pandemic ? Here we investigate how national output and international collaboration on coronavirus-related research (CRR) correlate with prior activity in the health sciences, pandemic-related factors, and the broader socio-economic context. We find that prior CRR experience is influential in national CRR, particularly in the first three months of the pandemic. Subsequently, more general health science capacity becomes the dominating factor of CRR output. National COVID-19 incidence rates, national confinement measures, and broader socio-economic conditions turn out to be only weakly correlated with national CRR. However, they do play a role in the context of international collaboration to some extent. Existing projects have been redirected from addressing global health challenges to specifically tackle COVID-19 related issues. Dependence on external funding is detrimental to productivity at various stages of the COVID-19 response, both domestically and in the collaboration aspect. Finally, the rapid expansion of global CRR mostly followed the structure laid out by the global health science system. However, the international CRR Network experienced a significant decrease in hierarchy accompanied by an increased collaboration within pre-established regional science communities.

The paper at hand treats the outbreak of the novel coronavirus Sars-CoV-2 in January 2020 as an exogenous shock to the international health science system. The subsequent COVID-19 pandemic has been global in the sense that, within a relatively short period, most countries worldwide have been directly concerned. Consequently, coronavirus-related research (CRR) became a top priority for health science on national and international levels worldwide.

The burst of CRR across countries and the associated emergence of a large international CRR network have been documented since the early phases of the pandemic [Aviv-Reuven and Rosenfeld, 2021, Cai et al., 2021, Chahrour et al., 2020, Fry et al., 2020, Haghani and Bliemer, 2020, Radanliev et al., 2020, Zhang et al., 2020]. In the rankings of national scientific output as well as network centrality, the usual suspects tend to score high; the United States takes the lead, followed by China and the most developed countries. A potential element influencing the adjustment of CRR is the pre-existing global health science capacity. Yet, quantitative evidence is lacking because most studies focus exclusively on CRR. Therefore, existing research captures system dynamics mostly by considering pre-pandemic CRR as initial conditions from which the CRR network strides away during the pandemic as it expands. The existing broader health science system in which the international CRR Network expands has not been explicitly accounted for as a potential attractor. So far, the mirroring of global CRR and health science capacity is merely hypothesized. Further note that the hypothesized mirroring is unlikely to be perfect. The global science system is shaped by an interwoven web of processes playing inside and outside the science system, and the COVID-19 pandemic shifted some of the relevant parameters.

This paper investigates the dynamics of global CRR during the COVID-19 pandemic within the global health science system, taking into account initial socio-economic conditions, pandemic-related factors, and funding-specific characteristics such as exterior funding dependence. We conceive the global science system as a network of interconnected national science systems. The analysis proceeds in three steps. In a first step, we model national CRR conditional on initial conditions and pandemic development at the national level. The second analysis models bi-national collaboration on CRR. A third analysis looks at the convergence of global CRR towards global health science in terms of national scientific output, network centrality of countries, and the overall network structure. Thus, whereas analyses one and two reveal driving factors of national and bi-national CRR respectively, analysis three deals with global CRR dynamics at the macro level.

These analyses aim at a better understanding of the dynamics of the global science system, with a focus on internal science dynamics triggered by the external COVID-19 shock. How the global science system responds to external shocks is certainly relevant given the global social and environmental challenges ahead [Schot and Steinmueller, 2018]. Knowledge about the patterns of response to global crises may guide the organization of science in the future. For one, scientific capacity needs to be built on national and international levels that facilitate changes in the direction of scientific development. Another issue is how the existing scientific capital can be efficiently orchestrated during crises. Our paper does not provide definite answers to these questions. However, we hope that our empirical description contributes to the rationalization of global science dynamics during the pandemic.

Our empirical analyses show that the pre-pandemic, broader health science system is the dominating factor in the dynamics of global CRR during the pandemic. Other factors that are currently debated, and we explicitly account for in our analysis, have more limited influence on the national output but appear to be significant for collaboration dynamics. Furthermore, we observe a detrimental impact of funding dependence on both international and national CRR outputs, although these effects seem to manifest during distinct phases of the pandemic. Consistently, at the macro level, we find that global CRR rapidly converged towards the global structure of the broader health sciences. But convergence is not perfect. In the first two months, network hierarchy overshoots such that it significantly exceeds the 'typical' levels of global health science. Subsequently, network hierarchy decreases significantly below the hierarchy of global health whereas collaboration activity within world regions increases significantly. Thus, the science world responds to the global crises with increased regionalization.

For science policy, our findings provide an empirical argument for a strategy that emphasizes generic scientific capability. Furthermore, active participation in global knowledge flows during 'normal' times will be the key to handling global crises in the future. How exactly global science efforts should be orchestrated remains an open issue. Future research could investigate to what extent our observations on the health system dynamics — global concentration followed by global diffusion and regional interaction — constitute an efficient response to global crises. Another question is whether the observed regionalization is due to (transient) pandemic circumstances or accentuates a long-term trend in the sciences and other socio-economic spheres. The final potential pathway could be the exploration of the type of research conducted

and funded (in terms of originality, impact, and topic) before and after the pandemic on both CRR and non-CRR papers.

The paper follows a conventional structure: In section 1.1, we introduce the significance of the national context in science, outline the structure of scientific collaboration, and highlight the characteristics of the health and coronavirus research system. Next, in section 1.2, we provide information on the data sources and analytical methods used. Moving forward, section 1.3 unveils the outcomes of the three analyses. Lastly, section 1.4 presents a discussion of the findings and their limitations and offers concluding remarks.

1.1 Background

1.1.1 Overview

This section starts with a bird's-eye view of the national context of science systems. The second part of the section presents the characteristics and structure of the scientific system. We then conclude this section by examining the specificity of the global health science system and coronavirus research.

1.1.2 The 'national' in global science

Science does not exist in isolation. The global science system evolves within the overall social and economic system with strong mutual feedback. On the one hand, science contributes to the global pool of knowledge that drives social and economic development [see e.g. Kuznets, 1973]. On the other hand, society strongly conditions scientific activity.

The national context is particularly influential [Ben-David, 1971]. In Europe from around 1800 recognition of the utility of science by national governments and industry led to increased support and autonomy of the scientific community [Beaver and Rosen, 1978]. The establishment of formal and informal institutions has been mostly a national effort [Beaver and Rosen, 1979]. Also, the expansion of science in the late 19th century and the first half of the 20th century has been born out of an explicitly national rational. Building on arguments of militaristic and economic competition among nations, and national reputation, governments of advanced economies expanded and shaped purposefully their national science systems. This gave rise notably to the foundation of today's national research institutions in ad-

vanced economies such as the CNRS in France, the Kaiser-Wilhelm-Gesellschaft (succeeded by Max-Planck-Gesellschaft after the Second World War) in Germany, or NSF in the US [Ben-David, 1971].[1] In the second half of the 20th century, the contribution of science to national economic growth has become a major justification for national investments in higher education and scientific research [Pavitt, 1991].

To this end, national governments set strategic targets for Gross Domestic Expenditure on R&D (GERD) of around three percent. Government funding of science specifically may be proxied best by research expenditures of higher education institutions, i.e. higher education R&D expenditure (HERD) Stephan [2010].[2] In the past 20 years, HERD in OECD countries has been between 0.2 to 0.6 percent of GDP, fairly stable within countries and across, with an average of 0.4 percent. For non-OECD countries, HERD is systematically lower, around 0.2 percent of GDP (OECD, 2019).[3]

Public science expenditures also respond to national and global events. In the USA, for example, the Sputnik shock in the late 1950s led to a considerable increase in government expenditures. During the Vietnam War, relative expenditures decreased again [Stephan, 2010]. In 1998, the National Institutes of Health (NIH) doubled its funding to strategically support the high growth of the biotechnology industry [Zucker et al., 1994]. The American Recovery and Reinvestment Act of 2009 increased considerably public science expenditures as a countercyclical measure [Stephan, 2010]. Most recently, the US Congress discussed the proposition to increase the NSF budget by 100 billion US$ — the largest increase in the agency's history — with the explicit goal of maintaining global innovation leadership as a response

[1] President Roosevelt's letter to the director of the 'Office of Scientific Research and Development' in 1945, initiating the foundation of the National Science Board in the US, is a good example on that point: *"DEAR DR. BUSH: The Office of Scientific Research and Development, of which you are the Director, represents a unique experiment of team-work and cooperation in coordinating scientific research and in applying existing scientific knowledge to the solution of the technical problems paramount in war. [...] its tangible results can be found in the communiques coming in from the battlefronts all over the world. [...] There is, however, no reason why the lessons to be found in this experiment cannot be profitably employed in times of peace. The information, the techniques, and the research experience developed by the Office of Scientific Research and Development and by the thousands of scientists in the universities and in private industry, should be used in the days of peace ahead for the improvement of the national health, the creation of new enterprises bringing new jobs, and the betterment of the national standard of living."* https://www.nsf.gov/about/history/nsf50/vbush1945_roosevelt_letter.jsp

[2] Note that not all government-funded science is performed in the higher education system, not all research in higher education is science, and the government is the main but not the only funding source of university research HERD.

[3] Supra-national entities such as the EU also define science targets, but funding remains national to a large extent.

to China's national efforts in science [Mervis, 2021, Remmel, 2021]. An integral part of China's science strategy in the past 30 years has been scientific collaboration — bolstered through large-scale student and scientist exchange programs — notably with the US but also Europe [Wang and Wang, 2017]. In parallel, European-wide research programs, notably Horizon 2020, now Horizon Europe (95 billion Euro), are relevant for national science systems and support their integration on a regional level. Similar tendencies can be observed in practically all world regions.

The phenomenon of increasing international research collaboration (IRC) is one aspect of globalization in science. It accelerated in the early 1980s [Adams, 2013] and may have reached a saturation level in some more advanced economies by now [Ponds, 2009]. International collaboration is observed in particular among productive researchers from top-tier universities located in advanced national scientific systems [Pan et al., 2012, Jones et al., 2008]. The gain is more excellent research [Adams, 2013, Pan et al., 2012]. The tendency of 'excellence-attracting-excellence', however, entails the risk of increasing hierarchical stratification not only within but also between national science systems [Beaver and Rosen, 1979, Jones et al., 2008, Horlings and Van den Besselaar, 2011]. In order to catch up scientifically, or at least not to fall behind, being well connected to the global knowledge flows has become a science policy imperative in most countries.

Examining the global science system by considering both national and international scientific endeavors collectively, while also examining funding dynamics, becomes crucial for achieving a deeper understanding of the evolution of this scientific ecosystem.

1.1.3 Structure and processes in global science

The global science system is characterized by two salient features: First, worldwide scientific activity is highly concentrated in some places. In other words, the global science system is highly hierarchical. Second, countries exhibit a certain 'preference' to collaborate with certain other countries. Such national tendencies in IRC create clusters of scientific activity. We discuss each in the following.

The world's uneven distribution of scientific capacity has been discussed early on in the literature. Davidson Frame et al. [1977] counted country affiliations of scientific articles published in an early 1973 ISI collection, and found that the concentration of science production exceeds economic concentration in the world. Subsequent studies repeated and varied the exercise. The extent of scientific concentration is immedi-

Table 1.1: World's largest science countries.

	1973 (Frame et al., 1977)	1981-1994 (May, 1997)	1997-2001 (King, 2004)	2008-2018 (Allik, 2020)
1	US (38.2)	US (34.6)	US (34.6)	US (19.9)
2	UK (9.2)	UK (8.0)	UK (8.5)	China (11.7)
3	USSR (9.0)	Japan (7.3)	Japan (8.0)	UK (6.0)
4	West Germany (6.0)	Germany (7.0)	Germany (7.4)	Germany (5.3)
5	France (5.5)	France (5.2)	France (5.6)	Japan (4.1)
6	Japan (5.2)	Canada (4.5)	Canada (4.6)	France (3.7)
7	Canada (4.4)	Italy (2.7)	Italy (3.3)	Canada (3.3)
8	India (2.5)	India (2.4)	Russia (3.3)	Italy (3.2)
9	Australia (1.9)	Australia (2.1)	India (2.8)	India (2.8)
10	Italy (1.7)	Netherlands (2.0)	Netherlands (2.3)	Australia (2.8)
Total share	84%	76%	80%	63%

Frame et al. (1977) covers 2,300 journals indexed by SCI, May (1997) covers 4,000 journals indexed by ISI, King (2004) covers 8,000 journals indexed by ISI, Alik (2020) covers 12,000 journals indexed by ESI.

ately grasped by looking at the global share of the most prolific countries; shown in Table 1.1 for four time periods [based on Davidson Frame et al., 1977, May, 1997, King, 2004, Allik et al., 2020]. In the period from 1970 to 2000, the ten most productive countries contributed around 80 percent to worldwide scientific output ('Total share' in Table 1.1). In that period, we see the fall of the USSR in the 1980s and 1990s in the data.[4] Somewhat less dramatic, but still noticeable, is the scientific rise and fall of Japan. Besides the movements of individual countries, worldwide concentration remained relatively stable throughout the fourth quarter of the last century.

In the most recent period, i.e. 2008 to 2018 in the last column of Table 1.1, the global share of the top ten countries dropped to around 60 percent. The lower share may be partly explained by the larger sample of journals which includes more non-English and less established journals. However, the drop reflects a significant real-world development. In the last twenty years, in particular Eastern European and Asian emerging economies, notably China, experienced high growth rates that have

[4]May [1997] excluded the USSR (and Russia) due to issues in assigning research output.

been much higher than growth rates in the economic and scientifically more advanced countries[Horlings and Van den Besselaar, 2011]. This results in lower concentration. On the other hand, less developed countries had very little or no growth during that period, which increases the global divide [Horlings and Van den Besselaar, 2011]. In summary, over the last two decades, there has been a trend towards increased equality between advanced and developing economies, yet this parity has not been achieved globally.

Patterns in IRC are coupled to the size and dynamics of national science systems and are tied to the same geo-political and economic processes. For example, the fall of the USSR has been associated with a significant West orientation in international scientific collaboration of the former east-block countries between 1985 and 1995 [Braun and Glänzel, 1996]. China's growth of scientific output has been coupled with an increase of internationally co-authored papers [Niu and Qiu, 2014]. More accurately, China's IRC intensity, i.e., internationally co-authored papers over total papers, remained stable at around one-fourth [Niu and Qiu, 2014]. The same pattern of proportional growth of IRC and total scientific output has also been observed for other larger emerging economies (China, India, South Korea, Brazil) over the growth period 1980 to 2010 by Adams [2013]. During the same period, advanced economies have grown much more through internationally co-authored papers rather than through domestic papers, implying an increasing IRC intensity over time. Another observation is that IRC intensity decreases with the overall size of the national science system. In other words, the absolute number of internationally co-authored papers tends to increase with the total number of papers with an elasticity below one [Davidson Frame and Carpenter, 1979, Luukkonen et al., 1992, Pan et al., 2012]. Thus, larger science nations tend to have more international collaborations in total but fewer international collaborations per paper than smaller science nations.

National science output together with variations in national IRC intensity, translate into a natural order in the hierarchy of the worldwide IRC network. Gui et al. [2018] investigate hierarchy in the IRC network by looking at the countries' network centrality, i.e., degree, closeness, betweenness, and eigenvector centrality.[5] The different centrality measures are highly correlated, which is typical for real-world networks.

We note that the ranking by network centrality corresponds largely to the ranking

[5]The study is based on papers in the Web of Science Core Citation Database in the years 2000 and 2015.

by countries' science production as seen in Table 1.1. This is due to the coupling of national size and the number of international collaborations. In the few cases where rankings disagree, national IRC intensity is the obvious explanation. Given overall science production, Japan is less central in the IRC network than expected. The reason is Japan's relatively low IRC intensity, which in turn may be explained by geography, culture, and language [Yonezawa, 2011]. Russia tends to be less central in the IRC network than in terms of national science production. On the other hand, small and advanced Western countries, in particular the Netherlands and Switzerland, have high IRC intensity and are consequently among the ten most central countries in the network. The question of partner choice, i.e., national tendencies to collaborate with certain other countries, is most likely a secondary effect that is dominated by overall collaboration activity in the creation of the hierarchy.[6]

The hierarchical structure of the IRC network suggests that positive feedback drives the local accumulation of science capacity. Such rich-get-richer, or accumulating advantage effects, are central in economic and sociological theories. The core-periphery theory introduced by Friedmann [1967] describes positive feedback in the structuring of international economic and political relations. This theory puts forward, among others, that the system's core is marked by a high density of creative potential leading to high creative interaction compared to the periphery, which attracts further inflow of creative potential from the periphery. Observed migration flows of scientists are in line with that theory [Stephan and Levin, 2001], and the literature on international migration of scientists provides a close connection to international research collaboration that is arguably causal [Jonkers and Tijssen, 2008, Kato and Ando, 2017]. Accumulating advantages in science through social effects of scientific collaboration have been emphasized by Beaver and Rosen [1979]. They noted that scientific collaboration is not only about resolving the inter-dependence of physical or cognitive resources but also instrumental for gaining recognition, reputation, and status. As a result, differential access to resources drives international scientific status and vice versa. The same logic goes through in particular for IRC [Melkers and

[6]One argument is the strong alignment between network centrality and IRC activity (number of collaborations). Another argument is that the different network measures applied by Gui et al. [2018] produce essentially the same ordering. Consider India as the exception where partner choice makes a difference. India became a regional science leader in 2015. As a hub in the (semi-)periphery, India has high closeness centrality (well connected to all other countries worldwide) but low betweenness and eigenvector centrality (many knowledge flows in the network sidestep India). The fact that for most countries these measures closely align, indicates that differences in partner choice (which are arguably there, see below) do not heavily influence the hierarchy dimension of the network position.

Kiopa, 2010], susceptible of creating a rich-get-richer phenomenon [Hâncean et al., 2021, Katz, 2000, Wagner and Leydesdorff, 2005, Wagner et al., 2019].

Once country size effects are accounted for, national preferences in IRC partner choice are revealed. Collaboration preferences are shaped along multiple dimensions that are often not purposefully created at all (e.g. geographical distance), somehow inherited (e.g. culture and language, common cognitive basis, and historical connections from colonial times or past migration flows), or subject to (inter-)national policy (e.g. efforts to structure the European Research Area, exchange agreements in higher education, migration policies to attract scientists, joint research mega-projects such as CERN) [Davidson Frame and Carpenter, 1979, Luukkonen et al., 1992, Zitt et al., 2000, Frenken et al., 2009, and the literature cited therein]. Most of these factors can be conveniently couched in terms of dyadic distance [Frenken et al., 2009]. Hence, countries that are close in some dimensions of space tend to form mostly regional science clusters [Wagner and Leydesdorff, 2005].

The role of the evolving technological regime, notably the ICT revolution, for IRC is still somewhat ambiguous in the empirical literature [Gui et al., 2018, and the literature cited therein]. ICT developments facilitate fast and massive information exchange. This has probably contributed to the rise of international scientific collaboration since 1980. However, ICT impacts are not likely to be uniform because developed countries still have preferential access to ICT, and ICT may help more to overcome geographical than other kinds of distances.

Let us note the key ideas on the global science system laid out so far. First, national scientific activity is highly concentrated in advanced economies and emerging economies. The international research collaboration network exhibits essentially the same hierarchy. Countries tend to collaborate depending on intra- and extra-science factors, creating regional 'science clusters'. Dynamics in terms of national scientific activity and network density and structure tend to be on a time scale of decades, which can be explained by the stability of underlying factors but also the process of developing international science capacity characterized by positive feedback effects.

1.1.4 COVID-19 and global health sciences

Global *health* science has essentially the same structure and is subject to the same processes as global sciences [Cantner and Rake, 2014, Wagner et al., 2017, Gazni et al., 2012], but some specificities are to be noted. First, health sciences include a variety of scientific fields, and IRC intensity varies substantially across these fields:

Most international is Molecular Biology and Genetics with 25 percent of internationally co-authored publications, while in Clinical Medicine IRC is least common with an IRC share of around 15 percent [Gazni et al., 2012]. Second, countries differ in their scientific profiles. In terms of investments, while public GERD expenditures are in general highly correlated with GDP, wealthier countries tend to spend a higher share of their GERD on health sciences.[7] Looking at scientific publications — some exceptions allowed — Eastern countries (former USSR and Asia) tend to focus more on engineering and technologies. In contrast, Western countries (USA, Western Europe) tend to be particularly strong in health sciences [Glänzel, 2001]. Third, health sciences are embedded in the national social and economic system. Medical practice is carried out within the broader health infrastructure (e.g., hospitals). And health science is highly connected to the research-intensive pharmaceutical industry [Zucker et al., 1994]. All this suggests that global health sciences may exhibit a pronounced hierarchical structure and strong inertia in (inter-)national development.

Prior epidemics caused by a coronavirus (MERS and SARS) led to the formation of CRR-specific national capacity and IRC network structures [Haghani and Bliemer, 2020, Mendes and Carvalho, 2020, Zhang et al., 2020]. These epidemics took place mainly in the developing world and have been regionally confined. Countries that have been directly concerned had a strong incentive to increase relevant knowledge production and to shape research priorities and the research agenda. One way forward has been research collaboration among concerned countries as well as developed countries to which strong ties existed before, primarily due to historical, initially non-scientific relationships, or who had strong competencies in a given field [Zhang et al., 2020, Haghani and Bliemer, 2020]. This resulted in regional CRR networks in the Middle East and Asia connecting to advanced economies. In particular, China (after SARS in 2002), Saudi Arabia (after the MERS epidemic in 2012), and some developed countries (US, UK, Germany, and Netherlands involved in both) built CRR-specific competencies [Mendes and Carvalho, 2020, Zhang et al., 2020]. These regional efforts have been continued after respective epidemics and resulted in regional specialization patterns. Notably, Saudi Arabia, particularly affected by the MERS crisis in 2012, still had a strong focus on CRR in 2019; contributing 6% of CRR compared to 0.6% of non-CRR research output in our sample (Section 1.2.2).

CRR before the COVID-19 pandemic naturally yields CRR relevant scientific

[7] Own calculation based on the 'Research and Development' dataset of UNESCO Institute of Statistics (UIS), release date March 2021.

capacity. This footprint may influence the dynamics of international CRR during the COVID-19 pandemic. However, health research on any pandemic necessarily deals with a vast array of research fields and topics [Zhang et al., 2020]. Countries therefore may still differ in their relative research focus (e.g. public health in the US and biochemistry in China) [Zhang et al., 2020]. Hence, many countries that did not research specifically the coronavirus before the pandemic nevertheless will have CRR capacity [Lee et al., 2020]. This suggests a strong alignment of global CRR during the global COVID-19 pandemic with the existing scientific, technological, and human capital in the broader international health sciences.

The global spreading of the coronavirus implies that observed cases and, hence, national research needs and opportunities varied over time. The spreading of the pandemic is well documented by the Johns Hopkins Coronavirus Resource Center (CRC) [Dong et al., 2020]. The origin of the pandemic has been in China where the number of cases strongly increased until mid-February. The first cases in neighboring Asian countries were confirmed mid January, closely followed by first confirmations in the US, Europe, and Australia at the end of January. In Europe, Italy has been particularly concerned by the end of February 2020. At around the same time, the virus has been confirmed in the Middle East and North Africa. By the end of March 2020, the virus had been confirmed all around the world with varying intensity from then on. The global infection process provides a compelling narrative for explaining certain aspects of the observed dynamics, in particular the early centrality of USA, China, and later Italy in CRR research [as put forward in Fry et al., 2020].

Existing bibliometric studies on the impact of the COVID-19 pandemic on global health science tend to focus only on coronavirus-related research. Several empirical studies show that — within the first three months of the pandemic — global scientific output increased significantly through a highly uneven contribution of individual countries; often framing it as a scientific race [Aviv-Reuven and Rosenfeld, 2021, Chahrour et al., 2020, Haghani and Bliemer, 2020, Radanliev et al., 2020, Zhang et al., 2020]. A closer look at the content of CRR papers shows remarkable differences across countries in terms of disciplinary focus. Zhang et al. [2020] shows the ten most prolific countries that they contribute in the scientific field in which they 'specialized'. Guleid et al. [2021] show for African countries that only one percent of CRR focuses on therapeutics or vaccines, while one-fourth of publications assess countries' preparedness and response to the pandemic, and another twenty percent describe indirect health impacts of the pandemic. The empirical methodology of

CRR is mostly based on prospective observational studies and case series (clinical series), whereas non-CRR papers are mainly composed of randomized controlled trials, at least in the top medical journals [based on 402 papers published in 2019-2020, Gai et al., 2021].

International scientific collaboration and cooperation to address the global pandemic has been emphasized immediately after the outbreak of the global COVID-19 pandemic in January 2020 by the scientific community, policy, and the public at large. The WHO actively coordinated international efforts, and various consortia leveraged their international networks for CRR [Kinsella et al., 2020].

The timely and widespread diffusion of relevant research results has been one pillar of IRC. One example is the public release of the 2019-nCoV viral sequence by Lu et al. [2020], a Chinese research group, in January 2020. The upsurge of preprint papers on COVID-19-related papers marks the shift to more open science by the scientific community as a whole [Aviv-Reuven and Rosenfeld, 2021, Homolak et al., 2020]. Due to travel restrictions, the health science community met at large digital conferences. Also, publishers contributed to the fast dissemination of research by considerably shortening the time from submission to publication; something already seen in prior pandemics [Aviv-Reuven and Rosenfeld, 2021, Palayew et al., 2020].

The other pillar has been joint research on an international level. Scientific breakthroughs on CRR, as documented in highly cited scientific papers, have often been achieved by international teams [Aviv-Reuven and Rosenfeld, 2021]. Travel restrictions made physical exchange more difficult, and perhaps the need to reduce transaction costs due to urgency led to smaller team size in IRC on CRR compared to pre-pandemic levels [Cai et al., 2021, Fry et al., 2020, Lee et al., 2020]. Aviv-Reuven and Rosenfeld [2021] compare coronavirus-related research papers with other health science papers published during the pandemic, finding that coronavirus research teams tend to be smaller and less international. Despite that, international research teams account for around one-third of the overall publication output on coronavirus-related research [Aviv-Reuven and Rosenfeld, 2021, Cai et al., 2021, Lee and Haupt, 2021]. That share remained stable during the exponential upscaling of coronavirus-related research in 2020, despite travel restrictions and other impediments to international collaboration during the pandemic [Cai et al., 2021]. This has been made possible by a high influx of scientists new to coronavirus-related research. Around 80 percent of all co-authors on relevant papers did not cooperate before on CRR (but potentially on related topics) [Liu et al., 2021].

Zhang et al. [2020], Haghani and Bliemer [2020], Fry et al. [2020] provide an account — up to April 2020 — of the transformation of the IRC network on CRR during the global COVID-19 pandemic. A pertinent finding is that the network expanded rapidly (within the first two to three months) worldwide. Another common finding is that countries are highly heterogeneous not only in terms of individual science production but also in their network centrality. Fry et al. [2020] and Cai et al. [2021] note that the IRC network on CRR has become more 'elitist' with the pandemic. The central role of two countries, USA and China, and their strong IRC interaction has been emphasized [e.g Fry et al., 2020]. Interestingly, clinical medicine has been a long-standing core subject of US-Chinese IRC [shown for the period 2000 to 2010 by Niu and Qiu, 2014]. Cai et al. [2021] document for the second half of 2020 network position dynamics of some central countries, and in particular a relative weakening in the USA-China interaction on CRR.

In summary, previous empirical research on the development of international CRR during the COVID-19 pandemic mainly focused on CRR in isolation. Contextual factors put forward are rarely accounted for in systematic empirical analyses of (inter-)national CRR dynamics. Ideas on the role of CRR capacity built prior to the pandemic, global virus spreading dynamics, travel restrictions, and ICT solutions for science dissemination are part of the discussion. The relevance of the global health science system — in which CRR is embedded — is visible across all studies but somewhat remains 'the elephant in the room'. Therefore we ask, how does the structure of global health sciences relate to the scientific response to the pandemic?

1.2 Data and methods

1.2.1 Overview

The empirical analysis consists of three interrelated parts. The first and the second analyses investigate factors driving national CRR output and international collaboration on CRR respectively. The third analysis investigates how national and international CRR development relates to the worldwide distribution of health science. Scripts to reproduce the analyses are available on GitHub `https://github.com/P-Pelletier/Global-health-sciences-response-to-COVID-19` and the final data used for regression can be found here `https://zenodo.org/record/8238355`. This section continues with a description of the data before we turn to the empirical methods used in the analyses.

1.2.2 Data

We measure scientific output and international research collaboration on scientific articles.[8] Our main dataset is a collection of peer-reviewed articles in journals indexed by MEDLINE. The restriction to MEDLINE-indexed journals ensures that papers in the sample fall into our scope of biomedical research and are of minimum scientific quality. We collect papers from the pre-COVID-19 period (Jan.–Dec.2019) and from the COVID-19 period (Jan.–Dec.2022).[9] Furthermore, we distinguish CRR papers from non-CRR papers. The analysis is based on the papers' submission dates to stay close to the actual research activity.

The dataset has been constructed as follows. We downloaded papers appearing in MEDLINE journals from the PubMed database as of June 2023. Coronavirus-related papers are identified through a text search query suggested by PubMed Central Europe on the papers' title, abstract, and MESH terms.[10] Countries are identified in paper affiliations through regular expressions and subsequent manual cleaning.[11] PubMed API also provides information regarding research funding. The specific information we focus on relates to the country of origin for the grants.[12] Additionally, we are interested in the identification numbers associated with these grants.

To assess the data, one has to keep in mind the existence of several time lags. First, there is an unknown time lag from research to submission. Then, there is a time lag from submission to acceptance by the journal. Finally, there is a time lag from journal acceptance to entry on PubMed. For the papers in our dataset, time lags from submission to acceptance to entry in PubMed are known. Appendix 1.5.1 provides respective distributions for CRR and non-CRR papers. Looking at the time lag from journal acceptance to PubMed entry, some underreporting becomes likely

[8]This is common practice in the literature. The pros and cons are discussed elsewhere [Katz and Martin, 1997, is the seminal reference]

[9]We chose to confine our pre-COVID-19 timeframe to 2019 due to our primary interest in examining the convergence between the CRR and non-CRR scientific systems. Upon conducting an investigation, it became evident that the CRR pattern remained consistent from 2015 to 2019, prompting us to simplify our analysis.

[10]In detail, the search query is: ("2019-nCoV" OR "2019nCoV" OR "COVID-19" OR "SARS-CoV-2" OR "COVID19" OR "COVID" OR "SARS-nCoV" OR ("wuhan" AND "coronavirus") OR "Coronavirus" OR "Corona virus" OR "corona-virus" OR "corona viruses" OR "coronaviruses" OR "SARS-CoV" OR "Orthocoronavirinae" OR "MERS-CoV" OR "Severe Acute Respiratory Syndrome" OR "Middle East Respiratory Syndrome" OR ("SARS" AND "virus") OR "soluble ACE2" OR ("ACE2" AND "virus") OR ("ARDS" AND "virus") or ("angiotensin-converting enzyme 2" AND "virus")).

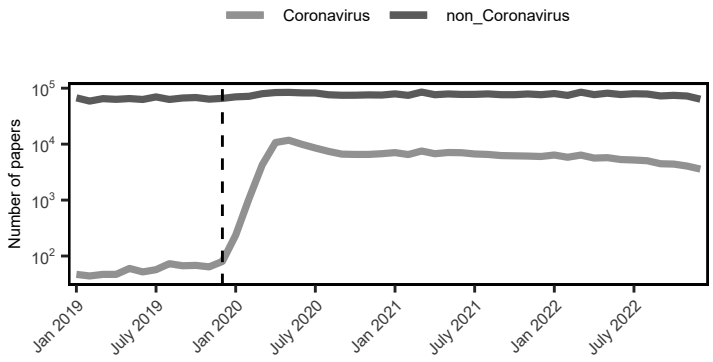
[11]In cases where the author had two affiliations we took only the first one

[12]In cases where the funder is an international agency, the country is labeled as "International"

from December 2022.

Our bibliometric sample (2019-2022) consists of 3,786,964 papers. We distinguish CRR from non-CRR, and pre-COVID-19 period (Jan.–Dec.2019) from COVID-19 period (Jan.–Dec.2022). This yields four categories: 779,787 non-CRR, pre-COVID-19 papers, 706 CRR, pre-COVID-19 papers, 2,784,382 non-CRR, COVID-19 papers, and 222,089 CRR, COVID-19 papers.

Figure 1.1: Coronavirus and non-coronavirus papers by submission month (log-scale).



Research output in CRR is extremely dynamic, particularly in spring 2020, while it is relatively stable in non-CRR over the analysis period (see Fig 1.1). In the pre-COVID-19 period, CRR output is relatively stable, at roughly 50 papers per month. Starting with the January 2020 outbreak, CRR grows exponentially to about 11,700 submissions in May 2020 and then decreased again to about 5,000 submissions in December 2020. Non-CRR output is stable throughout, at about 67,000 papers, and even increases slightly with the pandemic

The distribution of national health science output is highly skewed, with a Gini coefficient of around 0.9 before and during the pandemic in non-CRR and CRR.[13] The ten most prolific countries in CRR generate 65 percent of output during COVID-19. In the order of CRR output the top ten countries are (CRR papers in 2019; CRR papers between 2020 and 2022): USA (217; 61,942), China (193; 26,322), Italy (17; 18406), UK (54; 17,354), India (11; 14,445), Spain (14; 8,851), Canada (29; 9,190),

[13]Gini coefficients are for CRR pre-pandemic 0.90, CRR pandemic 0.87, non-CRR pre-pandemic 0.88, non-CRR pandemic 0.88.

Germany (35; 9,675), France (20; 6,797), Australia (25; 7463). All these countries contributed considerably to CRR during the pandemic, but not consistently to CRR before the pandemic. Some scholars explained the (early) high CRR output of China and Italy by the global infection process. On the other hand, rankings based on CRR during the pandemic correspond surprisingly well to rankings based on overall scientific production in the last two decades (see Table 1.1); supporting the idea of the relevance of slowly-accumulating science capacity discussed in the 'Background' section.

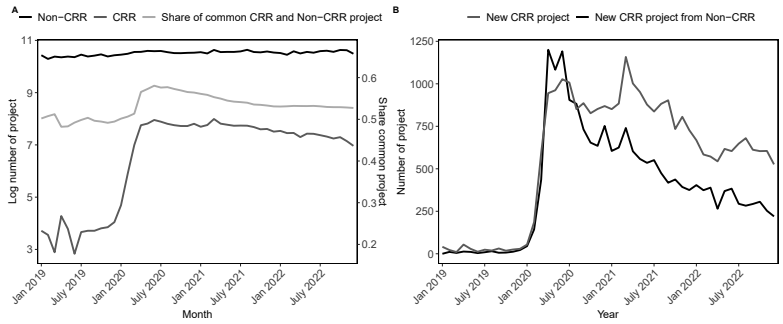
National scientific output is closely related to IRC in our sample. Around one-fourth of the papers in our sample are internationally co-authored. This holds true for CRR as well as non-CRR papers, before as well as during the pandemic (see Appendix 1.5.1, Table 1.7). Thus, CRR papers during the pandemic are, on average, as international as other papers.

Looking at the country level, our data confirms a strong positive relationship between national scientific output and national IRC intensity. In a simple OLS regression of the number of internationally co-authored papers on the total number of papers (both in logs), we obtain an elasticity of 0.9 for health science papers in 2019, health science papers in 2020, and also for CRR papers in 2020. This exceeds well the estimate of the elasticity of 0.7 by [Davidson Frame and Carpenter, 1979, Luukkonen et al., 1992], but seems reasonable if one takes into account the positive trends in IRC observed by Adams [2013]. Elasticity for CRR papers in 2019 is slightly lower; around 0.8. Thus IRC may have become somewhat more relevant to the national output of CRR during the pandemic, but neither more nor less as expected on IRC in non-CRR.

Figure 1.2 illustrates the progression of the count of unique grant IDs associated with CRR and non-CRR papers. As depicted in Figure 1.2A, the log count of distinct projects for CRR papers follows a similar pattern to Figure 1.1, experiencing an exponential increase from January 2020 until April 2020. The yellow line indicates the proportion of grant IDs present in non-CRR and CRR papers during that month or prior. The notable jump suggests an increased usage of already existing projects in the Global Health science system, aimed at addressing COVID-19 challenges. This assumption gains further support from Figure 1.2B. The surge in the blue line indicates a significant rise in new projects dedicated to COVID-19. Meanwhile, existing projects from broader global health were used towards CRR. While fresh projects were initiated, in the initial months, CRR papers heavily relied

on pre-existing health science projects until July. We interpret the black line as a reflection of researchers' adaptability to the ongoing pandemic shock, while the blue line represents the institutions' flexibility.

Figure 1.2: Project Dynamics in CRR and Non-CRR Research: Tracking Uniqueness and Cross-Utilization



Notes: Panel A displays the log count of unique projects within a specific month for both CRR and Non-CRR research. The yellow line, linked to the second y-axis, depicts the proportion of CRR projects utilized in past Non-CRR research over the entirety of CRR projects. In Panel B, the blue line illustrates new projects that have not been observed before in both CRR and non-CRR research. Meanwhile, the black line represents projects that have not been seen in CRR before but have previously appeared in non-CRR research.

We create international science networks for CRR as well as non-CRR research for each month, from January 2019 to December 2022. Networks are weighted following a full-count assignment, i.e. we increment the weight of the edge between two countries for each paper where both countries appear in the affiliations. Networks are accumulated over time by adding up all edge weights from the beginning of the analysis, January 2019, up to the focal month. Table 1.2 provides basic statistics on the accumulating networks.

In total, we identified 205 countries in our papers' affiliations. We aggregate the 2019 period to indicate international scientific activity before the pandemic. At the end of 2019, the accumulated CRR network included 65 countries, with about 500 international co-author ties (edge weights). The accumulated non-CRR network includes in Dec. 2019 nearly all countries, 201, connected through 490k ties. The non-CRR network was stable in 2020. Its decreasing growth rate of IRC papers (Weight % growth) is due to the addition of a relatively constant number of around 50k IRC papers each month.

Looking at CRR network growth in Table 1.2, one may distinguish four phases during which the CRR network expands in 2020. The first month, January 2020, may be considered the first phase in which the CRR network grows mostly in terms

of joint papers (edge weights) and less in terms of collaborating countries entering the network (nodes). The second phase, from February to April 2020, is characterized by high growth in terms of both network entrants, around 30 percent growth rate, and joint collaborations, 119 to 278 percent. In the third phase, from May to July 2020, network entry slows down considerably. The growth rate of joint collaborations is very high at the beginning of the phase but slows down from 60 to 22 percent during this period. In the fourth phase, from August to December 2020, few latecomers entered the network, and collaboration growth rates stabilized at around ten percent growth per month.

The CRR growth rate continues to slow down in 2021 and seems to converge to the non-CRR growth rate in 2022 (see Appendix 1.5.4, Table 1.9)

Table 1.2: International science networks, CRR and non-CRR, accumulated over months.

CRR network (accumulated), 2019												
	Jan.	**Feb.**	**Mar.**	**Apr.**	**May**	**June**	**July**	**Aug.**	**Sept.**	**Oct.**	**Nov.**	**Dec.**
Countries	18	26	30	34	40	44	53	53	57	59	64	65
Country % growth		44	15	13	18	10	20	0	8	4	8	2
Edge weights	23	38	74	94	120	142	235	264	300	368	415	481
Weight % growth		65	95	27	28	18	65	12	14	23	13	16
non-CRR network (accumulated), 2019												
	Jan.	**Feb.**	**Mar.**	**Apr.**	**May**	**June**	**July**	**Aug.**	**Sept.**	**Oct.**	**Nov.**	**Dec.**
Countries	186	193	193	197	198	200	201	201	201	201	201	201
Country % growth		3	0	2	0	1	0	0	0	0	0	0
Edge weights	40.7k	77.6k	120k	157k	199k	238k	283k	322k	365k	411k	451k	490k
Weight % growth		91	54	31	27	20	19	14	13	13	10	9
CRR network (accumulated), 2020												
	Jan.	**Feb.**	**Mar.**	**Apr.**	**May**	**June**	**July**	**Aug.**	**Sept.**	**Oct.**	**Nov.**	**Dec.**
Countries	69	96	116	150	166	171	175	178	181	182	183	183
Country % growth	6	39	21	29	11	3	2	2	2	1	1	0
Edge weights	641	1.4k	5.3k	16.8k	26.9k	36.9k	44.9k	50.5k	55.9k	62.3k	69.4k	77.4k
Weight % growth	33	119	278	215	60	37	22	12	11	11	11	12
non-CRR network (accumulated), 2020												
	Jan.	**Feb.**	**Mar.**	**Apr.**	**May**	**June**	**July**	**Aug.**	**Sept.**	**Oct.**	**Nov.**	**Dec.**
Countries	201	201	201	201	202	202	202	202	202	202	202	202
Country % growth	0	0	0	0	0	0	0	0	0	0	0	0
Edge weights	534k	578k	627k	677k	730k	784k	836k	882k	928k	979k	1028k	1078k
Weight % growth	9	8	8	8	8	7	7	6	5	5	5	5

The first two analyses, Analysis 1 and Analysis 2, aim to shed some light on the role of different factors in driving national and international CRR output throughout the pandemic. Contextual factors are obtained from several supplementary datasets: The national situation during the pandemic is captured through COVID-19 incidences and governmental measures. The COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University provides COVID-19 cases and deaths [Dong et al., 2020].[14] Social restrictions and international travel restrictions are obtained from Hale et al. [2021].[15] National economic data is from the 'Penn World Table' [Feenstra et al., 2015]. Finally, national socio-economic development is proxied by the human development index obtained from the Human Development Report Office of the United Nations (see hdr.undp.org). Not all countries observed in the publication data are included in the supplementary datasets. Due to missing observations, the number of countries was reduced from 205 to 156. All the countries dropped from the sample are small, except for Taiwan which is dropped because the UN does not provide statistics for Taiwan separately from China.

The third analysis, Analysis 3, focuses on how the development of global CRR during the pandemic relates to global non-CRR. That analysis is based only on publication data and therefore includes all 205 countries.

1.2.3 Analysis 1: National scientific output

1.2.3.1 Variables

CRR. Our dependent variable is the number of CRR papers accumulated from January 2020 to a given month t', which we shorthand by $(c_{i,t'})$. Accumulating output over months seems reasonable because we are ultimately interested in the dynamics of scientific activity — not submissions. For example, a paper submitted in April 2020 may well rely on research conducted from January 2020 to March 2020 and may have been influenced by events during that period (e.g. by the infection process for which we control; see below). Alternatively, one may consider the outcome variable as a proxy for the formation of national scientific capacity in CRR during the pandemic.

[14]The data has been downloaded from https://github.com/owid/covid-19-data/ in June 2023.
[15]We use the two data files 'international-travel-covid.csv' and 'stay-at-home-covid.csv'

Initial scientific capacity. Initial scientific capacity is captured by the accumulated number of CRR (non-CRR) papers up to December 2019, denoted $c_{i,t0}$ ($n_{i,t0}$).

Pandemic context. The accumulated number of confirmed deaths associated with the coronavirus, $deaths_{i,t'-1}$, may be interpreted as a proxy for research needs and opportunities. The number of hospitalized cases would be closer to what we have in mind but is a less reliable statistic. Pandemic policy restrictions may also be constraining but also inciting research. We include the accumulated number of days with a requirement not to leave the house (with minimal exceptions), $lock\text{-}down_{i,t'-1}$, and the number of days with total border closure, $border\text{-}closed_{i,t'-1}$.

Socio-economic context. GDP per capita in 2019, $gdp_{i,t0}$, (expenditure-based PPP in 2017 US Dollars), and the human development index in 2019, $hdi_{i,t0}$, serve to control for country size, economic wealth, and state of development respectively.

International Dependence. The accumulated number of CRR and non-CRR funded by an International agency or/and other country but not by country i and normalized by the number of papers of country i that have at least one grant, $IntDep_{i,t0}$.

We take logs (i.e. $\log(x+1)$) of all variables except for hdi and International Dependence. The pragmatic argument is that our variables are highly skewed to the right. Table 1.3 provides basic descriptive statistics of the variables used in the analysis on national scientific output.

Table 1.3: Descriptive statistics (Analysis 1)

	mean	sd	min	q1	median	q3	max	obs	NAs
$border_{t'}$	3.675	2.142	0	2.565	4.644	5.112	6.830	5,616	156
$c_{i,t'}$	4.434	2.434	0	2.708	4.304	6.286	11.034	5,616	0
$c_{i,t0}$	0.795	1.153	0	0	0	1.171	5.384	5,616	0
$deaths_{i,t'}$	4.437	2.641	0	2.418	4.864	6.821	8.806	5,616	156
$IntDep_{i,t0}$	0.548	0.230	0	0.371	0.540	0.717	1	5,616	0
$gdp_{i,t0}$	9.404	1.227	5.531	8.514	9.488	10.412	11.635	5,616	0
$hdi_{i,t0}$	0.731	0.152	0.394	0.609	0.749	0.851	0.957	5,616	0
$locked_{i,t'}$	4.143	2.061	0	3.611	4.820	5.629	6.898	5,616	156
$n_{i,t0}$	6.311	2.306	1.386	4.595	6.100	8.120	12.216	5,616	0

1.2.3.2 Empirical model

The estimating equation is the following linear model:

$$\begin{aligned} c_{i,t'} =& \beta_{0,t'} + \beta_{1,t'}\, n_{i,t0} + \beta_{2,t'}\, c_{i,t0} + \\ & \beta_{3,t'}\, deaths_{i,t'} + \beta_{4,t'}\, lock\text{-}down_{i,t'} + \beta_{5,t'}\, border\text{-}closed_{i,t'} + \\ & \beta_{6,t'}\, IntDep_{i,t0} + \beta_{7,t'}\, gdp_{i,t0} + \beta_{8,t'}\, hdi_{i,t0} + \epsilon_{i,t'} \end{aligned}$$

where the $\epsilon_{i,t'}$ denotes the error term. The model is estimated with simple OLS separately for each month t' in 2020. This allows for varying coefficients over time to uncover dynamics. Time effects common to all countries, for example, augmenting the number of papers on open-source platforms, are then naturally accounted for.

1.2.4 Analysis 2: International research collaboration

1.2.4.1 Variables

Analysis 2 focuses on drivers of bi-national collaboration on CRR. Some variables describe a single country i (or j) that is part of the dyad, others characterize the dyad ij. Variables measured before the shock form the initial conditions and are indicated by a time subscript $t0$. Variables that vary over the period of the pandemic are all accumulated over the pandemic (indicated by t'). The reason, again, is that a mapping from research context to research output within each month is less likely than a mapping of past and current research context to past and current research output.

Joint CRR. The dependent variable is the number of CRR papers signed by two countries i and j between January 2020 up to a given month t', denoted $c_{ij,t'}$.

Initial scientific capacity. The number of joint publications in 2019 between both countries on CRR ($c_{ij,t0}$) and non-CRR ($n_{ij,t0}$) are part of the initial conditions. For each country, say i, we indicate scientific capacity by the total number of CRR papers ($c_{i,t0}$) and non-CRR papers ($n_{i,t0}$). In the link formation context, these measures may be interpreted as factors that help to attract or initiate collaborations. In addition, the combination of health science competence is indicated by the sum ($n_{i,t0} + n_{j,t0}$) and absolute difference ($|n_{i,t0} - n_{j,t0}|$). These two factors indicate whether pooled health science capacity in the dyad is high ($n_{i,t0} + n_{j,t0}$) and to what extent scientific capacity is equally distributed among the partners ($|n_{i,t0} - n_{j,t0}|$).

Pandemic context. The relative pandemic situation of both countries is indicated by taking into account the accumulated number of COVID-19-related deaths ($death_{i,t'-1}$), number of days under lock-down ($locked_{i,t'-1}$), and number of days with closed borders ($closed_{i,t'-1}$). Lock-down and closed borders of a country may hinder the initiation or maintenance of international collaboration. Therefore, we keep these variables as individual factors. The number of deaths approximates the number of COVID-19 cases in the country on which research may be conducted. Therefore, we sum cases over the dyad ($death_i + death_j$), and take into account their absolute difference $|death_i - death_j|$. If for example the sum is positively correlated with joint CRR, while the absolute difference is negatively correlated with CRR, then countries

with few cases would interact with countries having many cases. Potential explanations for such an observation could be international solidarity but also resource interdependence.

Socio-economic context. Economic wealth and development are indicated by gdp per capita ($gdp_{i,t0}$) and the Human Development Index ($hdi_{i,t0}$) respectively. The discussion on core-periphery processes creates an interest in understanding to what extent developed (developing) countries interact among and with each other. Again, the sum and absolute difference of the partners' characteristics provide some indication; leading to the four composite variables $gdp_i + gdp_j$, $|gdp_i - gdp_j|$, $hdi_i + hdi_j$, $|hdi_i - hdi_j|$.

Geographical space. The literature is clear in that there is a geographical bias in IRC link formation (see Section Background above). Therefore we control for the geographic distance between two countries, $distance_{ij}$, measured by the distance (in km) of the countries' geographic centers. Another control $same\_region_{ij}$ indicates whether two countries are located in the same world region, i.e. continent as defined by the World Bank Development Indicators.

Country dependence. We calculate the normalized absolute difference between the number of papers co-authored by both i and j, which are exclusively funded by i (and other countries/institutions but not j) and exclusively funded by j (and other countries/institutions but not i) in 2019. This difference is divided by the total number of papers in 2019 coauthored by i and j with at least one grant, $Dep_{ij,t0}$. [16]

We take logs (i.e. $\log(x + 1)$) of all independent variables except for hdi and Country dependence.

Table 1.4 provides basic descriptive statistics of the variables used in the analysis of international scientific output.

[16] One limitation is that we do not differentiate between equal partnership and its absence.

Table 1.4: Descriptive statistics (Analysis 2)

	mean	sd	min	q1	median	q3	max	obs	NAs
$closed_{i,t'}$	5.858	4.759	0	0	8.603	9.946	13.657	876,096	0
$c_{ij,t'}$	13.146	252.097	0	0	0	0	47,259	876,096	0
$c_{i,t0}$	0.613	1.447	0	0	0	0	10.760	876,096	0
$c_{ij,t0}$	0.023	0.176	0	0	0	0	5.204	876,096	0
$Dep_{ij,t0}$	0.015	0.093	0	0	0	0	1	876,096	0
$\lvert gdp_i - gdp_j \rvert$	9.388	1.530	0	8.722	9.686	10.444	11.632	876,096	0
$gdp_i + gdp_j$	10.396	0.849	6.222	9.831	10.556	11.018	12.328	876,096	0
$Distance$	8.613	1.030	0	8.276	8.820	9.218	9.898	876,096	0
$\lvert hdi_i - hdi_j \rvert$	0.173	0.127	0	0.068	0.148	0.257	0.563	876,096	0
$hdi_i + hdi_j$	1.461	0.215	0.788	1.313	1.472	1.628	1.914	876,096	0
$locked_{i,t'}$	7.161	4.665	0	0	9.222	10.755	13.793	876,096	0
$n_{i,t0}$	12.596	3.285	2.303	10.205	12.489	14.854	24.432	876,096	0
$n_{ij,t0}$	1.427	1.678	0	0	1.099	2.303	11.965	876,096	0
$Sameregion$	0.247	0.431	0	0	0	0	1	876,096	0
$\lvert death_i - death_j \rvert$	7.549	3.480	0	5.784	8.321	9.947	13.885	876,096	0
$death_i + death_j$	8.047	3.412	0	6.621	8.857	10.335	14.578	876,096	0

1.2.4.2 Empirical model

The empirical model is essentially a gravity model, where the interaction intensity between two countries (here the number of joint CRR papers) depends on the countries' distance and further relational factors. The model is common in spatial scientometrics [Frenken et al., 2009] and has been applied in the same form by Hoekman et al. [2009].

In detail, we estimate a zero-inflated negative binomial model in order to take into account the fact that many countries do not have any joint CRR paper. The

model combines the negative binomial count density with a binary process[17] to model excess zeros in the outcome [see e.g. Cameron and Trivedi, 2005, p.681].[18]

The first part of the model, termed the zero-model, models the binary process of two countries i and j having *no* joint CRR papers. The linear predictor of the zero-model includes the typical ingredients of a gravity model:

$$\begin{aligned} p = Pr\left[c_{ij,t'} = 0|\boldsymbol{\gamma}\right] = \text{logit}(&\gamma_0 + \gamma_1 \log(n_{i,t0} + n_{j,t0}) + \\ &\gamma_2 \log(|n_{i,t0} - n_{j,t0}|) + \\ &\gamma_3 \log(|c_{i,t0} - c_{j,t0}|) + \gamma_4 \log(|c_{i,t0} - c_{j,t0}|) + \\ &\gamma_5 \log(distance_{ij}) + \gamma_6 same\_region_{ij}) \ , \end{aligned}$$

where $(n_{i,t0} + n_{j,t0})$ captures the joint weight of the two countries in 'health sciences', and $(|n_{i,t0} - n_{j,t0}|)$ their absolute difference. Geographical proximity is captured by $distance_{ij}$ and the $same\ region_{ij}$ dummy. With probability $(1 - p)$ the count density applies with expectation

$$\begin{aligned} E\left[c_{ij,t'}|\boldsymbol{\beta}\right] = \exp(\beta_0)\, c_{ij,t0}^{\beta_1}\, n_{ij,t0}^{\beta_2}\, (c_{i,t0}c_{j,t0})^{\beta_3}\, (n_{i,t0}n_{j,t0})^{\beta_4} \times \\ (d_{i,t'-1} + d_{j,t'-1})^{\beta_5}\ (|d_{i,t'-1} - d_{j,t'-1}|)^{\beta_6} \times \\ (gdp_{i,t0} + gdp_{j,t0})^{\beta_7}\ \left(|gdp_{i,t0} - gdp_{j,t0}|\right)^{\beta_8} \times \\ locked_{i,t'-1}^{\beta_9}\, closed_{i,t'-1}^{\beta_{10}}\, distance_{ij}^{\beta_{11}}\ \exp(\beta_{12} same\_region_{ij}) \times \\ \exp(\beta_{13}(hdi_{i,t0} + hdi_{j,t0}))\ \exp(\beta_{14}\,(|hdi_{i,t0} - hdi_{j,t0}|))\ \exp(\beta_{15} Dep_{ij,t0})\ , \end{aligned}$$

where $\exp(\beta_0)$ is a scaling factor (the intercept). Individual variables are introduced above but some notes on the model structure are in order. First, the model allows for a lasting impact of an established relation in non-CRR and CRR through $(n_{ij,t0},\ c_{ij,t0})$. Second, countries may 'attract' IRC through their research competencies $(n_{i,t0}, c_{i,t0})$. Because this effect is symmetric (the relationship ij is the same as ji), we enforce the same coefficient for i and j. Control variables are entered such that they capture the joint 'mass' of the partners as well as their absolute differences. For example scientific interaction may be driven by countries having many

[17] We chose a logit model.

[18] The choice of the negative binomial density over a Poisson density is supported by a strong and significant estimate of the variance related parameter.

COVID-19 cases in sum ($d_{i,t'-1} + d_{j,t'-1}$), but also if one country has many more cases than the other ($|d_{i,t'-1} - d_{j,t'-1}|$). The former can be interpreted as a common pool of resources (or incentives), and the second captures dyadic interdependence due to inequality in resources.[19] Third, we introduce geographic distance and same region indicators in the count model because geographic proximity is susceptible to driving whether and how much research is conducted jointly.

The zero model is kept relatively light, in the form of a basic gravity model, for the following reasoning. First, one may interpret the zero model as the potential to collaborate in the health sciences as such. This is mostly an issue of mutual (or one-sided) awareness driven by a combination of global visibility and geographic distance. Given awareness, the intensity of joint CRR may then be determined by various factors capturing the needs and benefits of collaboration. There is also a more pragmatic argument. The two-step process is an artificial interpretation of the model. In fact, we deal here with one convolution of densities determined by all the factors we consider at once. In this sense, separating factors into different parts helps clarify individual factors' overall contribution to the outcome.

The link formation model is estimated on dyadic data which by construction may result in network correlation of errors. In general, correlated errors maintain unbiasedness and consistency of coefficient estimates, but create a downward bias of the estimated standard deviations of coefficient estimates. The reason is essentially that network dependence of observations reduces the information content compared to independent observations (which is assumed by standard estimators). Most of the dependence across dyad observations can be expected from repeated observations for the same individual countries. The Multiple Regression Quadratic Assignment Procedure (MRQAP) therefore creates a Null distribution through random permutation of rows and columns of the adjacency matrix (in effect a random re-labeling of nodes in the network). In principle, there are different ways to implement the idea of MRQAP (Decker et al., 2007). In our case, the preferred choice is to permute one right-hand-side factor, keeping everything else fixed, to generate the distribution of the z-statistic (coefficient estimate divided by its standard deviation) under the Null hypothesis that the right-hand-side factor is not systematically related to the outcome controlling for all other factors.[20] In the results, we report the estimated

[19] We tried various ways of introducing country-specific factors into the model and found this formulation to be the most compelling (in terms of fit and reasoning).

[20] There are pros and cons to the different ways of implementing MRQAP. The simplest way would be to permute the outcome but that has the disadvantage to create a distribution under

z-value together with significance levels of one-sided tests based on the Null distribution of z-values from permutation.

1.2.5 Analysis 3: Convergence to global health science

The COVID-19 shock changed dramatically the global distribution of CRR. Because the global health science system responded in a very short time, CRR during the pandemic must have built essentially on resources developed before the outbreak. Therefore it seems reasonable that the global structure of CRR during the pandemic converges to the pre-pandemic global structure of health sciences. We investigate this hypothesis by looking at i) national scientific output, ii) countries' network centrality in the IRC network, and iii) IRC network structure.

1.2.5.1 National scientific output

For each month in our period, we create a ranking of countries based on CRR output ($c_{i,t}$) and non-CRR output ($n_{i,t}$) respectively. We speak of convergence in national scientific output when the two rankings become more similar over time. Similarity is captured by the rank correlation coefficient τ_X, as described in Emond and Mason [2002]. This statistic is similar to Kendall's τ, except that it handles ties the same as dominant relationships (entering 1 and not 0 in the dominance matrix). In principle, this is favorable in case of many ties in the rankings, as we have in Corona pre-COVID-19 research; but does not affect the results. A 90 percent confidence interval around τ_X is then obtained by a traditional jackknife, or leave-one-out, approach as described for example in Abdi [2013].

1.2.5.2 Network centrality

The international science network is highly hierarchical, and network centrality captures the position of a country within that hierarchy. Therefore, we ask 'How does

the null that no factor is relevant. Furthermore, our (non-linear) model fails to converge if the model does not fit the data, which is almost always the case when the outcome is permuted. Permutation of error terms after partialing is a common alternative but is only valid under relatively strict assumptions on the error term which are unlikely to be met in our model as it is a convolution of two different distributions. Our preferred alternative of permuting individual right-hand side factors has the disadvantage of breaking existing correlations with other right-hand side factors. The effect seems however to be bearable if a pivotal statistic is used, as we do [Dekker et al., 2007].

network centrality of countries in the CRR network align with their centrality in the overall health science network?'

The international science network can be safely said to exhibit a core-periphery structure. Therefore, network centrality is appropriately captured through s-core decomposition [Eidsaa and Almaas, 2013]. The s-core decomposition identifies sets of nodes that are heavily connected to each other. Together they form the network core(s). Roughly, the algorithm starts from the complete network and proceeds by iterative removal of the least connected node in the remaining network. The s-core is the strength of the node in the remaining network at the time of removal. To compare different networks over time, we normalize the s-core by the maximum s-core in the network. The s-core ranges from 0 for isolates in the network, to 1 for highest core members. S-cores are measured for each country every month on the CRR network and the non-CRR network respectively. As the difference between the s-cores decreases, hierarchies converge.

1.2.5.3 Network structure

Convergence in network structure is first investigated by correlating the CRR adjacency matrix with the non-CRR adjacency matrix every month. The weights in the adjacency matrix are in logs because they are highly skewed — many countries do not collaborate (zero entry) and some do heavily (many joint papers). Statistical significance of correlations is obtained through the Quadratic Assignment Procedure (QAP). The QAP test creates a null distribution through the re-labeling of nodes in one network; maintaining the structure of the networks. The correlation analysis tells us whether the CRR and non-CRR networks become more similar over time.

In a second step, we investigate in which aspects the CRR network differs from the non-CRR network. The background section highlights network hierarchy and communities as two salient features in IRC networks. This is what we focus on.

Hierarchy is captured by the largest absolute eigenvalue of the adjacency matrix, say λ. More hierarchical networks tend to have a higher largest eigenvalue. It has been shown that the largest eigenvalue is maximal for nested-split graphs. A nested-split graph is a specific type of hierarchical network, in which the most central node connects to all other nodes, and less central nodes connect to subsets of alters of more central nodes. Interestingly, nested-split graphs emerge in network games where payoffs are strategic complements in effort levels [König et al., 2014], which is a reasonable assumption for science networks.

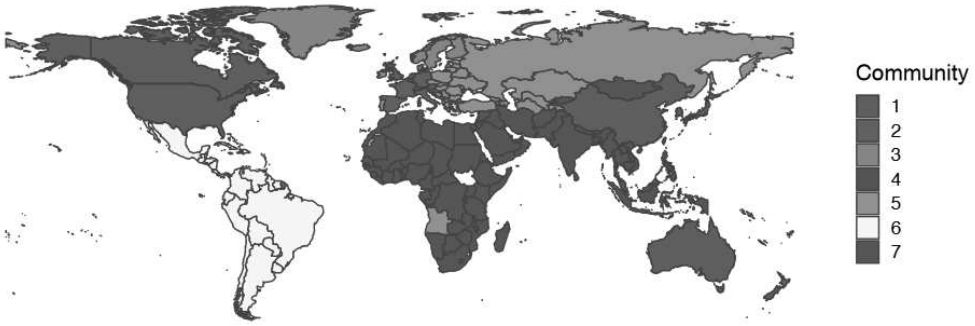
We measure λ on the accumulating CRR network for each month from January 2019 to December 2020. The value of λ in itself is not very telling as it depends on various features of the network; most importantly network size. For interpretation, it is, therefore, useful to compare this statistic to a network null model to see whether the network of interest is more or less hierarchical as the null network. Most studies create a null distribution of networks by fixing some aspects of the focal network, e.g. network size and degree distribution, and randomize the rest. Our interest is in how the CRR network structurally deviates from the non-CRR network. Therefore we use the non-CRR network formed in 2019 to generate our null distribution. In detail, one realization s is obtained by drawing links uniformly at random with replacement from the non-CRR network until a network is created of the same size (same number of links, and hence same total strength) as the CRR network. Each realization yields a statistic λ_s. We draw 100 realizations to obtain the null distribution for our statistics. Based on the null distribution, we calculate the z-value for the largest eigenvalue λ, as $z_\lambda = \frac{\lambda_{crr} - \hat{E}[\lambda_s]}{\widehat{sd}(\lambda_s)}$. A statistic z_λ above (below) zero indicates that the CRR network is more (less) hierarchical than would be expected by the structure of the prior non-CRR network.

Network communities are commonly thought of as subsets of nodes with relatively strong interaction. IRC networks feature communities due to scientists' tendency to collaborate with other scientists that are somewhat close in space; with space broadly defined as scientific, geographical, cultural, etc. (see Section Background). The hypothesis we wish to test is that during the pandemic the accumulating CRR network converges towards the same community structure as the prior non-CRR network. The outline of our empirical approach is as follows: In a first step, we detect communities in the prior non-CRR network. This becomes our reference community structure, say benchmark. Then, for each month in the observation period, we measure how well the accumulating CRR network 'fits' that benchmark. As in the hierarchy analysis, we account for varying network sizes over time and across networks by creating a null distribution through resampling from the non-CRR network. We discuss the details in the following.

Community detection in the prior non-CRR network follows closely the procedure of Fitzgerald et al. [2021]. The main idea is to find a network partition that maximizes some network modularity statistic, say Q. Newman [2004b] proposed a statistic that measures to what extent nodes belonging to the same community form

ties beyond what would be expected based on their link strength alone.[21] Reichardt and Bornholdt [2006] proposed multiple generalizations known as the spin-glass algorithm. Among others, they introduce a parameter γ for tuning the resolution of the network partition. A smaller (higher) γ tends to yield less (more) clusters. For $\gamma = 1$ the objective function coincides with Newman's network modularity measure [Newman, 2004b]. We use that extension. The optimization is done through simulated annealing (as implemented in the R-package 'igraph'). Simulated annealing is a stochastic optimization algorithm and hence may provide different partitions for the same data and parameter settings. This is actually useful, because the robustness of communities for a given parameter γ across multiple optimizations signals whether communities are well identified. We search a robust and informative community partitioning through a grid search on γ. Appendix 1.5.3 details how.

Figure 1.3: Communities in 2019



Notes: Communities in the 2019 non-CRR network from spin-glass algorithm with tuning parameter $\gamma = 1.2$; the reference community structure for Analysis 3.3.

Figure 1.3 displays a partition of the 2019 non-CRR network for $\gamma = 1.2$. It is remarkable how similar that partitioning is to the partitioning obtained by Fitzgerald et al. [2021, Fig.5b] based on all Scopus publications in 2015. The map displays seven communities with numbers ordered by the average network strength of their members (Appendix 1.5.3 lists the countries belonging to each community). Community 1 is a global community that includes most importantly the US, China, Great Britain, and Japan. All other communities cover world regions. Roughly, Community 2 covers Central Europe, Community 3 Northern Europe, Community 4 includes only Israel,

[21] In some sense, the null model here is a simple gravity model based on only node size. In the prior IRC literature, a similar normalization has been applied by calculating Salton's measure, i.e. observed strength of interaction between two nodes divided by the product of the nodes network strength. Salton's measure revealed in particular country preferences of interaction, revealing the role of distance in IRC.

Jersey, and Montserrat, Community 5 spans East Europe and Russia, Community 6 corresponds to South America, and Community 7 covers Africa and the Middle East.

This partition is used in the analysis as the benchmark for the accumulating CRR network. For each month in the observation period, we calculate the modularity statistic on the accumulating CRR network (Q_{crr}) on the benchmark partitioning obtained in the previous step on the non-CRR network. In addition, multiple resamples are obtained from the non-CRR network such that each resample s has the same size (w.r.t. total edge weights) as the current CRR network. As for the CRR network, we calculate for each resample of the non-CRR network a modularity statistic, Q_s. This provides us a z-value, i.e. $z_Q = \frac{Q_{crr} - \hat{E}[Q_s]}{\widehat{sd}(Q_s)}$. If the z-value is high (low) we can say that the CRR network fits well (badly) the community structure of the 2019 non-CRR network.

1.3 Results

1.3.1 National scientific output

Tables 1.5, 1.10, and 1.11 present the regression analyses concerning the cumulative count of national CRR papers for each month in 2020, 2021, and 2022 respectively.[22] The correlation between the pre-pandemic CRR (c_{t0}) and our dependent variables is notably positive and significant, while the pre-pandemic non-CRR (n_{t0}) shows no effect. However, this trend undergoes a reversal within the initial pandemic months as the CRR gains traction. During this phase, the broader health science capacity (n_{t0}) emerges as the predominant factor influencing CRR output. The significance of CRR-specific experience gradually wanes, giving way to the prominence of pre-pandemic non-CRR (n_{t0}) and the count of COVID-19-related deaths ($deaths_{t'}$). Remarkably, these patterns persist through 2021 and 2022, but a negative aspect arises in terms of international dependence on accumulated CRR output ($IntDep_{i,t0}$). While this variable exhibited a positive influence only in March 2020, potentially questioning the

[22]We conducted robustness checks by excluding both the US and China, and also by introducing and eliminating variables such as HDI and Population size. Our rationale for excluding China lies in the ongoing debate surrounding reported death numbers, as we aimed to prevent potential bias in our findings. Simultaneously, due to their substantial research output, the USA and China could introduce a bias towards nullifying the impact of variables. Although the detailed outcomes are not presented here to maintain clarity, it's important to note that the results remained stable throughout all the conducted tests, confirming the robustness of our findings.

robustness of the result, it consistently assumes a negative and significant role in 2021 and 2022. This phenomenon can be attributed to the initial urgency of the situation, during which pre-existing funding was redirected in response to the pressing demands, thereby mitigating the country's international dependence on output. Subsequently, in 2021 and 2022, as the initial funds that were shifted disappeared, larger countries reduced collaborative research funding with foreign actors. This shift implies that countries dependent on international funding exhibit lower output due to diminished opportunities for funding support once the system relied heavily on new COVID-19-specific projects. Other Socioeconomic context variables seem to have no effect on scientific output

Table 1.5: Accumulated number of CRR papers ($c_{t'}$) in 2020 (all variables in logs, standardized to zero mean and one std.dev., 156 countries).

	Jan. '20	**Feb. '20**	**Mar. '20.**	**Apr. '20**	**May '20**	**June '20**	**July '20**	**Aug. '20**	**Sept. '20**	**Oct. '20**	**Nov. '20**	**Dec. '20**
n_{t0}	0.021	0.172***	0.501***	0.713***	0.795***	0.860***	0.874***	0.902***	0.916***	0.925***	0.924***	0.935***
	(0.021)	(0.044)	(0.041)	(0.060)	(0.067)	(0.060)	(0.058)	(0.062)	(0.062)	(0.061)	(0.053)	(0.053)
c_{t0}	0.601***	0.725***	0.550***	0.284**	0.194	0.111	0.107	0.072	0.058	0.035	0.025	0.015
	(0.061)	(0.085)	(0.079)	(0.135)	(0.149)	(0.128)	(0.131)	(0.120)	(0.114)	(0.113)	(0.120)	(0.121)
$Int - Dep_{t0}$	−0.162	−0.075	0.068***	0.079	0.076	0.177	0.082	0.102	0.120	0.134	0.036	0.005
	(0.197)	(0.139)	(0.022)	(0.095)	(0.073)	(0.126)	(0.149)	(0.141)	(0.130)	(0.106)	(0.157)	(0.163)
$deaths_{t'}$				0.024	0.029	0.066*	0.073***	0.078***	0.079***	0.080***	0.062***	0.045**
				(0.021)	(0.024)	(0.034)	(0.021)	(0.009)	(0.008)	(0.010)	(0.011)	(0.019)
$border - closed_{t'}$				0.035	0.004	−0.003	0.002	0.005	0.006	−0.0001	−0.005	−0.003
				(0.042)	(0.023)	(0.019)	(0.021)	(0.022)	(0.025)	(0.024)	(0.020)	(0.019)
$lock - down_{t'}$				0.046*	0.070	0.072*	0.060	0.061	0.057	0.055	0.058	0.053
				(0.027)	(0.043)	(0.041)	(0.044)	(0.050)	(0.048)	(0.045)	(0.040)	(0.036)
gdp_{t0}	−0.054	0.003	0.026	0.155	0.094	0.133	0.154	0.202	0.226	0.229	0.214	0.220
	(0.051)	(0.150)	(0.191)	(0.238)	(0.226)	(0.261)	(0.275)	(0.270)	(0.271)	(0.272)	(0.240)	(0.253)
hdi_{t0}	−0.032	−0.830	−0.931	−1.357	−1.337	−2.026	−2.285	−2.616	−2.866	−2.825	−2.687	−2.717
	(0.088)	(1.212)	(1.599)	(2.015)	(1.898)	(1.999)	(2.180)	(2.151)	(2.154)	(2.162)	(1.961)	(2.131)
R^2	0.772	0.817	0.892	0.923	0.923	0.934	0.934	0.941	0.945	0.947	0.948	0.946

Notes: Recall from Section 1.2.3 that c_{t0} and n_{t0} proxy initial scientific capacity in CRR and non-CRR respectively. The main finding is that correlation of c_{t0} with CRR output starts off high but decreases as n_{t0} takes over the main explanatory power. There is some support that national pandemic severity ($deaths_{t'}$, $locked_{t'}$) are positively, and economic development (hdi_{t0}) is negatively associated with CRR by the second half of 2020.

1.3.2 International research collaboration

Estimation results presented in Tables 1.6, 1.12, and 1.13 provide information on the dynamics of joint CRR over the pandemic, taking into account various factors.[23] Regressions of the accumulated number of joint CRR papers start in January 2020 and are estimated separately for each month up to December 2022. The discussion below focuses on coefficient estimates.

The zero model (lower part of the table) estimates the probability of having *no* joint CRR. The sum of non-CRR papers in 2019 ($n_{i,t0} + n_{j,t0}$) is negative and highly significant, meaning that chances increase to have at least one joint CRR paper during our period. Absolute difference in non-CRR papers is positive and highly significant, inequality makes it less likely to work together. These 2 results can be explained by a reputation-based mechanism, wherein prominent countries tend to collaborate with other well-established and visible counterparts. The same result is found for CRR showing an IRC dependence on both global health capacity and Coronavirus related research. Geographic distance plays a significant role as such, but being in the same region is negatively related to having no joint CRR. Inter-regional collaboration is one lever used by countries in order to tackle the urgency of the pandemic. Looking at the count model of the intensity of dyadic CRR. Past joint non-CRR papers ($n_{ij,t0}$) are one the main factor that drives future CRR collaboration and is consistent across all of our period. Inversely, past joint CRR papers ($c_{ij,t0}$) had a negative and significant impact starting in April demonstrating a willingness of countries that participated in past CRR to either work with new collaborators or to share knowledge with new entrants. Once past dyadic interaction is taken into account, large health science capacity ($n_{i,t0}$) plays no role in the collaboration formation CRR. Although there seems to be a small positive effect of Coronavirus capacities ($c_{i,t0}$) at the end of 2020, this effect dies out really quickly. The positive effect of the sum of the number of deaths ($death_i + death_j$) can be explained by a mutual national incentive to do research on CRR while the, sometimes significant, negative effect of the absolute difference ($|death_i - death_j|$) shows collaboration between a heavily affected country and relatively not affected country meaning that other reason than national incentive plays a role. We posit that the observed positive impact linked to the cumulative number of days of lockdown ($locked_{i,t'}$), along with the lack of significance of border closure ($border_{i,t'}$), might indicate that domestic restrictions

[23]The same robustness check was done as for analysis 1 and the results were consistent across specification

can hinder in-person interactions with colleagues. Consequently, individuals might opt to collaborate remotely with international coauthors when presented with the choice. The negative impact of interdependence between two countries ($dep_{ij,t0}$) during the initial phases of the pandemic might indicate that collaborations are less probable when funding is disproportionately skewed in one direction. Future CRR collaborations are more likely to occur when there is a balanced partnership. Such circumstances could potentially have an adverse effect on research output if a nation is accustomed to being the partner with less funding originally. The initial positive impact of the absolute difference of GDP per capita ($|gdp_i - gdp_j|$) during the early months of the pandemic, which subsequently shifts to a positive effect of the sum of GDP per capita ($gdp_i + gdp_j$), might indicate an inclination of wealthier nations towards smaller and emerging economies and their research efforts. This preference could exist until larger countries adjust to the impact of the COVID-19 shock, at which point the "rich-get-richer" phenomenon becomes the prevailing pattern within this research ecosystem. Interestingly there is some sign that developing countries collaborate with each other (negative $hdi_i + hdi_j$), but not particularly with more advanced countries (negative $|hdi_i - hdi_j|$). Finally, Distance and Region seem to play no role in the intensity of IRC. Combined with the zero model result this could indicate that distance (outside of the regional aspect) plays a negative role in their first interaction but is then subsequently insignificant once the relation is already existing.

Table 1.6: Zero-inflated negative binomial model of (accumulated) joint coronavirus related papers ($c_{ij,t'}$) during the pandemic, est. coefficient (z-value, p-value)

	Jan. 20	Feb. 20	Mar. 20	Apr. 20	May 20	Jun. 20	Jul. 20	Aug. 20	Sep. 20	Oct. 20	Nov. 20	Dec. 20
Count-model												
$n_{ij,t0}$	1.066***	1.256***	0.871***	0.965***	1.042***	1.036***	1.051***	1.041***	1.038***	1.022***	1***	0.973***
	(5.875, 0)	(13.385, 0)	(19.551, 0)	(34.568, 0)	(39.285, 0)	(41.406, 0)	(41.993, 0)	(41.521, 0)	(41.765, 0)	(40.329, 0)	(39.669, 0)	(40.28, 0)
$c_{ij,t0}$	0.178	0.173	-0.07	-0.182*	-0.219***	-0.215***	-0.23***	-0.218***	-0.222***	-0.199***	-0.174***	-0.149*
	(1.133, 0.1)	(1.617, 0.1)	(-1.279, 0.333)	(-4.056, 0.033)	(-4.746, 0)	(-4.869, 0)	(-5.149, 0)	(-4.939, 0)	(-5.236, 0)	(-4.76, 0)	(-4.31, 0)	(-3.874, 0.033)
$n_{i,t0}$	-0.343***	-0.235	0.25***	0.104	0.053	0.017	0.015	0.01	0	-0.013	-0.025	-0.012
	(-2.079, 0)	(-2.928, 0.067)	(6.152, 0)	(4.371, 0.067)	(2.507, 0.133)	(0.885, 0.6)	(0.808, 0.733)	(0.525, 0.8)	(0.014, 1)	(-0.745, 0.667)	(-1.437, 0.4)	(-0.721, 0.6)
$c_{i,t0}$	0.236	0.036	-0.017	0.009	0.011	0.018	0.027	0.033	0.04*	0.042*	0.045*	0.037*
	(2.447, 0.067)	(0.883, 0.6)	(-0.91, 0.567)	(0.714, 0.767)	(0.976, 0.6)	(1.737, 0.4)	(2.661, 0.1)	(3.343, 0.1)	(4.257, 0.033)	(4.578, 0.033)	(5.032, 0.033)	(4.316, 0.033)
$death_i + death_j$		1.024	-0.053	0.032	0.015	0.045	0.06	0.047	0.088*	0.094	0.119***	0.13***
		(0.073, 0.967)	(-0.342, 0.833)	(1.056, 0.6)	(0.519, 0.7)	(1.795, 0.267)	(2.442, 0.2)	(2.209, 0.233)	(3.833, 0.033)	(4.342, 0.067)	(5.489, 0)	(6.207, 0)
$\lvert death_i - death_j \rvert$		-0.888	0.03	-0.041	-0.016	-0.023	-0.031	-0.016	-0.036	-0.03	-0.043	-0.056
		(-0.064, 1)	(0.196, 0.867)	(-1.673, 0.6)	(-0.674, 0.667)	(-1.165, 0.6)	(-1.563, 0.3)	(-0.975, 0.633)	(-1.972, 0.167)	(-1.802, 0.4)	(-2.549, 0.133)	(-3.486, 0.133)
$locked_{i,t'}$			0.046***	0.022	0.033***	0.033***	0.035***	0.031***	0.036***	0.036***	0.033***	0.03***
			(0.648, 0)	(3.266, 0.1)	(6.768, 0)	(8.107, 0)	(8.979, 0)	(7.68, 0)	(9.086, 0)	(9.72, 0)	(9.38, 0)	(8.975, 0)
$closed_{i,t'}$				-0.001	-0.001	-0.003	0.003	0.003	-0.001	-0.002	-0.006	-0.003
				(-0.133, 0.9)	(-0.121, 1)	(-0.68, 0.767)	(0.822, 0.7)	(0.737, 0.833)	(-0.302, 0.867)	(-0.655, 0.733)	(-1.739, 0.4)	(-1.124, 0.533)
$Dep_{ij,t0}$	-0.846*	-0.421	-0.669***	-0.599***	-0.499*	-0.567***	-0.607***	-0.548*	-0.589***	-0.633***	-0.65***	-0.619***
	(-2.017, 0.033)	(-1.587, 0.133)	(-5.384, 0)	(-6.364, 0)	(-4.978, 0.033)	(-6.112, 0)	(-6.712, 0)	(-6.243, 0.033)	(-6.972, 0)	(-7.653, 0)	(-8.003, 0)	(-7.938, 0)
$gdp_i + gdp_j$	-0.571	0.416	0.363	0.233	0.185	0.209	0.283	0.33***	0.347***	0.358***	0.388***	0.421***
	(-0.923, 0.407)	(1.512, 0.267)	(2.595, 0.133)	(2.611, 0.367)	(2.309, 0.167)	(2.852, 0.167)	(3.94, 0.067)	(4.639, 0)	(5.031, 0)	(5.319, 0)	(5.886, 0)	(6.621, 0)
$\lvert gdp_i - gdp_j \rvert$	0.023	0.119	0.053	0.067***	0.072***	0.066***	0.064***	0.06	0.056***	0.05	0.043	0.045***
	(0.191, 0.833)	(2.037, 0.1)	(2.106, 0.367)	(3.823, 0)	(4.324, 0)	(4.354, 0)	(4.29, 0)	(4.06, 0.1)	(3.927, 0)	(3.645, 0.067)	(3.186, 0.133)	(3.49, 0)
$hdi_i + hdi_j$	-1.662	-4.18*	-2.843***	-1.881***	-1.857***	-1.873***	-2.293***	-2.55***	-2.491***	-2.437***	-2.542***	-2.688***
	(-0.668, 0.6)	(-3.962, 0.033)	(-5.232, 0)	(-5.326, 0)	(-5.772, 0)	(-6.331, 0)	(-7.904, 0)	(-8.866, 0)	(-8.952, 0)	(-8.984, 0)	(-9.594, 0)	(-10.566, 0)
$\lvert hdi_i - hdi_j \rvert$	1.147	-1.514	-1.52	-2.166***	-1.918***	-1.957***	-2.047***	-2.11***	-1.952***	-1.825***	-1.771***	-1.825***
	(0.621, 0.6)	(-1.776, 0.267)	(-3.528, 0.067)	(-7.451, 0)	(-7.129, 0)	(-7.923, 0)	(-8.444, 0)	(-8.78, 0)	(-8.363, 0)	(-7.97, 0)	(-7.94, 0)	(-8.511, 0)
distance	1.21***	0.278***	0.009	0.067	0.081	0.077	0.069	0.068	0.043	0.042	0.046	0.059
	(4.136, 0)	(2.992, 0)	(0.217, 0.9)	(2.444, 0.3)	(3.222, 0.067)	(3.389, 0.067)	(3.111, 0.067)	(3.109, 0.133)	(2.038, 0.3)	(2.047, 0.233)	(2.288, 0.233)	(3.09, 0.067)
sameregion	1.298*	0.055	-0.033	-0.059	-0.051	-0.107*	-0.116*	-0.111*	-0.106*	-0.083*	-0.084	-0.065
	(2.908, 0.033)	(0.323, 0.7)	(-0.436, 0.633)	(-1.176, 0.3)	(-1.09, 0.333)	(-2.483, 0.033)	(-2.761, 0.033)	(-2.698, 0.033)	(-2.637, 0.033)	(-2.119, 0.033)	(-2.212, 0.067)	(-1.779, 0.133)
$log(\theta)$	10.678	0.819***	1.695***	1.718***	1.491***	1.544***	1.505***	1.527***	1.623***	1.671***	1.736***	1.819***
	(0.148, 0.882)	(3.568, 0)	(12.922, 0)	(19.936, 0)	(20.076, 0)	(21.726, 0)	(21.98, 0)	(22.085, 0)	(23.231, 0)	(23.356, 0)	(25.254, 0)	(27.879, 0)
Zero-model												
$n_{i,t0} + n_{j,t0}$	0.018	-1.94***	-5.765***	-4.861***	-4.667***	-4.339***	-4.505***	-4.412***	-4.332***	-4.02***	-3.671***	-3.512***
	(0.031, 1)	(-3.073, 0)	(-3.989, 0)	(-11.221, 0)	(-13.786, 0)	(-16.201, 0)	(-16.998, 0)	(-17.756, 0)	(-17.896, 0)	(-18.67, 0)	(-19.667, 0)	(-20.38, 0)
$\lvert n_{i,t0} - n_{j,t0} \rvert$	-0.198	0.062	3.301*	1.734***	1.796***	1.749***	1.853***	1.811***	1.794***	1.678***	1.498***	1.41***
	(-0.677, 0.533)	(0.188, 0.767)	(2.758, 0.034)	(5.74, 0)	(8.256, 0)	(10.721, 0)	(11.958, 0)	(12.971, 0)	(13.347, 0)	(14.052, 0)	(14.432, 0)	(14.745, 0)
$c_{i,t0} + c_{j,t0}$	-2.29*	0.13	-0.106	0.318	-0.264	-0.643	-0.731***	-0.705***	-0.722***	-0.813***	-0.899***	-0.899***
	(-3.093, 0.038)	(0.162, 0.862)	(-0.157, 0.9)	(1.017, 0.633)	(-1.093, 0.6)	(-3.229, 0.1)	(-3.894, 0)	(-4.006, 0)	(-4.246, 0)	(-5.13, 0)	(-6.064, 0)	(-6.355, 0)
$\lvert c_{i,t0} - c_{j,t0} \rvert$	1.149***	-1.15	-0.209	0.218	0.613***	0.873***	0.93***	0.893***	0.873***	0.951***	1.051***	1.096***
	(2.548, 0)	(-1.723, 0.2)	(-0.378, 0.7)	(0.827, 0.567)	(2.998, 0)	(5.204, 0)	(5.954, 0)	(6.162, 0)	(6.289, 0)	(7.398, 0)	(8.856, 0)	(9.765, 0)
distance	2.34***	-0.222	0.149	0.757***	0.936***	0.888***	0.938***	0.881***	0.886***	0.857***	0.769***	0.754***
	(3.324, 0)	(-0.379, 0.767)	(0.496, 0.467)	(4.319, 0)	(6.928, 0)	(8.351, 0)	(9.211, 0)	(9.143, 0)	(9.499, 0)	(9.913, 0)	(9.717, 0)	(10.06, 0)
sameregion	2.76***	-5.785***	-5.45***	-2.396***	-2.371***	-2.093***	-2.034***	-2.081***	-2.024***	-1.793***	-1.637***	-1.546***
	(2.801, 0)	(-3.567, 0)	(-4.52, 0)	(-6.067, 0)	(-7.77, 0)	(-8.479, 0)	(-8.67, 0)	(-9.253, 0)	(-9.066, 0)	(-8.947, 0)	(-9.398, 0)	(-9.822, 0)
obs.	12090	12090	12090	12090	12090	12090	12090	12090	12090	12090	12090	12090
loglik	-364	-1214	-2946	-5213	-6232	-6955	-7217	-7350	-7354	-7548	-7670	-7870

— P-values are based on MRQAP (1000 permutations) and one-sided (because null distributions are not symmetric). One, two, and three stars signal significance values below 5%, 1%, and 0.1% respectively.

— Recall from Section 1.2.4 that n and c stands for non-CRR and CRR respectively. Indices i indicate a country, and ij a country dyad. Country-level variables are joined to capture the country dyad's sum ($x_i + x_j$) and their absolute difference ($|x_i - x_j|$). $log(\theta)$ captures over-dispersion in the count model.

— The main finding is that in February initial CRR competence ($c_{i,t0}$) is positively and non-CRR ($n_{i,t0}$) is negatively associated with collaboration. Prior collaborations in non-CRR ($n_{ij,t0}$) are relevant throughout. In the second half of the year, countries with strong non-CRR competence ($n_{i,t0}$) attract collaborations, and less developed countries tend to collaborate more among each other ($hdi_i + hdi_j$ and $|hdi_i - hdi_j|$). Same region effect is present in all months.

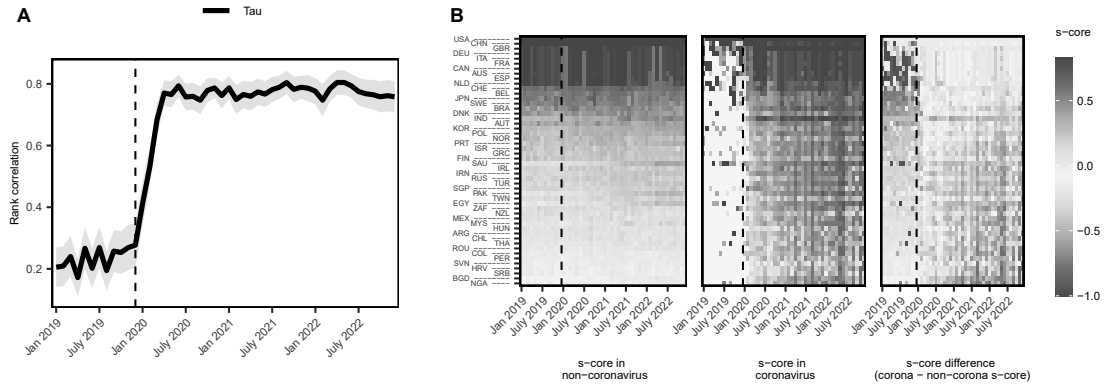
1.3.3 Convergence to global health sciences

1.3.3.1 National scientific output

Countries take rapidly very similar positions in rankings on coronavirus papers as they do in rankings on other health papers. This result is evident in Fig 1.4A which provides the rank correlation coefficient of country rankings in CRR and non-CRR research by month. Until the outbreak in January 2020 (vertical dashed line

in Fig 1.4A) rank correlations were rather low at around 0.2. After the outbreak, the monthly scientific output of countries in Corona aligns with non-CRR output attaining a correlation of 0.80 in April 2020. From then on, correlations stayed consistent throughout our period.

Figure 1.4: Ranking and convergence of hierarchy among nations



Notes: Countries take on the same role in coronavirus research as in the global health sciences. (A) Correlation of country rankings by coronavirus and non-coronavirus research by month. (B) Country centrality based on s-core decomposition of the coronavirus and non-coronavirus network by month.

This trend is consistent with estimation results on the accumulated national CRR output (Table 1.5), where pre-pandemic health science output ($n_{i,t0}$) becomes the dominant factor from March 2020 on.

1.3.3.2 Network centrality

Similarly, countries' network centrality in the coronavirus research network aligns with their centrality in the overall health science network.

Fig 1.4B provides the monthly s-cores on the non-CRR network (left panel), CRR network (middle panel), and the difference of s-cores in CRR and non-CRR networks (right panel). The figure shows the 60 most central countries in the non-CRR network and applies that same ordering across all three panels. The remaining countries are highly peripheral in either network.

The left panel of Fig 1.4B shows that the global network hierarchy is very stable. The core is formed by developed countries and China. Centrality in the coronavirus network is more dynamic (middle panel). Pre-COVID-19, most countries are not involved in CRR collaborations and, hence, are found in the extreme periphery (white). The core of the CRR network includes only a few of the leading countries in health

sciences. On the other hand, Saudi Arabia is part of the core in the CRR network, but peripheral in the overall health science network. The presence of Saudi Arabia in the pre-pandemic CRR network may be explained by previous regional MERS-CoV outbreaks (Variations in core membership over time may be explained by lower research activity overall which leads to more erratic signals.) After the shock, the structure of the CRR network shifted rapidly towards the hierarchy in health science at large. This is easily seen in the right panel of Fig 1.4B which shows the difference between the s-core centrality in the CRR and non-CRR network. Prior to the shock, s-core differences range from -1 in dark blue (for countries at the extreme periphery in the coronavirus network and in the core in the other network), over 0 in white (same s-core in both networks), up to 1 in red (for countries in the CRR network core and peripheral in the non-CRR network). After the shock, the global core rapidly takes its role in CRR, and so does the global periphery (all countries appear in light colors with an s-core difference of around zero from April 2020).

Closer inspection further reveals that the normalized s-core of more peripheral countries tends to be somewhat higher in the CRR network as in the non-CRR network (Panel 's-core difference' turns rather red than blue in 2020). This means that the distance to the highest core in the network is somewhat reduced. Thus, while the *ranking* by network centrality of countries in the CRR network is maintained as in the non-CRR network, hierarchy has become less steep (at least among the top 60 countries).

1.3.3.3 Network structure

Fig. 1.5 shows how CRR network structure compares to non-CRR network structure over the course of the pandemic.

Fig. 1.5A provides the correlation coefficient of the accumulated CRR and 2019 non-CRR adjacency matrix (in logs) over the observation period. During the pre-pandemic period (2019) the two adjacency matrices were increasingly but only weakly correlated with a coefficient of around 0.25. After the shock in January 2020, within three months (April 2020), the correlation coefficient jumps to 0.75 and increases further to around 0.9 until the end of the analysis period, December 2022. All correlations, pre- and post-pandemic, are highly significant based on a QAP test. Thus, with the COVID-19 pandemic, the CRR network structure rapidly approached the prior non-CRR network structure.

However, the correlation coefficient is not approaching one either, which implies

that some structural difference remains between the two networks. So, what is the difference?

This can be seen in the middle panel (Fig. 1.5B), which provides statistics on network hierarchy (z_λ) and network modularity (z_Q) of the accumulated CRR network relative to the prior non-CRR network.

First consider the hierarchy measure z_λ, the line in red. We start with a z_λ-value of 1.3 in January 2019 and increase as more CRR collaborations are accumulated up to 7.7 by December 2019. Thus, before the shock, the CRR network is more hierarchical than the non-CRR network. This is because only very few countries actually have some CRR activity making worldwide CRR collaboration highly unequal. In other words, in the land of the blind, the one-eyed man is king. Immediately after the shock, in January 2020 hierarchy increased to $z_\lambda = 11.7$. That level of hierarchy is mostly kept in February 2020 but then decreases constantly, indicating that by the end of 2022, the CRR network is much less hierarchical than the non-CRR network.

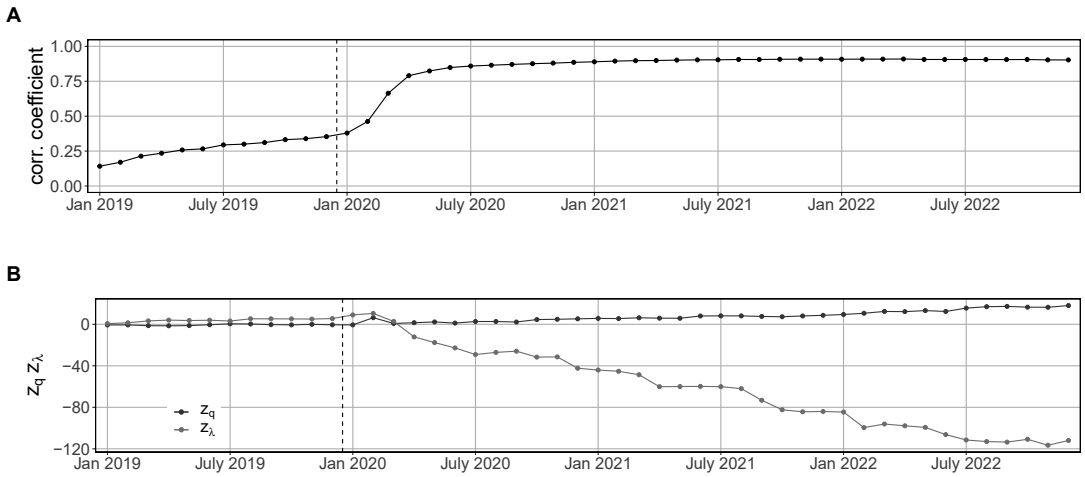
The development of network modularity (z_Q) is different. During the pre-pandemic period, 2019, modularity in the CRR network is comparable to modularity in the non-CRR network with a z_Q around zero. This means that pre-pandemic international collaborations on CRR tend to follow the same community structure as pre-pandemic international collaborations on non-CRR.[24] The first reaction of the community statistic is observed in February 2020 where it jumps to a high level of 6.3 and immediately drops thereafter. Then, from March to December 2022, network modularity increases constantly but at a slower pace than the decreasing hierarchy.

For interpretation, we considered z_Q for each community individually (numbers not reported here). The bump of z_Q in February 2020 originates in particular from within cluster interaction in community 1 (in particular USA, China) and to some extent also in community 6 (South America) and community 7 (Middle East and Africa). In contrast, the subsequent more continuous rise of z_Q is driven only by regional communities, i.e. all communities except community 1. In particular community 2 (Central Europe), community 3 (Norther Europe), 6 (South America) and 7 (Middle East and Africa) contribute to this trend.

In sum, by the end of the observation period, the accumulated CRR network has become very similar to the non-CRR network but less hierarchical and more modular along regional communities.

[24]Recall that by construction the modularity statistic takes only into account countries that have (any) CRR collaboration.

Figure 1.5: Convergence of CRR network to non-CRR network.



Notes: In all plots, the CRR network is accumulated from Jan. 2019 up to a given month, and the non-CRR network, accumulated from Jan. 2019 to Dec. 2019, serves as the reference network. (A) Pearson's correlation coefficient of (log of) CRR adjacency matrix and non-CRR adjacency matrix. All correlations are highly significant, according to QAP. (B) Development of z_q (z-value of modularity statistic Q of CRR network with null from non-CRR) and z_λ (z-value of the largest eigenvalue of CRR network with null from non-CRR network).

1.4 Discussion and Conclusion

Based on the population of Medline-indexed journal articles, we investigated the global scientific response to the COVID-19 pandemic. Taking an (inter-)national perspective, our analysis focused on three interrelated aspects: national scientific output, international research collaboration, and network formation. The aim has been to describe the dynamics of CRR during the pandemic and to put them into context. The following first summarizes our main results. Then we turn to the limitations of our study in order to provide policy implications and directions for future research.

Our main results may be summarized in four points: First, our three analyses on three different levels yield highly consistent results. Regressions of national CRR output (Analysis 1) and regressions of international collaboration on CRR (Analysis 2), show that national and inter-national scientific activity are closely coupled. Both follow similar dynamics but are driven by different factors. Analysis 3 on the international CRR network describes the order of global CRR that emerges as a consequence. The close coupling of national and international positioning in CRR is in strong agreement with the broader literature on (inter-)national science systems [see

e.g. Adams, 2013, Davidson Frame et al., 1977, Luukkonen et al., 1992, Pan et al., 2012, Gui et al., 2018, in combination].

Second, regression analyses 1 and 2 show that CRR-specific science capacity accumulated before the pandemic has been influential for national CRR output and international CRR collaboration in particular during the first three months of the pandemic. Broader health science capacity has become the dominant factor for (inter-)national CRR at least three months after the beginning of the pandemic. Contextual factors related to the pandemic exhibit varying impacts. Starting from the end of 2020, COVID-19-related deaths positively influenced international research collaboration (IRC) and national CRR output (though with more fluctuation in significance). In contrast, lockdown measures seem only to affect IRC, while border closures do not affect either. Likewise, socioeconomic indicators like GDP per capita and Human Development Index (HDI) do not provide meaningful explanations for Analysis 1, but they play a more substantial role in shaping international collaboration dynamics throughout the pandemic.

We thus complement prior, more qualitative, studies. Haghani and Bliemer [2020] and Zhang et al. [2020] show that pre-pandemic CRR and non-CRR both provide relevant knowledge for the COVID-19 pandemic. Our regressions show a clear order in time of how coronavirus-specific knowledge and more general health science capacity have been leveraged. The role of pandemic-related factors has been put forward to explain observations on (inter-)national CRR in several studies [Haghani and Bliemer, 2020, Fry et al., 2020, Zhang et al., 2020]. For example, the dynamic of the infection process has been related to dynamics in CRR output in particular for China and Italy [e.g. Fry et al., 2020]. While pandemic-related explanations may well hold for individual countries, we find that such effects are *on average* relatively small for national CRR output with only the number of deaths being significant and positive but also volatile. This can be explained by the fact that the global scientific community rapidly felt some urgency. Similarly, the absence of effect of GDP and HDI suggests that, in the short run, science capacity is largely fixed. Our estimates on the relevance of regional collaboration comply with the general literature on IRC [Frenken et al., 2009] as well as observations on regional epidemics [Haghani and Bliemer, 2020, Zhang et al., 2020].

Thirdly, funding plays a pivotal role in shaping the field of science and holds significant importance in both national and international contexts. Our research has uncovered a notable phenomenon where existing grants are being used to address ur-

gent issues. This phenomenon seems to have been at the forefront of early research efforts and can be attributed to researchers' ability to repurpose existing resources to tackle pressing problems. Subsequently, the creation of new grants followed, peaking in April 2021, which indicates a delay in results for new funding projects. Our first analysis demonstrates that international funding dependence lacks significance during the initial year of the pandemic but exerts a negative influence on national CRR in 2021 and 2022. This shift can be explained by existing grants, that researchers were utilizing to focus on addressing the challenges posed by COVID-19, reaching their conclusion within the first year. Our second analysis reveals a negative effect of funding dependence between countries on collaboration during the first and half years of the pandemic. This may be attributed to an early search for new partners or driven by the emergence of new funding opportunities that encourage exploration of the collaboration space.

Finally, consistent with regression results, Analysis 3 shows that global CRR during the pandemic rapidly converges towards the global order of broader health science capacity. Within three months, country rankings by national CRR output approach rankings by pre-pandemic non-CRR output (Analysis 3.1), countries take the same centrality position in the international CRR network as in the pre-pandemic broader health science network (Analysis 3.2), and the CRR network converges to the structure of the pre-pandemic health science network (Analysis 3.3). However, the alignment is not perfect. Global CRR deviates systematically from the attracting global distribution of broader health science before the pandemic in that the global CRR network is significantly less hierarchic and more regional than global health science (Analysis 3.3).

These results are all consistent with prior empirical studies on the expansion of the international CRR network [Cai et al., 2021, Fry et al., 2020, Haghani and Bliemer, 2020, Zhang et al., 2020]. Our study clarifies however that the pre-existing global health science system systematically structures the expansion of global CRR, and add that global CRR systematically deviates in relevant dimensions from the inherited global structure.

One major limitation of this study is the arguably rough division of research papers into two categories; CRR and non-CRR. The following discussion takes this limitation into account to delineate policy implications and to outline potential avenues for future research.

The COVID-19 shock opened a scientific race. In order to start off, most scientists

had to re-orient or adapt ongoing research. Our observation that countries with prior experience in CRR occupied pole positions but countries of the global health science core rapidly took the lead, suggests that a broad science base provides the ability to re-orient research effectively. This is an argument for a more autonomous development of the science system because it deliberates policy and society from the burden and impossibility to specify a detailed scientific agenda to meet future challenges.

However, having no notion of distance between non-CRR and CRR, we can say nothing about the degree of flexibility in research orientation, nor what exactly underlies such flexibility. In particular, we would like to know to what extent CRR research during the pandemic benefited from very generic scientific capacity obtained through basic science, from closely related to CRR scientific capacity or their interaction. This would give some indication for science policy on how to balance basic science against applied research to guard against future challenges. To achieve a better understanding of the phenomenon, one could employ topic modeling and unsupervised learning techniques.

Moreover, the funding aspect explored in our study presents notable limitations. Funding schemes vary significantly depending on the country of operation, introducing complexities into our analysis. While we have taken steps to test the robustness of our models by excluding data related to the United States, it is evident that additional research in this domain is required. Questions arise regarding whether the phenomenon of repurposing research resources is unique to the United States or extends to other regions. There is also a need for further investigation into the nature of research conducted within existing projects, the characteristics of research within repurposed projects, and the attributes of research within entirely new projects. Additionally, it is essential to assess whether brand-new projects contribute to more novelty in the field compared to existing ones and what implications this holds for their overall impact. These critical questions demand deeper exploration in subsequent research endeavors to gain a comprehensive understanding of the dynamics of research funding and its impact on scientific outcomes in times of crisis.

Immediately after the shock the need for international collaboration has been emphasized by all stakeholders. Our results show that national CRR output and international CRR collaboration have been so closely aligned that they should be considered as two dimensions of the same activity. Furthermore, we found that the international collaboration network on CRR largely followed the prior international

health science network. The implication is that any national strategy to develop science capacity must take a global systems perspective — the embedding of the national science system in the international collaboration network is part of the slowly accumulating and path-dependent national science base.

Expressing that insight without prompting two subsequent questions is challenging: What defines an "appropriate" network structure, and who is it appropriate for? If CRR converges toward an existing structure we must scrutinize the effectiveness of that structure. How can we establish an "efficient" network structure under such circumstances?

According to our analysis, the international CRR network exhibits a less steep hierarchy and stronger regionalization than the international broader health science network. Yet, what exactly that signifies remains unclear. Recall that we do not differentiate papers by quality, we only measure quantities. COVID-19 has been a hot science topic in 2020 and the following years, and it may be that many papers have been hastily crafted but add little knowledge. A higher propensity of low-quality papers in countries with lower scientific capacity could contribute to our finding of a less steep global hierarchy as measured by output quantity, while global hierarchy in terms of research excellence may be in fact maintained. Similarly, increased regionalization, as we observe it, is consistent with more independent science in the different world regions and hence more independent accumulation of science capacity on a regional level. But it is also consistent with maintained but unresolved resource (inter-)dependence and increasing vertical stratification by scientific excellence as well as horizontal stratification by research topics. Thus, the paper at hand clearly describes structural changes but leaves the question of individual network gains and global network efficiency for future research.

The underlying cause of structural changes in global CRR remains unclear. On the one hand, stronger regionalization may be largely due to the pandemic context. Our regressions include relevant parameters such as national pandemic policy measures and virus-spreading dynamics, but these are certainly no perfect controls. Given those controls, our results are consistent with the idea that we observe the accentuation of a long-term trend towards regionalization; a trend already observed by Fitzgerald et al. [2021] for the sciences in general. In case the pandemic context is the underlying cause, the organization of global science during the pandemic may be as transitory as the pandemic itself. If the organization of global CRR during the pandemic accentuates a general global trend, it may not only cast the shadow of a

future multi-polar world but even accelerate that trend.

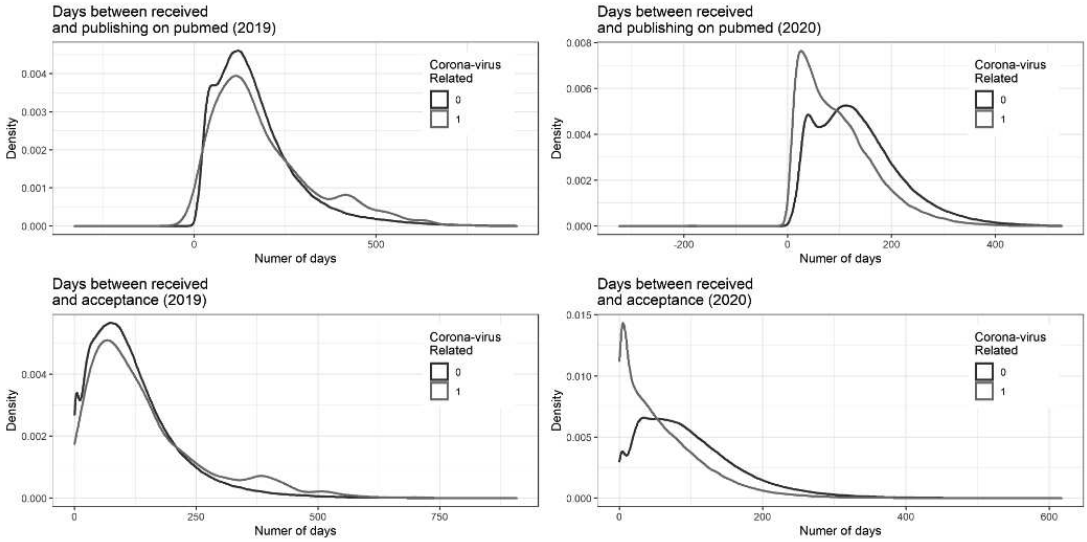
In the limited time frame we had, we opted for specifications that could be improved as the period increased. There is still exploration to be done regarding the convergence, resilience, and adaptability of the science network after exogenous shock.

1.5 Appendix

1.5.1 Sample

1.5.2 Time-lags from research over acceptance to entry into the PubMed dataset

Figure 1.6: Sample time-lags



Notes: Time-lags from submission date (data received) to (clockwise, starting with the upper left panel): i) publishing on PubMed in 2019, ii) publishing on PubMed in 2020, iii) journal acceptance in 2020, iv) journal acceptance in 2019. Our sample is from PubMed. Therefore, the analysis period is based on the time-lag distribution of the upper right panel.

Table 1.7: Descriptive statistics on scientific production (at submission date)

	Coronavirus related		*Others Documents*		*All Documents*	
	2019	2020	2019	2020	2019	2020
Nb of authors	6/7.36 (4.86)	5/6.86 (7.05)	6/6.66 (17.51)	6/6.65 (14.67)	6/6.66 (17.5)	6/6.66 (14.2)
Solo authors	2.83	6.72	3.38	3.22	3.38	3.5
International collab.	25.44	23.06	23.57	23.63	23.57	23.58
Nb of country	1/1.36 (0.92)	1/1.35 (1.29)	1/1.31 (1.05)	1/1.32 (1.01)	1/1.31 (1.05)	1/1.33 (1.03)
Days received-accept	110.04/143.31 (119.52)	45/60.87 (58.53)	99.96/118.97 (93.82)	84.96/97.96 (73.24)	99.96/119 (93.85)	81.96/94.88 (72.86)
Share aff. captured	1/0.95 (0.2)	1/0.92 (0.26)	1/0.93 (0.24)	1/0.95 (0.21)	1/0.93 (0.24)	1/0.94 (0.22)
# Document	743(0.1%)	74250(8.2%)	775738(99.9%)	830914(91.8%)	776481	905164

Notes: Binary indicators in [%], for continuous measures [median/mean (s.d.)].

1.5.3 Community detection in the 2019 non-CRR network

The procedure is the same as in Fitzgerald et al. [2021], but relevant details are repeated here for readers' convenience.

We identify communities with the spin glass community detection algorithm of Reichardt and Bornholdt [2006]. The objective criterion to be maximized is

$$Q = \frac{1}{m}\left(\sum_{ij} c_{ij} - \gamma \frac{k_i k_j}{m}\right) d(\sigma_i, \sigma_j),$$

where m is the sum over all edges, c_{ij} is the number of joint papers between country i and j, k_i (k_j) is the strength of country i (j) (i.e. the sum over all weighted edges of the country), σ_i (σ_j) indicates the community of i (j) in a given network partition, with the function $d()$ evaluating to one if both countries belong to the same community (i.e. $\sigma_i = \sigma_j$) and zero else. Finally, the tuning parameter γ trades-off the two objectives of having high edge weights within clusters against having few edge weights between clusters. A smaller (higher) γ tends to yield less (more) clusters.

The optimization is done through simulated annealing (as implemented in the

R-package 'igraph'). Simulated annealing is a stochastic optimization algorithm and hence may provide different partitions for the same data and parameter settings. In order to find robust communities, we run 100 times the optimization algorithm over a grid of γ, starting from 0.8 to 1.6. The resulting community partitions for a given γ are then compared, each partition against all others. Similarity of two partitions is measured by the 'Variation of Information' (VI). The following definition is paraphrased from [Fitzgerald et al., 2021, Appendix 4]:

"*Consider two partitions X and Y of a set A into disjoint subsets, $X = X_1, X_2, \ldots, X_k$ and $Y = Y_1, Y_2, \ldots, Y_l$, VI is defined as follows. Let $n = \sum_i |X_i| = \sum_j |Y_j| = |A|$, $p_i = |X_i|/n$, $q_j = |Y_j|/n$, $r_{ij} = |X_i \cap Y_j|/n$. Then the normalized variation of information between the two partitions is:*

$$VI(X,Y) = -1 \log N \sum_{i,j} r_{ij} \left[\log(r_{ij}/p_i) + \log(r_{ij}/q_j)\right] .$$"

Table 1.8 provides the VIs for 100 optimizations for each γ on the grid. The most robust partitioning is obtained for $\gamma = 0.8$. The resolution of that partitioning however is very low, dividing the world into one cluster of less developed economies (consisting of countries in Africa, Middle East, South Asia, India, Mongolia, and Kazakhstan) and another cluster including all other countries. The partitioning with $\gamma = 1.2$ is slightly less robust but at a much higher resolution. Figure 1.3 shows the partitioning with the lowest average VI under $\gamma = 1.2$. We chose that partition as our benchmark. Interestingly, the partitioning that we obtain on the 2019 non-CRR network is not identical but very similar to the partitioning obtained by Fitzgerald et al. [2021, Fig.5b] based on all Scopus publications in 2015.

Table 1.8: Variation of Information for varying γ over 100 optimizations.

γ	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6
VI	0.099	0.302	0.465	0.161	0.102	0.349	0.255	0.158	0.232

The partitioning displayed in Figure 1.3 consists of the following communities (ordered by the average degree of their members):

Cluster 1: Australia, Canada, China, French Polynesia, Gibraltar, Grenada, Japan, Macao SAR China, New Zealand, Singapore, Solomon Islands, South Korea, Taiwan, United Kingdom, United States

Cluster 2: Austria, Belgium, Curacao, France, Germany, Italy, Liechtenstein, Luxembourg, Martinique, Monaco, Netherlands, Réunion, Spain, Switzerland, Vatican City

Cluster 3: Denmark, Faroe Islands, Finland, Greenland, Guinea-Bissau, Iceland, Isle of Man, Norway, Sweden

Cluster 4: Israel, Jersey, Montserrat

Cluster 5: Albania, Angola, Armenia, Azerbaijan, Belarus, Bosnia & Herzegovina, Bulgaria, Croatia, Cyprus, Czechia, Estonia, Georgia, Greece, Hungary, Ireland, Kazakhstan, Latvia, Lithuania, Malta, Moldova, Montenegro, North Macedonia, Poland, Portugal, Romania, Russia, San Marino, Serbia, Sint Maarten, Slovakia, Slovenia, Tajikistan, Turkey, Turkmenistan, Ukraine, Uzbekistan

Cluster 6: Andorra, Antigua & Barbuda, Argentina, Belize, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Dominica, Dominican Republic, Ecuador, El Salvador, French Guiana, Guatemala, Guyana, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, St. Kitts & Nevis, Suriname, Uruguay, Venezuela

Cluster 7: Afghanistan, Algeria, Aruba, Bahamas, Bahrain, Bangladesh, Barbados, Benin, Bermuda, Bhutan, Botswana, Brunei, Burkina Faso, Burundi, Cape Verde, Cambodia, Cameroon, Cayman Islands, Central African Republic, Chad, Congo - Kinshasa, Egypt, Eritrea, Ethiopia, Fiji, Gabon, Gambia, Ghana, Guadeloupe, Guinea, Haiti, India, Indonesia, Iran, Iraq, Côte d'Ivoire, Jamaica, Jordan, Kenya, Kuwait, Kyrgyzstan, Laos, Lebanon, Lesotho, Liberia, Libya, Madagascar, Malawi, Malaysia, Maldives, Mali, Mauritania, Mauritius, Mongolia, Morocco, Mozambique, Myanmar (Burma), Namibia, Nepal, New Caledonia, Niger, Nigeria, Oman, Pakistan, Palau, Palestinian Territories, Philippines, Qatar, Congo - Brazzaville, Rwanda, St. Lucia, St. Vincent & Grenadines, Samoa, Saudi Arabia, Senegal, Seychelles, Sierra Leone, Somalia, South Africa, Sri Lanka, Sudan, Syria, Tanzania, Thailand, Togo, Tonga, Trinidad & Tobago, Tunisia, Turks & Caicos Islands, Uganda, United Arab Emirates, Vietnam, Yemen, Zambia, Zimbabwe

No cluster: Djibouti, Äland Islands, British Virgin Islands, Sao Tome and Principe, Tuvalu

1.5.4 Extended tables

Table 1.9: International science networks, CRR and non-CRR, accumulated over months For the year 2021 and 2022.

CRR network (accumulated), 2021												
	Jan.	**Feb.**	**Mar.**	**Apr.**	**May**	**June**	**July**	**Aug.**	**Sept.**	**Oct.**	**Nov.**	**Dec.**
Countries	185	187	187	189	191	191	191	192	192	193	195	195
Country % growth	1	1	0	1	1	0	0	1	0	1	1	0
Edge weights	84.7k	92.8k	99.2k	107.4k	114.8k	121.6k	128.5k	135.1k	145.7k	153.2k	159.8k	166.2k
Weight % growth	9	10	7	8	7	6	6	5	8	5	5	4
non-CRR network (accumulated), 2021												
	Jan.	**Feb.**	**Mar.**	**Apr.**	**May**	**June**	**July**	**Aug.**	**Sept.**	**Oct.**	**Nov.**	**Dec.**
Countries	202	203	203	203	203	204	204	204	204	204	204	205
Country % growth	0	0	0	0	0	0	0	0	0	0	0	0
Edge weights	1131k	1178k	1236k	1286k	1341k	1393k	1444k	1495k	1542k	1591k	1644k	1695k
Weight % growth	5	4	5	4	4	4	4	4	3	3	3	3
CRR network (accumulated), 2022												
	Jan.	**Feb.**	**Mar.**	**Apr.**	**May**	**June**	**July**	**Aug.**	**Sept.**	**Oct.**	**Nov.**	**Dec.**
Countries	195	195	195	195	195	195	196	196	200	200	200	200
Country % growth	0	0	0	0	0	0	1	0	2	0	0	0
Edge weights	172.3k	179.3k	185.9k	192.3k	201.6k	209.2k	215k	219.7k	224.2k	227.9k	234k	237.7k
Weight % growth	4	4	4	3	5	4	3	2	2	2	3	2
non-CRR network (accumulated), 2022												
	Jan.	**Feb.**	**Mar.**	**Apr.**	**May**	**June**	**July**	**Aug.**	**Sept.**	**Oct.**	**Nov.**	**Dec.**
Countries	205	205	205	205	205	205	205	205	205	205	205	205
Country % growth	0	0	0	0	0	0	0	0	0	0	0	0
Edge weights	1745k	1795k	1849k	1896k	1948k	1998k	2048k	2096k	2143k	2190k	2233k	2273k
Weight % growth	3	3	3	3	3	3	3	2	2	2	2	2

Table 1.10: Accumulated number of CRR papers ($c_{t'}$) in 2021 (all variables in logs, standardized to zero mean and one std.dev., 156 countries).

	Jan. '21	**Feb. '21**	**Mar. '21.**	**Apr. '21**	**May '21**	**June '21**	**July '21**	**Aug. '21**	**Sept. '21**	**Oct. '21**	**Nov. '21**	**Dec. '21**
n_{t0}	0.936***	0.944***	0.944***	0.950***	0.952***	0.957***	0.961***	0.965***	0.957***	0.961***	0.955***	0.959***
	(0.021)	(0.044)	(0.041)	(0.060)	(0.067)	(0.060)	(0.058)	(0.062)	(0.062)	(0.061)	(0.053)	(0.053)
c_{t0}	0.010	−0.006	−0.013	−0.027	−0.035	−0.043	−0.052	−0.062	−0.053	−0.059	−0.051	−0.057
	(0.061)	(0.085)	(0.079)	(0.135)	(0.149)	(0.128)	(0.131)	(0.120)	(0.114)	(0.113)	(0.120)	(0.121)
$Int - Dep_{t0}$	−0.078	−0.111	−0.230***	−0.235**	−0.252***	−0.220*	−0.296**	−0.286**	−0.310**	−0.305***	−0.249	−0.255
	(0.197)	(0.139)	(0.022)	(0.095)	(0.073)	(0.126)	(0.149)	(0.141)	(0.130)	(0.106)	(0.157)	(0.163)
$deaths_{t'}$	0.034	0.024	0.022	0.013	0.011	0.010	0.009	0.009	0.021***	0.026***	0.035***	0.038*
	(0.024)	(0.024)	(0.025)	(0.021)	(0.024)	(0.034)	(0.021)	(0.009)	(0.008)	(0.010)	(0.011)	(0.019)
$border - closed_{t'}$	−0.005	−0.005	−0.011	−0.014	−0.014	−0.013	−0.013	−0.013	−0.012	−0.013	−0.016	−0.017
	(0.020)	(0.021)	(0.018)	(0.042)	(0.023)	(0.019)	(0.021)	(0.022)	(0.025)	(0.024)	(0.020)	(0.019)
$lock - down_{t'}$	0.051	0.050	0.046	0.053**	0.056	0.057	0.053	0.057	0.052	0.048	0.049	0.047
	(0.03)	(0.039)	(0.029)	(0.027)	(0.043)	(0.041)	(0.044)	(0.050)	(0.048)	(0.045)	(0.040)	(0.036)
gdp_{t0}	0.227	0.225	0.203	0.196	0.182	0.178	0.185	0.197	0.194	0.192	0.180	0.178
	(0.051)	(0.150)	(0.191)	(0.238)	(0.226)	(0.261)	(0.275)	(0.270)	(0.271)	(0.272)	(0.240)	(0.253)
hdi_{t0}	−2.815	−2.702**	−2.584	−2.497	−2.326	−2.278	−2.350	−2.399	−2.471	−2.516	−2.474	−2.496
	(0.088)	(1.212)	(1.599)	(2.015)	(1.898)	(1.999)	(2.180)	(2.151)	(2.154)	(2.162)	(1.961)	(2.131)
R^2	0.947	0.947	0.950	0.951	0.950	0.949	0.949	0.950	0.950	0.951	0.954	0.954

Table 1.11: Accumulated number of CRR papers ($c_{t'}$) in 2022 (all variables in logs, standardized to zero mean and one std.dev., 156 countries).

	Jan. '22	**Feb. '22**	**Mar. '22.**	**Apr. '22**	**May '22**	**June '22**	**July '22**	**Aug. '22**	**Sept. '22**	**Oct. '22**	**Nov. '22**	**Dec. '22**
n_{t0}	0.960***	0.966***	0.968***	0.975***	0.976***	0.980***	0.974***	0.978***	0.972***	0.967***	0.966***	0.966***
	(0.021)	(0.044)	(0.041)	(0.060)	(0.067)	(0.060)	(0.058)	(0.062)	(0.062)	(0.061)	(0.053)	(0.053)
c_{t0}	−0.057	−0.066	−0.071	−0.082	−0.087	−0.096	−0.088	−0.091	−0.086	−0.079	−0.078	−0.079
	(0.061)	(0.085)	(0.079)	(0.135)	(0.149)	(0.128)	(0.131)	(0.120)	(0.114)	(0.113)	(0.120)	(0.121)
$Int - Dep_{t0}$	−0.252	−0.249*	−0.245***	−0.256***	−0.270***	−0.276**	−0.259*	−0.263*	−0.253*	−0.233**	−0.222	−0.218
	(0.197)	(0.139)	(0.022)	(0.095)	(0.073)	(0.126)	(0.149)	(0.141)	(0.130)	(0.106)	(0.157)	(0.163)
$deaths_{t'}$	0.039	0.038	0.039	0.037*	0.039	0.035	0.041**	0.042***	0.046***	0.052***	0.054***	0.055***
	(0.012)	(0.013)	(0.013)	(0.021)	(0.024)	(0.034)	(0.021)	(0.009)	(0.008)	(0.010)	(0.011)	(0.019)
$border - closed_{t'}$	−0.017	−0.017	−0.019	−0.019	−0.019	−0.021	−0.020	−0.020	−0.020	−0.021	−0.021	−0.022
	(0.014)	(0.014)	(0.015)	(0.042)	(0.023)	(0.019)	(0.021)	(0.022)	(0.025)	(0.024)	(0.020)	(0.019)
$lock - down_{t'}$	0.046	0.045	0.045	0.045*	0.045	0.044	0.044	0.045	0.045	0.044	0.044	0.045
	(0.019)	(0.018)	(0.018)	(0.027)	(0.043)	(0.041)	(0.044)	(0.050)	(0.048)	(0.045)	(0.040)	(0.036)
gdp_{t0}	0.182	0.181	0.177	0.176	0.176	0.183	0.177	0.178	0.179	0.173	0.168	0.169
	(0.051)	(0.150)	(0.191)	(0.238)	(0.226)	(0.261)	(0.275)	(0.270)	(0.271)	(0.272)	(0.240)	(0.253)
hdi_{t0}	−2.571	−2.575**	−2.548	−2.533	−2.560	−2.623	−2.595	−2.622	−2.642	−2.611	−2.590	−2.605
	(0.088)	(1.212)	(1.599)	(2.015)	(1.898)	(1.999)	(2.180)	(2.151)	(2.154)	(2.162)	(1.961)	(2.131)
R^2	0.953	0.953	0.953	0.953	0.953	0.953	0.953	0.953	0.953	0.954	0.954	0.955

Table 1.12: Zero-inflated negative binomial model of (accumulated) joint coronavirus related papers ($c_{ij,t'}$) during the pandemic, est. coefficient (z-value, p-value)

	Jan. 21	**Feb. 21**	**Mar. 21**	**Apr. 21**	**May 21**	**Jun. 21**	**Jul. 21**	**Aug. 21**	**Sep. 21**	**Oct. 21**	**Nov. 21**	**Dec. 21**
Count-model												
$n_{ij,t0}$	0.95***	0.885***	0.879***	0.854***	0.834***	0.826***	0.81***	0.8***	0.759***	0.759***	0.758***	0.75***
	(40.966, 0)	(40.954, 0)	(41.31, 0)	(41.316, 0)	(41.224, 0)	(41.557, 0)	(40.846, 0)	(40.942, 0)	(40.293, 0)	(41.059, 0)	(41.388, 0)	(40.937, 0)
$c_{ij,t0}$	-0.131***	-0.085	-0.081	-0.065	-0.058	-0.05	-0.04	-0.029	-0.01	-0.002***	0.003***	0.008***
	(-3.53, 0)	(-2.466, 0.2)	(-2.4, 0.067)	(-1.985, 0.1)	(-1.787, 0.2)	(-1.546, 0.2)	(-1.225, 0.233)	(-0.909, 0.267)	(-0.314, 0.233)	(-0.081, 0)	(0.1, 0)	(0.252, 0)
$n_{i,t0}$	-0.005	0.02	0.021	0.028	0.037	0.03	0.034	0.04	0.056***	0.056*	0.051***	0.054
	(-0.275, 0.867)	(1.318, 0.633)	(1.408, 0.333)	(1.913, 0.3)	(2.538, 0.133)	(2.062, 0.267)	(2.317, 0.2)	(2.766, 0.167)	(4.013, 0)	(4.097, 0.033)	(3.767, 0)	(3.943, 0.067)
$c_{i,t0}$	0.034***	0.028	0.029	0.028*	0.024	0.025	0.025	0.022	0.018	0.015	0.016	0.017
	(4.083, 0)	(3.513, 0.133)	(3.673, 0.067)	(3.702, 0.033)	(3.227, 0.133)	(3.423, 0.233)	(3.449, 0.2)	(2.978, 0.133)	(2.576, 0.1)	(2.169, 0.3)	(2.311, 0.267)	(2.458, 0.233)
$death_i + death_j$	0.122***	0.114***	0.108***	0.097***	0.102***	0.102***	0.112***	0.119***	0.113***	0.119***	0.116***	0.11***
	(6.323, 0)	(6.077, 0)	(5.856, 0)	(5.675, 0)	(5.88, 0)	(5.82, 0)	(6.367, 0)	(6.698, 0)	(6.708, 0)	(7.435, 0)	(7.072, 0)	(6.708, 0)
$\lvert death_i - death_j \rvert$	-0.058*	-0.062*	-0.06	-0.05	-0.05	-0.045	-0.05*	-0.052***	-0.046	-0.052***	-0.05	-0.046*
	(-3.956, 0.033)	(-4.353, 0.033)	(-4.264, 0.067)	(-3.927, 0.067)	(-3.767, 0.1)	(-3.4, 0.167)	(-3.747, 0.033)	(-3.891, 0)	(-3.851, 0.067)	(-4.644, 0)	(-4.316, 0.1)	(-4.064, 0.033)
$locked_{i,t'}$	0.028***	0.028***	0.027***	0.027***	0.026***	0.025***	0.025***	0.025***	0.024***	0.023***	0.022***	0.023***
	(8.904, 0)	(9.78, 0)	(9.832, 0)	(10.203, 0)	(10.158, 0)	(9.971, 0)	(10.142, 0)	(10.082, 0)	(10.078, 0)	(9.952, 0)	(9.599, 0)	(9.741, 0)
$closed_{i,t'}$	-0.001	0.001	0	-0.001	0.002	0.001	0.002	0.003	0.003	0.004	0.004	0.005
	(-0.331, 0.967)	(0.212, 0.967)	(-0.07, 0.933)	(-0.225, 0.933)	(0.714, 0.833)	(0.577, 0.733)	(0.861, 0.8)	(1.452, 0.5)	(1.318, 0.4)	(1.732, 0.367)	(1.825, 0.367)	(2.123, 0.233)
$Dep_{ij,t0}$	-0.546***	-0.459***	-0.415***	-0.418***	-0.38*	-0.34***	-0.34***	-0.329	-0.306	-0.31*	-0.289	-0.274
	(-7.178, 0)	(-6.346, 0)	(-5.841, 0)	(-6.066, 0)	(-5.621, 0.033)	(-5.108, 0)	(-5.094, 0)	(-4.989, 0.067)	(-4.765, 0.067)	(-4.886, 0.033)	(-4.573, 0.067)	(-4.335, 0.067)
$gdp_i + gdp_j$	0.468***	0.477***	0.477***	0.45***	0.435***	0.432***	0.442***	0.444***	0.427***	0.458***	0.463***	0.488***
	(7.523, 0)	(8.05, 0)	(8.229, 0)	(7.977, 0)	(7.797, 0)	(7.808, 0)	(7.95, 0)	(8.017, 0)	(7.871, 0)	(8.57, 0)	(8.73, 0)	(9.222, 0)
$\lvert gdp_i - gdp_j \rvert$	0.039	0.036	0.031*	0.031	0.026	0.021	0.019	0.017	0.016	0.015	0.015	0.015
	(3.1, 0.067)	(3.043, 0.167)	(2.716, 0.033)	(2.775, 0.133)	(2.35, 0.2)	(1.866, 0.2)	(1.747, 0.3)	(1.592, 0.567)	(1.505, 0.4)	(1.421, 0.533)	(1.455, 0.433)	(1.42, 0.533)
$hdi_i + hdi_j$	-2.841***	-2.812***	-2.88***	-2.808***	-2.671***	-2.615***	-2.592***	-2.602***	-2.486***	-2.629***	-2.675***	-2.772***
	(-11.448, 0)	(-11.924, 0)	(-12.491, 0)	(-12.516, 0)	(-12.103, 0)	(-12.003, 0)	(-11.848, 0)	(-11.956, 0)	(-11.695, 0)	(-12.548, 0)	(-12.853, 0)	(-13.334, 0)
$\lvert hdi_i - hdi_j \rvert$	-1.817***	-1.772***	-1.725***	-1.624***	-1.503***	-1.408***	-1.372***	-1.33***	-1.27***	-1.26***	-1.266***	-1.29***
	(-8.717, 0)	(-8.992, 0)	(-8.958, 0)	(-8.653, 0)	(-8.123, 0)	(-7.687, 0)	(-7.451, 0)	(-7.282, 0)	(-7.105, 0)	(-7.136, 0)	(-7.236, 0)	(-7.372, 0)
$distance$	0.053	0.036	0.027	0.031	0.017	0.009	0.01	0.003	-0.005	0.001	0.003	0.001
	(2.836, 0.1)	(2.061, 0.367)	(1.574, 0.533)	(1.867, 0.3)	(1.038, 0.6)	(0.571, 0.8)	(0.636, 0.767)	(0.198, 0.9)	(-0.285, 0.9)	(0.052, 1)	(0.208, 0.933)	(0.081, 1)
$sameregion$	-0.075*	-0.07	-0.073	-0.067***	-0.067	-0.056	-0.051	-0.053	-0.042	-0.043	-0.041	-0.037
	(-2.134, 0.033)	(-2.095, 0.067)	(-2.237, 0.1)	(-2.111, 0)	(-2.157, 0.067)	(-1.835, 0.133)	(-1.652, 0.067)	(-1.746, 0.3)	(-1.408, 0.267)	(-1.48, 0.267)	(-1.403, 0.333)	(-1.29, 0.4)
$log(\theta)$	1.87***	1.99***	2.031***	2.08***	2.102***	2.113***	2.092***	2.103***	2.141***	2.169***	2.18***	2.175***
	(30.275, 0)	(34.183, 0)	(35.963, 0)	(38.048, 0)	(39.499, 0)	(40.794, 0)	(41.236, 0)	(42.265, 0)	(44.132, 0)	(45.32, 0)	(46.019, 0)	(46.358, 0)
Zero-model												
$n_{i,t0} + n_{j,t0}$	-3.412***	-3.282***	-3.251***	-3.18***	-3.156***	-3.149***	-3.14***	-3.138***	-3.127***	-3.134***	-3.123***	-3.117***
	(-20.941, 0)	(-21.662, 0)	(-22.035, 0)	(-22.447, 0)	(-22.727, 0)	(-23.079, 0)	(-23.249, 0)	(-23.437, 0)	(-23.794, 0)	(-23.887, 0)	(-23.933, 0)	(-23.964, 0)
$\lvert n_{i,t0} - n_{j,t0} \rvert$	1.338***	1.269***	1.251***	1.194***	1.186***	1.176***	1.17***	1.168***	1.16***	1.158***	1.158***	1.162***
	(14.96, 0)	(15.381, 0)	(15.683, 0)	(15.839, 0)	(16.186, 0)	(16.528, 0)	(16.762, 0)	(17.005, 0)	(17.447, 0)	(17.569, 0)	(17.741, 0)	(17.898, 0)
$c_{i,t0} + c_{j,t0}$	-0.921***	-0.941***	-0.909***	-0.95***	-0.973***	-0.982***	-0.994***	-1.003***	-1.023***	-1.048***	-1.084***	-1.087***
	(-6.707, 0)	(-7.149, 0)	(-7.017, 0)	(-7.459, 0)	(-7.717, 0)	(-7.856, 0)	(-7.989, 0)	(-8.103, 0)	(-8.353, 0)	(-8.542, 0)	(-8.84, 0)	(-8.871, 0)
$\lvert c_{i,t0} - c_{j,t0} \rvert$	1.166***	1.223***	1.214***	1.27***	1.297***	1.32***	1.338***	1.353***	1.382***	1.408***	1.431***	1.42***
	(10.767, 0)	(11.844, 0)	(12.006, 0)	(12.798, 0)	(13.23, 0)	(13.6, 0)	(13.877, 0)	(14.122, 0)	(14.582, 0)	(14.855, 0)	(15.112, 0)	(15.045, 0)
$distance$	0.759***	0.763***	0.767***	0.796***	0.776***	0.778***	0.78***	0.787***	0.789***	0.812***	0.823***	0.834***
	(10.509, 0)	(11.087, 0)	(11.338, 0)	(11.964, 0)	(11.825, 0)	(11.972, 0)	(12.073, 0)	(12.241, 0)	(12.4, 0)	(12.74, 0)	(12.909, 0)	(13.079, 0)
$sameregion$	-1.438***	-1.31***	-1.247***	-1.174***	-1.164***	-1.142***	-1.107***	-1.079***	-1.056***	-1.047***	-1.041***	-1.02***
	(-9.866, 0)	(-9.823, 0)	(-9.66, 0)	(-9.445, 0)	(-9.579, 0)	(-9.572, 0)	(-9.407, 0)	(-9.269, 0)	(-9.239, 0)	(-9.194, 0)	(-9.181, 0)	(-9.023, 0)
obs.	12090	12090	12090	12090	12090	12090	12090	12090	12090	12090	12090	12090
loglik	-8005	-8190	-8261	-8367	-8449	-8528	-8593	-8655	-8761	-8776	-8804	-8837

— P-values are based on MRQAP (1000 permutations) and one-sided (because null distributions are not symmetric). One, two, and three stars signal significance values below 5%, 1%, and 0.1% respectively.
— Recall from Section 1.2.4 that n and c stands for non-CRR and CRR respectively. Indices i indicate a country, and ij a country dyad. Country-level variables are joined to capture the country dyad's sum ($x_i + x_j$) and their absolute difference ($|x_i - x_j|$). $log(\theta)$ captures over-dispersion in the count model.
— The main finding is that in February initial CRR competence ($c_{i,t0}$) is positively and non-CRR ($n_{i,t0}$) is negatively associated with collaboration. Prior collaborations in non-CRR ($n_{ij,t0}$) are relevant throughout. In the second half of the year, countries with strong non-CRR competence ($n_{i,t0}$) attract collaborations, and less developed countries tend to collaborate more among each other ($hdi_i + hdi_j$ and $|hdi_i - hdi_j|$). Same region effect is present in all months.

Table 1.13: Zero-inflated negative binomial model of (accumulated) joint coronavirus related papers ($c_{ij,t'}$) during the pandemic, est. coefficient (z-value, p-value)

	Jan. 22	Feb. 22	Mar. 22	Apr. 22	May 22	Jun. 22	Jul. 22	Aug. 22	Sep. 22	Oct. 22	Nov. 22	Dec. 22
Count-model												
$n_{ij,t0}$	0.75***	0.743***	0.738***	0.738***	0.728***	0.716***	0.713***	0.719***	0.716***	0.716***	0.714***	0.708***
	(40.632, 0)	(40.686, 0)	(40.405, 0)	(40.445, 0)	(40.363, 0)	(39.752, 0)	(39.549, 0)	(39.507, 0)	(39.359, 0)	(39.185, 0)	(39.109, 0)	(38.581, 0)
$c_{ij,t0}$	0.008***	0.014	0.017	0.014	0.021	0.03	0.032	0.029	0.032	0.033	0.035	0.04
	(0.27, 0)	(0.465, 0.233)	(0.543, 0.267)	(0.454, 0.4)	(0.693, 0.3)	(0.979, 0.167)	(1.046, 0.233)	(0.972, 0.4)	(1.067, 0.467)	(1.081, 0.333)	(1.168, 0.1)	(1.326, 0.3)
$n_{i,t0}$	0.051	0.05	0.049	0.049***	0.048	0.051	0.052	0.042	0.04	0.038	0.039	0.038
	(3.71, 0.167)	(3.663, 0.067)	(3.564, 0.133)	(3.571, 0)	(3.493, 0.2)	(3.694, 0.067)	(3.743, 0.067)	(3.034, 0.067)	(2.84, 0.133)	(2.713, 0.1)	(2.712, 0.167)	(2.677, 0.133)
$c_{i,t0}$	0.02	0.021	0.022	0.021	0.02	0.019	0.018	0.021	0.021	0.022	0.021	0.023*
	(2.881, 0.2)	(3.006, 0.167)	(3.13, 0.067)	(2.984, 0.167)	(2.927, 0.167)	(2.665, 0.3)	(2.521, 0.167)	(2.967, 0.1)	(3.037, 0.1)	(3.127, 0.067)	(3.059, 0.1)	(3.196, 0.033)
$death_i + death_j$	0.099***	0.102***	0.104***	0.093***	0.09***	0.092***	0.087***	0.091***	0.095***	0.097***	0.092*	0.091***
	(6.062, 0)	(6.073, 0)	(6.063, 0)	(5.214, 0)	(5.16, 0)	(5.086, 0)	(4.738, 0)	(4.945, 0)	(5.14, 0)	(5.166, 0)	(4.951, 0.033)	(4.885, 0)
$\lvert death_i - death_j \rvert$	-0.039	-0.042*	-0.044*	-0.038	-0.034	-0.037	-0.032	-0.035	-0.035	-0.038	-0.038	-0.037
	(-3.46, 0.2)	(-3.562, 0.033)	(-3.751, 0.033)	(-3.231, 0.133)	(-3.083, 0.067)	(-3.154, 0.133)	(-2.771, 0.1)	(-3.023, 0.067)	(-3.106, 0.2)	(-3.271, 0.167)	(-3.306, 0.167)	(-3.235, 0.1)
$locked_{i,t'}$	0.023***	0.023***	0.022***	0.022***	0.022***	0.022***	0.022***	0.022***	0.021***	0.021***	0.021***	0.02***
	(9.904, 0)	(9.803, 0)	(9.728, 0)	(9.703, 0)	(9.636, 0)	(9.462, 0)	(9.478, 0)	(9.439, 0)	(9.09, 0)	(9.048, 0)	(9.005, 0)	(8.773, 0)
$closed_{i,t'}$	0.005	0.005	0.005	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.007
	(2.398, 0.367)	(2.202, 0.267)	(2.318, 0.333)	(2.521, 0.367)	(2.72, 0.333)	(2.654, 0.333)	(2.664, 0.333)	(2.677, 0.2)	(2.955, 0.167)	(2.822, 0.1)	(2.881, 0.167)	(3.14, 0.167)
$Dep_{ij,t0}$	-0.25	-0.232	-0.209	-0.197	-0.192	-0.183	-0.173	-0.161	-0.144	-0.139	-0.124	-0.105
	(-3.91, 0.167)	(-3.652, 0.167)	(-3.286, 0.241)	(-3.086, 0.25)	(-3.034, 0.138)	(-2.883, 0.077)	(-2.733, 0.276)	(-2.541, 0.138)	(-2.275, 0.276)	(-2.187, 0.333)	(-1.949, 0.5)	(-1.642, 0.367)
$gdp_i + gdp_j$	0.504***	0.5***	0.484***	0.488***	0.49***	0.497***	0.503***	0.523***	0.508***	0.513***	0.519***	0.509***
	(9.435, 0)	(9.412, 0)	(9.105, 0)	(9.172, 0)	(9.279, 0)	(9.37, 0)	(9.465, 0)	(9.788, 0)	(9.477, 0)	(9.541, 0)	(9.626, 0)	(9.361, 0)
$\lvert gdp_i - gdp_j \rvert$	0.015	0.017	0.018	0.019	0.018	0.02	0.019	0.02	0.019	0.02	0.02	0.022
	(1.374, 0.4)	(1.587, 0.467)	(1.721, 0.333)	(1.793, 0.367)	(1.758, 0.3)	(1.844, 0.2)	(1.828, 0.433)	(1.869, 0.267)	(1.82, 0.4)	(1.907, 0.133)	(1.91, 0.367)	(2.055, 0.2)
$hdi_i + hdi_j$	-2.88***	-2.883***	-2.846***	-2.879***	-2.858***	-2.895***	-2.926***	-3.006***	-2.945***	-2.958***	-3.018***	-2.987***
	(-13.739, 0)	(-13.829, 0)	(-13.641, 0)	(-13.805, 0)	(-13.818, 0)	(-13.967, 0)	(-14.11, 0)	(-14.447, 0)	(-14.148, 0)	(-14.14, 0)	(-14.391, 0)	(-14.111, 0)
$\lvert hdi_i - hdi_j \rvert$	-1.309***	-1.311***	-1.307***	-1.322***	-1.314***	-1.334***	-1.363***	-1.4***	-1.368***	-1.395***	-1.431***	-1.452***
	(-7.416, 0)	(-7.478, 0)	(-7.452, 0)	(-7.557, 0)	(-7.563, 0)	(-7.66, 0)	(-7.828, 0)	(-8.027, 0)	(-7.85, 0)	(-7.988, 0)	(-8.19, 0)	(-8.224, 0)
distance	0.002	0	-0.007	-0.008	-0.012	-0.014	-0.021	-0.023	-0.025	-0.027	-0.027	-0.033
	(0.131, 0.967)	(-0.024, 0.967)	(-0.434, 0.667)	(-0.535, 0.733)	(-0.766, 0.8)	(-0.942, 0.533)	(-1.379, 0.433)	(-1.49, 0.4)	(-1.647, 0.3)	(-1.743, 0.5)	(-1.796, 0.4)	(-2.145, 0.167)
sameregion	-0.035	-0.025	-0.028	-0.035	-0.036	-0.038	-0.037	-0.039	-0.043	-0.047	-0.052	-0.052
	(-1.186, 0.4)	(-0.879, 0.667)	(-0.967, 0.433)	(-1.2, 0.3)	(-1.272, 0.233)	(-1.335, 0.433)	(-1.277, 0.233)	(-1.342, 0.367)	(-1.484, 0.233)	(-1.618, 0.1)	(-1.811, 0.133)	(-1.8, 0.133)
$log(\theta)$	2.15***	2.16***	2.155***	2.153***	2.166***	2.161***	2.16***	2.166***	2.165***	2.163***	2.167***	2.161***
	(46.331, 0)	(47.013, 0)	(47.423, 0)	(47.832, 0)	(48.585, 0)	(48.83, 0)	(49.048, 0)	(49.274, 0)	(49.439, 0)	(49.45, 0)	(49.619, 0)	(49.53, 0)
Zero-model												
$n_{i,t0} + n_{j,t0}$	-3.125***	-3.136***	-3.124***	-3.083***	-3.079***	-3.102***	-3.093***	-3.098***	-3.087***	-3.061***	-3.022***	-3***
	(-24.036, 0)	(-24.152, 0)	(-24.189, 0)	(-24.06, 0)	(-24.109, 0)	(-24.207, 0)	(-24.187, 0)	(-24.178, 0)	(-24.095, 0)	(-23.94, 0)	(-23.657, 0)	(-23.425, 0)
$\lvert n_{i,t0} - n_{j,t0} \rvert$	1.165***	1.166***	1.164***	1.137***	1.132***	1.141***	1.135***	1.139***	1.137***	1.12***	1.091***	1.07***
	(18.024, 0)	(18.146, 0)	(18.241, 0)	(17.998, 0)	(18.046, 0)	(18.227, 0)	(18.204, 0)	(18.283, 0)	(18.268, 0)	(18.04, 0)	(17.583, 0)	(17.217, 0)
$c_{i,t0} + c_{j,t0}$	-1.072***	-1.077***	-1.077***	-1.104***	-1.117***	-1.1***	-1.085***	-1.079***	-1.102***	-1.113***	-1.152***	-1.183***
	(-8.75, 0)	(-8.799, 0)	(-8.816, 0)	(-9.047, 0)	(-9.166, 0)	(-9.017, 0)	(-8.903, 0)	(-8.841, 0)	(-9.013, 0)	(-9.081, 0)	(-9.36, 0)	(-9.552, 0)
$\lvert c_{i,t0} - c_{j,t0} \rvert$	1.409***	1.422***	1.414***	1.437***	1.453***	1.449***	1.439***	1.435***	1.444***	1.447***	1.473***	1.495***
	(14.957, 0)	(15.108, 0)	(15.068, 0)	(15.335, 0)	(15.519, 0)	(15.502, 0)	(15.423, 0)	(15.384, 0)	(15.459, 0)	(15.486, 0)	(15.703, 0)	(15.856, 0)
distance	0.836***	0.836***	0.839***	0.831***	0.832***	0.839***	0.835***	0.85***	0.851***	0.837***	0.827***	0.838***
	(13.117, 0)	(13.137, 0)	(13.205, 0)	(13.117, 0)	(13.157, 0)	(13.259, 0)	(13.207, 0)	(13.41, 0)	(13.406, 0)	(13.199, 0)	(13.008, 0)	(13.121, 0)
sameregion	-1.019***	-1.025***	-1.023***	-1.002***	-1.012***	-0.984***	-0.981***	-0.957***	-0.947***	-0.957***	-0.958***	-0.951***
	(-9.03, 0)	(-9.109, 0)	(-9.121, 0)	(-8.96, 0)	(-9.076, 0)	(-8.832, 0)	(-8.818, 0)	(-8.606, 0)	(-8.506, 0)	(-8.584, 0)	(-8.572, 0)	(-8.478, 0)
obs.	12090	12090	12090	12090	12090	12090	12090	12090	12090	12090	12090	12090
loglik	-8874	-8921	-8967	-8997	-9041	-9072	-9083	-9063	-9061	-9041	-9002	-8957

— P-values are based on MRQAP (1000 permutations) and one-sided (because null distributions are not symmetric). One, two, and three stars signal significance values below 5%, 1%, and 0.1% respectively.

— Recall from Section 1.2.4 that n and c stands for non-CRR and CRR respectively. Indices i indicate a country, and ij a country dyad. Country-level variables are joined to capture the country dyad's sum ($x_i + x_j$) and their absolute difference ($|x_i - x_j|$). $log(\theta)$ captures over-dispersion in the count model.

— The main finding is that in February initial CRR competence ($c_{i,t0}$) is positively and non-CRR ($n_{i,t0}$) is negatively associated with collaboration. Prior collaborations in non-CRR ($n_{ij,t0}$) are relevant throughout. In the second half of the year, countries with strong non-CRR competence ($n_{i,t0}$) attract collaborations, and less developed countries tend to collaborate more among each other ($hdi_i + hdi_j$ and $|hdi_i - hdi_j|$). Same region effect is present in all months.

Table 1.14: Marginal effects of zero-inflated negative binomial regression of (accumulated) joint CRR papers (in percent).

	Jan. '20	**Feb. '20**	**Mar. '20**	**Apr. '20**	**May '20**	**June '20**	**July '20**	**Aug. '20**	**Sept. '20**	**Oct. '20**	**Nov. '20**	**Dec. '20**
Count-model												
$c_{ij,t0}$	12.94	12.47	-4.60	-11.57	-13.77	-13.51	-14.38	-13.65	-13.90	-12.50	-11.03	-9.51
$n_{ij,t0}$	43.31	54.92	32.88	37.71	41.92	41.60	42.42	41.85	41.64	40.76	39.54	38.07
$n_{i,t0}$	14.29	2.03	-0.91	0.49	0.62	1.01	1.50	1.85	2.28	2.39	2.56	2.11
$c_{i,t0}$	-0.34	-0.24	0.25	0.10	0.05	0.02	0.01	0.01	-0.00	-0.01	-0.03	-0.01
$Dep_{ij,t0}$	-54.51	-32.66	-46.53	-43.02	-37.53	-41.32	-43.44	-40.21	-42.46	-44.70	-45.51	-43.92
$death_i + death_j$		101.04	-3.11	0.28	0.05	0.08	0.11	0.07	0.11	0.12	0.14	0.16
$\|death_i - death_j\|$		-44.85	1.82	-0.39	-0.05	-0.05	-0.06	-0.02	-0.05	-0.04	-0.05	-0.07
$closed_{i,t'}$				-0.04	-0.02	-0.07	0.07	0.06	-0.03	-0.06	-0.13	-0.08
$locked_{i,t'}$			3.27	0.77	0.64	0.63	0.61	0.52	0.57	0.58	0.53	0.47
$gdp_i + gdp_j$	-0.56	0.42	0.36	0.23	0.19	0.21	0.28	0.33	0.34	0.36	0.38	0.42
$\|gdp_i - gdp_j\|$	0.02	0.12	0.05	0.07	0.07	0.06	0.06	0.05	0.05	0.04	0.04	0.04
distance	1.21	0.28	0.01	0.07	0.08	0.08	0.07	0.07	0.04	0.04	0.05	0.06
region	202.33	4.18	-2.49	-4.36	-3.80	-7.72	-8.32	-7.98	-7.62	-6.02	-6.09	-4.75
$hdi_i + hdi_j$	-2.41	-5.94	-4.07	-2.71	-2.68	-2.70	-3.29	-3.66	-3.57	-3.50	-3.64	-3.85
$\|hdi_i - hdi_j\|$	0.59	-0.41	-0.41	-0.54	-0.49	-0.49	-0.51	-0.52	-0.49	-0.47	-0.45	-0.47
zero-model												
$c_{i,t0} + c_{j,t0}$	-1.20	0.12	-0.27	0.34	-0.22	-0.44	-0.43	-0.38	-0.36	-0.38	-0.38	-0.37
$\|c_{i,t0} - c_{j,t0}\|$	0.71	-1.40	-0.64	0.36	0.92	1.18	1.14	0.98	0.89	0.94	0.96	0.97
$n_{i,t0} + n_{j,t0}$	0.00	-0.64	-2.16	-1.63	-1.38	-1.13	-1.10	-1.00	-0.93	-0.81	-0.68	-0.62
$\|n_{i,t0} - n_{j,t0}\|$	-0.03	0.02	1.81	0.62	0.57	0.49	0.48	0.44	0.41	0.36	0.29	0.26
distance	0.30	-0.08	0.06	0.26	0.28	0.23	0.23	0.20	0.19	0.18	0.14	0.13
region	22.65	-50.03	-52.18	-31.91	-28.53	-24.46	-22.50	-21.74	-20.49	-18.17	-16.13	-14.94

Notes: Average marginal effects have been obtained by comparing for each variable predictions based on observations with predictions on observed increased by one percent. In case observed value is zero, we compare it with having one. Marginal effects of the dummy 'same region' are obtained by comparing zero with one outcomes.

Table 1.15: Marginal effects of zero-inflated negative binomial regression of (accumulated) joint CRR papers (in percent).

	Jan. '20	**Feb. '20**	**Mar. '20**	**Apr. '20**	**May '20**	**June '20**	**July '20**	**Aug. '20**	**Sept. '20**	**Oct. '20**	**Nov. '20**	**Dec. '20**
Count-model												
$c_{ij,t0}$	-8.42	-5.56	-5.31	-4.28	-3.81	-3.27	-2.60	-1.91	-0.62	-0.13	0.25	0.56
$n_{ij,t0}$	36.86	33.53	33.19	31.99	31.01	30.61	29.84	29.37	27.44	27.45	27.40	27.03
$n_{i,t0}$	1.95	1.57	1.61	1.57	1.35	1.42	1.43	1.22	1.03	0.85	0.90	0.95
$c_{i,t0}$	-0.00	0.02	0.02	0.03	0.04	0.03	0.03	0.04	0.06	0.06	0.05	0.05
$Dep_{ij,t0}$	-40.02	-35.05	-32.32	-32.56	-30.10	-27.48	-27.48	-26.75	-25.15	-25.40	-23.93	-22.87
$death_i + death_j$	0.14	0.12	0.11	0.10	0.10	0.10	0.11	0.12	0.11	0.11	0.11	0.11
$\lvert death_i - death_j \rvert$	-0.07	-0.07	-0.07	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05
$closed_{i,t'}$	-0.02	0.01	-0.00	-0.01	0.04	0.03	0.04	0.07	0.07	0.08	0.09	0.10
$locked_{i,t'}$	0.44	0.44	0.42	0.41	0.39	0.38	0.38	0.36	0.35	0.34	0.31	0.31
$gdp_i + gdp_j$	0.47	0.48	0.48	0.45	0.43	0.43	0.44	0.44	0.43	0.46	0.46	0.49
$\lvert gdp_i - gdp_j \rvert$	0.04	0.04	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.01	0.01	0.01
distance	0.05	0.04	0.03	0.03	0.02	0.01	0.01	0.00	-0.00	0.00	0.00	0.00
region	-5.49	-5.10	-5.34	-4.92	-4.94	-4.16	-3.76	-3.92	-3.09	-3.20	-3.02	-2.78
$hdi_i + hdi_j$	-4.07	-4.02	-4.12	-4.02	-3.83	-3.74	-3.71	-3.72	-3.57	-3.76	-3.83	-3.96
$\lvert hdi_i - hdi_j \rvert$	-0.47	-0.46	-0.45	-0.43	-0.40	-0.38	-0.37	-0.36	-0.35	-0.35	-0.35	-0.36
zero-model												
$c_{i,t0} + c_{j,t0}$	-0.36	-0.36	-0.34	-0.34	-0.34	-0.33	-0.33	-0.33	-0.33	-0.33	-0.34	-0.33
$\lvert c_{i,t0} - c_{j,t0} \rvert$	1.01	1.04	1.01	1.04	1.04	1.05	1.05	1.06	1.08	1.09	1.10	1.08
$n_{i,t0} + n_{j,t0}$	-0.58	-0.53	-0.51	-0.48	-0.47	-0.46	-0.45	-0.44	-0.43	-0.43	-0.42	-0.42
$\lvert n_{i,t0} - n_{j,t0} \rvert$	0.23	0.21	0.20	0.19	0.18	0.17	0.17	0.17	0.16	0.16	0.16	0.16
distance	0.13	0.12	0.12	0.12	0.12	0.11	0.11	0.11	0.11	0.11	0.11	0.11
region	-13.62	-12.14	-11.42	-10.50	-10.22	-9.88	-9.44	-9.12	-8.78	-8.56	-8.41	-8.16

Notes: Average marginal effects have been obtained by comparing for each variable predictions based on observations with predictions on observed increased by one percent. In case observed value is zero, we compare it with having one. Marginal effects of the dummy 'same region' are obtained by comparing zero with one outcomes.

Table 1.16: Marginal effects of zero-inflated negative binomial regression of (accumulated) joint CRR papers (in percent).

	Jan. '20	**Feb. '20**	**Mar. '20**	**Apr. '20**	**May '20**	**June '20**	**July '20**	**Aug. '20**	**Sept. '20**	**Oct. '20**	**Nov. '20**	**Dec. '20**
Count-model												
$c_{ij,t0}$	0.60	1.01	1.17	0.98	1.46	2.06	2.20	2.04	2.24	2.27	2.45	2.79
$n_{ij,t0}$	27.05	26.70	26.46	26.48	26.03	25.48	25.35	25.60	25.47	25.49	25.40	25.12
$n_{i,t0}$	1.14	1.18	1.23	1.17	1.14	1.04	0.98	1.16	1.19	1.23	1.20	1.27
$c_{i,t0}$	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.04
$Dep_{ij,t0}$	-21.07	-19.75	-18.00	-17.05	-16.68	-15.94	-15.17	-14.18	-12.81	-12.36	-11.10	-9.48
$death_i + death_j$	0.10	0.10	0.10	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09
$\lvert death_i - death_j \rvert$	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
$closed_{i,t'}$	0.12	0.11	0.11	0.12	0.13	0.13	0.13	0.13	0.14	0.14	0.14	0.15
$locked_{i,t'}$	0.32	0.31	0.31	0.31	0.30	0.30	0.30	0.30	0.29	0.28	0.28	0.27
$gdp_i + gdp_j$	0.50	0.50	0.48	0.49	0.49	0.50	0.50	0.52	0.50	0.51	0.52	0.51
$\lvert gdp_i - gdp_j \rvert$	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
distance	0.00	-0.00	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.03	-0.03	-0.03	-0.03
region	-2.59	-1.91	-2.09	-2.59	-2.72	-2.85	-2.73	-2.87	-3.17	-3.46	-3.86	-3.86
$hdi_i + hdi_j$	-4.12	-4.12	-4.07	-4.11	-4.08	-4.14	-4.18	-4.30	-4.21	-4.23	-4.31	-4.27
$\lvert hdi_i - hdi_j \rvert$	-0.36	-0.36	-0.36	-0.36	-0.36	-0.37	-0.37	-0.38	-0.37	-0.38	-0.39	-0.39
zero-model												
$c_{i,t0} + c_{j,t0}$	-0.33	-0.32	-0.32	-0.32	-0.32	-0.32	-0.31	-0.31	-0.31	-0.30	-0.30	-0.30
$\lvert c_{i,t0} - c_{j,t0} \rvert$	1.06	1.06	1.05	1.06	1.07	1.06	1.04	1.03	1.03	1.01	1.01	1.01
$n_{i,t0} + n_{j,t0}$	-0.42	-0.42	-0.41	-0.40	-0.40	-0.40	-0.39	-0.39	-0.39	-0.38	-0.37	-0.37
$\lvert n_{i,t0} - n_{j,t0} \rvert$	0.16	0.16	0.16	0.15	0.15	0.15	0.15	0.15	0.14	0.14	0.14	0.13
distance	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.10	0.10	0.10
region	-8.09	-8.08	-8.02	-7.82	-7.83	-7.55	-7.48	-7.24	-7.13	-7.15	-7.10	-6.95

Notes: Average marginal effects have been obtained by comparing for each variable predictions based on observations with predictions on observed increased by one percent. In case observed value is zero, we compare it with having one. Marginal effects of the dummy 'same region' are obtained by comparing zero with one outcomes.

Chapter 2

Novelpy: A *Python* Package To Measure Novelty And Disruption Of Bibliometric And Patent Data

This chapter was co-authored with

Pierre Pelletier

Summary of the chapter

Novelpy (v1.2) is an open-source *Python* package designed to compute bibliometric indicators. The package aims to provide a tool for the scientometrics community that centralizes various measures of novelty and disruptiveness, enables their comparison, and fosters reproducibility. This paper offers a comprehensive review of the different indicators available in *Novelpy* by formally describing these measures (both mathematically and graphically) and presenting their advantages and limitations. We then compare the different measures on a random sample of 1.5M articles drawn from the Pubmed Knowledge Graph to demonstrate the module's capabilities. We encourage anyone interested to participate in the development of future versions.

2.1 Introduction

Identifying and tracking relevant pieces of knowledge remains a core issue in the Science of Science research. A better understanding of knowledge flow dynamics, mechanisms behind the emergence of new ideas, and identification of novel or impactful documents are crucial for fostering effective science, which will, in turn, help address future societal challenges [Fortunato et al., 2018, Foster et al., 2021, OECD, 2021]. This article proposes integrating various bibliometric indicators within a *Python* package. It assembles within a single module novelty or, more broadly, creativity measurements through combinatorial novelty indicators [Uzzi et al., 2013, Foster et al., 2015, Lee et al., 2015, Wang et al., 2017, Shibayama et al., 2021], as well as several impact measures, including disruption metrics [Wu et al., 2019, Wu and Yan, 2019, Wu and Wu, 2019, Bu et al., 2019, Bornmann et al., 2019a].

This module is intended for researchers in the emerging and multidisciplinary field of Science of Science. There is an increasing tendency to create new scientometric indicators, but fewer initiatives exist to design reproducible experiments. For novelty indicators, there is minimal reference to prior approaches when creating a new indicator; thus, the flexibility in the choice of measures raises the temptation to choose the measure that produces the intended outcome [Foster et al., 2021]. Only a few studies attempt to establish a conceptual background of creativity and the formalization of the indicators [Foster et al., 2021]. This article provides a mathematical and graphical description of these indicators. To the best of our knowledge, it is the first tool that enables the computation of these metrics.

Two macro types of analysis can describe Scientometrics: performance analysis and Science Mapping Analysis (SMA) [Moral Muñoz et al., 2020]. Performance analysis aims to assess the activities of scientific actors and their impact. Its purpose is to assign a value to the productivity and pervasiveness of research conducted by a unit (article, author, institution). SMA "is mostly directed at monitoring a scientific field to determine its (cognitive) structure, its evolution, and main actors within" [Noyons et al., 1999]. It captures a snapshot of a part of the scientific system at a given moment to analyze its structure. The present package allows analysis through disruption measures and assesses papers' originality potential using novelty indicators. Both metrics require science mapping analysis to be measured since they are generated through maps of the structure of science. Inputs, outputs, and impacts of these scientific activities are the three perspectives used in bibliometric analysis

[Sugimoto and Larivière, 2018].[1] Entities involved in most combinatorial novelty indicators use only the output part of documents to compute their measures [Uzzi et al., 2013, Foster et al., 2015, Lee et al., 2015, Shibayama et al., 2021], except for Wang et al. [2017], which uses references from future articles to control for re-utilization. Disruption indicators [Wu et al., 2019, Bu et al., 2019, Bornmann et al., 2020] take the outputs and impacts of a given document to construct their metrics. They are based on both the references and citations of a given document. This module focuses on metrics using outputs (references/keywords) and impact features (citations/references and keywords from future articles).

While citation is an invaluable source of information, several limitations exist when using the sheer number of citations to evaluate impact. Inter-field (and even intra-field) comparisons can be challenging, as the sheer number of scientists and the way science is performed vary significantly depending on the research domain (methodology, solo author vs. team publication, citation habits). The gap in the number of citations is mainly due to the field's structure and does not necessarily represent the documents' quality. This phenomenon becomes an issue when raw numbers are used to measure the importance of research [Purkayastha et al., 2019]. The same problem arises with self-citation, comparing national and international journals, or document languages [Van Leeuwen et al., 2001].

Network effects have been observed in citation dynamics. Wallace et al. [2012] showed that scholars tend to cite researchers with whom they have a deeper social connection. They also found that researchers are more likely to cite collaborators of collaborators, thereby creating a citation continuum. Articles with international collaborations are more cited due to network effects [Wagner et al., 2019]. Other negative citation behaviors arise in Bornmann and Daniel [2008]; scholars tend to cite papers to satisfy editors and reviewers, showing an apparent disconnection between citation and actual importance during the creation process. Field-specific issues can be addressed using normalization methods or different counting methods of citations (see Waltman [2016] for a comprehensive review). One family of normalized indicators is disruptiveness [Wu et al., 2019, Wu and Yan, 2019, Wu and Wu, 2019, Bu et al., 2019, Bornmann et al., 2019a]. These measures analyze how a focal article

[1] Input refers to human and financial resources and captures the different interactions of agents in the system at various levels (authors/institutional/country levels). Output results from the research process, the different entities that characterize a document. Finally, impact measures knowledge dissemination generated by an article through citations, attention by the general public, or re-utilization of a document's components.

acts as a bottleneck between future papers and the references of the focal papers. They capture whether a document consolidates a domain (i.e., future papers rely on the same pieces of knowledge as the focal paper) or constitutes a starting point for documents from various areas (i.e., future papers only use information from the document).

Scientific advancement is the result of individuals' creativity, where *creativity* is defined as "*held to involve the production of high-quality, original, and elegant solutions to complex, novel, ill-defined, or poorly structured problems*" [Hemlin et al., 2013]. Scholars have proposed measurements to complement these impact indicators with creativity indicators, usually called "atypicality", originality", or "novelty" indicators. The need for quantifying novelty comes from its position as an essential component of the structure of the scientific and economic system. Novelty is at the origin of peer recognition, which acts as a "reward system" for individuals. The "priority rule" grants recognition to the first person making the discovery [Merton, 1957, Carayol et al., 2019]. Novelty is also at the core of the theory developed in evolutionary economics, in which technological progress and creativity influence the cyclical nature of the economy [Schumpeter et al., 1939, Nelson, 1985, Amendola et al., 2014]. Scientific progress remains elusive, and novelty indicators are intended to approach creativity, as making relevant novel combinations is perceived as innovative [Burt, 2004, Rodríguez-Navarro, 2016, Bornmann et al., 2019b]. The earliest novelty indicators focused mainly on past information (i.e., using an entity created the same year) or the distance between articles from a given year, based on their references' overlapping [Dahlin and Behrens, 2005].

More recently, scholars have integrated the conceptual framework of knowledge recombination (a combination of pre-existing ideas that leads to invention) into novelty indicators. This concept was already developed by Poincaré [1910]. Although he refers to the specific science case, it can be extended to any non-scientific creative process where combinations can be both material and conceptual [Winter and Nelson, 1982]. Weitzman [1998] discussed how knowledge could be generated through a combinatorial process of past ideas and how this can generate economic growth as long as potential new ideas are exploitable. At the same time, an invention does not necessarily arise from combining two components for the first time. Indeed, it can also occur from creating a new relationship between two already linked components [Schumpeter et al., 1939, Henderson and Clark, 1990]. This deepens the idea brought by Jacob [1977] that scientific advancement emerges from looking at something from

a new angle rather than incorporating a new instrument. Scientists have proposed a more probability-based approach to capture this combinatorial process. Instead of focusing solely on the degree of novelty of a combination, they look at how unlikely this combination is to happen. The more distant the items in the combination, the more complex and unlikely it is to make this combination. Therefore, the combination is more novel. To solve mathematical problems, Poincaré used the knowledge he found in another field [Poincaré, 1910]. The more distant the fields were, the more insight he gained. However, novel documents exhibit higher variance in citation performance. Academics adopting an exploration strategy face a higher risk of failure [Fleming, 2001, Foster et al., 2015, Wang et al., 2017, OECD, 2021]. Scientific documents that have a fair mix of novel and conventional ideas are more likely to be "sleeping beauties" than other documents (see Ke et al. [2015] and Wang et al. [2017]). The idea of March [1991] that organizations that explore and consolidate existing processes/technologies are more likely to survive can also be applied in the scientific realm.[2] Novelty indicators can be applied to different entities (patents, papers, webpages, etc.) using various units of knowledge (references, keywords, MeSH terms, text, and others).

Most packages available in *R* and *Python* deal with performance or SMA. Moral Muñoz et al. [2020] carried out a detailed and up-to-date review of the tools and libraries that help researchers in their daily work. Although much work has been done to study citation or co-authorship, novelty and disruption indicators are still unavailable, and researchers have to code these metrics themselves. Concerning the reproducibility of novelty studies, only Shibayama et al. [2021] shared their code on Github to calculate their new novelty indicator, but this is still an isolated event. This tool, therefore, ensures that indicators of novelty and disruption used in future studies will be replicable.

The rationale for incorporating novelty and disruption indicators in a single package comes from the fact that they both capture different aspects of the documents: the former aims at quantifying the risky profile of research, looking at the balance between exploitation and exploration [March, 1991] of the knowledge space. At the same time, the latter analyzes how impactful an article is for science. The link between novelty and citation count has been of interest in previous research [Uzzi et al., 2013, Wang et al., 2017], and more recently, Lin [2021] studied the relationship between novelty and disruption indicators. The different studies only look at specific

[2]Here, survival can be expressed as a high citation count.

novelty indicators; a complete benchmark is still missing. This paper contributes to an ongoing effort to systematically benchmark and compare multiple indicators of impact and novelty by proposing an open-source tool to the community.

This article contributes to the Science of Science literature by providing an open-source *Python* package, *Novelpy*, to compute Novelty and Disruption measurements. It unifies the existing indicators in a common framework using a formalization based on graph theory and provides some hands-on experience. We hope that *Novelpy* will contribute to homogenizing our practice in the science of science and support researchers in their work. The package will be available in *Python*, one of the most popular open-source programming languages (hence with the most prominent community support), and will be maintained long-term. The package currently works with a specific and documented data structure, but tools to easily use well-known data sources are under development. The package will be hosted on *PyPI* and also on *Github*, which allows the creation of bug reporting and/or proposition of development.[3] The rest of the paper is structured in the following way. Section 2.2 contains the formalization of the indicators implemented in *Novelpy*. Section 2.3.2 demonstrates the package's capabilities on a random sample drawn from PubMed. We close the paper with a discussion on the remaining limitations of novelty indicators' usages and the purpose of the package.

2.2 Supported indicators

This section details the content of *Novelpy*, describes the computation for each indicator, and the data required. The *Novelpy* Python package provides a set of functions to perform quantitative analysis in scientometrics. The structure of the module is divided between novelty and disruption indicators. Novelty indicators are also separated between indicators based on co-occurrence matrices and ones based on text embedding techniques, as represented in figure 2.1.

Practically, disruption indicators are all calculated through the same function, while novelty indicators have a function for each measure. All functions are explained in the module's documentation (`https://novelpy.readthedocs.io/`).

Different data types can be employed depending on the indicator, as shown in 2.1. All indicators working with a co-occurrence matrix can use references, journals, or keywords, and disruption indices rely on the citation network. Shibayama et al.

[3]Documentation is available here `https://novelpy.readthedocs.io/en/latest/usage.html`

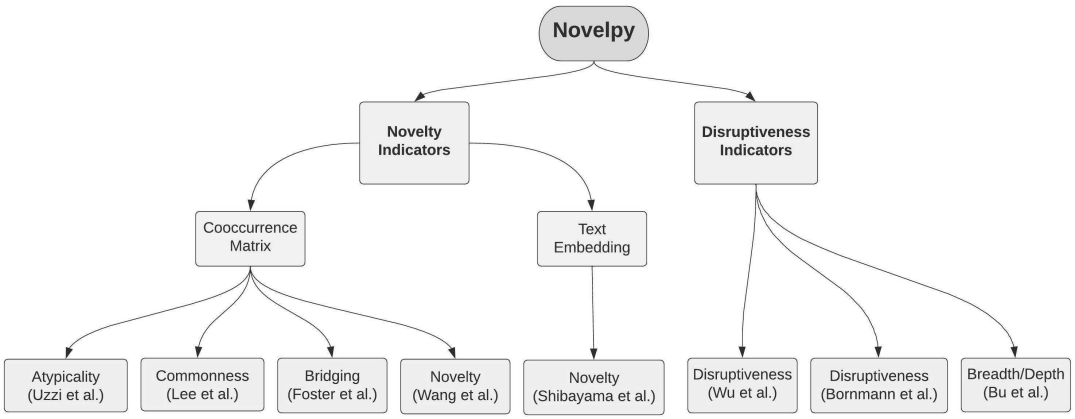


Figure 2.1: *Novelpy*'s module structure

[2021]'s indicators use the citation network and title or abstract to represent the article's semantics in a vector space. Various tools to preprocess bibliometric data are also included within the package to simplify the computation of proposed measures (e.g., co-occurrence matrix construction, text embedding, citation and co-authorship network creation).[4] Table 2.1 summarizes the indicators available in the module, their strengths and weaknesses, and the possible variables to compute them.

The module supports a wide range of data sources as long as they are in the proper format; note that transforming data to the expected structure is relatively simple. Helper functions are available to transform PubMed Knowledge Graph data into the desired structure.[5] For other databases, further backend to OpenAlex, Web of Science, Scopus, and PATSTAT are under construction. The package currently works with documents in JSON or MongoDB format. Mongo will be preferred for large databases to avoid overflowing the RAM.

2.2.1 Novelty Indicators

We focus on novelty indicators in the package based on the combinatorial idea. As discussed in section 2.1, novelty indicators can be differentiated into two groups regarding how they compute the distance between items. The first group uses a combination of items, such as keywords and journals, to create a co-occurrence matrix. Algorithms make use of this matrix to compute the distance. The more distant, the

[4] see `https://novelpy.readthedocs.io/en/latest/utils.html`

[5] Expected structure is presented here: `https://novelpy.readthedocs.io/en/latest/usage.html#format-supported`

Type	Indicator	Pros	Cons	Variables used			
				Ref. Journals	*Keywords*	*Citation net.*	*Title/Abs.*
Novelty	Uzzi et al. [2013]	Conserve dynamical citation structure	Computationally intensive	X	X		
	Lee et al. [2015]	Computationally lightweight Data-saving	Conceptually less advanced	X	X		
	Foster et al. [2015]	Consider undirect link Computationally lightweight	Discret distances	X	X		
	Wang et al. [2017]	Computationally lightweight	Data-Intensive	X	X		
	Shibayama et al. [2021]	High granularity	Computationally and data-intensive			O	O
Disruptiveness	Wu et al. [2019]	Normalized	Data-intensive Issue with term K_{FP}			X	
	Bornmann et al. [2019a]	Normalized	Data-intensive			X	
	Bu et al. [2019]	Normalized	Data-intensive			X	

Table 2.1: *Novelpy*'s available indicators. X means that you can run the indicator on either variable. O Means you need both variables to run it

more unexpected and, therefore, novel the combination. The second type of indicator maps items in an Euclidean space with text embedding techniques like word2vec [Mikolov et al., 2013]. The distance is then computed in this semantic space. As shown in Figure 2.1, novelty indicators are split between those using co-occurrence of entities such as journals or keywords and those using word embedding techniques. For the first group of indicators, we first need to create a co-occurrence matrix for each year of the given dataset. While some indicators only use the focal year to compute the score for each combination [Uzzi et al., 2013, Lee et al., 2015, Carayol et al., 2019], others take into account past combinations in the score calculation [Foster et al., 2015] and future re-utilization [Wang et al., 2017].

Atypicality [Uzzi et al., 2013], Commonness [Lee et al., 2015], and Novelty [Wang et al., 2017] are all indicators that use references of an article at a journal level. Previous studies usually focused on one type of knowledge unit, but as long as one can create a co-occurrence matrix between items, it becomes trivial to generalize. Carayol et al. [2019] reformulate Lee et al. [2015] and apply it to keywords and construct the indicator accounting for inter-field heterogeneity by splitting the analysis. Fleming [2001] computes a combination of patent subclasses, a prevalent practice

in patentometrics. Dahlin and Behrens [2005] propose a novelty measure based on the overlapping between documents' references that was reused by Trapido [2015]. Based on this work Matsumoto et al. [2021] propose an extension that computes the average share of references that are shared between a focal paper and all other documents in the same field. These indicators are not present in *Novelpy* (v1.2) but will be added in future versions.

Although the co-occurrence matrix can be considered an adjacency matrix, only a handful of indicators use graph theory to compute the distance between items. Indeed, indicators *à la* Uzzi et al. [2013], or Lee et al. [2015] take into account only the direct neighborhood during distance calculation. If items A and B are close, items B and C are close, and D is unrelated to any of them, then the combination of A and C is more likely to happen than A and D. This logic is completely ignored if one considers the direct neighbors. Wang et al. [2017] integrated this into their indicator by considering the cosine similarity between nodes' neighbors, which considers common friends (A and C in the example above). Using community detection as in Foster et al. [2015], one can better represent the distance between two units by using the global structure of the network. However, the discrete nature of the novelty score can be argued. Using text embedding, one can have a continuous representation of the distance between items. This distance is related to the text's structure since word similarity depends on their neighborhood. Some initiatives used these techniques with different purposes but could be used to create a novelty score. Hain et al. [2020] create a similarity measure between patents using word2vec [Mikolov et al., 2013]. Shibayama et al. [2021] was the first to apply word embedding techniques in a novelty context. They embed references in a Euclidean space using spaCy and then compute a distribution of cosine distances between documents present in the references for a given document.

We propose a mathematical formalization of these indicators. Setting up this framework offers a basis for defining future new indicators. These indicators are formulated based on graph theory, where the network's nodes are units of knowledge (journals, keywords, or references), and edges represent the co-occurrence of these units in entities (documents or patents).

- Co-occurrence matrix can be written as a graph $G = (V, E, w)$.
- Set of nodes V of dimension v represent here the entities (e.g. keywords, journals), a given entity is defined as V_i.

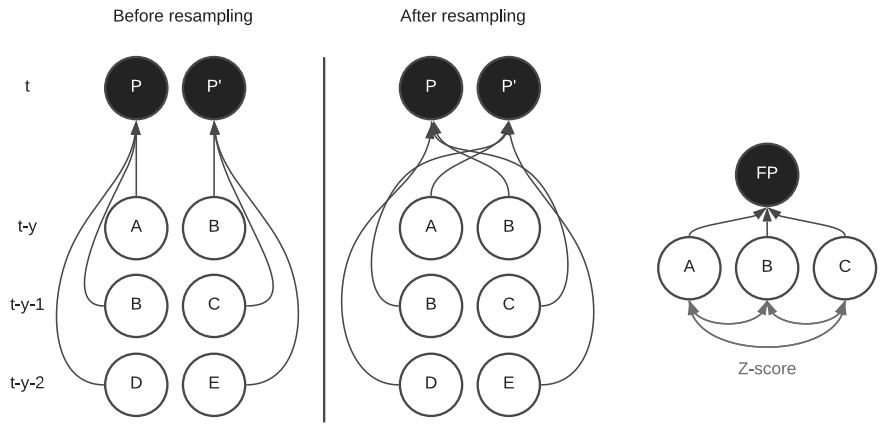
- Set of edges is noted E.
- Number of combinations between V_i and V_j is the weight for the edge (V_i, V_j) and is written $w(V_i, V_j)$.
- Degree of a node V_i is written k_i. N is the sum of the weighted edges in G without self-loops, $N = \Sigma_{i=1}^{v-1}\Sigma_{j=i+1}^{v} w(V_i, V_j)$.

D defines our set of documents of dimension n. Each focal paper, FP, has its network, which can be defined as G_{FP}, E_{FP} is the subset of edges present in document FP. G_{FP} uses the same set of nodes V as G and can be express as $G_{FP} = (V, E_{FP}, w_{FP})$. In some cases, G_{FP} is an unweighted network and will be written then $G_{FP} = (V, E_{FP})$. The number of links, $w(V_i, V_j)$, is then defined as the sum of all combinations of two given entities overall document in D, $w(V_i, V_j) = \Sigma_{d=1}^{n} w_d(V_i, V_j)$ where $w_d(V_i, V_j)$ is binary if the graph is unweighted at the document level. $G(V, E, w)$ can be defined at a year level. For example, in year t, and the associated network will be noted $G_t(V, E_t, w_t)$. Uzzi et al. [2013], Lee et al. [2015], use only the subgraph G_t for calculation. Foster et al. [2015] use the accumulation of past networks. For Wang et al. [2017], several subgraphs are involved in computing the indicator. The novelty indicators *à la* Wang et al. [2017] deal with four subgraphs of G. One needs to consider two different past sets of documents (noted P and B) and a set of future documents (noted F).

2.2.1.1 Uzzi et al. [2013]: Atypicality

The goal of the measure proposed by Uzzi et al. [2013], called "Atypicality", is to compare an observed network with a random network. The network is shuffled, preserving the temporal distribution of references at the paper level. As shown in Figure 2.2, a document citing two articles from, for example, 1985 and one from 1987 will still cite articles published the same year, but the journal can change. The frequency of the combination (V_i, V_j) at time t is defined as $w_t(V_i, V_j)$, and we extract the adjacency matrix of observed frequencies. The idea is basically to compute the frequency Z-score for each journal combination. The Z-score is defined as $z = (obs - exp)/\sigma$; an observed frequency is compared with a theoretical one.

The theoretical frequency is generated through Markov chain Monte Carlo simulation, preserving the dynamical structure of citations. In the case of Atypicality, we are dealing with $s + 1$ different networks for the year t, the existing network and



(a) Resampling strategy (b) Score calculation

Figure 2.2: Uzzi et al. [2013] [6]

s resampled ones. The existing network is G_t, as defined above. The others are generated by preserving an article's temporal distribution of references. For each document FP, we want to keep the number of references published in year $t - y$ stable for all y to ensure that the global age distribution of the pieces of knowledge used at time t remains stable.

One needs to generate s random networks G_t. After re-sampling, the publishing year of references is no longer considered. Edges' weights are then aggregated to fit with G_t edge structure E_t by summing over all combinations. The observed frequency for each sample is computed for each edge (V_i, V_j). We write the set of frequencies for the combination of V_i and V_j in the s samples $w_t^s(V_i, V_j)$. One can then compute the mean and standard deviation for each edge's frequency and compute a z-score.

$$Z - score_{ijt} = \frac{w_t(V_i, V_j) - mean(w_t^s(V_i, V_j))}{std(w_t^s(V_i, V_j))}$$

For each paper, taking all combinations made (E_{FP}), a distribution of z-score written Z_{FP} is computed, and the 10th percentile (P_{10}) of this distribution (the novelty) and the median (P_{50}) (the conventionality). The novelty and conventionality for document FP are then written:

$$Novelty_{FP} = P_{10}(Z_{FP})$$

[6](a): P and P' are two distinct papers, P cites journals A, B, and D. P' cites journals B, C, and E. The goal is to shuffle the network by conserving the dynamic structure of citations at the paper level. P no longer cites A from $t - y$ but cites B from year $t - y$. (b): Comparing the observed and resampled networks, we can compute a z-score for each journal combination.

$$Conventionality_{FP} = P_{50}(Z_{FP})$$

While this indicator only requires data from a specific year, it is still computationally greedy. Generating the s samples and the computation of the average and the standard deviation for each possible combination is expensive. On the contrary, this indicator allows for keeping the temporal structure stable, which is more in line with the reality of the availability of the knowledge pieces.

2.2.1.2 Lee et al. [2015]: Commonness

Lee et al. [2015] compares an observed network with a theoretical network (Observed vs Expected frequency of edges) at a year level. The observed number of combinations (V_i, V_j) at time y_t is the number of edges $w_t(V_i, V_j)$, the theoretical number of combinations is $\frac{k_i * k_j}{N_t}$, the degree for entity i and j multiplied together and divided by the total number of combinations made in year t.

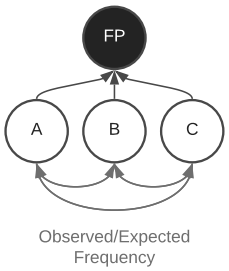


Figure 2.3: Lee et al. [2015]

$$Commonness_{ijt} = \frac{w_t(V_i, V_j) * N_t}{k_i * k_j}$$

For each paper, taking all combinations made in document FP (E_{FP}), a distribution of commonness-score written C_{FP} is computed. The commonness for document FP is the 10th percentile (P_{10}) of this distribution and is written as:

$$Commonness_{FP} = -log(P_{10}(C_{FP}))$$

The main advantage of the commonness indicator is its speed of calculation; it is the least demanding indicator in terms of the execution time of the package. The indicator only requires data from a specific year. Note that this indicator is very close to Uzzi et al. [2013]'s one. Both would be equal if Uzzi et al. [2013] resampling method would not consider the references' publishing year.

2.2.1.3 Foster et al. [2015]: Bridging

Foster et al. [2015] propose a novelty indicator based on community detection algorithms. It captures the distance between two entities taking into account undirected edges. The goal of the measure is to identify the network's community studied and capture proximity through the community in which the combined entities are clustered.

Any community algorithm can be applied to this indicator. We rely on the Louvain algorithm in *Novelpy* following Foster et al. [2021], but we intend to add further options. After applying the community algorithm on $G(V, E, w)$, we are left with multiple clusters of entities. C_i is the community to which the entity i belongs.

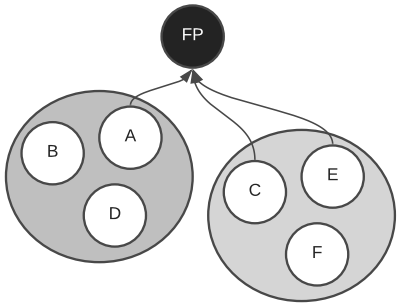


Figure 2.4: Foster et al. [2015] [7]

$$Novelty_{FP} = \frac{\sum_{(i,j)\in E_{FP}} 1 - \delta(C_i, C_j)}{|E_{FP}|}$$

Where $\delta(C_i, C_j) = 1$ if $C_i = C_j$ (i.e., both entities, i and j, are in the same community), $\delta(C_i, C_j) = 0$ otherwise. The novelty score of an entity is the proportion of pairwise combinations that are not in the same community.

This indicator brings into the field algorithms that capture the global network structure and only require data from a specific year. At the same time, this indicator does not allow measuring distances between communities and proposes only a binary distinction.

2.2.1.4 Wang et al. [2017]: Novelty

Wang et al. [2017] propose a measure of difficulty for pairs of references that have

[7] FP cites different journals which belong to different communities. The novelty is the number of journal combinations from two different communities. Communities of journals are computed through a community detection algorithm.

never been made before. These new pairs need to be reused after the given publication's year (scholars do not have to cite directly the paper that creates the combination, but only the combination itself). The idea is to compute the cosine similarity for each journal combination based on their co-citation profile b years before t. The cosine similarity between W_i^B and W_j^B is defined:

$$COS(W_i^B, W_j^B) = \frac{W_i^B . W_j^B}{\|W_i^B\| \|W_j^B\|}$$

where W_i^B represent all links of entity i, B years before year t.

Novelty *à la* Wang et al. [2017] relies on four subgraphs of G constructed using two different past sets of documents, a set of future documents, and the set of documents for the focal year. These different subgraphs are defined as follows (note the first year of the dataset y_0 and the last as y_n):

- $G_t = (V, E_t, w_t)$ is a subgraph of G from year t (documents published year t)
- $G_P = (V, E_P, w_P)$ is a subgraph of G from year t_0 to $t-1$ (documents published before year t)
- $G_B = (V, E_B, w_B)$ is a subgraph of G from year $t-b$ to $t-1$ is used to measure the cosine similarity between nodes. This set is a subgraph of G_P (documents are published in a given window before year t)
- $G_F = (V, E_F, w_F)$ is a subgraph of G from year $t+1$ to $t+f$ (documents published in a given window after year t)

This indicator focuses on new combinations reused afterward and not achieved before the given year y_t. One needs to keep all elements of $E_t \notin E_P$ and $E_t \in E_F$. More precisely, edges belonging to the following subset (that we call E_N) are the only edges used to compute this indicator $E_N = (E_t \cap E_F) \cap \overline{E_P}$

Cosine similarities are calculated using G_B. For each document, we compute an undirected and unweighted network. The novelty is the sum of all edges from $E_{FP} \in E_N$, that is:

$$Novelty_{FP} = \sum_{(i,j) \in E_N} 1 - COS(W_i^B, W_j^B)$$

[8]For a given article at time t, we check if the journal combined were already combined in the past (G_P). We then check if the combination is reused in the future (G_F). If the combination is new and reused, the difficulty of making such a combination is calculated on the recent past (G_B)

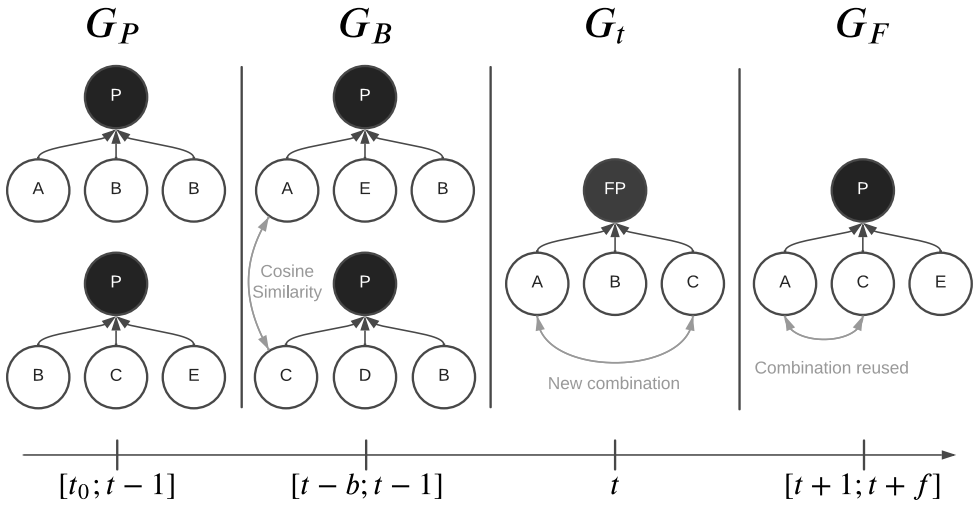


Figure 2.5: Wang et al. [2017] [8]

The main issue with this indicator is the amount of data needed to compute the measure. One needs as much data as possible before the focal year to ensure that the combination has never been made. At the same time, some hyperparameters involved in this measurement can drastically modify the results. For example, the time window to capture the re-utilization of a combination or the number of times reused needed to be novel is very arbitrary.

2.2.1.5 Shibayama et al. [2021]: Novelty

Shibayama et al. [2021] propose to incorporate semantic distances to capture diversity in the set of references from a given article following Hain et al. [2020] and their similarity measure between patents. Document centroids are computed by summing all word representations for each document.

Consider a directed unweighted graph $G(V, E)$ containing the citation network. For a given document FP, a referenced document is denoted by r, and the set of nodes that are cited by FP is then $In_{FP} = r : (FP, r) \in E$. Shibayama et al. [2021] compute all distances between each document's centroids ($C^2_{|In_{FP}|}$ combinations). All documents have a vectorial representation in a semantic space of length 200. Distances between two references $i, j \in In_{FP}$ are calculated through cosine similarity: $n_{ij} = 1 - COS(T_i, T_j)$, where T_i is the dense vector text representation for a document i. A distribution of novelty scores $N_{FP} = n_{ij} : i, j \in Out_{FP}$ is then computed, and for each document, the final score is a percentile of N_{FP}.

[9] For a given article, each reference's abstract (or title) is represented in a semantic space through

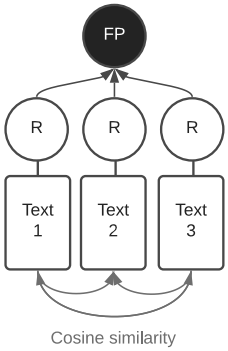


Figure 2.6: Shibayama et al. [2021] [9]

Shibayama et al. [2021]'s indicator is both data-intensive and computationally intensive. One needs all references' titles/abstracts for a given set of articles. The package currently works with a pre-trained Word2Vec model, *en_core_sci_lg* from *spacy*, to compute the dense representation of a document. Future versions will incorporate a back-end to use any pre-trained model.

2.2.2 Disruption Indicators

Disruption indicators offer alternative measures of impact to the number of citations. They allow understanding if a given article behaves as a bottleneck between the knowledge mobilized in a given article and the articles that will cite it. Disruptiveness was introduced in scientometrics by Wu et al. [2019] and was previously proposed for patents by Funk and Owen-Smith [2017]. Following Azoulay [2019]'s definition, a paper can either consolidate or disrupt existing knowledge. If future papers cite a focal paper and its references, then the focal paper consolidates the existing knowledge space but does not disrupt it. On the other hand, if future papers cite only the focal paper and not its references, then the focal paper is considered disruptive. Quoting Bornmann et al. [2019a], "[...] many citing documents not referring to the FP's cited references indicate disruptiveness. In this case, the FP is the basis for new work which does not depend on the context of the FP, i.e., the FP gives rise to new research." All presented measures normalize citation and give a relative perspective on a publication's impact [Bu et al., 2019]. Disruption indicators consider the importance of pieces of knowledge (references) in a given article for other articles. In contrast, Depth and Breadth, as proposed in Bu et al. [2019],

text embedding techniques. The distance between two references is then computed through cosine similarity.

capture how the knowledge generated by that given item is reused and whether it allows for the consolidation of a domain or is instead used in a disparate manner.

Consider a directed unweighted graph $G(V, E)$ containing the citation network.

- For a given document FP, we note a document cited by FP, r. The set of nodes that are cited by FP is then $In_{FP} = \{r \in V | (FP, r) \in E\}$
- For a given document FP, we note a document citing FP, c. The set of nodes that are citing FP is then $Out_{FP} = \{c \in V \ | (c, FP) \in E\}$
- The number of citations for FP is then $deg^{-}(FP) = |Out_{FP}|$ and number of references $deg^{+}(FP) = |In_{FP}|$
- The set of references for an article citing FP is then noted In_c

2.2.2.1 Wu et al. [2019]: Disruptiveness

By adapting Wu et al. [2019] notation, we called I_{FP} the set of nodes with FP as a parent that does not have FP's parents as parents. More formally $I_{FP} = \{c \in Out_{FP} | In_c \notin In_{FP}\}$. The set of J^l_{FP} is the set of nodes with FP as a parent that share at least l parents with FP. We note $J^l_{FP} = \{c \in Out_{FP} || \{In_c \in In_{FP}\}| > l\}$. Finally, K_{FP} is the set of nodes that share parents with FP but that do not have FP as a parent: $K_{FP} = \{v \in V | v \in In_{FP}\}$.

The disruptiveness *à la* Wu et al. [2019] is then noted :

$$DI_1 = \frac{|I_{FP}| - |J^1_{FP}|}{|I_{FP}| + |J^1_{FP}| + |K_{FP}|}$$

Some variants that consider only paper sharing at least l references have been proposed:

$$DI_5 = \frac{|I_{FP}| - |J^5_{FP}|}{|I_{FP}| + |J^5_{FP}| + |K_{FP}|}$$

2.2.2.2 Bornmann et al. [2019a]: Disruptiveness

A variant that removes the term $|K_{FP}|$ has been proposed by Wu and Yan [2019] because the number of documents that cite references from the focal documents without citing the focal documents is often too large compared to the paper from other

[10]For a given article FP, we retrieve: (a): Articles citing FP and references from FP (named J). (b): Articles citing FP but no references from FP (named I). (c): Articles citing references from FP but do not cite FP (named K).

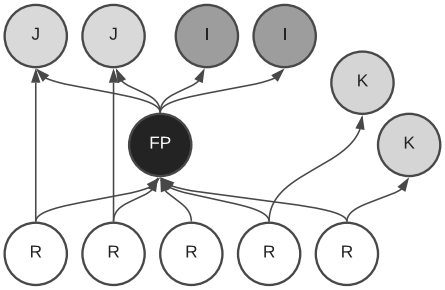


Figure 2.7: Wu et al. [2019], Bornmann et al. [2019a] [10]

sets. Wu and Wu [2019] show how considering the set K_{FP} can lead to a decrease in disruptiveness when the term $|I_{FP}| - |J^1_{FP}|$ is negative. In that configuration, more papers that do not cite FP ($|K_{FP}|$) lead to higher disruptiveness, which is different from how the indicators conceptually work. Defined as DI^{nok}_l by Bornmann et al. [2019a], we note:

$$DI^{nok}_l = \frac{|I_{FP}| - |J^l_{FP}|}{|I_{FP}| + |J^l_{FP}|}$$

2.2.2.3 Bu et al. [2019]: Breadth and Depth

Bu et al. [2019] propose an alternative to the above disruption indicators. It calculates the proportion of articles citing the focal paper that also cites other articles citing it. The indicator allows us to understand whether the document contributes to a restricted research domain; the documents citing the focal paper are interdependent and cite each other. On the contrary, the documents using the focal paper's research may also be unconnected and belong to a more extensive research space.

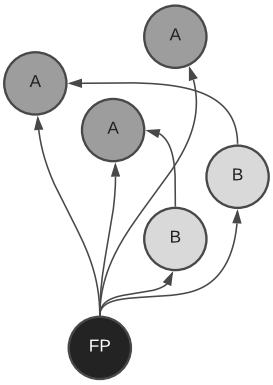


Figure 2.8: Bu et al. [2019] [11]

[11] For all articles citing FP, we check if they also cite papers citing FP.

Let FP be the focal paper, the articles citing it the set Out_{FP}. We are interested in the articles cited by the documents of the set Out_{FP}. For each element c of Out_{FP}, we observe a set of associated references named In_c. The proportion of documents citing document FP and also citing documents that are citing FP is then written as:

$$Depth_{FP} = \frac{|\{c \in Out_{FP} : |In_c \in Out_{FP}| > 0\}|}{|Out_{FP}|}$$

On the contrary, the breadth, the proportion of papers citing FP that do not cite other publications also citing FP, is written:

$$Breadth_{FP} = \frac{|\{c \in Out_{FP} : |In_c \in Out_{FP}| = 0\}|}{|Out_{FP}|} = 1 - Depth_{FP}$$

Bu et al. [2019] also propose a measure of dependence. It captures the average number of references shared between the focal paper FP and documents citing it. In_{FP} is the set of references of FP. For all document c that cite FP (Out_{FP}), we want to know the number of references shared:$|\{In_c \in In_{FP} : c \in Out_{FP}\}|$. The average number of references shared between document FP and all documents citing it ($c \in Out_{FP}$) is then:

$$Dependence_{FP} = \frac{\sum\limits_{c \in Out_{FP}} |In_c \in In_{FP}|}{|Out_{FP}|}$$

Our package does not compute two other indicators from Bu et al. [2019]: Independence and Dependence. However, they represent the proportion of publications citing a focal paper that also cites references from the focal paper. Using notation from 2.2.2.1: $\frac{|I_{FP}|}{|I_{FP}|+|J^1_{FP}|}$ one can easily derive this value from disruption indicators DI^{nok}_1. Indeed from $DI^{nok}_1 = \frac{|I_{FP}|-|J^1_{FP}|}{|I_{FP}|+|J^1_{FP}|}$ we can compute the independence, the proportion of articles citing the focal paper that do not cite articles cited by the focal paper

$$\frac{|I_{FP}|}{|I_{FP}| + |J^1_{FP}|} = \frac{DI^{nok}_1 + 1}{2} \quad \text{if} \quad |Out_{FP}| > 0$$

All these measures are rather demanding in terms of data requirements. Indeed, for each given article, we need to access the references, the articles citing the focal paper, and the articles citing the references of the focal paper.

2.3 Sample analysis

2.3.1 Descriptive statistics

This section provides examples of applications that could be performed with *Novelpy*. We use a Pubmed Knowledge Graph (PKG) sample [Xu et al., 2020], which stores research articles published on Pubmed and offers metadata for all papers. This analysis is proposed as an example to demonstrate our module features after computing the indicators.[12] All figures and tables can be found in the appendix. The sample is restricted from 1995 to 2015; the focal period is 2000-2010. The sample is composed of 1,469,352 papers and 2,959,650 distinct authors. Authors are disambiguated in PKG using advanced heuristics and algorithms. The sample was chosen to ensure that articles had the attributes needed to run the indicators. Each paper has references, mesh terms, titles, and abstracts. Table 2.2 and Figure 2.9 summarize the statistics of the sample. On average, the number of references used in a paper is 23, consistent with typical citation behavior [Abt and Garfield, 2002]. The number of papers almost doubled in 10 years, which aligns with the literature [Fortunato et al., 2018].

2.3.2 Results

As discussed in previous sections, research on novelty indicators still needs to be conducted across multiple dimensions. *Novelpy* will facilitate computing different indicators on various entities. Researchers can then use the novelty scores provided by the package to perform their analyses. Individual-level analysis can be conducted by examining the distribution of novelty scores, as shown in Figure 2.10. Comparing indicators and studying the evolution of novelty over the years are the primary motivations for this package. Only a few studies examine the dynamics of novelty over time. Nevertheless, understanding the evolution of creativity in papers, patents, or other entities can offer insights into the trade-off between exploration and exploitation of the research space in a given field. Figure 2.11 displays the evolution of the mean novelty score for each indicator, given the variable (references, mesh terms). We cannot draw conclusions since the sample is random and aggregated across all fields within Pubmed. The pattern of trends varies significantly depending on the indicator and variable. This heterogeneity might be evidence that further investiga-

[12] Interested readers will find code and resources to create tables, plots, and indicators here `https://novelpy.readthedocs.io/en/latest`

tion is required to understand precisely what these indicators capture and in which cases they best predict novelty. This question is even more relevant, considering the lack of correlation between indicators in Figure 2.12.

2.4 Discussion

This paper aims to demonstrate the capabilities of the new *Python* package *Novelpy*. We presented a sample analysis using the functions within this package to showcase how it can assist interested readers in computing and analyzing existing indicators or addressing current challenges related to novelty measurement. Several critiques can be made on current novelty measurements, and addressing these points is crucial for solidifying our understanding and usage of these indicators.

The diversity and convergence in how novelty indicators are created raise questions about what they measure. As observed in our sample analysis, the results are highly dependent on the indicator used, which confirms previous concerns about cherry-picking the indicator [Shibayama et al., 2021, Foster et al., 2021]. Simultaneously, indicators often focus on the same entity (keywords or reference journals). Recent measures like Shibayama et al. [2021] and Arts et al. [2021] broaden this domain by utilizing text information from references. Novelty indicators are rarely conceptualized and often require a qualitative background. Qualitative studies like Tahamtan and Bornmann [2018] question the significance of literature in authors' creative processes. The link between references and creativity is debated, and further investigation is needed to determine if references can be reliably used as a proxy variable for creativity.

Research evaluation was once performed solely by experts in the scientometric field and specialists working for public institutions. The availability of open-access data has recently extended the responsibility of conducting this evaluation to a more diverse group of researchers and members of the public. These new actors need the necessary tools to compute scientometric indicators and some understanding of their relevance. Using software creates a gap between the user and the actual data, which may lead to issues if the assumptions necessary for the indicators' relevance are overlooked. Data-driven decisions can become inefficient if the algorithm used is a black box and is misused. A solid background in how and why these indicators are created is necessary to limit bias in selecting indicators when used in research. As seen in Section 3, every indicator has its pros and cons, different hyperparameters (time win-

dow, re-utilization, number of samples, and others), and is highly dependent on the database used. The coverage varies greatly depending on the database (language, fields, nationality, and others) [Sugimoto and Larivière, 2018]. These aspects and the increasing number of novelty indicators create arbitrary decision-making when using them. Sugimoto and Larivière [2018] suggests that indexing and classification of documents differ between databases, making it challenging to reproduce studies on other databases. Constructing a general indicator applicable to all scientific disciplines is difficult, as citation habits are heterogeneous, making comparisons between fields risky [Carayol et al., 2019]. Depending on the country, methods and standards may differ within a discipline, and the historical practice of a field may change the representations.

Improving novelty measurement is essential for supporting innovative research. Highly novel documents are less likely to be cited in the short run and are less likely to be published in high-impact factor journals [Wang et al., 2017, Mairesse et al., 2021]. Due to the pressure from citation count evaluation, the exploration of science is less likely to occur. Researchers may tend to conform to conventional references within their field, which is already accentuated during submission. Documents already highly cited, considered stepping stones in the field, will thus receive even more citations, creating a vicious circle. This vicious circle has the consequence of narrowing research, where only those who agree with the existing paradigm are rewarded with citations.

This phenomenon is already observed in AI research, where topics become increasingly less diverse [Klinger et al., 2020]. The goal of science is not to persist with merely satisfactory solutions but to explore a range of possibilities, even those that may prove fruitless. Citation indicators typically do not emphasize researchers who take risks by attempting novel approaches. Various funding methods exist to support high-risk, high-reward (i.e., highly novel) research [OECD, 2021]. Experts are not free from bias when evaluating novelty, funding processes are not uniform, and many decisions remain arbitrary. Currently, none of them uses novelty indicators to evaluate proposals. Novelty measurement might be relevant in providing reliable information when awarding grants to research proposals.

We conclude this discussion with a roadmap and our aspirations for *Novelpy*. The primary feature we aim to develop in future versions is automatic execution using well-known databases (PATSTAT, Microsoft Academic Knowledge Graph, Arxiv, etc.). At present, users must pre-process data to match our format. Although we

provide a comprehensive example and make the sample available here `https://novelpy.readthedocs.io/en/latest/usage.html#id5`, we believe that expanding the accepted inputs will aid researchers in working on improving novelty indicators. The second feature we plan to add is a time complexity analysis. To conduct a proper benchmark between indicators, we need to compare their computing speeds. Users can currently perform this manually, but we intend to streamline the process and add plots to address this gap. Finally, we will selectively add new and past indicators. Anyone interested in contributing to the module can visit GitHub `https://github.com/Kwirtz/novelpy` and create a pull request.

2.5 Appendix

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
n papers	49,872	52,046	54,721	58,439	62,241	67,361	70,501	75,717	81,228	84,496	89,168
mean of cited paper	27.3871	27.3672	27.9704	28.5654	28.8562	29.3572	30.0423	30.3297	31.0576	31.5393	32.3128
var of cited paper	707.008	708.596	742.314	709.619	807.733	758.342	809.695	795.216	845.84	944.337	896.461
mean of meshterms per paper	13.3097	13.4067	13.2431	13.3788	13.2862	13.1364	12.8499	12.8425	12.8575	12.9128	12.8867
var of meshterms per paper	26.5811	27.6517	26.0774	26.9045	27.2265	26.4599	22.9795	23.3933	23.3725	24.6855	24.9734

Table 2.2: Sample Statistics

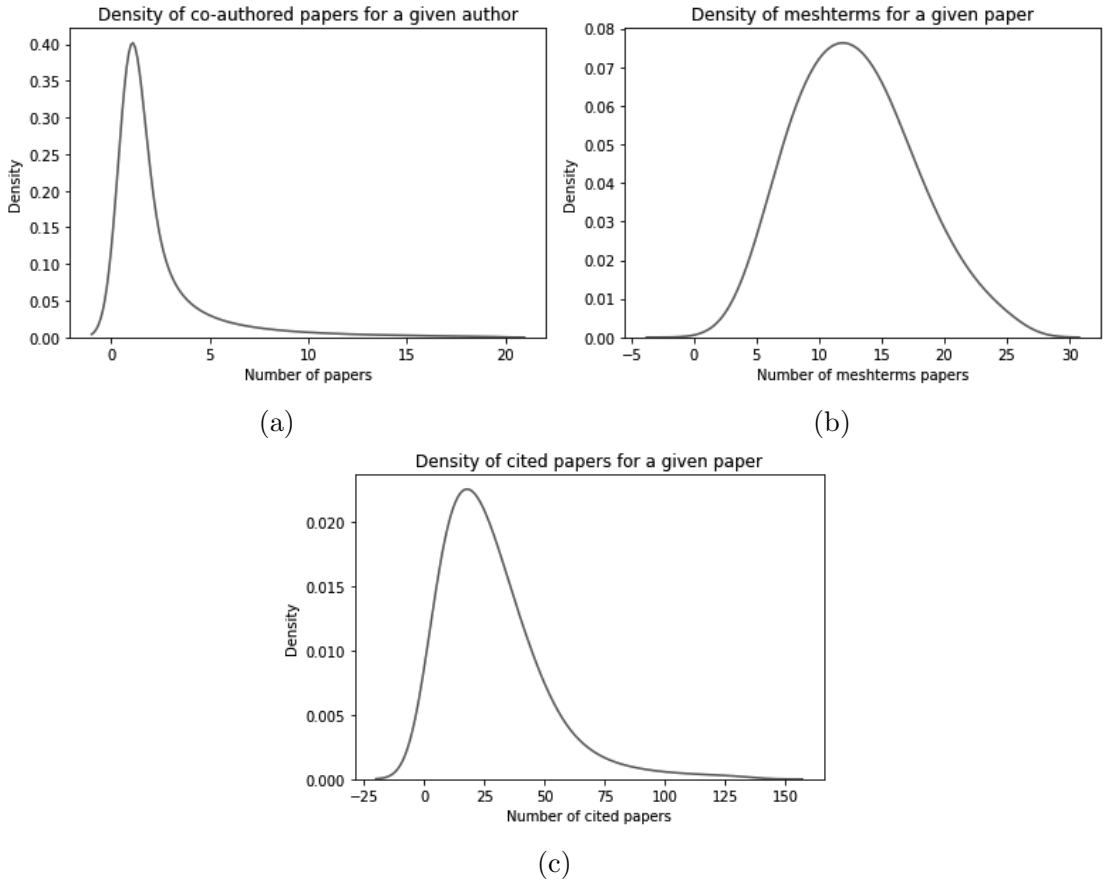


Figure 2.9: (a) Density of contribution of authors. On average an author has 2.6 publications (solo or co-authored) in 10 years. (b) Density of the number of mesh terms between 2000-2010. On average, a paper is labelled with 13 mesh terms. (c) Density of references between 2000-2010. On average, a paper has 23 references

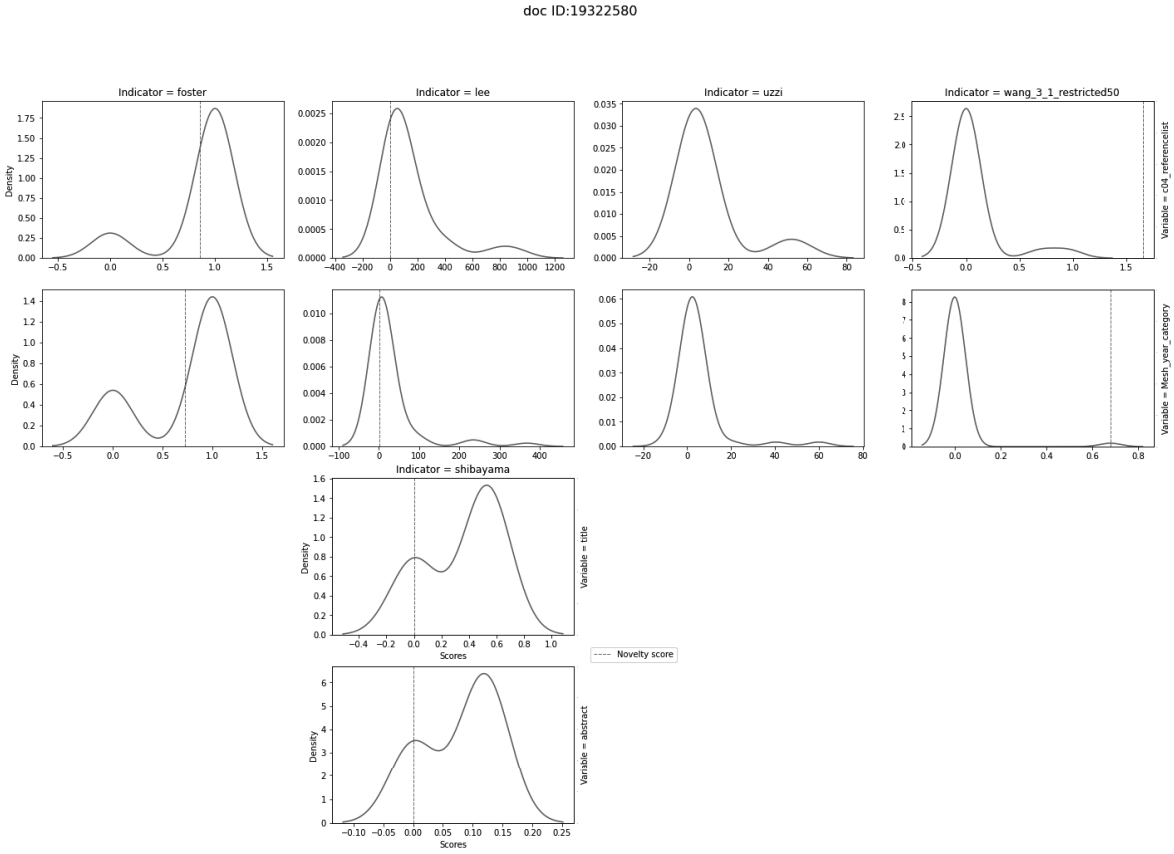


Figure 2.10: Each combination has a novelty score. A single plot represents the density of score combinations for a specific paper (PMID 10698680) for a specific indicator and entity (i.e. mesh terms, journals, title, abstract). The scores for the first row were computed using a combination of cited journals. The scores for the second row were computed on mesh terms combinations. Each column represents an indicator. The last two rows are for text embedding-based indicators (i.e. Shibayama et al. [2021]: Novelty, Author proximity) on the paper's title or abstract.

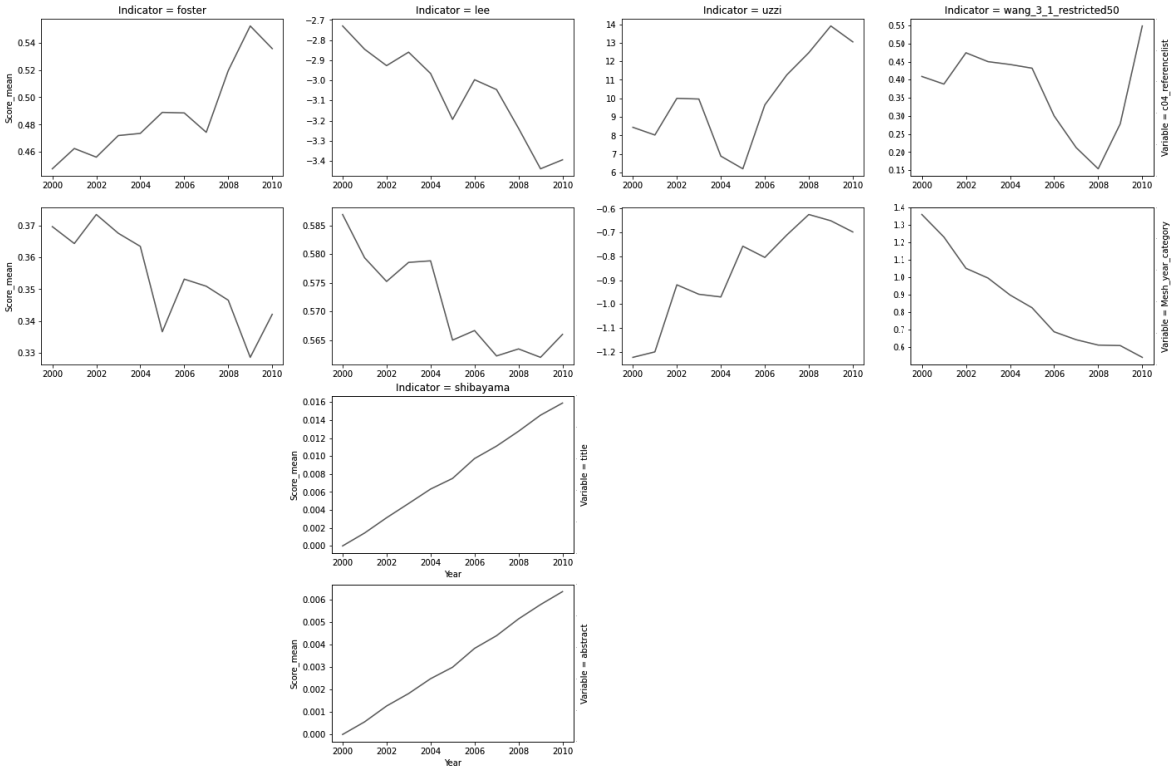


Figure 2.11: The mean novelty score on every document for a given year. Columns and rows represent respectively indicators and variables.

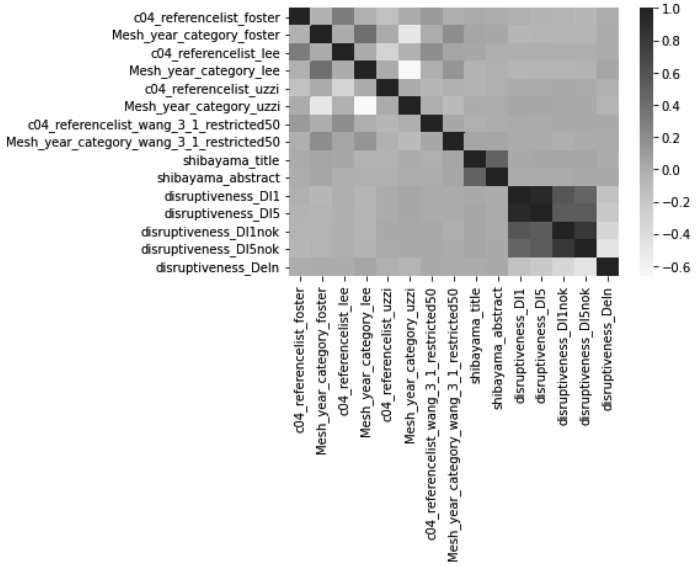


Figure 2.12: The correlation between the novelty score for each indicator, given the entity, for the period 2000-2010

Chapter 3

Unpacking Scientific Creativity: A Team Composition Perspective

This chapter was co-authored with

Pierre Pelletier

Summary of the chapter

This paper investigates the relationship between cognitive diversity within scientific teams and their ability to generate innovative ideas and gain scientific recognition. We propose a novel author-level metric based on the semantic representation of researchers' past publications to measure cognitive diversity at individual and team levels. Using PubMed Knowledge Graph (PKG), we analyze the impact of cognitive diversity on novelty, as measured by combinatorial novelty indicators and peer labels on Faculty Opinion. We assessed scientific impact through citations and disruption indicators. Cognitive diversity between team members appears to be always beneficial to combine more distant knowledge. We show that while the effect is positive, it is marginally decreasing. Our findings also reveal that within-team average exploratory profiles follow an inverse U-shaped relationship with combinatorial novelty and citation impact. We show that the presence of highly exploratory individuals is profitable to generate distant knowledge combinations only when balanced by a significant proportion of highly exploitative individuals. Also, teams with a high share of exploitative profiles consolidate science, while those with a high share of both profiles disrupt it. These results emphasize the implication of team composition in scientific creativity, suggesting that combining these two types of individuals leads to the most disruptive and distant knowledge combinations.

3.1 Introduction

Creativity is a crucial driving force in fostering the production of new knowledge in an ever-growing landscape of scientific research and technological innovation [Geuna, 1999, Amendola et al., 2014, Witt, 2016]. A broadly accepted definition of creativity assumes a bipartite composition involving a combination of novelty and effectiveness [Runco and Jaeger, 2012]. As science moves towards a team-based model [Wuchty et al., 2007], the creativity of scientific publications should be studied from a social perspective. The cognitive dimension (i.e., differences in thinking, problem-solving approaches, and perspectives among individuals) plays a crucial role in enabling the exchange of information and creation of new knowledge [Nooteboom, 2000, Nooteboom et al., 2007]. It is induced by individuals' characteristics and the trade-off carried out between *exploration* and *exploitation* [March, 1991] of the knowledge space during their career. In the context of science, exploration involves actively pursuing the expansion of one's understanding and curiosity across various areas of knowledge. On the other hand, exploitation refers to individuals specializing in a specific field and continuously building upon their expertise in that area. The presence of individuals with exploratory profiles appears to facilitate communication among team members who are cognitively distant and foster creativity as the intersection of different perspectives is commonly required to solve complex scientific problems [pag, 2007].

This paper aims to study the extent to which the exploratory nature of scholars and the cognitive diversity of scientific teams shape their ability to generate innovative ideas and obtain scientific recognition. We propose a new author-level measure of cognitive diversity based on the semantic representation of their past papers; this metric allows us to proxy both intra-individual and inter-individual cognitive dimensions and their consequences on creativity in science.

In scientific creativity, originality and success emerge as two essential components [Runco and Jaeger, 2012]. However, the focus has predominantly shifted towards success. The excessive emphasis on success through measures such as citation counts for articles or authors has been found to constrain novelty and originality by providing limited incentives for researchers, ultimately leading to suboptimal research choices. The reliance on an impact metric to reward and evaluate researchers created a harmful behavior whereby scientists maximize the metric to become more appealing to funding agencies or institutions. As Goodhart's law states, "*when a measure becomes a target, it ceases to be a good measure*" [Goodhart, 1984]. The h-index, although

heavily criticized, became a central evaluation instrument of researchers [Costas and Franssen, 2018]. This negatively impacts novelty as innovative research tends to be less cited in the short run [Wang et al., 2017]. Researchers are discouraged from opting for a more exploratory approach when developing a research question, as work that is too innovative tends to be rejected when it deviates too much from the established paradigm [Carayol and Dalle, 2007, Trapido, 2015].

Career choices are directly affected by this phenomenon. Given the heterogeneity in terms of the impact of novel research, researchers have less incentive to produce highly innovative work because of the uncertainty linked with novel research. In the short term, individuals might turn to more conventional research to maximize their h-index while minimizing the risk associated with novel research. The bias toward maximizing the h-index already has a tangible impact on limiting novelty in various research fields. The imbalance between growth in the scientific workforce and research funding has led to 'hyper-competition' in the medical sciences; the scientific system favors individuals who can ensure outcomes over those with potentially groundbreaking ideas that might disrupt the field [Alberts et al., 2014]. Such a focus of the researchers on their impact is done at the expense of their novelty, showing a clear disconnection between the goal of science and its operationalization.

One of science goals is to advance the boundary of the knowledge space [Shi et al., 2015, Witt, 2016, Veugelers and Wang, 2019]. Novelty (also referred to as originality or invention) lies at the cornerstone of innovative research, bridging existing knowledge and unexplored scientific territories. Effectiveness, on the other hand, refers to the recognition attributed to this novelty. Novelty is at the foundation of peer recognition and acts as a "reward system" wherein the individual credited with the initial discovery garners recognition. [Merton, 1957, Stephan, 1996, Carayol et al., 2019]. Novelty is crucial for scientists to develop new solutions to the grand challenges of the century (climate change, poverty, global pandemics, and others) [Petersen et al., 2021]. Highly innovative research is frequently referred to as "High-Risk High-Reward" (HRHR) to reflect its high volatility of outcomes (i.e., novelty does not imply effectiveness). In particular, highly novel research receives more citations on average, but the uncertainty is also more considerable [Wang et al., 2017]. Funding opportunities are limited for innovative research due to its risky nature [Ayoubi et al., 2021, OECD, 2021, Franzoni et al., 2022]. Multiple grant initiatives try to support HRHR research, and funding decisions are all based on expert judgment [OECD, 2021]. But there is a direct bias towards novelty when scholars evaluate a

peer's work [Wang et al., 2017, Ayoubi et al., 2021] and the effect is accentuated by the cognitive distance with the examiner [Boudreau et al., 2016]. Measures such as novelty indicators attempt to estimate the originality of a document and might guide experts to support innovative research. Yet, these novelty indicators are still relatively recent and understudied as it is mostly intended to explain success. As a result, it is essential to explore and validate new methods to understand better how to detect potential innovative and impactful research based on different criteria than past novelty or previous success.

Not all idea combinations are worth exploring, hence the challenge of distinguishing between novel and impactful ones. [March, 1991] distinguishes two different strategies for invention in organizations: "Exploration and exploitation". Exploitation focuses on a combination of ideas that are closely related to each other, thus representing a low-risk strategy. On the other hand, exploration represents the navigation through the knowledge space to combine more distant ideas, inducing more volatile results. March [1991] supports the idea that a mix of exploitation and exploration is the key to an organization's survival. Put differently, producing a valuable invention would require a proper mix of typical and atypical combinations of knowledge, as seen in Uzzi et al. [2013]. This dichotomy has been studied in different domains, as mentioned in Foster et al. [2015] (e.g., "conformity" versus "dissent" in the philosophy of science), and can also be applied to research. As the body of knowledge in science expands, researchers increasingly specialize their competencies [Jones et al., 2008, Jones, 2009] and thus are better able to recombine information locally in the knowledge space, facing incentives to collaborate [Fleming, 2001, Boudreau et al., 2016]. Science is seen as a social phenomenon [Fleck, 2012]. Indeed, agents that recombine knowledge are individuals embedded in a social context, and cognitive and social phenomena strongly influence the invention process [Fleming, 2001]. Team size has been shown to impact creativity [Paulus and Nijstad, 2003, Shin and Zhou, 2007, Wuchty et al., 2007, Falk-Krzesinski et al., 2011, Erren et al., 2017, Mueller, 2019]; however, the authors' characteristics have not been adequately considered in the process as current novelty indicators primarily focus on the information *within* a document.[1] We argue here that the cognitive distance between co-authors and the team composition of a research paper may be among the most critical factors influencing knowledge creation. Based on the concept of exploration and exploitation, we

[1] E.g., references, text, keywords. A detailed review of classical re-combinatory novelty indicators can be found in Pelletier and Wirtz [2022].

propose an indicator that serves as a proxy for exploratory *vs.* exploitative trade-off at both the individual and team levels through past publications. In a nutshell, our indicator measures the cognitive distance between team members as well as the individual propensity to work on various subjects.

We are unaware of previous studies that have used individuals' past research experiences to investigate how the cognitive dimension influences the novelty and recognition of the resulting articles. Note that we do not consider our indicator a replacement for current novelty indicators but rather as a tool that could enhance our understanding of the mechanisms behind creativity. In fact, by incorporating the cognitive dimension into novelty studies, we can develop a more comprehensive understanding of the complex relationship between cognitive aspects, interdisciplinary efforts, and the nature of scientific innovation. Furthermore, examining these questions enables us to provide valuable insights and guidance for researchers and institutions striving to enhance scientific progress while avoiding potentially misleading interpretations of research performance measurement.

Using PubMed Knowledge Graph (PKG), we empirically investigate the role of these cognitive diversities in the production of novel research outcomes and the ability to obtain scientific recognition. We performed the analysis on novelty on five combinatorial novelty indicators [Uzzi et al., 2013, Lee et al., 2015, Foster et al., 2015, Wang et al., 2017, Shibayama et al., 2021], both on references and MeSH terms, as well as on perceived novelty, using labels submitted by researchers to qualify the contribution of an article (Faculty Opinion).[2] For scientific recognition, we rely on the traditional number of citations and six indicators of disruption and consolidation [Wu et al., 2019, Bu et al., 2019, Bornmann et al., 2019a].

Our findings emphasize the crucial role of cognitive dimensions in creativity, significantly impacting originality and success. We show that cognitive diversity always seems beneficial to combine more distant knowledge. In contrast, the within-team average exploratory profile follows an inverse U-shaped relation with combinatorial novelty (i.e., there is a turning point where it is no longer beneficial). The same relation can be found with citation counts, but we show that the cognitive dimension also strongly influences the nature of citations. Teams with more exploitative profiles consolidate science, while those with high exploratory profiles disrupt it only if they are associated with exploitative researchers. The union of those two types of individuals leads to the most disruptive and distant knowledge combinations. To maximize

[2]More information can be found here: https://facultyopinions.com/

the relevance of these combinations, maintaining a limited number of highly exploratory individuals is essential, as highly specialized individuals must question and debate their novel perspectives. These specialized individuals are the most qualified to extract the full potential from novel ideas and situate them within the existing scientific paradigm.

The remainder of the paper is organized as follows. In section 3.2 we review the existing literature. Section 3.3 details the creation of our metrics and the methodology for addressing our research questions. Section 3.4 presents the results of our analysis. Section 3.5 concludes the paper and outlines future directions for developing novelty indicators.

3.2 Background and literature review

This section highlights the team's relevance in fostering creativity in science and emphasizes how team size can influence this process. We also underscore the importance of identifying the social dimensions of the team, a crucial factor in generating new knowledge. Finally, we propose a new approach based on the semantic representation of authors' past publications that allows studying the role of the cognitive dimension in a team's ability to produce new and impactful knowledge.

3.2.1 Team science as an engine of creativity

Over the past two decades, there has been a significant increase in interest surrounding the Science of Team Science (SciTS) [Falk-Krzesinski et al., 2011].[3] Since the 1950s, the average number of authors per paper has risen across all scientific disciplines [Wuchty et al., 2007]. Research collaborations have also become more diverse. Inter-institutional collaborations in science and engineering and social science grew by 32.8% and 34.4%, respectively, between 1975 and 2005 [Jones et al., 2008]. In addition, international collaboration has also expanded, with one in five research projects now involving multiple countries [Xie and Killewald, 2012].

Teamwork has proven to be a practical approach to producing impactful scientific results. Articles written by teams tend to have a higher impact, receiving more citations on average, and are more likely to become influential than articles authored solely [Wuchty et al., 2007, Whitfield, 2008]. Researchers benefit from collaboration

[3]For an up-to-date and comprehensive review, see Wang and Barabási [2021].

in various ways. Collaborative efforts can enhance rigor through co-authors' verification [Leahey, 2016] and facilitate the dissemination of their work beyond their immediate networks [Leahey, 2016]; this effect is further amplified when collaborations are international or inter-institutional [Adams, 2013, Jones et al., 2008]. Additionally, teams have better access to resources, as projects executed by groups are more likely to apply for funding and succeed in obtaining it [Rawlings and McFarland, 2011]. Teams are more likely to produce novel articles than solo-authored publications [Carayol et al., 2019, Uzzi et al., 2013, Wagner et al., 2019]. As highly cited work is often associated with a combination of novel and conventional ideas [Uzzi et al., 2013], teams of researchers may be more adept at generating novel ideas or striking a balance between novel and traditional concepts than individual authors.

Successful team performances put individuals and their interactions at the heart of the creative process. Over recent decades, the perception of teamwork has undergone significant changes. In the early 1990s, the prevailing belief was that groups should not be used for creativity because of inherent process loss in the creative process. This perspective has shifted dramatically, and team collaboration is now considered a critical factor in promoting creativity [Paulus and Nijstad, 2003]. Creativity relies on an individual's existing knowledge base: "*Creative thinking cannot happen unless the thinker already possesses knowledge of a rich and/or well-structured kind*" [Boden, 2001]. Knowledge exists on a continuum, ranging from explicit to tacit [Nonaka, 1994]. The generation of new knowledge occurs through interactions between explicit and tacit knowledge via a process known as the socialization, externalization, combination, and internalization (SECI) spiral. Tahamtan and Bornmann [2018] highlighted various approaches reported by researchers for fostering creativity. Engaging in conversations with colleagues seems to remain central to problem-solving and generating new, practical ideas. New ideas are becoming more challenging to discover as the idea space expands linearly while scientific publications grow exponentially [Bloom et al., 2020, Milojević, 2015]. As scientific knowledge increases, team sizes grow, and agents increasingly specialize their competencies [Jones et al., 2008, Jones, 2009].

The burst of possible combinations in the knowledge space suggests that agents can more effectively recombine information locally [Fleming, 2001]. "Local search" for an inventor involves exploiting existing combinations or using standard technological components. Agents tend to direct their research towards familiar subjects, focusing on topics related to their expertise or that of their co-authors (local search/-

exploitation) [Fleming, 2001, Nelson, 1985, March, 1991]. Conversely, exploration (or "distant search") is characterized by using new components or testing novel combinations [Fleming, 2001, March, 1991]. The nature of the new combinations realized depends on agents' trade-offs between exploiting and exploring the knowledge landscape. Exploitation reduces the risk of failure, as researchers draw from experience with combinations and architectures that have previously failed [Vincenti, 1990]. Researchers must then collaborate with others to explore the knowledge space more efficiently, and the team's composition might determine this balance between exploration and exploitation.

3.2.2 Team characteristics in the creative process

We review here some dimensions of the team composition that affect the scientific process.

Size dimension: The importance of co-authors during the process of creativity has been debated in the literature, and the effect of team size and composition on creativity has been the focus of multiple studies [Paulus and Nijstad, 2003, Shin and Zhou, 2007, Wuchty et al., 2007, Falk-Krzesinski et al., 2011, Erren et al., 2017, Mueller, 2019]. Team size shapes and is shaped by the nature of the work carried out. Large teams tend to be more risk-averse and consolidate a field rather than introducing new opportunities [Christensen and Christensen, 2003, Paulus et al., 2013, Lakhani et al., 2013, Wu et al., 2019]. Larger teams use more up-to-date and influential research in their work, consequently fostering greater engagement within their scientific community and further increasing their impact [Wu et al., 2019]. However, large teams are more prone to coordination and communication failures as the entire team must have faith in the project to succeed. The agreement and communication between team members can be challenging and time-consuming [Bikard et al., 2015]. In fact, the number of people involved in a project can have heterogeneous effects on creativity, and no optimal team size fits every project. A small team may be more useful in the conceptualization phase, while a larger team might be beneficial in the implementation and testing phase of the project [Wang and Barabási, 2021]. Shin and Zhou [2007] highlight the organization's importance for creativity. Using evidence from Cambridge and AT&T's Bell Laboratories (home to numerous Nobel Prize winners), they discuss researchers' ideal context for fostering creativity and conclude that the presence of a healthy environment for a small group of people (up to seven) promotes creativity. These results are further confirmed by Lee et al.

[2015] and Carayol et al. [2019], indicating that the relationship between team size and novelty appears U-shaped and is highly heterogeneous across disciplines.

Structural and relational social capital: Nahapiet and Ghoshal [1998] conceptualize three dimensions of social capital that impact intellectual capital development: structural, relational, and cognitive. Though primarily used to understand intellectual capital development in organizations and firms, the dimensions of social capital presented in Nahapiet and Ghoshal [1998] can be applied to the context of knowledge production in science due to their intrinsic relevance to relationship and network dynamics [Liao, 2011]. Structural capital examines the links between individuals, and structural distances have been widely studied through collaboration networks (see Kumar [2015] for an extensive review on network collaborations). Relational capital represents the nature and intensity of the connections between team members. A critical factor in intellectual development is the ability to communicate with each other, and the actors' experience reinforces the phenomena [Taylor and Greve, 2006, Liao, 2011, Kelchtermans et al., 2020]. For instance, McFadyen and Cannella Jr [2004] emphasize the role of the intensity of past relationships between scientists in fostering new knowledge. Indeed, members with strong relationships, norms, obligations, and mutual trust tend to communicate more easily [Liao, 2011]. Other relational aspects, such as hierarchical or geographical dimensions, also impact the knowledge space exploration. For example, supervising doctoral students is not only associated with entering new areas but also extending towards more distant fields [Kelchtermans et al., 2020].

Cognitive social capital: The cognitive capital remains challenging to measure as it is linked to the shared background between coauthors and their common language. Cognitive diversity is often encouraged through interdisciplinary projects as the intersection of different perspectives is commonly required to solve complex scientific problems [pag, 2007]. Indeed, people from outside a domain may have some advantage to offer fresh ideas through their distinct knowledge [Jeppesen and Lakhani, 2010, Kuhn, 1962]. The effectiveness of generating new knowledge is impacted by factors such as variations in background, belief, and reasoning styles among scientists, all of which contribute to cognitive diversity. The cognitive distance between team members is expected to display an inverted U-shaped correlation with both learning and innovation [Nooteboom et al., 2007], as people being too distant will

face difficulty in communicating, and those being cognitively too similar benefit less from distinct perspectives in the knowledge creation process.

Cognitive distances between individuals can be studied through various metrics. Kumar et al. [2017] used, for example, citations networks and citations context in full text. Boudreau et al. [2016] represented the cognitive distance between funding evaluators and the proposal through MeSH terms similarity. Similarly, Ayoubi et al. [2017] represent the distance between the focal scientist and her team by comparing cosine similarities of referenced journals from scientists' past publications. Other measurements, without being explicit, may relate to cognitive dimensions, Wagner et al. [2019] discovered that international collaborations negatively affect novelty and produce more conventional knowledge combinations, highlighting barriers and transaction costs that influence the production of creative work. Finally, measures of cognitive distance strongly relate to interdisciplinarity. Petersen et al. [2021] represent author diversity using the discipline of the institution. Using authors' disciplinary diversity, Abramo et al. [2018] show that more distant coauthors produce articles with more diverse references.

Exploratory profile: Individual characteristics and the ability to interact with individuals from different fields are essential to efficiently managing cognitive diversity in a team. When the distance between disciplines is too high, a "Renaissance" individual [Jones, 2009] can ease their connection [Wu et al., 2022]. The presence of a scientist with a multifaceted profile bridges the gap between the different backgrounds of other team members. This is crucial as a shared knowledge base between researchers streamlines the socialization process and facilitates knowledge recombination, fostering creativity. Shin and Zhou [2007] focused on the relationship between diversity (interdisciplinarity) and creative ideas in groups. Shin and Zhou [2007]'s idea is that the presence of a "transformational leader", whose role is to mediate between individuals, each specialized in a different field, leads to greater team creativity. Xu et al. [2022] provided a first answer to this hypothesis by examining the share of team members engaged in the conceptual work, the L-ratio, which was deduced from the analysis of author contribution reports. The findings suggest that hierarchical teams generate less novelty than egalitarian teams and tend to develop existing ideas more frequently.[4] We argue that the notion of transformational leader

[4]Through Louvain algorithms, they identified clusters of co-occurring research activities in their first dataset. They then built a neural network to infer author roles based on their characteristics and predicted it for 16 million articles on Microsoft Academic Graph (MAG).

or renaissance individual is connected to exploratory profile *à la* March [1991], individuals enabled to link others in the knowledge space due to their ability to navigate in different spaces.

3.2.3 Exploring the cognitive dimension

We investigate scientific impact through citation networks and recent indicators of disruption and breath and depth [Wu et al., 2019, Wu and Wu, 2019, Bu et al., 2019, Bornmann et al., 2019a]. These indicators determine whether a document consolidates a domain or constitutes a founding step. To explore its influence on novelty, we use two approaches, one based on combinatorial novelty indicators [Uzzi et al., 2013, Lee et al., 2015, Foster et al., 2015, Wang et al., 2017, Shibayama et al., 2021] and one based on external validation via Faculty Opinion (previously called F1000) following Bornmann et al. [2019b]. Faculty Opinion is a website hosting reviews of papers tagged as presenting "New Results", "Novel Drug target", "Technical advancement", "Interesting hypothesis", and "Controversial results", among other categorizations labeled by experts in the field. It allows us to empirically assess the capacity of novelty indicators and our indicators to predict the novelty as perceived by other researchers in the community.

Novelty indicators have been compared and evaluated based on citation count [Uzzi et al., 2013, Lee et al., 2015, Foster et al., 2015, Wang et al., 2017]. Fontana et al. [2020] compared Wang et al. [2017],Uzzi et al. [2013], and Lee et al. [2015] using randomized citation networks and demonstrated the ability of the Uzzi et al. [2013], Lee et al. [2015] indicators to better track novelty. Their findings are supported by using some Nobel Prize winners' articles and a list of APS milestone articles. Other studies have evaluated these indicators based on surveys, such as Shibayama et al. [2021] and Matsumoto et al. [2021], whereas Bornmann et al. [2019b] have evaluated them based on labels collected on Faculty Opinion and found similar results as in Fontana et al. [2020]. However, only a few indicators have been compared and tested simultaneously. This study intends to validate the effect of the cognitive dimension on a large variety of metrics.

Our indicator is not a substitute for other novelty indicators. It does not represent the novelty of an article as it is based upon previous information and would be similar even without the focal article. Instead, it provides an understanding of team composition that would benefit creativity in science. We can think of our measure as a measure of *potential novelty*, i.e., opportunities for new knowledge recombina-

tion available through the diversity of background in the team and the capacity of individuals to bridge the gap between other team members. In comparison, combinatorial novelty indicators would capture then the *realized novelty*, i.e. the output of the research conducted by this team in terms of pieces of knowledge used. Finally, Faculty Opinion labeling and other external validation methods can describe the *perceived novelty*, i.e., the peers' perception of this study. Hence, in these terms, we ask whether potential novelty contributes to realized and perceived novelty and its scientific recognition. Two research questions can be drawn regarding the effect of the cognitive dimension on creativity. Do teams with higher cognitive diversity are more likely to approach a subject creatively, demonstrating originality (*perceived* and *realized*) and recognition? Does the presence of *explorative* individuals within a team enhance communication among members and facilitate their exploration of the knowledge space to develop new and relevant solutions to research problems? Studying the cognitive dimension of creativity in science is of great interest, especially as it can help identify how to improve collaboration and communication among researchers with diverse cognitive profiles. Through our metric, we also offer a different approach to resource allocation decisions, giving another picture of teams with a high potential for creative output.

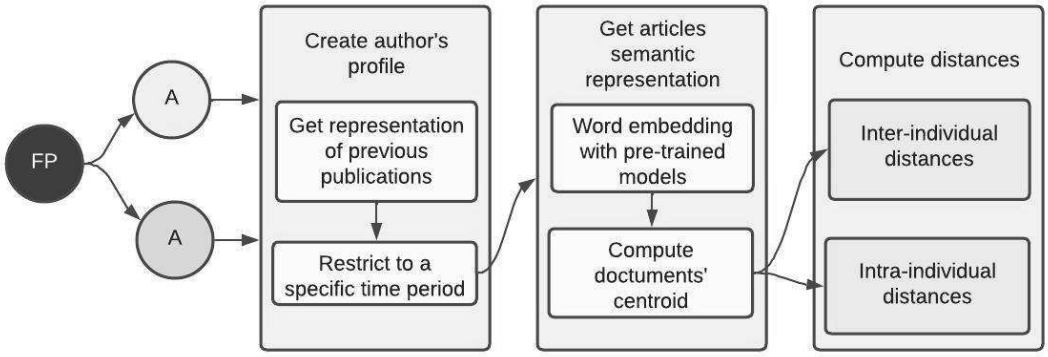
3.3 Data and methods

3.3.1 Measuring cognitive diversity and exploratory profile

The proposed metric examines the semantic heterogeneity of researchers' work as a proxy for their cognitive diversity. Thus, it offers an alternative to using categories, keywords, or citation networks, more complex to be monitored directly by the researchers themselves. Following Hain et al. [2020] and Shibayama et al. [2021], we can embed this list of documents in a vectorial space to apply a distance measure such as cosine similarity [Mikolov et al., 2013]. We assume that an author of a paper in a specific position within the semantic space possesses knowledge embedded around that position. Our indicator has two properties: it offers a measure of researchers' profiles at the individual level and a measure of distances between them. Consequently, we can proxy the trade-off between exploitation and exploration that a researcher undergoes throughout their career (intra-individual) and the trade-off materializing during the formation of a team (inter-individual) within the

same mathematical space.

Figure 3.1: Construction of the indicator



As explained in Figure 3.1, we track authors to create a list of authors' past publications. Then, we can create a cognitive profile for each author at a given time t; each publication is embedded in the semantic space and represents the cognitive landscape of the author. We restrict to publications up to b years before t to account for researchers' current topics of interest and difficulty retaining information [Argote et al., 1990]. We can finally define a researcher's exploratory profile at time t by calculating pairs of cosine distances between past papers published. This will create a density of cosine distances which, using the taxonomy of March [1991], can be interpreted the following way: the fatter the right (left) tail is, the more exploratory (exploitative) the researcher. The same holds for the team. A sizeable right tail indicates cognitively distant researchers within the team. This provides us with information on how distant their knowledge base is from others. The greater the distance, the less likely their respective knowledge space can be combined, thus affecting the probability of combining novel ideas. An intra-author and inter-author distribution enables a comprehensive exploration of the relationship between novelty, creativity, and teams.

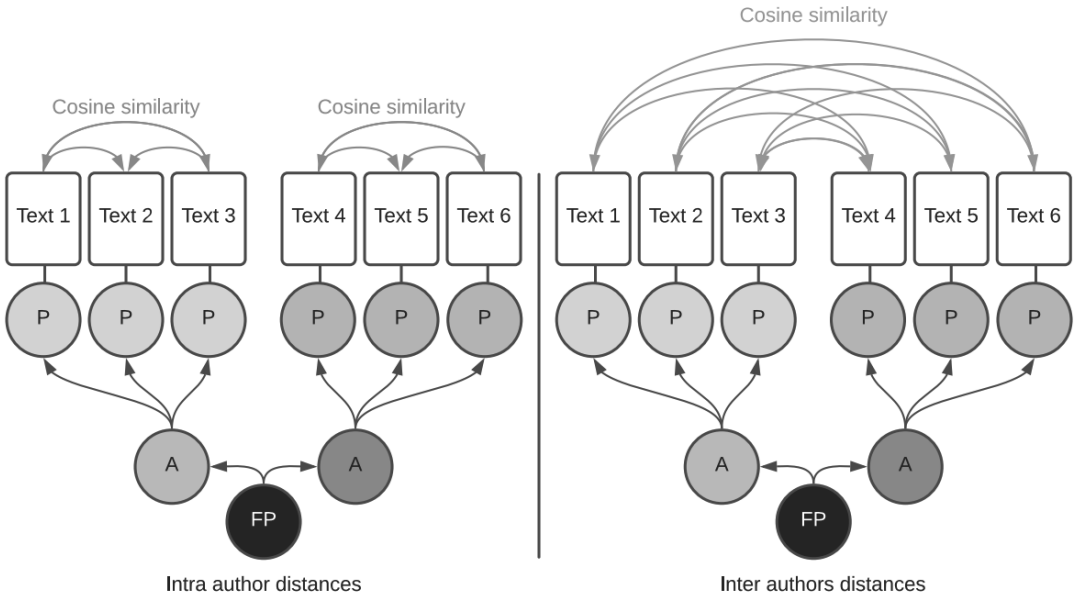


Figure 3.2: Exploratory profile and cognitive diversity

We model our measures on two different perspectives as represented in Figure 3.2: intra-author distances, which asses exploratory profile, and inter-authors distances, to capture cognitive diversity. A given focal paper (FP) is written by two authors named 'A'. We retrieve each author's past production, named 'P'. On the one hand, we can then calculate the distance between all publications from a given author (intra-author distances). On the other hand, we can also compare past publications from two authors (inter-author distances). We can build our framework using directed bipartite networks, defined as $G(U, V, E)$. U represents the set nodes for authors, V is the set for articles, and E is the set of links between authors and articles.

We only consider collaborations between authors when looking at a given article; collaboration is implicit since the set of parents of a given document $FP \in V$ corresponds to the set of authors that collaborate. For a given document FP, an author that has contributed to FP is noted a, and the set of nodes that contributes to FP is then $In_{FP} = \{a \in U : (a, FP) \in E\}$.

We want to retrieve all past publications for all authors in In_{FP}. The global set of publications before FP is noted V_{FP}^{t-b}, the set of articles published b years before the document FP. The set of past publications for author $a \in In_{FP}$ is noted $Out_a^{t-b} = \{v \in V_{FP}^{t-b} : (a, v) \in E\}$. For a document FP and an author $a \in In_{FP}$, we retrieve

the set of past publications Out_a^{t-b}. All Out_a^{t-b} elements vectorial representations are compared, from which distribution of cosine distances are calculated. The distance between two documents $i, j \in Out_a^{t-b}$ is $d_{ij} = 1 - COS(T_i, T_j)$ where T_i is the dense vector text representation for document i.

Intra-author semantic distances: A distribution of semantic distance score D_a is computed through cosine similarity using all document $i, j \in Out_a^{t-b}$, the process is repeated for each authors $a \in In_{FP}$. The intra-author distance for a given author a is the q-th percentile (P_q) of this distribution and is written as:

$$Intra_a = P_q(D_a)$$

A general distribution of the intra-authors publication distances is constructed using the set of distances for all authors $A_{FP} = \{D_a : a \in In_{FP}\}$, the individual trade-off between exploitation/exploration is then captured through the average of the exploratory profiles in a given team.

$$Intra_{FP} = \frac{\sum_a (P_q(D_a))}{|In_{FP}|}$$

Inter-authors semantic distances: A distribution of semantic distance score between authors' previous work is constructed by comparing different authors' publications. For two given authors $a, e \in In_{FP}$, $|Out_a^{t-b}| \times |Out_e^{t-b}|$ distances are used to construct the distribution of distances $D_{a,e}$ between a and e. The final distribution then groups together all distances between authors' previous works $B_{FP} = \{D_{a,e} : a, e \in In_{FP}\}$, the trade-off between exploitation/exploration in team composition is captured through the percentile of B_{FP}:

$$Inter_{FP} = P_q(B_{FP})$$

Current techniques for large-scale author disambiguation allow the investigation of individual trajectories in science. However, the use of this information comes with a computational cost. This indicator pushes towards a massive use of data because one needs all authors' past publications for a given set of documents. Structuring the data to compute the measure is time-consuming and data-intensive. All papers' text from all authors in a given database are required. However, using pre-trained embedding models allows direct computing indicators without the requirement of

complete database access. Therefore, measures are not dependent on the study sample as indicators of novelty based on cooccurrence matrices but rather on the sample used to train the model. Also, by processing titles and abstracts through embedding techniques, the authors' background is represented with greater granularity than through the keywords or the journals where the authors have been published.

3.3.2 Data

Scripts to reproduce the analyses are available on GitHub `https://github.com/Kwirtz/Unpacking-scientific-creativity` and data for the regressions is available here `https://zenodo.org/record/8382881`. Our analysis relies on two databases. The first, PubMed Knowledge Graph (PKG), allows us to test the effect of the cognitive dimension on scientific impact and *realized* novelty of articles. In contrast, the second, Faculty Opinion, verifies whether the cognitive dimension affects the *perceived* novelty by peers.

We use Pubmed Knowledge Graph (PKG), a collection of 35 million scientific papers and books from life science and biomedical journals provided by the National Library of Medicine (NLM) at the National Institutes of Health (NIH). Authors are disambiguated by leveraging Natural Language Processing (NLP) and online data, as outlined by Xu et al. [2020]. We based our analysis on all the 3.5M articles written by 3,276,250 authors and published in 9,348 journals between 2000 and 2005. We selected fairly old data due to the nature of the process studied. Indeed, novel articles are more likely to become "sleeping beauties" and accumulate citations in the long run [Lin et al., 2021]. Also, to compute novelty indicators, we require information about references. We rely both on abstracts of references to embed their semantics and calculate the distance as in Shibayama et al. [2021]. Also, we use past publication references' journals to build past cooccurrence matrices used to capture combination existence and difficulty for other novelty indicators. For this purpose, we used the database between 1980 and 2005 to get all the information needed, representing 11,261,955 documents.

To test if our indicators affect the novelty perceived by peers, we used Faculty Opinion following Bornmann et al. [2019b]. Faculty Opinion is a database featuring papers tagged as presenting 'New Results', 'Novel Drug target', 'Technical advancement', 'Interesting hypothesis', and 'Controversial results', among other categorizations determined by the platform users. The platform hosts reviews of the most significant research in Biology and Medicine. This makes it easy to match the arti-

cles in the database with PKG. Indeed, from the 190k articles in Faculty Opinion, we found 27,122 in our sample (2000-2005).

3.3.3 Empirical strategy

To explore the relationship between the team's cognitive dimension and its ability to recombine pieces of knowledge in novel ways and achieve recognition, we start with a basic exploratory data analysis followed by three econometric analyses to test our hypotheses.

The first two analyses aim to understand how a team's cognitive diversity and the exploratory profiles of its members impact *perceived* novelty (i.e., peer labeling on Faculty Opinion) and *realized* novelty (i.e., indicators of combinatorial novelty). Then our analysis seeks to comprehend the effect of the cognitive dimension on scientific recognition using citation and disruption measures.

Realized novelty and scientific impact connections with cognitive dimension are both investigated through PKG, the normalization performed at the field and year levels of this measure provides a value ranging between 0 and 1, which we model using linear models with cluster robust standard errors at the journal level. Lastly, we examine how the presence of highly exploratory and exploitative individuals influences the team's creativity. This analysis will help determine if cognitive diversity and the presence of exploratory profiles are explicitly visible in an article's knowledge composition.

For the analysis of *perceived* novelty, we employ the Faculty Opinion database and model, through Logit and Poisson regressions, the likelihood of an article being labeled with "novel" categories ("Technical Advance", "Interesting Hypothesis", "Novel Drug Target"). In our sample, 80% of the observations are labeled as 'New Findings', and 95% of the total sample would be considered new using the top 4 most represented categories (22,216 novel articles versus 1,750 not-novel). The fact that most articles are labeled as new findings makes this category less informative; therefore, we decided to exclude it and remove articles solely labeled with this category. As a result, our prediction is based on a more balanced sample (8,950 novel articles versus 3,605 not-novel). This will enable us to understand whether the cognitive dimension is associated with *perceived* novelty. We do not expect a direct effect but rather hypothesize that cognitive diversity influences a latent variable representing the article's actual contribution. This actual contribution of the paper may or may not be visible in the *realized* novelty measured by novelty indicators but might be

then reflected in labeling made by peers.

3.3.4 Variables

Variables used in our empirical analysis can be separated into four categories: novelty indicators, scientific impact, cognitive, and control variables. For control variables, aside from data from PKG, we use journals listed in Scimago to control for scientific domains and measure of the impact associated with the journal. Each of our variables is at the paper level. For the empirical strategy, novelty, impact and cognitive measures will be field weighted by year using the percentile rank procedure – noted (FW). We use the first category of the journal from Scimago to approximate the field.

Novelty indicators

The indicators used in our analysis are Uzzi et al. [2013], Lee et al. [2015], Foster et al. [2015], Wang et al. [2017], Shibayama et al. [2021]. A formal mathematical description of them can be found in Chapter 2 of this thesis. Note that we have inversed the sign of the measures related to Uzzi et al. [2013] for simplicity and comparison with other indicators. The computation is done with *Novelpy*.[5]

Scientific impact variables

For impact measures, we use citation counts and disruption indicators, also described in Chapter 2. We used all available indicators in *Novelpy*, namely: Wu et al. [2019], Bu et al. [2019] and Bornmann et al. [2019a].

Cognitive variables

Team cognitive diversity: The mean of the inter-authors semantic distance as defined in Section 3.3.1 with q=90 for a given paper. It measures to what extent a team is composed of highly cognitively distant authors (i.e. Author 1 background is vastly dissimilar to Author 2 background). Furthermore, we suppose the relation between the team's cognitive diversity and other measures is not linear. We take the square of the team's cognitive diversity to test this.

[5] *Novelpy* is a python package that allows computing novelty and disruption indicators. More details can be found here: https://novelpy.readthedocs.io/

Average exploratory profile: The mean of the intra-authors semantic distance as defined in Section 3.3.1 with q=90 for a given paper. It captures to what extent a team comprises authors with distant past publications (i.e. Author 1 worked on diverse subjects). As for team cognitive diversity, we add a square term in the regressions.

Number of highly exploratory authors: To have more information on the team structure, we decided to define a threshold to identify highly exploratory authors. Looking at the intra-author's semantic distance as defined in Section 3.3.1. An author is considered highly exploratory if its 90^{th} percentile is in the top 10% of all $Intra_{FP}$ in our sample.

Number of highly exploitative authors: We expect highly exploratory authors to work best with highly exploitative authors (i.e. Novelty is probably most successful with a combination of typical and atypical individuals). We construct this measure following the same procedure as exploratory authors. Looking at the intra-author's semantic distance as defined in Section 3.3.1. An author is considered highly exploitative if its 90^{th} percentile is below our sample's median of all $Intra_{FP}$.

Interaction term between highly exploratory and highly exploitative authors: We added an interaction term between the two types of profiles as both competencies might complement each other.

Control variables

We included as control variables the number of authors, references and MeSH terms. We also controlled for the year and information related to the journal of publication.

Scimago Journal Ranking (SJR): An indicator of a journal's prestige based on weighted citation and eigenvector centrality derived from Scopus' citation networks by Scimago [González-Pereira et al., 2009].

Scimago Journal Category: Scimago provides a classification of journals based on various fields. We used the first category linked to a journal; our database contains journals from 271 categories.

3.3.5 Descriptive statistics and preliminary evidence

We further clean our database and restrict it to papers with at least 2 references/ MeSHterms/ authors and with a journal ISSN. Our final dataset represents approximately 2.1M articles.

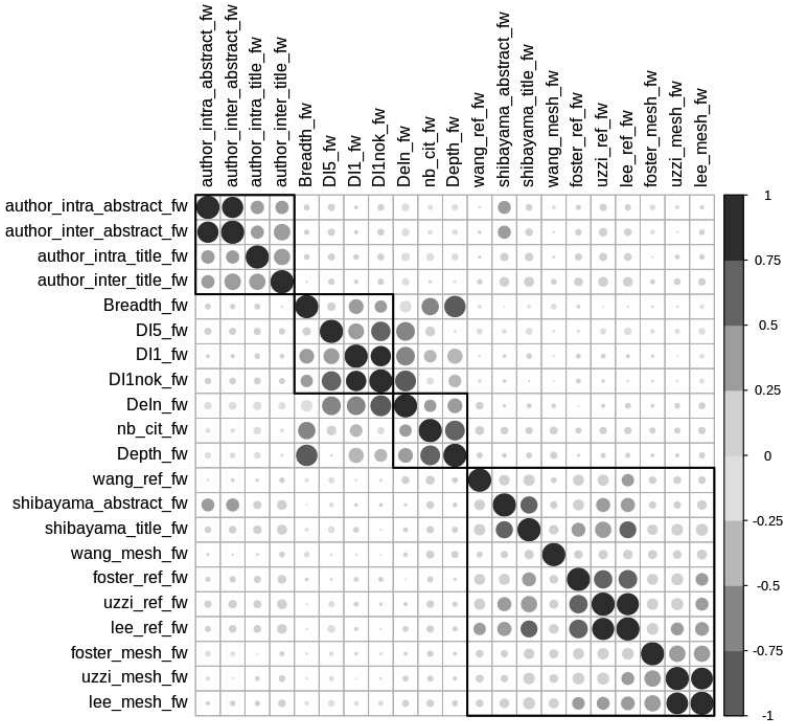
Table 3.1 presents the descriptive statistics for the variables in our sample. Examining their distribution, it is worth noting that some indicators concentrate novelty around a small number of articles, as in Foster et al. [2015] or in Wang et al. [2017], merely 21% of the articles possess non-zero values (measured on references). Also, indicators such as citation count or Uzzi et al. [2013] among others, display relatively extreme values. Specifically for Uzzi et al. [2013], it is highly dependent on the z-score computation, when the variance of the journal combination is minimal, the z-score can rapidly become substantial. These disparities in distribution prompted us to apply a percentile rank procedure by field and year, as explained in the previous subsection.

Table 3.1: Descriptive statistics

Statistic	Min.	Pctl(25)	Median	Mean	Pctl(75)	Max	St. Dev.	N
# References	2	12	22	27.37	36	2690	25.76	2108280
# Meshterms	2	9	13	13.25	16	51	5.19	2108280
# Authors	2	3	4	5	6	282	2.94	2108280
# Citations	0	9	22	46.99	50	81577	129.47	2108280
SJR	0.1	0.627	1.130	1.787	2.035	39.946	2.22	2094669
Disruption_1	-1	-0.007	-0.001	0.003	05179	1	0.06	2108280
Disruption_{1noK}	-1	-0.588	-0.269	-0.192	0.111	1	0.51	2108280
Disruption_5	-1	0	0.001	0.018	0.009	1	0.07	2108280
Disruption_{DeIn}	0	0.79	1.662	2.067	2.875	92.5	1.81	2108280
Breadth	0	0.307	0.5	0.517	0.714	1	0.26	2108280
Depth	0	0.258	0.5	0.458	0.672	1	0.26	2108280
Share Exploratory	0	0	0	0.063	0	1.0	0.14	2108280
Share Exploitative	0	0	0.333	0.365	0.6	1	0.32	2108280
Author intra $_{abs}$	0	0.22	0.29	0.29	0.36	1.02	0.09	1837749
Author inter $_{abs}$	0	0.26	0.33	0.33	0.40	1.02	0.09	1837748
Shibayama $_{abs}$	0	0.222	0.274	0.275	0.327	0.991	0.07	2081854
Uzzi_{Ref}	-62396.32	-7.34	3.66	-18.03	14.02	199.49	206.82	1891079
Lee_{Ref}	-17.581	0.145	0.840	0.567	1.466	6.006	1.45	2092283
Foster_{Ref}	0	0.117	0.4	0.366	0.583	1	0.25	2092283
Wang_{Ref}	0	0	0	0.583	0	2872.106	4.79	2092283
Uzzi_{Mesh}	-287.0	-1.1	0.9	2.7	4.5	189.1	8.19	765751
Lee_{Mesh}	-7.996	0.4562	0.807	0.794	1.174	4.717	0.60	2105186
Foster_{Mesh}	0	0.274	0.476	0.424	0.591	1	0.22	2105186
Wang_{Mesh}	0	0	0	0.299	0.307	28.668	0.76	2105186

The correlogram in Figure 3.3 illustrates the various indicators' interconnection. A hierarchical clustering algorithm is applied to the correlation matrix and several clusters emerge. It includes citation and consolidation indicators, novelty indicators, cognitive dimension indicators, and disruption indicators. Regardless of whether MeSH terms or references are used to derive the indicators, the novelty indicators group remains consistent, suggesting that combinatorial novelty indicators capture a shared underlying dimension of innovation in scientific research. The correlation between Lee et al. [2015] and Uzzi et al. [2013] is particularly robust since both measures are nearly identical except for the incorporation of the reference's publication year in Uzzi et al. [2013]'s resampling process. It should be noted that a negative correlation is expected since low values signify atypicality in Uzzi et al. [2013], while high values represent novelty in Lee et al. [2015], this is why we inverse the sign of Uzzi et al. [2013] to get positive correlation between indicators. A strong correlation is observed between Shibayama et al. [2021] and our indicators, as it employs the same measurement on references, and some elements may overlap. Specifically, self-citation increases the correlations between Shibayama et al. [2021] and our indicator since the same combinations are calculated in the author and reference parts. Moreover, the clustering differentiates between citation count, consolidation indicators (Depth, DeIN), and disruption indicators (DI1, DI5, DI1nok, and Breadth). These distinctions emphasize how consolidation indicators are more closely related to citation count and demonstrate how disruption indicators capture other dimensions of scientific impact.

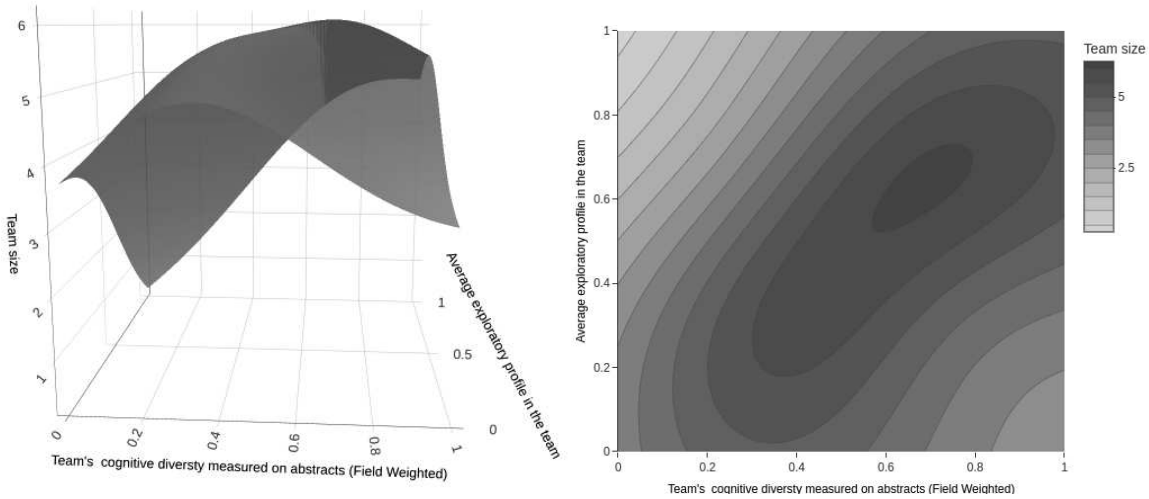
Figure 3.3: Correlogram with hierarchical clustering



The development of an author-level indicator necessitates examining its relationship with team size. Figure 3.4 illustrates how intra- and inter-individual cognitive indicators are strongly associated with team size. Although it is unclear whether cognitive diversity generates a specific team size or if team size produces this diversity, it is visible that as the cognitive diversity within a team increases, the average exploratory profile must also rise to maintain a comparable team size. The U-shape relationship on both sides is easily observable, suggesting that the more diverse the team and/or the more exploratory the individuals, the smaller the team. Conversely, highly homogeneous teams typically imply smaller average team sizes, even if the average exploratory profile is high. This pattern is partially attributable to the construction of our indicator, which averages distance. In larger teams high distance between members might be compensated by other members that are close to each other. This counterbalancing is less pronounced in smaller teams, resulting in more extreme values. However, several explanations for this phenomenon can be offered. For instance, substantial cognitive diversity might create communication barriers among team members, particularly when individuals are less explorative. Consequently, smaller teams are formed due to potential coordination and knowl-

edge exchange difficulties. In cases with a high average exploratory profile combined with high cognitive diversity, forming smaller teams may be more convenient, as researchers might explore the knowledge space too broadly. Smaller teams could help prevent efforts from dispersing in various directions. As for teams with low cognitive diversity, the absence of cognitive diversity and exploratory profiles could relate to niches where individuals possess similar knowledge and expertise. As a result, many team members might not be necessary, as they can efficiently navigate the local knowledge space. The same argument can be made for individuals with comparable skills and exploratory profiles, as they may represent teams that regularly collaborate on diverse topics. The distinct skill requirements for these teams may be lower, leading to smaller team sizes.

Figure 3.4: Team size, exploratory profiles and cognitive diversity

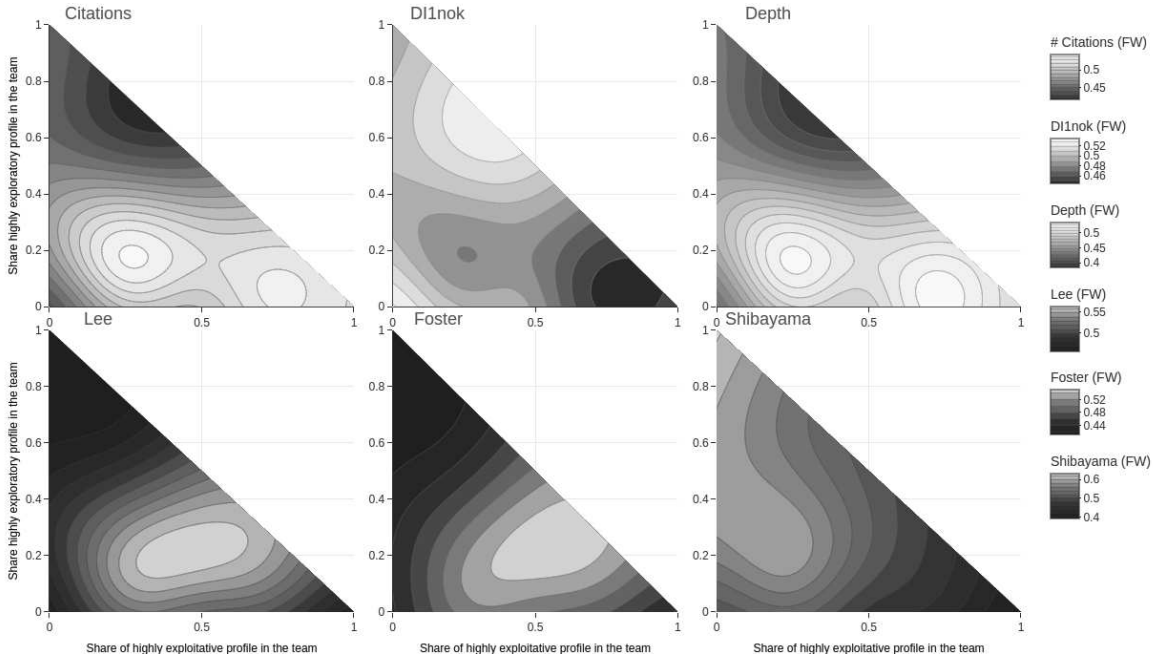


Interestingly, when comparing the analysis of team size with disruption, we confirm the findings of Wu and Wu [2019]. As illustrated in Figure 3.6 in the Appendix, peripheral observations are more disruptive (represented with DI1nok), corresponding to the location of our smaller teams in Figure 3.4 right panel. Teams consolidating science, as indicated by the Depth variable, are also, on average, the most prominent teams. Small teams that disrupt science tend to have exploratory profiles and/or diverse team compositions. Science disruption seems to occur through small teams with either highly distinct skills or very exploratory profiles. What seems essential is the ability to access a broader knowledge space, regardless of whether this space is reached through the team's highly explorative profiles or the team's diversity. Teams

composed of individuals who are, on average, highly exploratory but with low team cognitive diversity represent teams with similar skills that cover the knowledge space effectively. In contrast, highly diverse teams with specialized individuals also span the knowledge space to propose disruptive ideas, although they may face communication challenges. The combination of these two factors also appears to contribute to disruption, albeit less prominently, suggesting the detrimental effect of excessive diversity. Another inverted U-shaped relationship exists between a team's average exploratory profile and novelty indicators. When balanced by a relatively explorative average profile, cognitive diversity appears beneficial without showing a saturation point.

The relationship differs when we adopt an alternative perspective and consider the proportion of highly exploratory and exploitative individuals within scientific teams. A dome is visible in each indicator, signifying successful trade-offs between exploitation and exploration. Figure 3.5 offers insight into the relationship between these two aspects and scientific recognition and combinatorial novelty. Teams with fewer highly exploratory individuals and a higher proportion of highly exploitative individuals typically contribute to consolidating the field (Depth metric). Conversely, groups with a higher proportion of highly exploratory individuals and a smaller proportion of highly exploitative individuals are more likely to initiate disruptions in their fields (DI1nok metric). These observations complement the findings of Uzzi et al. [2013], which suggest that a balance between conventional and atypical knowledge combinations produces the most impactful research. Moreover, this analysis enables us to examine how the balance between exploratory and exploitative individuals affects knowledge creation.

Figure 3.5: Relation between the share of highly exploitative and highly exploratory profile in a team with and Novelty/ Scientific Impact



Teams featuring a fair proportion of exploratory individuals and a more sustained level of exploitative individuals seem to be most likely to generate compelling new combinations of knowledge. Figure 3.5 suggests that an optimal team composition would consist of approximately 50% highly exploitative and 20% highly exploratory individuals to increase the likelihood of combining distant knowledge. The situation is less clear for Shibayama et al. [2021], where a high proportion of highly exploratory individuals appears to be beneficial.[6] Exploratory individuals contribute to the team by introducing fresh and innovative ideas from their extensive knowledge. These individuals can challenge conventional thinking and steer the team in new directions. Simultaneously, they might foster communication among group members with distant knowledge. In contrast, highly exploitative individuals are crucial for refining and optimizing these novel ideas. Their specialized expertise allows the team to identify feasible and effective solutions, ensuring the creative potential of the exploratory individuals is appropriately channeled into tangible outputs. Additionally,

[6]This might be connected with the relationship between our measure and the measure of Shibayama et al. [2021] as it is measured in a similar manner. Self-citation also directly impacts the relationship between these two metrics as the same combination of articles will be calculated in both metrics.

their deep understanding of a specific field facilitates effective communication. The highly exploratory profile complements the specialized knowledge and proficiency of the highly exploitative team members. This dynamic enables the team to capitalize on the full potential of their diverse cognitive abilities, optimizing the innovation process and yielding scientific advancements.

3.4 Results

3.4.1 Cognitive dimension and novelty

3.4.1.1 Realized novelty

This subsection examines the relationship between the team's cognitive dimension and novelty indicators. To this end, we report the results of an OLS to identify the joint impact of authors' intra-diversity and inter-diversity on the indicators. The outcomes of these models are presented in Table 3.2.

First, we confirm that cognitive diversity in a scientific team fosters realized novelty. Team cognitive diversity (Row 1-2) reveals a significant positive effect on combinatorial novelty. This suggests distant individuals can ease the combination of distant journals in the references. The squared term has negative coefficients. However, the turning point is higher than 1, meaning the relationship is strictly increasing (See Table 3.13 in Appendix). However, it means that the marginal benefit of cognitive distance is decreasing. When interpreting the coefficients, it is important to remember that the independent and dependent variables are expressed in percentile rank within a given field and year. A one percentage point increase in the independent variable's percentile rank implies a β percentage point increase in the dependent variable. In our case, the marginal effect of a quadratic term depends on the value of the independent variable. We can calculate marginal effects at the mean values of the independent variable. For example, in Uzzi et al. [2013] (model 1), the marginal effect of Author inter$_{abs}$ (FW) at the mean value is calculated this way: $\frac{\Delta y}{\Delta(inter)} = 0.169 - 2 * (-0.031) * Mean(Inter)$. Since variables are expressed in percentile rank, the mean and the median are 0.5. The marginal effect can be then calculated easily, $\frac{\Delta y}{\Delta(inter)} = 0.169 - *(-0.031) = 0.2$. This means that by increasing one percentage point on the ranking of team diversity in a given field and year, one can increase by 0.2 percentage points in the ranking of the most novel articles in the field and year.

Table 3.2: Combinatorial Novelty: cognitive diversity and average exploratory profile (Field-Weighted/ References)

	Dependent variable:				
	Uzzi	Lee	Foster	Wang	Shibayama
	(1)	(2)	(3)	(4)	(5)
Author inter $_{abs}$ (FW)	0.169***	0.166***	0.116***	0.098***	0.284***
	(0.008)	(0.007)	(0.010)	(0.006)	(0.007)
Author inter $^{2}_{abs}$ (FW)	-0.031***	-0.034***	-0.023**	-0.028***	-0.118***
	(0.007)	(0.007)	(0.009)	(0.006)	(0.007)
Author intra $_{abs}$ (FW)	0.056***	0.043***	0.041**	-0.002	0.188***
	(0.014)	(0.013)	(0.019)	(0.008)	(0.009)
Author intra $^{2}_{abs}$ (FW)	-0.088***	-0.094***	-0.084***	-0.026***	-0.047***
	(0.011)	(0.010)	(0.015)	(0.006)	(0.010)
# References	0.002***	0.002***	0.001***	0.005***	0.002***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
# Meshterms	0.004***	0.006***	0.005***	-0.001***	0.004***
	(0.0004)	(0.0004)	(0.0004)	(0.0002)	(0.0004)
# Authors	0.008***	0.007***	0.007***	0.001***	0.007***
	(0.0004)	(0.0004)	(0.0005)	(0.0003)	(0.0003)
SJR	-0.012***	-0.011***	-0.014***	-0.008***	-0.011***
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
Year	Yes	Yes	Yes	Yes	Yes
Journal Cat.	Yes	Yes	Yes	Yes	Yes
Observations	1,647,430	1,815,603	1,815,603	1,815,603	1,809,155
R^2	0.055	0.062	0.039	0.122	0.130
Adjusted R^2	0.055	0.062	0.039	0.122	0.130
Residual Std. Error	0.281	0.278	0.310	0.345	0.267
F Statistic	406.544***	512.283***	315.079***	1,065.575***	1,143.840***

Notes: This table reports coefficients of the effect of cognitive diversity and average exploratory profile on combinatorial novelty using PKG. Standard errors are cluster robust at the journal level: $***$, $**$ and $*$ indicate significance at the 1%, 5% and 10% levels, respectively. The effects are estimated with an OLS. Variables are field-weighted and constant term, scientific field (Scimago Journal Category), and time-fixed effects are incorporated in all model specifications.

On the contrary, the average exploratory profile must remain reasonable to maximize novelty. As visible in Table 3.13, the turning points are around 30% for all indicators except Shibayama et al. [2021], for which it is upper than one. This can mean two things, and this is what we will examine in the second part of this results section, either the researchers have a relatively moderate explorative profile,

or there is a balance between exploratory and exploitative individuals. A set of too exploratory profiles seems detrimental, as does a set of too exploitative profiles. As shown in Table 3.2, this holds for all indicators on references, except for Wang et al. [2017], for which the individual effect is negative, one explanation can be the fact that Wang et al. [2017] control for future reutilization of the novel combination. Indeed, this gives a 'scientific impact' dimension to the metrics. The presence of more specialized individuals may impact the relevance of the combination for the community, making it more likely to be reused.

On MeSH terms, as visible in Table 3.11 in the Appendix, individual exploratory aspects appear to have a direct negative impact. Indexers assign the MeSH terms and may be subject to bias or misinterpretation. In contrast, the references directly relate to the researchers' choices and reflect their interests and preferences. There are two possibilities, indexers may be unable to capture all the nuances and subtleties of research conducted by individuals with high-average exploratory profiles. Alternatively, the novelty of references could be induced by an author bias in citing previous works irrelevant to the contribution. Researchers' past publications do not directly impact indexers, so she might not need to qualify the article with distant MeSH terms because the novelty is not sufficiently explicit. This suggests that MeSH terms do not reflect the diversity of knowledge and ideas present in individual past work but rather the diversity of competencies between team members.

These relations remain consistent when regressions are not performed using percentage rank information, and indicator behavior with MeSH terms and references seems to be much more corroborated, as visible in Table 3.14 and 3.15 in the Appendix. The fact that the effect is nearly the same on most of the indicators of novelty demonstrates the robustness of this analysis - our measure captures something similar regardless of the construction of the novelty indicator and the information used.

The potential for novelty seems more apparent when looking at the exact composition in terms of exploratory profiles, i.e., the share of explorative individuals and the share of highly exploitative individuals. In Table 3.3, we replace the average exploratory profile variables with the exploitative and exploratory individual shares and the interaction of these two variables.

While cognitive diversity appears to be always beneficial to combine new knowledge, the presence of too many explorative individuals is harmful. Indeed, its presence only becomes beneficial when counterbalanced by a higher share of exploitative

Table 3.3: Combinatorial Novelty: Cognitive diversity, highly exploratory and exploitative profile (Field-Weighted/ References)

	Dependent variable:				
	Uzzi	Lee	Foster	Wang	Shibayama
	(1)	(2)	(3)	(4)	(5)
Author inter $_{abs}$ (FW)	0.168***	0.163***	0.107***	0.066***	0.400***
	(0.014)	(0.012)	(0.020)	(0.008)	(0.012)
Author inter $^{2}_{abs}$ (FW)	-0.007	-0.006	0.028	0.001	-0.160***
	(0.012)	(0.011)	(0.018)	(0.007)	(0.012)
Share exploratory	-0.166***	-0.173***	-0.214***	-0.084***	-0.022***
	(0.007)	(0.007)	(0.010)	(0.004)	(0.006)
Share exploitative	0.027***	0.053***	0.057***	0.002	-0.092***
	(0.003)	(0.003)	(0.005)	(0.002)	(0.004)
Share exploratory * Share exploitative	0.298***	0.273***	0.390***	0.080***	-0.112***
	(0.016)	(0.016)	(0.020)	(0.011)	(0.018)
# References	0.002***	0.002***	0.001***	0.005***	0.002***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
# Meshterms	0.004***	0.006***	0.005***	-0.001***	0.004***
	(0.0004)	(0.0003)	(0.0003)	(0.0002)	(0.0004)
# Authors	0.008***	0.007***	0.007***	0.001***	0.006***
	(0.0004)	(0.0004)	(0.0005)	(0.0002)	(0.0003)
SJR	-0.012***	-0.011***	-0.014***	-0.008***	-0.011***
	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)
Year	Yes	Yes	Yes	Yes	Yes
Journal Cat.	Yes	Yes	Yes	Yes	Yes
Observations	1,647,430	1,815,603	1,815,603	1,815,603	1,809,155
R^2	0.059	0.068	0.046	0.122	0.129
Adjusted R^2	0.059	0.068	0.046	0.122	0.129
Residual Std. Error	0.280	0.277	0.308	0.345	0.267
F Statistic	436.681***	556.617***	372.829***	1,065.763***	1,132.467***

Notes: This table reports coefficients of the effect of cognitive diversity and highly exploratory and exploitative profiles on combinatorial novelty using PKG. Standard errors are cluster robust at the journal level: ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively. The effects are estimated with an OLS. Variables are field-weighted and constant term, scientific field (Scimago Journal Category), and time-fixed effects are incorporated in all model specifications.

individuals. We can clearly see how this trade-off is necessary to create novelty in the regressions. In the same way as before, the coefficients can be interpreted directly, a percentage point increase in the share of highly explorative individuals increases by β percentage point in the ranking of the most novel articles in the field and year.

Exploratory individuals will develop new perspectives that specialized individuals will capitalize on to make them succeed. A larger share of specialized individuals facilitates communication among members if they are in the same field; otherwise, scientists with diverse backgrounds appear to facilitate communication among team

members who are cognitively distant [pag, 2007]. This mirrors the "Renaissance" individual of [Jones, 2009] or the "transformational leader" of Shin and Zhou [2007] who can ease connections between distant members and foster the team's creativity.

Too many such individuals would make the exploration less efficient, and the emerging ideas would potentially not be successfully implemented because the embedding of the conducted research in a scientific paradigm would not be sufficient. The results are similar across novelty indicators, except for Shibayama et al. [2021], in which the best team composition is made from non-exploitative, non-highly exploratory researchers. Table 3.12 in the Appendix shows that the results also hold for indicators based on MeSH terms.

The two sets of results on the impact of cognitive distance and researcher profile show that combining specialized and exploratory profiles is a good proxy for potential novelty as it enhances the realized novelty in the team.[7] While Uzzi et al. [2013] show that this trade-off between conventional and atypical combinations of knowledge is the most impactful, we demonstrate that this idea holds at the team level as well and that these configurations are most likely to achieve atypical combinations.

3.4.1.2 Perceived novelty

In this subsection, we examine the relationship between the cognitive dimension and novelty as assessed by experts. Specifically, we employ a Logit model to identify the impact of authors' intra-diversity and inter-diversity on the likelihood of being classified in at least one novel category. The results of these models are presented in Table 3.4. The effect of team cognitive diversity plays a positive role in *perceived* novelty, as seen in the first and second specifications. This effect is less clear when considering individual characteristics. The average exploratory profile has a negative impact. In model 3, we can see that our previous results on *realized* novelty (Table 3.2) only holds for the cognitive distance between individuals when tested on *perceived* novelty. In contrast, when examining the specifications with the share of highly exploratory and exploitative individuals, the results corroborate the regressions performed on *realized* novelty. The proportion of highly exploratory individuals has a negative effect. Instead, typical individuals play a positive role, and the intersection of both types of researchers is indeed positive for predicting novelty. Note that in this specification, cognitive diversity between members is no longer significant.

[7] Table 3.17 provided in the Appendix shows that the results are similar when considering unnormalized indicators

Table 3.4: Faculty Opinions: cognitive diversity and average exploratory profile, highly exploratory and exploitative profile (Field-Weighted)

	Dependent variable:				
	Novelty Perceived				
	(1)	(2)	(3)	(4)	(5)
Author inter $_{abs}$ (FW)	0.306**	0.715*			0.330
	(0.126)	(0.388)			(0.302)
Author intra $_{abs}$ (FW)	-0.532***	-0.196			
	(0.155)	(0.419)			
Author inter $_{abs}$^2 (FW)		-0.438			-0.270
		(0.376)			(0.325)
Author intra $_{abs}$^2 (FW)		-0.364			
		(0.379)			
Share exploratory			-0.675**	-1.233***	-1.238***
			(0.275)	(0.371)	(0.384)
Share exploitative			0.339***	0.317***	0.337***
			(0.117)	(0.118)	(0.115)
Share exploratory * Share exploitative				2.360**	2.289**
				(1.052)	(1.062)
Control variables	YES	YES	YES	YES	YES
Observations	12,555	12,555	12,555	12,555	12,555
Log Likelihood	-7,076.944	-7,073.965	-7,072.608	-7,070.408	-7,069.551
AIC	14,423.890	14,421.930	14,415.220	14,412.820	14,415.100

Notes: This table reports coefficients of the effect of cognitive diversity, average exploratory profile, highly exploratory and exploitative profiles on perceived novelty from Faculty Opinions. Standard errors are cluster robust at the journal level in parentheses: ***, ** and * indicate significance at the 1%, 5% and 10% level, respectively. The effects is estimated using a Logit model. Variables are field-weighted and constant term, scientific field (Scimago Journal Category) and time fixed effects are incorporated in all model specifications.

However, when examining Table 3.7 in the Appendix, we can see that the effects are heterogeneous across labels. We chose the four labels for which more than 1,000 papers had been classified to perform the regressions. The effect of cognitive distance between team members is visible in the "Technical Advance" category but not significant for the remaining labels. Conversely, in Table 3.9, we can see that the results in terms of exploratory profiles are mainly driven by the 'Interesting Hypothesis' label. Results here are a bit different since we observe a U shape, meaning that highly specialized or highly diverse teams often publish articles labeled as "Interesting hypotheses". Results are quite similar when using Poisson regression and modeling the number of times a paper is labeled in a given category, as visible in

Table 3.8 and Table 3.10 in the Appendix.

3.4.2 Cognitive dimension and impact

This subsection examines the relationship between the team's cognitive dimension and impact measures. To this end, we report the results of an OLS to identify the joint impact of authors' intra-diversity and inter-diversity on the indicators. The outcomes of these models are presented in 3.5 and 3.6.

Our analysis emphasizes the need to differentiate the forms of impact to better understand how the cognitive aspect influences scientific recognition. Indeed, we use the traditional indicator of the number of citations and indicators of disruption and consolidation. The composition of the teams has a significant influence on the type of impact of the studies conducted.

The Table 3.5 regression tables indicate a double inverse U-shaped relationship between the cognitive dimension and the number of citations. Table 3.13 shows that both turning points are around 45%. Following Uzzi, a too-conventional work might not be as impactful as the contribution is more marginal. Conversely, peers may not sufficiently consider a too-novel study. This phenomenon is reflected in the composition of the teams as we can see in the differences between consolidation and disruption indicators. Indeed, to consolidate, it is necessary to have a team with a low average exploratory profile and low average cognitive distance between members. The relationship is negative for consolidation indicators (DeIn and Depth) for both intra and inter-individual levels; the effect is sometimes captured via quadratic terms. This means that cognitive diversity is negatively related to the fact that papers citing the focal paper also cite each other or cite many of the references from the focal article. Specialized teams are the ones who consolidate the science.

For disruptive indicators, the picture is somewhat different (DI1, DI5, DI1nok, and Breadth). Cognitive distance still seems to be globally favourable for disruption. Then, the Breadth disruption indicator, which examines how often articles citing the focal paper also cite each other, seems to indicate a U-shaped relationship with a turning point at 0.33, i.e., if the individuals are very distant or if they are very close, this produces the most disruptive articles in the sense that the citations will be concentrated towards the focal paper.

Although not always significant, the intra-individual effect is more mixed; teams with higher average explorative profiles globally appear to have a higher disruption potential, but this does not hold for DI1. The DI1nok index follows the same pattern

Table 3.5: Scientific recognition: cognitive diversity and average exploratory profile (Field-Weighted)

	Dependent variable:						
	# cit.	DI1	DI5	DI1nok	DeIn	Breadth	Depth
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Author inter $_{abs}$ (FW)	0.031***	0.021***	0.034***	0.047***	-0.067***	-0.010*	0.002
	(0.007)	(0.006)	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)
Author inter $^{2}_{abs}$ (FW)	-0.036***	0.012**	0.005	0.002	0.008	0.015***	-0.012**
	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.006)
Author intra $_{abs}$ (FW)	0.070***	-0.057***	0.026***	-0.008	0.009	0.014**	-0.004
	(0.008)	(0.007)	(0.008)	(0.008)	(0.009)	(0.006)	(0.008)
Author intra $^{2}_{abs}$ (FW)	-0.072***	0.038***	0.009	0.024***	-0.030***	0.021***	-0.038***
	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)
# References	0.003***	-0.001***	-0.003***	-0.002***	0.004***	-0.0001*	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.00005)	(0.0001)
# Meshterms	0.008***	-0.002***	-0.003***	-0.003***	0.005***	-0.003***	0.006***
	(0.0004)	(0.0002)	(0.0003)	(0.0003)	(0.0003)	(0.0002)	(0.0004)
# Authors	0.012***	-0.006***	-0.002***	-0.005***	0.006***	-0.009***	0.012***
	(0.0004)	(0.0003)	(0.0003)	(0.0003)	(0.0004)	(0.0003)	(0.0004)
SJR	0.039***	-0.019***	0.002	-0.006***	0.008***	-0.026***	0.030***
	(0.005)	(0.002)	(0.002)	(0.001)	(0.002)	(0.003)	(0.004)
Year	Yes	Yes	Yes	Yes	Yes	Yes	
Journal Cat.	Yes	Yes	Yes	Yes	Yes	Yes	
Observations	1,826,207	1,826,207	1,826,207	1,826,207	1,826,207	1,826,207	1,826,207
R^2	0.173	0.029	0.069	0.034	0.137	0.051	0.075
Adjusted R^2	0.173	0.029	0.069	0.034	0.137	0.051	0.075
Residual Std. Error	0.266	0.281	0.281	0.280	0.270	0.270	0.291
F Statistic	1,621.946***	233.770***	575.115***	269.625***	1,227.699***	413.321***	629.396***

Notes: This table reports coefficients of the effect of cognitive diversity and average exploratory profile on scientific recognition using PKG. Standard errors are cluster robust at the journal level: ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively. The effects are estimated with an OLS. Variables are field-weighted and constant term, scientific field (Scimago Journal Category), and time-fixed effects are incorporated in all model specifications.

as DI5, with the exception that it is the quadratic term that takes over.

The articles that are consolidating science are articles with low team diversity and low average exploratory profiles. Here we can observe the notion of highly specialized individuals who conduct more confirmatory and, therefore, consolidating research. The opposite is true for disruption. The teams' diversity always seems beneficial for proposing disruptive ideas. Articles receiving the most citations are again a matter of a trade-off between a cognitively not-too-distant team and a somewhat reasonable average level of exploration.[8]

[8]For regressions without field-year normalization as presented in Table 3.16, the results are more mixed and less clear. The cognitive aspect seems to follow a U-shaped pattern, with teams that are very close or distant being the most disruptive. The results are more robust for breadth,

In Table 3.6, we specify the team's composition in terms of exploratory/exploitative profile and found that the relationship of the cognitive distance with the impact measures remains almost similar. For consolidation metrics and citation counts, the share of exploitative individuals is clearly beneficial. The exploitative profile reduces the risk of failure as researchers learn from experience and combinations that have failed [Vincenti, 1990]. Whereas highly exploratory profiles seem to affect the expected number of citations negatively, the effect appears mixed for consolidation since it is positive DeIn and insignificant for Depth. In both cases, combining the two types of profiles is harmful. At the same time, the share of exploitative individuals is positive, suggesting that combining these two types of profiles is not optimal for consolidating research. To achieve disruption, it is better to minimize the number of individuals who are too exploratory or too specialized, but combining both types of profiles seems once again essential. We can see how the impact of highly explorative profiles is always negative, and the impact of exploitative profiles is also negative. Still, the interaction between the two is always positive for all disruption measures.

In conclusion, the analysis shows how teams with a high share of specialized individuals or low average exploratory profiles are teams that consolidate science. In contrast, teams that get the most recognition in terms of disruption combine highly exploitative and highly exploratory individuals and have cognitively more distant members.[9]

3.5 Conclusion

This paper examines the effect of exploratory scholars and, in a broader way, team composition on creativity. Our findings suggest that the cognitive dimension plays a crucial role in the creative process, and significantly influences the two pillars of creativity: originality and success. We first show that the team's cognitive diversity strongly influences novelty (*realized* and *perceived*) of the research conducted. We also show that a double-inversed U-shaped relationship exists between cognitive dimensions (intra and inter) and the impact in terms of citations. Our study also highlights the strong connection between the cognitive dimension and the nature of these citations. Teams with more exploitative profiles tend to consolidate science,

with diversity consistently appearing to be beneficial.

[9]For regressions without field-year normalization (see Table 3.19), the results are less homogeneous for the cognitive distance aspect, but the combination of explorative and exploitative is robust. The interaction of the two consistently leads to disruption.

Table 3.6: Scientific recognition: cognitive diversity, highly exploratory and exploitative profile (Field-Weighted)

	Dependent variable:						
	# cit.	DI1	DI5	DI1nok	DeIn	Breadth	Depth
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Author inter $_{abs}$ (FW)	0.088***	-0.058***	0.020*	0.004	-0.019	-0.004	0.003
	(0.010)	(0.009)	(0.011)	(0.010)	(0.012)	(0.007)	(0.009)
Author inter $^{2}_{abs}$ (FW)	-0.073***	0.067***	0.026***	0.042***	-0.037***	0.025***	-0.028***
	(0.010)	(0.009)	(0.009)	(0.008)	(0.009)	(0.007)	(0.008)
Share exploratory	-0.023***	-0.055***	-0.041***	-0.056***	0.058***	-0.006	-0.003
	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)	(0.004)	(0.005)
Share exploitative	0.029***	-0.033***	-0.056***	-0.049***	0.058***	-0.024***	0.032***
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)
Share exploratory * Share exploitative	-0.023**	0.132***	0.047***	0.096***	-0.087***	0.059***	-0.034***
	(0.012)	(0.010)	(0.011)	(0.010)	(0.011)	(0.011)	(0.012)
# References	0.003***	-0.001***	-0.003***	-0.002***	0.004***	-0.0001*	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.00005)	(0.0001)
# Meshterms	0.008***	-0.002***	-0.003***	-0.003***	0.005***	-0.003***	0.006***
	(0.0004)	(0.0002)	(0.0003)	(0.0003)	(0.0003)	(0.0002)	(0.0004)
# Authors	0.012***	-0.007***	-0.003***	-0.005***	0.006***	-0.010***	0.013***
	(0.0004)	(0.0003)	(0.0003)	(0.0003)	(0.0004)	(0.0003)	(0.0004)
SJR	0.038***	-0.018***	0.003	-0.006***	0.007***	-0.026***	0.030***
	(0.005)	(0.002)	(0.002)	(0.001)	(0.002)	(0.003)	(0.004)
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Journal Cat.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,826,207	1,826,207	1,826,207	1,826,207	1,826,207	1,826,207	1,826,207
R^2	0.174	0.030	0.071	0.035	0.139	0.051	0.075
Adjusted R^2	0.174	0.030	0.071	0.035	0.139	0.050	0.075
Residual Std. Error	0.266	0.281	0.280	0.280	0.270	0.270	0.291
F Statistic	1,619.636***	239.244***	586.510***	281.922***	1,243.711***	410.608***	626.296***

Notes: This table reports coefficients of the effect of cognitive diversity and highly exploratory and exploitative profiles on scientific recognition using PKG. Standard errors are cluster robust at the journal level: ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively. The effects are estimated with an OLS. Variables are field-weighted and constant term, scientific field (Scimago Journal Category), and time-fixed effects are incorporated in all model specifications.

while those with more exploratory individuals disrupt it and propose more distant knowledge combinations only when associated with exploitative ones. Our research underscores how team composition in terms of profiles lies at the heart of scientific creativity.

Multiple limitations arise in our study. First, concerning data used, PKG is based on advanced heuristics and algorithms to disambiguate authors using affiliation and additional metadata Xu et al. [2020]. While there is a considerable amount of research on addressing noise in Knowledge Graphs [Fasoulis et al., 2020] and improvements in these methods may increase their reliability in the future, we cannot guarantee that errors or inconsistencies will not occur when dealing with author-level information in PKG.

Other shortcomings are directly related to the creation of our indicator. First,

many methods and hyper-parameters were chosen for the simplicity of computation. The embedding is a pre-trained model from SpaCy and is not state-of-the-art. One should compare the behavior of different embedding techniques but also on what kind of text they are applied and the distance measure used. We suspect that the two papers might be close given a specific embedding and distance measure but highly distant given other parameters. In addition, the distance between the two papers would vary depending on whether the distance metric is applied to the paper's title, abstract, or full text. The semantic distances between researchers can be influenced by biases inherent in the fields and journal practices. For example, if researchers publish in different journals, the structure and format of their abstracts may be affected even if their research topic or area of expertise remains unchanged. Another hyper-parameter we used is the time window for an author's past publication. We considered a time window of 5 years. This suggests that any paper published by the author before this point would not be captured. One could argue that past behavior influences current behavior, and a highly diverse background can be proxied by recent publications. Yet, no evidence supports this hypothesis. Another issue is how we define authors' cognitive aspect by considering only past publications. Although we do not try to approximate the skills of a researcher but only their disposition to do diverse research, we are not sure how working on a topic is enough to understand and manage this new knowledge. This raises the question of the exact competencies of a transformational leader and if the past paper is sufficient to proxy it. Also, a specialized author could have previously worked on distant papers but only on his topic/methodology. Our measure defines it as diverse, yet is it true? Although solo publications can be used to construct an author's profile, the increasing significance of teamwork in scientific research makes it uncertain whether a complete and precise profile can be established solely on this basis. Another option could be to incorporate external information, such as educational background, and assign greater weight to papers that align with the author's education. However, obtaining this information can be challenging as it often requires web scraping, which is not easily scalable. The last issue in our mind about using past publications is ghost and honorary authorship, as it is common that some authors contributed very little to the production of the article [Sugimoto and Larivière, 2018, Pruschak and Hopp, 2022]. Both are problems to consider while defining a coauthored paper as part of your knowledge space.

In our analysis, we solely focused on the cognitive diversity of researchers, but diversity encompasses various aspects as highlighted by prior research studies [Medin

and Lee, 2012, Hofstra et al., 2020]. According to Koopmann et al. [2021], there are four proximity dimensions among researchers, namely cognitive, institutional, social, and geographical. Relying solely on PKG to approximate all of these dimensions could be challenging. Still, alternative sources such as OpenAlex could provide more comprehensive information on a researcher's institutions, past institutions, and authors' characteristics. For instance, relying on PKG to construct a researcher's seniority could be biased because of the restriction on health sciences papers. Exploring these additional channels could lead to developing supplementary measures that complement cognitive diversity.

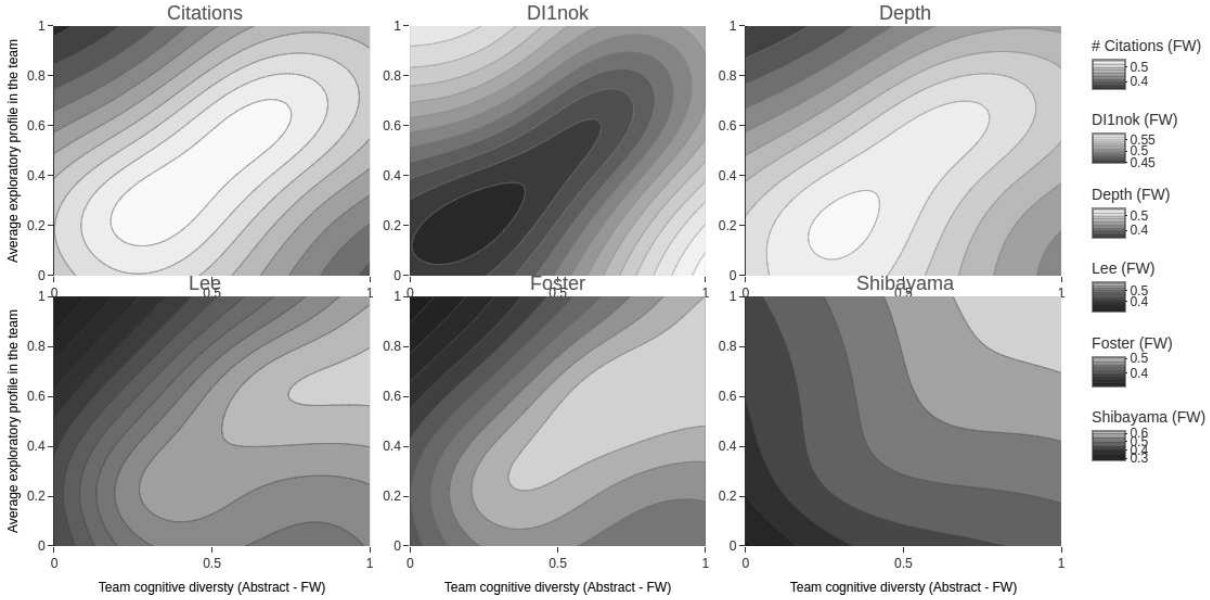
Another area worth exploring is the temporal dynamic between exploring new ideas and exploiting existing ones. As discussed earlier, discovering new concepts is essential for addressing significant challenges. However, there is often a pattern of moving through cycles of exploration and exploitation within a particular field. Similarly, authors may initially focus on a specific subject and then switch to a different area to gain a fresh perspective on the first one once they have developed sufficient expertise.

To increase the efficiency of the scientific system, it is necessary to conduct further research on the composition of research teams and their impact on creativity. Our preliminary results indicate that policymakers and grant evaluators should consider both individual and team-level characteristics and not only citations when making decisions about research funding and support. We have explored some research avenues to deepen our understanding of this phenomenon, and we encourage other researchers to build upon our work in this area. By continuing to investigate these factors, we can develop more effective strategies for supporting and fostering creativity within research teams, ultimately leading to more impactful and innovative scientific outcomes.

3.6 Appendix

Figures

Figure 3.6: Relation between cognitive diversity, average exploratory profile and Novelty/ Scientific Impact



Regressions

Novelty indicators and Faculty Opinion

Table 3.7: Faculty Opinions: Cognitive diversity and average exploratory profile (Field-Weighted)

	Dependent variable:			
	Logit Model			
	Interesting Hyp.	Technical Adv.	Confirmation	Controversial
Author inter $_{abs}$ (FW)	-0.625	1.485***	-0.427	-0.757
	(0.387)	(0.338)	(0.386)	(0.491)
Author inter $^{2}_{abs}$ (FW)	0.414	-1.101***	0.310	0.543
	(0.382)	(0.328)	(0.381)	(0.516)
Author intra $_{abs}$ (FW)	-0.191	0.209	0.001	0.278
	(0.388)	(0.336)	(0.384)	(0.580)
Author intra $^{2}_{abs}$ (FW)	-0.016	-0.465	0.199	0.016
	(0.383)	(0.324)	(0.365)	(0.602)
# References	0.006***	-0.012***	-0.0002	-0.001
	(0.001)	(0.002)	(0.001)	(0.002)
# Meshterms	0.019***	-0.040***	0.011***	0.004
	(0.004)	(0.006)	(0.004)	(0.005)
# Authors	-0.026***	0.018***	0.005	-0.017*
	(0.006)	(0.005)	(0.004)	(0.009)
SJR	0.065***	-0.019**	-0.008	0.006
	(0.011)	(0.010)	(0.005)	(0.008)
Year	Yes	Yes	Yes	Yes
Journal Cat.	Yes	Yes	Yes	Yes
Observations	12,555	12,555	12,555	12,555
Log Likelihood	-7,919.383	-7,202.326	-7,657.496	-3,866.333
Akaike Inf. Crit.	16,112.770	14,678.650	15,588.990	8,006.667

Notes: This table reports coefficients of the effect of cognitive diversity and average exploratory profile on perceived novelty from Faculty Opinions. Standard errors are cluster robust at the journal-level: ***, ** and * indicate significance at the 1%, 5% and 10% level, respectively. The effects are estimated with a Logit model. Variables are field-weighted and constant term, scientific field (Scimago Journal Category) and time fixed effects are incorporated in all model specifications.

Table 3.8: Faculty Opinions: Cognitive diversity and average exploratory profile (Field-Weighted)

	Dependent variable:			
	Poisson Model			
	Interesting Hyp.	Technical Adv.	Confirmation	Controversial
$Authorinter_{abs}$ (FW)	-0.410^{*}	1.225^{***}	-0.330	-0.828^{**}
	(0.209)	(0.204)	(0.287)	(0.413)
Author inter $^{2}_{abs}$ (FW)	0.291	-0.918^{***}	0.259	0.689
	(0.210)	(0.192)	(0.285)	(0.440)
Author intra $_{abs}$ (FW)	-0.009	0.320	0.129	0.271
	(0.184)	(0.217)	(0.248)	(0.495)
Author intra $^{2}_{abs}$ (FW)	-0.119	-0.492^{**}	-0.041	-0.108
	(0.181)	(0.219)	(0.223)	(0.521)
# References	0.003^{***}	-0.006^{***}	-0.0001	0.0005
	(0.001)	(0.002)	(0.001)	(0.001)
# Meshterms	0.013^{***}	-0.024^{***}	0.007^{*}	0.002
	(0.002)	(0.005)	(0.003)	(0.005)
# Authors	-0.017^{***}	0.009^{***}	0.003	-0.012
	(0.004)	(0.003)	(0.003)	(0.009)
SJR	0.039^{***}	-0.0004	0.002	0.014^{*}
	(0.007)	(0.007)	(0.004)	(0.007)
Year	Yes	Yes	Yes	Yes
Journal Cat.	Yes	Yes	Yes	Yes
Observations	12,555	12,555	12,555	12,555
Log Likelihood	-10,420.250	-9,880.221	-8,978.963	-4,358.803
Akaike Inf. Crit.	21,114.510	20,034.440	18,231.920	8,991.606

Notes: This table reports coefficients of the effect of cognitive diversity and average exploratory profile on perceived novelty from Faculty Opinions. Standard errors are cluster robust at the journal level: ***, ** and * indicate significance at the 1%, 5% and 10% level, respectively. The effects are estimated with a Poisson model. Variables are field-weighted and constant term, scientific field (Scimago Journal Category) and time fixed effects are incorporated in all model specifications.

Table 3.9: Faculty Opinions: Cognitive diversity, highly exploratory and exploitative profile (Field-Weighted)

	Dependent variable:			
	Logit Model			
	Interesting Hyp.	Technical Adv.	Confirmation	Controversial
Author inter $_{abs}$ (FW)	-0.859***	1.500***	-0.289	-0.476
	(0.277)	(0.317)	(0.309)	(0.377)
Author inter $^{2}_{abs}$ (FW)	0.590*	-1.244***	0.228	0.421
	(0.308)	(0.338)	(0.333)	(0.371)
Share exploratory	-0.754*	-0.644	0.868**	-0.193
	(0.450)	(0.443)	(0.441)	(0.607)
Share exploitative	0.304***	-0.014	-0.070	-0.097
	(0.083)	(0.118)	(0.097)	(0.160)
Share exploratory * Share exploitative	2.911***	0.015	-1.069	1.822
	(1.073)	(1.132)	(1.048)	(1.650)
# References	0.006***	-0.012***	-0.0001	-0.001
	(0.001)	(0.002)	(0.001)	(0.002)
# Meshterms	0.019***	-0.041***	0.012***	0.004
	(0.004)	(0.006)	(0.004)	(0.005)
# Authors	-0.024***	0.018***	0.005	-0.018*
	(0.006)	(0.005)	(0.004)	(0.009)
SJR	0.065***	-0.019*	-0.008	0.005
	(0.010)	(0.010)	(0.005)	(0.008)
Year	Yes	Yes	Yes	Yes
Journal Cat.	Yes	Yes	Yes	Yes
Observations	12,555	12,555	12,555	12,555
Log Likelihood	-7,910.400	-7,202.819	-7,655.698	-3,866.431
Akaike Inf. Crit.	16,096.800	14,681.640	15,587.400	8,008.863

Notes: This table reports coefficients of the effect of cognitive diversity and highly exploratory and exploitative profiles on perceived novelty from Faculty Opinions. Standard errors are cluster robust at the journal-level: ***, ** and * indicate significance at the 1%, 5% and 10% level, respectively. The effects are estimated with a Logit model. Variables are field-weighted and constant term, scientific field (Scimago Journal Category) and time fixed effects are incorporated in all model specifications.

Table 3.10: Faculty Opinions: Cognitive diversity, highly exploratory and exploitative profile (Field-Weighted)

	Dependent variable:			
	Poisson Model			
	Interesting Hyp.	Technical Adv.	Confirmation	Controversial
Author inter $_{abs}$ (FW)	-0.485***	1.336***	-0.171	-0.560*
	(0.157)	(0.224)	(0.248)	(0.322)
Author inter $^{2}_{abs}$ (FW)	0.318*	-1.106***	0.121	0.472
	(0.163)	(0.224)	(0.261)	(0.332)
Share exploratory	-0.718**	-0.550*	0.478*	-0.083
	(0.334)	(0.312)	(0.246)	(0.513)
Share exploitative	0.135***	-0.026	-0.020	-0.084
	(0.049)	(0.077)	(0.066)	(0.171)
Share exploratory * Share exploitative	2.112***	-0.106	-0.564	1.674
	(0.662)	(0.762)	(0.640)	(1.504)
# References	0.003***	-0.006***	-0.0001	0.0005
	(0.001)	(0.002)	(0.001)	(0.001)
# Meshterms	0.013***	-0.024***	0.007**	0.002
	(0.002)	(0.005)	(0.003)	(0.005)
# Authors	-0.016***	0.009***	0.003	-0.013
	(0.004)	(0.003)	(0.003)	(0.009)
SJR	0.039***	0.00003	0.002	0.014*
	(0.007)	(0.007)	(0.004)	(0.007)
Year	Yes	Yes	Yes	Yes
Journal Cat.	Yes	Yes	Yes	Yes
Observations	12,555	12,555	12,555	12,555
Log Likelihood	-10,415.010	-9,880.661	-8,977.912	-4,358.090
Akaike Inf. Crit.	21,106.010	20,037.320	18,231.830	8,992.180

Notes: This table reports coefficients of the effect of cognitive diversity and highly exploratory and exploitative profiles on perceived novelty from Faculty Opinions. Standard errors are cluster robust at the journal level: ***, ** and * indicate significance at the 1%, 5% and 10% level, respectively. The effects are estimated with a Poisson model. Variables are field-weighted and constant term, scientific field (Scimago Journal Category) and time fixed effects are incorporated in all model specifications.

Novelty indicators with Mesh Terms

Cognitive diversity and average exploratory profile effect on Novelty

Table 3.11: Combinatorial Novelty: cognitive diversity and average exploratory profile (Field-Weighted/ Meshterms)

	Dependent variable:			
	Uzzi	Lee	Foster	Wang
	(1)	(2)	(3)	(4)
Author inter $_{abs}$ (FW)	0.067***	0.114***	0.056***	0.062***
	(0.009)	(0.007)	(0.008)	(0.006)
Author inter $^{2}_{abs}$ (FW)	-0.016**	-0.050***	-0.025***	-0.008
	(0.008)	(0.006)	(0.008)	(0.006)
Author intra $_{abs}$ (FW)	-0.020*	0.025***	-0.029**	-0.055***
	(0.012)	(0.009)	(0.013)	(0.009)
Author intra $^{2}_{abs}$ (FW)	-0.047***	-0.062***	-0.042***	-0.010
	(0.010)	(0.008)	(0.011)	(0.007)
# References	0.001***	0.001***	0.001***	0.0002***
	(0.0001)	(0.00005)	(0.0001)	(0.00003)
# Meshterms	0.007***	0.014***	0.0004	0.029***
	(0.001)	(0.001)	(0.0004)	(0.0004)
# Authors	0.002***	0.008***	0.005***	0.001***
	(0.0004)	(0.0004)	(0.0004)	(0.0003)
SJR	-0.004***	-0.006***	-0.005***	0.003***
	(0.001)	(0.001)	(0.001)	(0.001)
Year	Yes	Yes	Yes	Yes
Journal Cat.	Yes	Yes	Yes	Yes
Observations	661,821	1,823,859	1,823,859	1,823,859
R^2	0.029	0.083	0.015	0.153
Adjusted R^2	0.029	0.083	0.015	0.152
Residual Std. Error	0.285	0.276	0.300	0.360
F Statistic	86.982***	699.050***	121.183***	1,390.452***

Notes: This table reports coefficients of the effect of cognitive diversity and average exploratory profile on combinatorial novelty using PKG. Standard errors are cluster robust at the journal level: ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively. The effects are estimated with an OLS. Variables are field-weighted and constant term, scientific field (Scimago Journal Category), and time-fixed effects are incorporated in all model specifications.

Share of Highly Exploratory Profile

Table 3.12: Combinatorial Novelty: Cognitive diversity, highly exploratory and exploitative profile (Field-Weighted/ Meshterms)

	Dependent variable:			
	Uzzi	Lee	Foster	Wang
	(1)	(2)	(3)	(4)
Author inter $_{abs}$ (FW)	0.045***	0.115***	0.007	0.002
	(0.011)	(0.010)	(0.013)	(0.008)
Author inter $^{2}_{abs}$ (FW)	0.013	-0.032***	0.030**	0.027***
	(0.010)	(0.010)	(0.012)	(0.007)
Share exploratory	-0.107***	-0.113***	-0.150***	-0.063***
	(0.006)	(0.006)	(0.007)	(0.005)
Share exploitative	0.068***	0.045***	0.056***	0.026***
	(0.003)	(0.003)	(0.004)	(0.002)
Share exploratory * Share exploitative	0.185***	0.189***	0.254***	0.106***
	(0.016)	(0.013)	(0.016)	(0.012)
# References	0.001***	0.001***	0.001***	0.0002***
	(0.00005)	(0.00005)	(0.0001)	(0.00003)
# Meshterms	0.007***	0.014***	0.0004	0.029***
	(0.001)	(0.001)	(0.0004)	(0.0004)
# Authors	0.003***	0.008***	0.005***	0.001***
	(0.0004)	(0.0004)	(0.0004)	(0.0003)
SJR	-0.004***	-0.006***	-0.004***	0.003***
	(0.001)	(0.001)	(0.001)	(0.001)
Year	Yes	Yes	Yes	Yes
Journal Cat.	Yes	Yes	Yes	Yes
Observations	661,821	1,823,859	1,823,859	1,823,859
R^2	0.033	0.086	0.019	0.152
Adjusted R^2	0.032	0.086	0.019	0.152
Residual Std. Error	0.284	0.275	0.299	0.360
F Statistic	97.014***	721.442***	149.772***	1,383.049***

Notes: This table reports coefficients of the effect of cognitive diversity and highly exploratory and exploitative profiles on combinatorial novelty using PKG. Standard errors are cluster robust at the journal level: ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively. The effects are estimated with an OLS. Variables are field-weighted and constant term, scientific field (Scimago Journal Category), and time-fixed effects are incorporated in all model specifications.

Turning points

Table 3.13: Turning Points for Combinatorial Novelty and Scientific Impact

Regression	Author intra *abs* (FW)	Author inter *abs* (FW)
Uzzi	0.318	2.725
Lee	0.229	2.441
Foster	0.244	2.521
Wang	0.038	1.75
Shibayama	2	1.203
# Cit.	0.486	0.43
DI1	0.75	0.875
DI5	-1.44	-3.4
DI1nok	0.166	-11.75
DeIn	0.15	-4.187
Breadth	-0.33	0.33
Depth	-0.052	0.083

Notes: This table reports the turning points of the effect of cognitive diversity and average exploratory profiles on combinatorial novelty and scientific recognition in Table 3.2 and 3.5.

Regression without field-year weighting

Table 3.14: Combinatorial Novelty: cognitive diversity and average exploratory profile (References)

	Dependent variable:				
	Uzzi	Lee	Foster	Wang	Shibayama
	(1)	(2)	(3)	(4)	(5)
Author inter $_{abs}$	183.520***	4.377***	0.940***	3.061***	0.268***
	(23.173)	(0.204)	(0.033)	(0.252)	(0.008)
Author inter $^{2}_{abs}$	-176.966***	-3.915***	-1.005***	-1.936***	-0.195***
	(31.732)	(0.270)	(0.043)	(0.335)	(0.012)
Author intra $_{abs}$	198.281***	3.825***	1.052***	0.095	0.226***
	(22.235)	(0.222)	(0.074)	(0.365)	(0.011)
Author intra $^{2}_{abs}$	-403.151***	-8.090***	-2.057***	0.130	-0.153***
	(38.759)	(0.381)	(0.107)	(0.619)	(0.018)
# References	0.518***	0.009***	0.001***	0.076***	0.0004***
	(0.072)	(0.0004)	(0.00004)	(0.007)	(0.00002)
# Meshterms	1.287***	0.025***	0.003***	-0.043***	0.001***
	(0.119)	(0.002)	(0.0003)	(0.003)	(0.0001)
# Authors	1.371***	0.025***	0.004***	0.005	0.002***
	(0.113)	(0.001)	(0.0004)	(0.004)	(0.0001)
SJR	-1.151***	-0.020***	-0.011***	-0.093***	-0.002***
	(0.264)	(0.004)	(0.002)	(0.021)	(0.0003)
Year	Yes	Yes	Yes	Yes	Yes
Journal Cat.	Yes	Yes	Yes	Yes	Yes
Observations	1,647,446	1,815,631	1,815,631	1,815,631	1,809,185
R^2	0.020	0.168	0.151	0.158	0.253
Adjusted R^2	0.020	0.168	0.151	0.158	0.253
Residual Std. Error	192.756	1.258	0.235	4.341	0.066
F Statistic	139.319***	1,504.472***	1,328.955***	1,399.846***	2,523.818***

Notes: This table reports coefficients of the effect of cognitive diversity and average exploratory profile on combinatorial novelty using PKG. Standard errors are cluster robust at the journal level: ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively. The effects are estimated with an OLS. The constant term, scientific field (Scimago Journal Category), and time-fixed effects are incorporated in all model specifications.

Table 3.15: Combinatorial Novelty: cognitive diversity and average exploratory profile (Meshterms)

	Dependent variable:			
	Uzzi	Lee	Foster	Wang
	(1)	(2)	(3)	(4)
Author inter $_{abs}$	14.010***	1.829***	0.399***	0.951***
	(1.109)	(0.067)	(0.023)	(0.070)
Author inter $^{2}_{abs}$	-16.049***	-2.002***	-0.495***	-0.915***
	(1.477)	(0.087)	(0.034)	(0.086)
Author intra $_{abs}$	11.644***	1.408***	0.405***	-0.177*
	(1.403)	(0.095)	(0.041)	(0.105)
Author intra $^{2}_{abs}$	-28.595***	-2.603***	-1.066***	-0.578***
	(2.138)	(0.140)	(0.063)	(0.154)
# References	0.038***	0.002***	0.001***	0.001***
	(0.002)	(0.0001)	(0.00004)	(0.0001)
# Meshterms	-0.022*	0.028***	0.001***	0.058***
	(0.013)	(0.001)	(0.0003)	(0.001)
# Authors	0.012	0.011***	0.002***	-0.0001
	(0.008)	(0.001)	(0.0003)	(0.001)
SJR	-0.052	-0.011***	-0.002***	0.015***
	(0.032)	(0.002)	(0.001)	(0.003)
Year	Yes	Yes	Yes	Yes
Journal Cat.	Yes	Yes	Yes	Yes
Observations	661,832	1,823,889	1,823,889	1,823,889
R^2	0.064	0.179	0.120	0.174
Adjusted R^2	0.063	0.179	0.120	0.174
Residual Std. Error	7.929	0.536	0.206	0.716
F Statistic	193.801***	1,638.907***	1,020.027***	1,586.294***

Notes: This table reports coefficients of the effect of cognitive diversity and average exploratory profile on combinatorial novelty using PKG. Standard errors are cluster robust at the journal level: ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively. The effects are estimated with an OLS. The constant term, scientific field (Scimago Journal Category), and time-fixed effects are incorporated in all model specifications.

Table 3.16: Scientific recognition: cognitive diversity and average exploratory profile

	Dependent variable:						
	# cit.	DI1	DI5	DI1nok	DeIn	Breadth	Depth
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Author inter $_{abs}$	39.622***	-0.019***	-0.043***	0.640***	-3.961***	0.044*	-0.080***
	(12.175)	(0.006)	(0.008)	(0.050)	(0.306)	(0.023)	(0.024)
Author inter $^2_{abs}$	-55.784***	0.034***	0.068***	-0.485***	3.848***	-0.023	0.057*
	(15.267)	(0.008)	(0.010)	(0.066)	(0.379)	(0.032)	(0.033)
Author intra $_{abs}$	99.833***	-0.064***	-0.067***	-0.090	-2.059***	0.111***	-0.033
	(12.635)	(0.007)	(0.008)	(0.073)	(0.385)	(0.032)	(0.033)
Author intra $^2_{abs}$	-130.168***	0.069***	0.094***	0.138	2.499***	-0.021	-0.141***
	(16.543)	(0.010)	(0.012)	(0.105)	(0.541)	(0.047)	(0.049)
# References	0.681***	-0.0002***	-0.0004***	-0.003***	0.023***	-0.0002***	0.001***
	(0.027)	(0.00001)	(0.00002)	(0.0001)	(0.001)	(0.00004)	(0.0001)
# Meshterms	0.338***	-0.0004***	-0.001***	-0.006***	0.019***	-0.003***	0.005***
	(0.080)	(0.00004)	(0.0001)	(0.0005)	(0.002)	(0.0002)	(0.0003)
# Authors	3.405***	-0.0004***	-0.0002***	-0.009***	0.023***	-0.008***	0.010***
	(0.365)	(0.00003)	(0.00005)	(0.0005)	(0.002)	(0.0003)	(0.0004)
SJR	16.482***	-0.001***	0.001***	-0.013***	0.025***	-0.023***	0.025***
	(1.269)	(0.0001)	(0.0002)	(0.002)	(0.009)	(0.003)	(0.003)
Year	Yes	Yes	Yes	Yes	Yes	Yes	
Journal Cat.	Yes	Yes	Yes	Yes	Yes	Yes	
Observations	1,826,237	1,826,237	1,826,237	1,826,237	1,826,237	1,826,237	1,826,237
R^2	0.116	0.042	0.077	0.133	0.238	0.107	0.159
Adjusted R^2	0.116	0.042	0.077	0.133	0.238	0.107	0.158
Residual Std. Error	126.203	0.056	0.061	0.467	1.591	0.250	0.241
F Statistic	984.300***	328.932***	626.917***	1,151.082***	2,343.227***	904.045***	1,416.198***

Notes: This table reports coefficients of the effect of cognitive diversity and average exploratory profile on scientific recognition using PKG. Standard errors are cluster robust at the journal level: ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively. The effects are estimated with an OLS. The constant term, scientific field (Scimago Journal Category), and time-fixed effects are incorporated in all model specifications.

Table 3.17: Combinatorial Novelty: Cognitive diversity, highly exploratory and exploitative profile (References)

	Dependent variable:				
	Uzzi	Lee	Foster	Wang	Shibayama
	(1)	(2)	(3)	(4)	(5)
Author inter $_{abs}$	280.445***	6.383***	1.533***	1.696***	0.430***
	(30.366)	(0.225)	(0.066)	(0.347)	(0.011)
Author inter $^{2}_{abs}$	-311.743***	-6.771***	-1.834***	-0.650	-0.350***
	(39.387)	(0.298)	(0.082)	(0.493)	(0.015)
Share exploratory	-22.978***	-0.416***	-0.087***	-0.328***	0.009***
	(3.028)	(0.029)	(0.005)	(0.034)	(0.001)
Share exploitative	8.714***	0.234***	0.048***	-0.468***	-0.021***
	(1.809)	(0.015)	(0.004)	(0.082)	(0.001)
Share exploratory * Share exploitative	29.023***	0.541***	0.186***	0.208	-0.052***
	(8.047)	(0.084)	(0.013)	(0.129)	(0.004)
# References	0.514***	0.009***	0.001***	0.076***	0.0004***
	(0.073)	(0.0004)	(0.00004)	(0.007)	(0.00002)
# Meshterms	1.292***	0.025***	0.004***	-0.042***	0.001***
	(0.119)	(0.002)	(0.0003)	(0.003)	(0.0001)
# Authors	1.472***	0.027***	0.005***	0.002	0.001***
	(0.118)	(0.001)	(0.0004)	(0.003)	(0.0001)
SJR	-1.206***	-0.022***	-0.011***	-0.090***	-0.002***
	(0.260)	(0.004)	(0.002)	(0.020)	(0.0003)
Year	Yes	Yes	Yes	Yes	Yes
Journal Cat.	Yes	Yes	Yes	Yes	Yes
Observations	1,647,446	1,815,631	1,815,631	1,815,631	1,809,185
R^2	0.020	0.167	0.150	0.158	0.252
Adjusted R^2	0.020	0.167	0.150	0.158	0.252
Residual Std. Error	192.763	1.258	0.235	4.340	0.067
F Statistic	138.223***	1,493.115***	1,310.908***	1,399.600***	2,503.790***

Notes: This table reports coefficients of the effect of cognitive diversity and highly exploratory and exploitative profiles on combinatorial novelty using PKG. Standard errors are cluster robust at the journal level: ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively. The effects are estimated with an OLS. The constant term, scientific field (Scimago Journal Category), and time-fixed effects are incorporated in all model specifications.

Table 3.18: Combinatorial Novelty: Cognitive diversity, highly exploratory and exploitative profile (Meshterms)

	Dependent variable:			
	Uzzi	Lee	Foster	Wang
	(1)	(2)	(3)	(4)
Author inter $_{abs}$	23.097***	2.721***	0.573***	0.687***
	(1.222)	(0.095)	(0.036)	(0.103)
Author inter $^{2}_{abs}$	-28.625***	-3.128***	-0.798***	-0.852***
	(1.611)	(0.118)	(0.047)	(0.125)
Share exploratory	-0.852***	-0.091***	-0.068***	-0.073***
	(0.100)	(0.007)	(0.004)	(0.007)
Share exploitative	1.977***	0.078***	0.045***	0.057***
	(0.092)	(0.007)	(0.003)	(0.006)
Share exploratory * Share exploitative	1.251***	0.083***	0.130***	0.135***
	(0.393)	(0.023)	(0.010)	(0.025)
# References	0.038***	0.002***	0.001***	0.001***
	(0.001)	(0.0001)	(0.00004)	(0.0001)
# Meshterms	-0.022*	0.028***	0.001***	0.058***
	(0.013)	(0.001)	(0.0003)	(0.001)
# Authors	0.028***	0.012***	0.003***	0.001
	(0.008)	(0.001)	(0.0003)	(0.001)
SJR	-0.064**	-0.012***	-0.003***	0.015***
	(0.031)	(0.002)	(0.001)	(0.003)
Year	Yes	Yes	Yes	Yes
Journal Cat.	Yes	Yes	Yes	Yes
Observations	661,832	1,823,889	1,823,889	1,823,889
R^2	0.065	0.179	0.119	0.174
Adjusted R^2	0.065	0.179	0.119	0.174
Residual Std. Error	7.923	0.537	0.206	0.716
F Statistic	197.576***	1,631.871***	1,014.063***	1,574.819***

Notes: This table reports coefficients of the effect of cognitive diversity and highly exploratory and exploitative profiles on combinatorial novelty using PKG. Standard errors are cluster robust at the journal level: ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively. The effects are estimated with an OLS. The constant term, scientific field (Scimago Journal Category), and time-fixed effects are incorporated in all model specifications.

Table 3.19: Scientific recognition: cognitive diversity, highly exploratory and exploitative profile

	Dependent variable:						
	# cit.	DI1	DI5	DI1nok	DeIn	Breadth	Depth
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Author inter $_{abs}$	114.316***	-0.088***	-0.115***	0.313***	-4.628***	0.114***	-0.095***
	(13.961)	(0.009)	(0.011)	(0.084)	(0.556)	(0.030)	(0.033)
Author inter $^{2}_{abs}2$	-122.966***	0.101***	0.143***	-0.193*	4.793***	-0.093**	0.041
	(16.903)	(0.011)	(0.013)	(0.103)	(0.669)	(0.039)	(0.043)
Share exploratory	-2.597**	-0.006***	-0.005***	-0.057***	0.176***	0.007**	-0.009***
	(1.100)	(0.001)	(0.001)	(0.005)	(0.018)	(0.003)	(0.003)
Share exploitative	3.759***	-0.005***	-0.008***	-0.085***	0.284***	-0.021***	0.027***
	(1.085)	(0.0005)	(0.001)	(0.004)	(0.015)	(0.002)	(0.002)
Share exploratory * Share exploitative	-23.770***	0.019***	0.014***	0.115***	-0.181***	0.031***	-0.017*
	(3.529)	(0.002)	(0.002)	(0.017)	(0.064)	(0.010)	(0.010)
# References	0.679***	-0.0002***	-0.0004***	-0.003***	0.023***	-0.0002***	0.001***
	(0.027)	(0.00001)	(0.00002)	(0.0001)	(0.001)	(0.00004)	(0.0001)
# Meshterms	0.337***	-0.0004***	-0.001***	-0.006***	0.019***	-0.003***	0.005***
	(0.080)	(0.00004)	(0.0001)	(0.0005)	(0.002)	(0.0002)	(0.0003)
# Authors	3.430***	-0.0004***	-0.0003***	-0.010***	0.025***	-0.009***	0.011***
	(0.364)	(0.00003)	(0.00005)	(0.0005)	(0.002)	(0.0003)	(0.0004)
SJR	16.427***	-0.001***	0.001***	-0.012***	0.024**	-0.023***	0.025***
	(1.268)	(0.0001)	(0.0002)	(0.002)	(0.009)	(0.003)	(0.003)
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Journal Cat.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,826,237	1,826,237	1,826,237	1,826,237	1,826,237	1,826,237	1,826,237
R^2	0.116	0.042	0.078	0.134	0.239	0.107	0.159
Adjusted R^2	0.116	0.042	0.078	0.134	0.239	0.107	0.159
Residual Std. Error	126.204	0.056	0.061	0.467	1.590	0.250	0.241
F Statistic	980.132***	328.824***	630.089***	1,160.693***	2,346.784***	901.029***	1,411.683***

Notes: This table reports coefficients of the effect of cognitive diversity and highly exploratory and exploitative profiles on scientific recognition using PKG. Standard errors are cluster robust at the journal level: ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively. The effects are estimated with an OLS. The constant term, scientific field (Scimago Journal Category), and time-fixed effects are incorporated in all model specifications.

Chapter 4

The Private Sector Is Hoarding AI Researchers

This chapter was co-authored with

Roman Jurowestki, Daniel S. Hain and Stefano Bianchini

Summary of the chapter

This study examines the migration of AI researchers from academia to industry and the potential impact on the advancement of AI technology and its diffusion into society. Using bibliometric data, we analyze the transition patterns between academia and industry and investigate the drivers and implications of this migration. Our results indicate a growing net flow of researchers from academia to industry, with those specializing in deep learning and high-impact research being more likely to transition while those engaged in highly novel research are less inclined to make the transition. Our analysis also reveals a decline in the novelty of researchers' work after transitioning to industry. These findings emphasize the importance of bolstering explorative public AI research to prevent the future of this powerful technology from being dominated solely by private interests and their focus on a potentially sub-optimal paradigm which is Deep Learning.

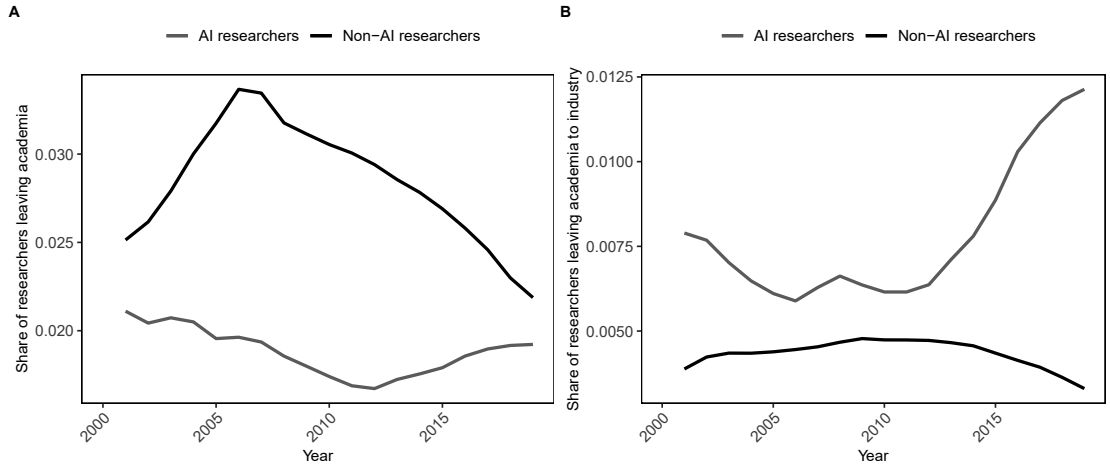
The case of Timnit Gebru in 2020, a renowned AI researcher and co-lead of the Ethical AI team at Google, ignited extensive discussions across academia, industry, and civil society [Allyn, 2020, Hao, 2020]. In her paper, Gebru and her coauthors raised concerns about the inherent biases and resource consumption of large language models [Bender et al., 2021], which subsequently led to a dispute with management and eventually resulted in her termination.

Fast forward to 2023 and it becomes apparent that AI development has accelerated with private companies leaping even more ahead of the public sector. Large language models have grown in size and complexity, culminating in the launch of OpenAI's GPT-3 and its derivative chatGPT. Throughout 2021 and 2022, we have seen organizations such as HuggingFace orchestrating large-scale research projects with broad academic involvement to develop advanced language models based on data foundations that consider and address bias and transparency issues [Scao et al., 2022]. Also in 2022, OpenAI has pushed AI techniques into mainstream discussions with the introduction of DALL-E 2 [Ramesh et al., 2022]. The open-source variant was shortly after developed by Stability AI [Rombach et al., 2021]. The private AI lab included open-source initiatives (LAION, EleutherAI) and academic collaborators (CompVis group at LMU Munich) in the creation. The trend persists: an increasing amount of AI techniques that define standards and directly impact society are designed and owned by private organizations.

Two key concerns drive the motivation behind this research: Firstly, as star scientists transition into private AI labs, there is a potential for constructive critique raised by community leaders to be silenced. This can have consequences such as a lack of discussions around potential issues surrounding the technology and a lack of transparency in its development. As seen with the release of chatGPT, these technologies are often developed behind closed doors and released to the public without adequate consideration of their potential impacts. This dynamic makes it increasingly challenging for regulators to develop suitable frameworks for AI. This can be seen in the ongoing challenges faced by the EU in its efforts to establish an AI Directive [Veale and Zuiderveen Borgesius, 2021]. Secondly, with more AI researchers transitioning from universities – which traditionally value freedom of thought, expression, and research exploration as core pillars – to industrial labs (Figure 4.1), questions arise about the potential risks of moving towards what some economists refer to as the "wrong kind of AI" [Sweeney, 2013, Hajian et al., 2016, Zou and Schiebinger, 2018, Acemoglu and Restrepo, 2019, Clark and Hadfield, 2019].

Although the ethical consequences are hard to measure, the risks associated with transition and the impact on research can be explored. The public sector might be losing its competency in the development of AI which could cause a technological lock-in. This privatization of AI knowledge and the dominance of private interests in its development is a concern that must be addressed to ensure that the societal benefits of AI are not constrained. Understanding the scope and magnitude of researcher migration from academia to industry as well as its impact on the type of research they do is crucial to better support public AI research. One risk is the potential decline in scientific Creativity[1] as researchers may respond to a different system of incentives and priorities in the new environment.

Figure 4.1: Transition of researchers



Notes: We calculate the proportion of researchers departing from academia to any destination (Panel A) and specifically to the industry sector (Panel B) using the complete OpenAlex database. Researchers are grouped into two categories: AI researchers (blue curve) and non-AI researchers (black curve). A higher share of non-AI researchers tends to leave academia for other destinations, particularly Hospitals and Public Research Organizations (PROs). This trend reverses drastically when we focus on the flow of researchers from academia to industry.

While some studies have raised alarms about the growing influence of industry in AI research [Ahmed et al., 2023], others have just begun to explore the possible implications of a brain drain of AI professors from universities [Gofman and Jin, 2022]. Here, we employ OpenAlex to estimate the number of AI Researchers transitioning from academia to industry and identify observable features that indicate such transitions. We specifically focus on explicitly observable transitions, reconstructing

[1] *Originality* and *impact* as decomposed by Runco and Jaeger [2012]

career trajectories from bibliographic data. We also investigate the consequences of industry transitions using a difference-in-differences setup. In this approach, we investigate impact and originality indicators[2] for publications before and after the transition to industry, in order to understand the impact of these transitions on the researcher's work and influence. This research provides a quantitative understanding of the phenomenon of AI researcher migration from academia to industry and its potential implications.

Our investigation reveals a growing pattern of AI researchers moving from academia to industry. Notably, those with significant impact and affiliated to top institutions are more inclined to make this transition. On the other hand, researchers who explore novel subjects and avoid deep learning tend to remain in academia. The private sector's attraction to high-impact researchers implies a selective recruitment process, which could potentially result in the decline of the talent pool available for public AI Research. Furthermore, the results from the difference-in-difference analysis indicate a decrease in academic creativity post-transition, with a short-term burst in impact immediately after the transition, which is followed by a decline in subsequent years. Additionally, there is an overall reduction in novelty observed over time.

The structure of this paper is organized as follows: section 4.1 provides an introduction to the phenomenon of AI researcher migration from academia to industry, along with relevant theoretical considerations. section 4.2 presents the data sources and methods used for the analysis, which are further explored in section 4.3 for the exploratory analysis and section 4.4 for the econometric analysis. Finally, section 4.5 presents a discussion of the findings and offers concluding remarks.

4.1 Background

4.1.1 The Rise of Private Sector Participation in AI Research

AI research has long coexisted in academia and industry. Yet, in recent years, the influence of industry is growing rapidly. The industry is becoming more influential in academic publications, cutting-edge models, and key benchmarks [Frank et al., 2019, Ahmed et al., 2023]. By way of example, until 2014, most of the significant Machine Learning (ML) models were released by academia. Since then, gradually

[2]Citation count but also Novelty and Disruptiveness indicators

the industry has taken over [Maslej et al., 2023].

Deep Learning (DL) algorithms have been particularly instrumental in fueling the growth of AI, as they have demonstrated remarkable performance in various applications, such as image recognition, natural language processing, and speech recognition [LeCun et al., 2015]. The increasing availability of large public and private datasets has enabled the development of these algorithms, with deep learning and more recently transformer models achieving state-of-the-art performance in a variety of domains.

The modern R&D trajectory of AI has seen private companies native to the digital economy, such as Google and Facebook, play an increasingly important role in basic research activities that used to be the domain of academia. In fact, the industry is now the dominant player in AI research, with significant contributions to research output, research tools, platforms, and frameworks [Ahmed et al., 2023]. For example, at the 2021 Neural Information Processing Systems (NeurIPS) conference, the largest annual AI and DL conference, Google and Facebook accounted for 11% (265) of all accepted papers and more than twice the number of accepted full papers compared to the second most represented institution, Stanford University.[3]

In addition to the growing amount of research output, many leading researchers in the field of deep learning have transitioned to full – or part – time positions in the tech industry. Notable examples include Geoffrey Hinton at Google, Yann LeCun at Facebook/Meta, Ian Goodfellow at Apple (via Google Brain), and Ruslan Salakhutdinov at Apple. Moreover, a study by Gofman and Jin [2022] found that from 2004 to 2018, Google, Microsoft, and Amazon hired 50 tenure-track AI professors from North American universities. This shift in the center of gravity of AI Research from academia to industry has raised concerns about an "AI brain drain" [Sample, 2017, Gofman and Jin, 2019].

Aside from established industry players like Google and Facebook, new companies such as Huggingface, Stability.ai, Cohere, and Anthropic are emerging in the market. These companies are focused on cutting-edge AI research, safety, and creating large language models to rival those from OpenAI and Google. These companies engage in collaborations with the public sector, particularly around topics with a societal impact. Cohere, for example, has launched a nonprofit research lab focused on fundamental AI research and contributing to the open-source community [Wiggers, 2022]. Huggingface orchestrated the "BigScience Workshop" involving over 1000 researchers

[3]Calculated from Scopus `https://www.scopus.com` records

worldwide in the generation of new large language models [Akiki et al., 2022]. This initiative led to the generation of several *side products*, the 1.6TB Composite Multilingual Dataset ROOTS [Laurençon et al., 2022] that can be used for other projects, a number of papers and research related to data ethics, governance, and engineering as well as a consortium funded by the European Commission working on multimodal (text+speech) models as well as two more extensive follow up projects. The activity of these new companies is expanding the scope of AI Research and development, further blurring the line between academia and industry.

There are several potential explanations for the increasing participation of private-sector companies in basic research activities. One reason is that modern AI methods require large datasets and computational infrastructures that are difficult to transfer to researchers in academia for technical, data protection and privacy reasons. Moreover, the development and deployment of modern AI systems poses an organizational and engineering challenge that universities or public research organizations may not be well-suited to execute, given their organizational structures and academic KPIs that primarily focus on publications. Another reason is the potential disconnect between the type of AI research undertaken in academia and the needs of the industry [Arora et al., 2020], which has led to innovation system failures [Gustafsson and Autio, 2011], leading private companies to take basic research activities "in their own hands".

Another reason for the increasing industry participation in basic research is the increasing integration of AI systems into private sector companies' cloud infrastructure, which helps them address their own needs and those of third-party customers. This integration enables companies to establish their AI systems as a *de facto* standard that increases the competitiveness of complementary platforms and cloud computing services. For instance, OpenAI's GPT and Cohere's NLP models are being increasingly integrated into various systems via their API, making it easier to deploy them correctly. Although similar open-source models are available, they require state-of-the-art MLOps (Machine Learning Operations) expertise, which can be challenging for organizations without in-house resources.

4.1.2 Challenges and Risks of AI Privatization

Private-sector opportunities may not align with social needs or take into account technology's externalities and broader socioeconomic impacts [Archibugi and Filippetti, 2018, Hain and Jurowetzki, 2017].

The increasing involvement of private-sector companies in AI research raises several concerns about the encroachment on public research agendas. Financial incentives from industry may lead academic researchers to prioritize commercialization over spillovers, leading to the homogenization of public and private research spheres [David, 2003, David and Hall, 2006]. Industry may induce researchers to prioritize less risky and exploratory (novel) research, instead favoring projects with immediate applications and commercial potential [Stephan, 2012].

Training deep neural networks requires enormous amounts of data and computing power, often exclusively available to large industry players and costly in terms of energy use and carbon emissions [Marcus, 2018, Russell, 2019, Strubell et al., 2019]. Although platforms and frameworks provided by industry, such as Tensorflow or PyTorch, decrease entry barriers and advance collective progress, the direction of search and effort along this trajectory reinforces the data and computation-hungry DL paradigm. The strong demand for data has led to the exploitation of large online corpora, which incorporate gender and racial biases that transmit into the trained models and their outputs [Paullada et al., 2020]. Studying the various modalities of release for large (mainly generative) AI models [Solaiman, 2023] shows that a growing trend since 2021 has been to keep the more powerful systems closed. These decisions not only indicate "closeness" but need to be seen in light of many challenges related to the need to understand potential impacts, develop appropriate models, and implement safety controls and ethical considerations when releasing them.

This emphasis on Deep Learning, favored by Industry, could potentially lead to a technological lock-in where other methodologies are disregarded in favor of concentrating on this particular branch of AI. This focus might result in a decline in AI research originality, as already highlighted by Klinger et al. [2020], resulting in a scenario where the Industry's influence leads to the acceptance of the potentially sub-optimal DL paradigm as the prevailing norm.[4] The widespread integration of AI technology across all sectors of society, as mentioned earlier, makes this an even more critical issue.

The goal of Academia would be to counter-balance this focus by increasing exploratory research. However, if star researchers exit the public sphere, this could reduce the innovative contributions of these researchers as they shift their focus to

[4]In a recent study conducted by Jiang et al. [2023] it was demonstrated that a straightforward k-nearest neighbors (kNN) approach utilizing a distance metric based on gzip compression outperforms BERT and other neural methods in the context of text classification. It's noteworthy that this research was conducted by scholars who are not affiliated with major tech companies.

company-based research. This could also pose a challenge for universities, making their tasks more intricate. Simultaneously, if researchers undergoing this transition gain more influence (measured by increased citations), it could lead to the propagation of the DL focus from the industry into academia. Another concern revolves around the insufficient funding for highly innovative initiatives [Ayoubi et al., 2021, OECD, 2021, Franzoni et al., 2022], which could further limit academic opportunities to break through this paradigm.

Despite the numerous concerns arising from this transition and brain drain, only a limited number of studies have examined this particular issue in AI research. Gofman and Jin [2022] conducted research on the migration of professors in AI-related fields from leading universities to the industry between 2004 and 2018. Their study revealed that this migration led to a decrease in the establishment of AI startups by students from those universities and a reduction in the amount of funding raised.

Our analysis aims to open a pathway for this literature. We focus on the attributes of scientists who have undergone this transition. We then examine the consequences of this transition on the nature of the research they undertake.

4.2 Data and Methods

4.2.1 Data

We collect data from OpenAlex `https://openalex.org`, an open and free scientific catalog built to replace Microsoft Academic Graph, that has been depreciated by the end of 2021. OpenAlex contains "metadata for 209M works (journal articles, books, etc.); 213M disambiguated authors; 124k venues (places that host works, such as journals and online repositories)[...]" [Priem et al., 2022].[5] We utilized the hierarchical structure of the concept taxonomy in OpenAlex, which organizes 65,000 unique concepts, to identify publications that have been categorized under sub-topics related to "artificial intelligence" and "machine learning", in total 402 concept-keywords.

We retrieved academic publications labeled with at least two concept-keywords of our list categorized under AI and ML research between 2000 and 2021. This collection includes metadata such as citation count, publication year and venue, title and abstract, author names, and affiliations. We obtained peer-reviewed academic

[5]Our current download, end of 2022, includes already 239.2M docs, 247.1M authors, 124k venues, 108k institutions

journal publications, conference proceedings, and preprint collections such as arXiv, which are popular media for disseminating knowledge in ML and AI research. In total, we collected 1.7M research papers. Our analysis revealed that 2.3M scholars have either used or developed AI methods in their research, which have been published in 21k journals and presented at 9,781 conferences.

To investigate the main research question of this paper, we constructed an affiliation history for all researchers who were identified as (co-)authors of the AI papers in our dataset. To ensure a complete historical overview, we extended our analysis to encompass non-AI publications of these authors, resulting in a dataset of 10.6 million focal papers. We specifically focused on researchers whose publications consisted of at least 50% (or at least 3) AI papers.[6]

We excluded scholars whose transition occurred prior to their initial AI publication and those who had no AI publications following their transition. Affiliation data was sourced from OpenAlex, offering disambiguated affiliations categorized as education, company, facility, government, healthcare, NGO, and others. To mitigate potential biases from short-term affiliations such as project-based co-affiliations, internships, visiting researcher programs, and random errors in extracting institutional information from paper metadata, we compute annual affiliations based on the institution found on most papers published by the researcher in that year. In the case of a tie, we prioritize affiliations in the order they are mentioned in the publication.

A transition is defined as the occurrence of the same affiliation for an author in years t and $t+1$, which is distinct from the affiliation in year $t-1$. For instance, if an author held the affiliation "Education" in the year 2000 and then transitioned to "Company" in 2001 and 2002, this would be classified as a transition from Academia to Industry in the year 2001.

Using this approach, we can distinguish two research-career trajectories over time: (i) those who remain affiliated solely with academia and (ii) those who experience a university-industry transition, defined as researchers who began their careers in academia but become primarily associated with the industry for at least one consecutive year. We do not differentiate between additional career paths such as "academia returnees" or "serial switchers", these are excluded in the econometric analysis.

Our analysis identified approximately 14k unique institutional affiliations, enabling us to construct an affiliation history for all authors. Of the total population

[6]As a robustness check, we performed the same analysis solely on AI papers, and the results can be found in Appendix 4.6.1.

of 192,885 researchers who met our criterias, 70,82% were affiliated only with education, while 1.52% were affiliated solely with industry. Additionally, 1.57% of the population transitioned from education to industry. To ensure the relevance of the information regarding researchers' career paths it should be noted that these numbers only pertain to researchers who were observed for at least three years, with at least 50% of the time range[7], and had seniority (first observation) after 1950 and left the sample after 2010.

Additionally, leveraging co-authorship information and citation data, we were able to compute co-authorship centrality indicators for authors and impact metrics such as received citations. Moreover, we used the Novelpy package in Python [Pelletier and Wirtz, 2022] to quantify impact and originality through metrics such as novelty and disruptiveness.

4.2.2 Analytical strategy

To explore the phenomenon of university-industry transitions in AI research, our analysis is structured into three steps. Firstly, a basic exploratory data analysis was performed to determine the magnitude, characteristics, patterns, and trends of these transitions.

Secondly, we aim to identify the drivers of university-industry transitions. We employ a survival analysis, specifically the Cox proportional hazard model, using the affiliation history of all AI researchers who remain in academia or transition to industry. The model estimates the probability of a researcher undergoing a university-industry transition in a particular year, considering the researcher's characteristics, research interactions, and pre-transition academic creativity as potential drivers for the transition.

Thirdly, we conducted a regression analysis to explore the consequences of university-industry transitions on research creativity. To address the endogenous selection of researchers who transition to industry, we employed a propensity-score matching (PSM) procedure. For each researcher who undergoes a university-industry transition, we identified the most similar counterpart among their peers who remained in academia throughout their careers.[8] As such, the PSM approach mimics a quasi-experimental setting (i.e., what would have happened to the researcher if they had

[7]For researchers active between 2000 and 2020, this translated to at least 10 years with identifiable affiliations

[8]we match these pairs based on concepts, average number of papers published, average number of co-authors, citations rank, and novelty rank

remained in academia?) and enables us to isolate the effect of the transition to industry on research output, as evaluated through received citations, novelty, and disruptiveness of the research conducted after the transition. To simulate a counterfactual scenario, we created an "artificial transition" point for each academic who remained in academia, occurring at the same point in their observed career as the actual transition of their industry-matched peer. Using this matched sample, we employed a difference-in-difference regression analysis to study the impact of transitions.

4.2.3 Variables

Dependent Variables

The **dependent variable** in the survival analysis of transition drivers is binary, taking the value of *zero* for years in which a researcher maintained their affiliation with academia in the current and previous year, and the value of *one* in the year they first transition to a corporate affiliation. OpenAlex provides the affiliation information used to measure this variable, derived from the researcher's published papers in the corresponding year.

When analyzing the effect of university-industry transitions on a researcher's career in a difference-in-difference regression, we used Nb. Citations$_{fw}$, Novelty$_{fw}$ and, Disruptiveness$_{fw}$, described below.

Independent Variables

We generate additional **independent variables** as follows:

Academic Age: We estimate the variable by the number of years since a researcher's first publication was observed in OpenAlex (i.e. This also includes papers outside of our sample).

DL Experience: We create a binary variable to indicate whether a researcher has published at least one paper in the current year with the OpenAlex concept label for "Deep Learning". This label encompasses sub-classes related to specific deep learning architectures, given the hierarchical structure of the concept taxonomy. Given that deep learning is a research field in which large data sets and computing power provide

a significant competitive advantage, we anticipate that researchers working in this area are more likely to transition from academia to industry.

Network Centrality (Academia): The degree centrality of authors in the co-publication network of papers published in the corresponding year was calculated. Edges in the network are weighted according to the number of researchers per paper, such that the edge weight assigned to a paper decreases as the number of authors on that paper increases.[9] This variable provides an estimate of the researcher's current level of embeddedness within the academic research community. Researchers who are more embedded in the community will either have more opportunities, or they will be less likely to transition because of their status in this network.

Network Centrality (Industry): This variable measures the degree centrality of the author in the co-publication network of papers published in the corresponding year, considering only the edges to co-authors who currently have industry affiliations. This variable serves as an approximation of the researcher's proximity to industry actors. We anticipate that researchers who are already actively collaborating with industry partners are more likely to transition to the industry themselves.

Top Institution : This is constructed from the Shanghai Ranking of 2010, which falls within the mid-period of our sample. Considering the number of unique institutions, we select the top 500 universities and are able to match 399 of them with researchers' affiliation profiles.[10]

typeswitcher: A binary variable indicating whether a researcher transitioned from academia to industry at some point in their observed career.

transited: This dummy variable takes a value of zero for researchers who have not undergone a university-industry transition throughout their observable career up to the corresponding year.

[9]Here we follow Newman [2004a] in assuming that a larger number of authors will lead to decreased interaction and general bonding between the authors.

[10]An alternative approach was to define the top 10% of institutions based on citations; however, this had a marginal impact on the Kaplan-Meier analysis and none on the regressions.

transited_t: This variable measures the number of years that have passed since a researcher underwent the university-industry transition and is assigned a value of zero for researchers who are still in academia.

Nb. Citations$_{fw}$: We use the mean citation count per year as the dependent variable in the difference-in-difference analysis, while also employing it as an explanatory variable in the survival analysis.

Novelty$_{fw}$: The Novelty indicator was calculated for each paper based on its associated concepts, following the methodology outlined in Lee et al. [2015]. This indicator relies on analyzing combinations of elements within a set of documents. The lower the frequency of occurrence of these combinations, the higher the Novelty$_{fw}$ score will be. For each AI researcher, we then calculate the average novelty score within a specific year for all of their papers.

Disruptiveness$_{fw}$: The Disruptiveness indicator was calculated for each paper using the methodology outlined in Bornmann et al. [2019b].[11] This indicator assesses the extent to which a paper cites a focal paper but not the focal paper's references. A greater recurrence of this pattern across different papers leads to a higher value of this indicator for the focal paper. Similar to the process for Novelty, we calculate the average disruptiveness score within a specific year for all papers authored by a researcher.

Control Variables

We additionally incorporated year and concepts as control variables in both econometric analyses.

Filter

We applied a normalization technique that involves field and year percentage ranks for variables denoted by the subscript *fw*. Percent Rank is commonly used to assess the relative position of individual values within an overall distribution, expressed as a percentage. In our context, a *zero* corresponds to researchers with the lowest citation rank, while a *one* designates those with the highest citation rank in the respective

[11] Various disruptiveness indicators provided by Novelpy were considered, and the results were robust. We selected the $DeIn$ indicator for reporting purposes.

year and field. This normalization approach is particularly suitable as it allows for the comparison of researchers making transitions with those remaining in academia. The definition of the field is established using the concept itself.

To address the sparsity present in the original data, we handled concepts as follows: For a given paper, we assigned concepts of level one, excluding those categorized as "Artificial Intelligence" or "Machine Learning". If no level one concept was identified, we considered level two concepts and traced their ancestors, i.e., Level 1 concepts related to the level 2 concept. If no concept was identified through this process, the paper was classified as "unknown". Subsequently, for each year, a researcher would have a list of concepts. We assigned to the researcher the concept that appeared most frequently in that year. In the event of a tie, the first concept to appear was selected. This methodology ensures a comprehensive representation of the researchers' topics and enables meaningful comparisons based on the defined criteria.[12]

We also calculated the moving average (with a window size of k=2) for all variables except DL, Top Institution, typeswitcher, transited, and transited$_t$. The rationale behind using the moving average is that the characteristics impacting the transition process are often accumulated over time, yet we hypothesize that only the most recent years hold a significant influence on this transition. Similarly, the effect of the transition on creativity might only become evident over multiple years.

For the survival analysis, we lagged all independent variables since the transition depends on characteristics present before the actual transition. In Table 4.1, you will find basic descriptive statistics of the variables used in the various analyses.

4.3 Exploratory Data Analysis

We first examine general patterns that can be conveyed through basic statistics and visual representations in Figure 4.2. Figure 4.2A illustrates the progression of paper output in the fields of AI and ML, which has exhibited a significant increase both in raw number and share of publications since 2015 attaining more than 150k papers and representing 1.5% of all the OpenAlex publications in 2020. In Figure 4.2B, it is evident that the proportion of papers featuring at least one author affiliated with an industry has been steadily increasing since 2015. This trend underscores

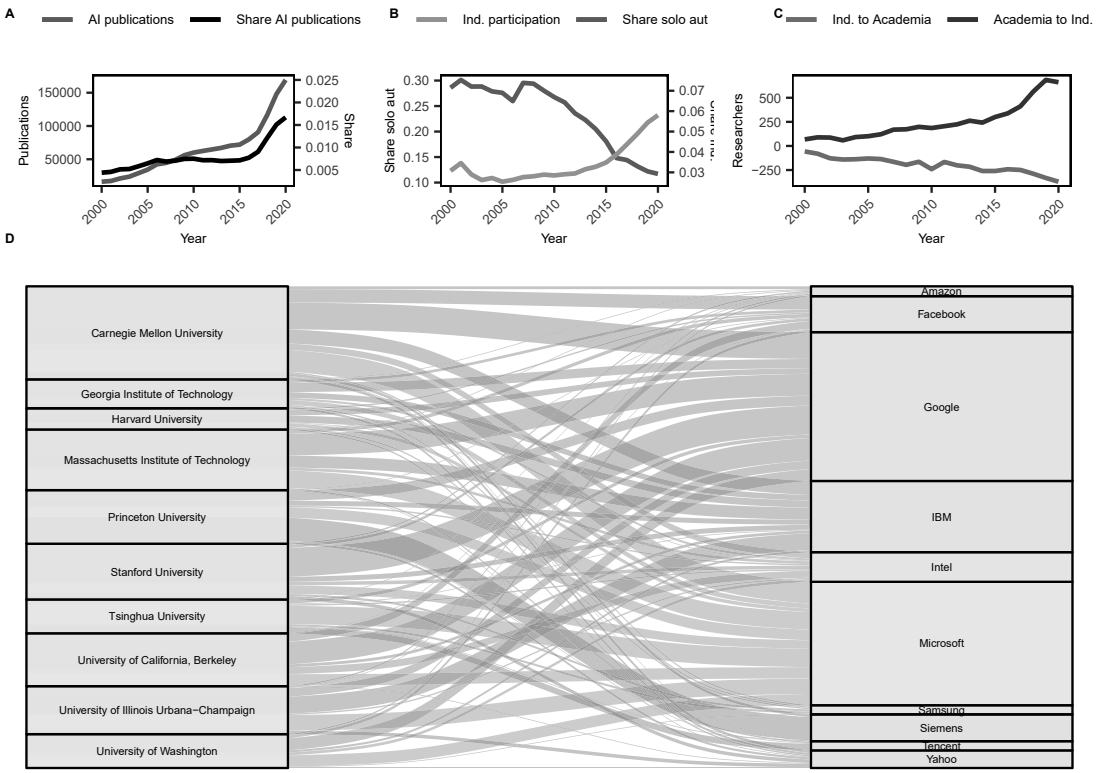
[12]We conducted experiments using three different approaches, and while the results remained consistent across all methods, we ultimately determined the proposed methodology to be the most logically coherent choice.

Table 4.1: Descriptive statistics after filtering, moving average,lag and before percent rank

	mean	sd	min	q1	median	q3	max	obs	NAs
Nb. Citations	18.880	88.913	0	2	6.500	17.158	28,782	1,703,305	202,927
Network Centrality (Academia)	2.331	3.115	0	0.517	1.367	2.948	172.282	1,703,305	230,341
Network Centrality (Industry)	0.182	1.572	0	0	0	0	158.815	1,703,305	230,341
Disruptiveness	1.638	1.965	0	0	1.167	2.500	289.824	1,703,305	0
DL Experience	0.110	0.313	0	0	0	0	1	1,703,305	0
Novelty	-0.138	0.578	-5.130	-0.474	-0.165	0.129	5.610	1,703,305	230,341
Nb. Papers	5.719	8.009	0	1.500	3.500	7	408	1,703,305	230,341
Academic Age	12.226	10.172	1	4	9	17	62	1,703,305	0
Top Institution	0.440	0.496	0	0	0	1	1	1,703,305	113,528

the growing inclination of companies to engage in research publications. A plausible explanation for this trend can be observed in the behavior represented by the blue line. As the percentage of solo-authored papers diminishes over time, the rise in the number of authors per paper implies a higher probability of authors originating from sources beyond the public sphere. Figure 4.2C starts to depict the story of the transition of AI researchers. As for the first two graphs, we see an increasing growth starting in 2015 concerning the movement of researchers departing academia and entering the private sector, with the count surpassing 600 individuals in 2020. Simultaneously, the number of individuals transitioning from industry to academia has been steadily increasing. This evolving pattern is highlighted in the alluvial plot presented in Figure 4.2D. The alluvial plot shows the top 20 institutions involved in these transitions, encompassing the foremost 10 universities alongside the leading 10 companies. Notably, the universities are predominantly renowned establishments based in the United States. Conversely, on the corporate front, it is discernible that prominent tech companies, some of which were previously mentioned in Section 4.1, are attracting these transitioning researchers. Given the growing trend and increasing importance of AI research, our study's motivations are well supported. The changing landscape further reinforces our perspective and highlights the relevance of our paper.

Figure 4.2: Exploratory Data Analysis

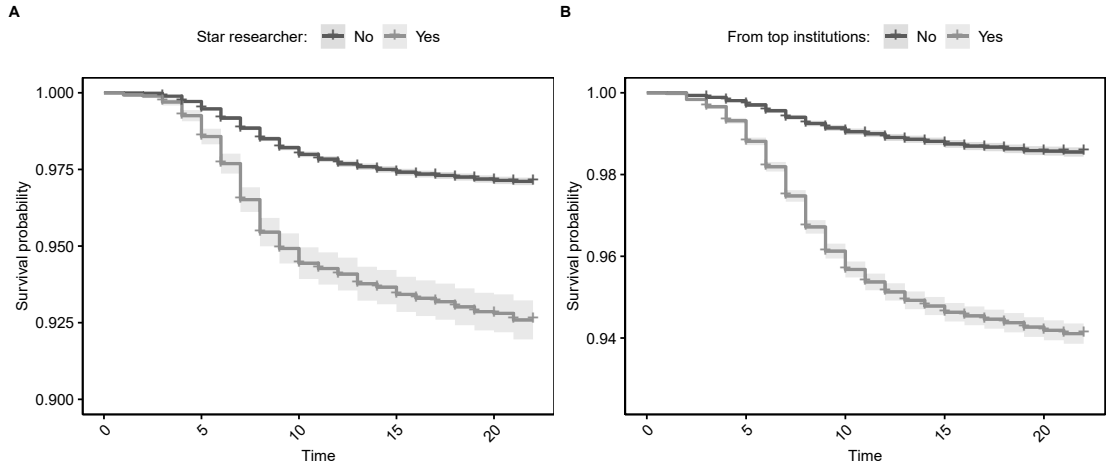


4.4 Econometric analysis

4.4.1 Drivers of switching - Survival analysis

This section examines the drivers and mechanisms of university-industry transition in AI research. We start by conducting a Kaplan-Meier analysis. This approach aims to provide a general understanding of the factors influencing transitions. Specifically, we scrutinized the "survival probability" of two distinct researcher groups: individuals classified as star researchers (representing the top 10 percent of field-weighted citations) and those affiliated with top institutions before their transition. Our analysis revealed a notable pattern. For both of these groups, the likelihood of transitioning to industry is notably higher during earlier time intervals when compared to the baseline reference point.

Figure 4.3: Kaplan-Meier estimation for star researchers and top institutions



We then use a survival analysis to estimate the likelihood of an academic AI researcher transitioning into the industry at a given time. Specifically, we employ a proportional hazard model [Cox, 1972], a multivariate regression technique that enables us to identify the joint impact of continuous and categorical variables on the probability of the transition event. The outcomes of this model are presented in Table 4.2, where Panel (1) comprises only the control variables and Academic age. Panel (2) introduces the Deep Learning dummy variable. Panel (3) includes network-related independent variables, Panel (4) consists of all previous variables and Top Institution. Finally, Panel (5) and (6) add impact and originality measures. In order to better understand the different effects, we separated the regression for the two impact measures, citations, and disruptiveness.

Model (1), which includes only the control variables, reveals a significant negative effect for *Academic*. This suggests that transitions to industry occur earlier in a researcher's career, possibly after postdoctoral projects or before achieving tenure. This effect remains significant in all subsequent panels. Model (2) incorporates the Deep Learning DL variable, which has a significant positive coefficient. This finding aligns with our prior expectations that the features of this research domain make an industry transition more appealing, as well as the earlier argumentation of strong industry interest in Deep Learning. Model (3) examines the researchers' integration into the wider AI research community, as indicated by their position in the AI co-

Table 4.2: Survival analysis

	(1)	(2)	(3)	(4)	(5)	(6)
Academic Age	-0.711^{***}	-0.711^{***}	-0.670^{***}	-0.672^{***}	-0.700^{***}	-0.681^{***}
	(0.016)	(0.016)	(0.015)	(0.016)	(0.016)	(0.016)
DL Experience		0.697^{***}		0.501^{***}	0.102^{**}	0.397^{***}
		(0.044)		(0.047)	(0.047)	(0.047)
Top Institution				1.365^{***}	0.938^{***}	1.034^{***}
				(0.044)	(0.046)	(0.045)
Network Centrality (Academia)			0.048^{***}	0.025^{***}	-0.159^{***}	-0.074^{***}
			(0.007)	(0.009)	(0.012)	(0.010)
Network Centrality (Industry)			0.089^{***}	0.082^{***}	0.108^{***}	0.107^{***}
			(0.003)	(0.004)	(0.004)	(0.004)
Novelty					-0.493^{***}	-0.583^{***}
					(0.084)	(0.085)
Disruptiveness					3.511^{***}	
					(0.085)	
Nb. Citations						2.625^{***}
						(0.077)
Control variables.	Yes	Yes	Yes	Yes	Yes	Yes
Observations	139,604	139,604	139,486	139,486	139,415	139,415
R^2	0.090	0.092	0.090	0.098	0.112	0.107

Note: *p<0.1; **p<0.05; ***p<0.01

author network, as a driver for industry transition. Our initial expectations were that researchers with more extensive connections would be more desirable for recruitment by industry and that those who collaborate with industry partners during their academic careers exhibit a particular affinity for industry and a willingness to transition. This is validated in this model, both network centrality shows a significant positive coefficient. Model (4) includes all previous variables and adds information about researchers' institutions. All the variables hold the same significance and direction of coefficients. The parameter associated with Top Institutions is positive further validating the result already seen in the Kaplan-Meier. Lastly, model (5) and (6) examines the Creativity of research performance on transitions, measuring both the past quality of research outputs and their novelty. The results indicate that although the quality of research outputs is a robust predictor of industry transitions, novelty has a significant negative effect. This may suggest that strong performance in rela-

tively established areas is more attractive to industry than entirely new approaches. This is further supported by the positive impact of DL experience and aligns with the hypothesis of technological lock-in, where the industry stands as the primary beneficiary. Interestingly after accounting for creativity measures the centrality associated with Academia exhibits a negative trend. This phenomenon can be attributed to the robust correlation between citation count and network size. We interpret this result in the following way, researchers with strong ties and co-authorship connections within Academia might experience hesitation when confronted with the decision between pursuing a path in Academia versus transitioning to Industry.

4.4.2 Consequences of switching - Difference-in-Difference analysis

We examine the impact of university-industry transition on research performance. We use a difference-in-difference approach, comparing the scientific performance development of researchers who transition to industry (treatment group) with those who remain in academia (control group). The dependent variables in this analysis are the two-year moving average of the researcher's annual citation rank (Nb. Citations$_{fw}$) and novelty rank (Novelty$_{fw}$). and disruptiveness rank (Disruptiveness$_{fw}$). The results of this set of regressions are presented in Table4.3. All models contain the dichotomous variable *typeswitcher*, indicating that the researcher experiences a transition at some point in their career. They also include the *transited* variable indicating that the transition has occurred, as well as the interaction term for those two variables.

The first two panels of the table investigate the impact of transitions on researcher output performance (Nb. Citations$_{fw}$). The interaction term Switcher/Year of transition in Panel (1) is not statistically significant, suggesting that industry transition does not affect research performance. This conclusion does not hold when we separate short-term and long-term effects. In Panel (2), a positive and significant coefficient for *Switcher/Year of transition* suggests that researchers initially experience a boost in their citation ranking after transitioning. However, there seems to be no sustained favorable impact in the long run. Instead, post-transition (*Switcher/Year since transition*), researchers experience a decline of 1.10% in their citation ranking per year, in comparison to their counterparts who remain in academia.

Panels (3) and (4) replicate the previous analysis, but now using the disruptiveness dependent variable (Disruptiveness$_{fw}$). In Panel (3), the average effect of the

transition is negative (*Switcher/Year of transition*) showing a potential loss of breakthrough papers done by migrated scholars. In Panel (4), the results follow a similar pattern as Nb. Citations$_{fw}$, although with a diminished effect after the transition and a slightly more pronounced negative impact (1.2%) over time. These differences can explain the negative significance observed in Panel (3).

Finally, Panels (5) and (6) explore the dynamics of transition and novelty. In Panel (5), the average effect of the transition is negative (*Switcher/Year of transition*), indicating a reduction in novelty. This decline, coupled with the observation that non-novel star researchers transition as revealed in the survival analysis, highlights a distinct lack of emphasis by companies on research originality. In Panel (6), the findings provide some moderation to the results. While there is a noticeable reduction in novelty immediately after the transition, there emerges a positive and statistically significant coefficient at the 5% threshold for the long-term effect. However, it's important to note that this long-term effect is modest (0.2% per year).

In summary, the outcomes of the difference-in-difference analysis indicate a positive impact of transition on impact metrics. However, this initial positive effect diminishes over time. In contrast, a negative influence on Novelty is observed. The process of absorption over time is found to be less clear for Novelty compared to impact metrics.

Table 4.3: Difference-In-Differences analysis

	Dependent variable:					
	Nb. Citations		Disruptiveness		Novelty	
	(1)	(2)	(3)	(4)	(5)	(6)
Switcher	0.021***	0.021***	0.006	0.006	−0.020***	−0.020***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)
Year of transition	0.111***	0.097***	0.063***	0.059***	−0.003	−0.001
	(0.004)	(0.005)	(0.004)	(0.005)	(0.003)	(0.004)
Years since transition		0.002***		0.0004		−0.0002
		(0.001)		(0.001)		(0.0005)
Interaction switcher/Year of transition	−0.008	0.046***	−0.023***	0.033***	−0.009**	−0.017***
	(0.005)	(0.006)	(0.005)	(0.006)	(0.004)	(0.006)
Interaction switcher/Years since transition		−0.011***		−0.012***		0.002**
		(0.001)		(0.001)		(0.001)
Observations	44,587	44,587	44,587	44,587	44,587	44,587
R^2	0.046	0.051	0.018	0.026	0.010	0.010

Note: *p<0.1; **p<0.05; ***p<0.01

4.5 Discussion and Conclusion

Our study examines the career trajectories of AI researchers, providing insights into the relationship between academic and corporate research in this field, as well as the drivers and consequences of potential brain drain from the public sector. Our goal is to contribute to science policy discussions regarding the development and

application of AI technologies and the preservation of a public research space for AI that prioritizes the creation of AI systems independently from short-term commercial interests.

Our findings suggest that the increasing involvement of the private sector in AI research has led to a growing migration of researchers from academia to industry, particularly to technology companies such as Google, Microsoft, and Facebook. The survival analysis provides evidence that researchers specializing in deep learning techniques, which have been the driving force behind recent advancements in AI systems, are more likely to transition to industry. These findings align with the growing capabilities of the private sector in state-of-the-art AI systems and raise concerns about the ability of public interest deep learning research to keep pace, particularly given the industry's tendency to recruit high-impact and influential researchers. Additionally, researchers who are more embedded in co-publication networks with companies, possibly through prior industry-funded projects, are also more likely to transition. These findings raise concerns about the potential impact of industry funding on academic research and suggest that universities should possibly be cautious when accepting funding from technology companies, as discussed by Abdalla and Abdalla [2021].

We also uncovered that novel researchers are less inclined to make a transition. This observation could be attributed to factors on either the researcher's side or the company's side. From the company's perspective, the reluctance towards novelty could be driven by their preference for remaining within the dominant Deep Learning (DL) paradigm. With ample resources at their disposal, including vast amounts of data and computational power, companies might show less interest in researchers who seek to push the boundaries of AI and ML algorithms. Their focus might lean towards impact metrics, such as h-index and citation count, which also holds true for institutions. As a result, they might selectively recruit individuals who align with these indicators. Furthermore, it's plausible to speculate that industries might have a shorter-term outlook, while novelty often implies a commitment to long-term research endeavors in the hope of achieving breakthroughs. Surprisingly, this hypothesis doesn't find support in the positive coefficient of disruptiveness. On the researcher's side, Deep Learning scientists who aspire to deepen their work in the field have more substantial incentives to leave academia than those who explore the broader knowledge landscape.

The fact that novel researchers stay in academia tells us about the kind of re-

search done in companies, but it is not detrimental in itself. Risks arise if the star researcher who leaves academia does even less novel research and if private research influences the research done in the public sphere. The second econometrical analysis gives us some preliminary answers to the first issue. In the diff-in-diff analysis, we observe indications of a decline in the academic creativity of researchers who make the transition to industry, particularly in terms of novelty. This finding is consistent with the outcomes of start-ups that are acquired and assimilated by large companies that may be more interested in exploiting existing technology rather than exploring entirely new paths. However, recent developments in natural language processing (NLP) suggest a different scenario where most breakthrough developments such as large language models originated from industrial labs. Despite this, there are arguments that such models align with the interests of large companies while less resource-intensive approaches in state-of-the-art language processing remain underexplored.

Subsequent investigations could also look at brain circulation instead of a brain drain. Examining how AI researchers who have transitioned continue to engage with the public is essential. Are they collaborating with the same community of individuals? Are they exploring new topics alongside different collaborators, yet within the same domain as their past work and within their existing network? Additionally, an investigation into serial switchers is warranted. Although a minority in our study, they constitute a group of considerable interest since they could be the pathway of Industry to influence public research.

The recent mainstreaming and hype surrounding generative AI have marked a crucial turning point in the evolution of this technology, making it applicable to a broader range of general-purpose applications. While industry investment is essential for advancing AI research and development, it is equally critical to maintain qualified academic capacity to explore alternative and potentially less-established approaches that may not be near commercialization. To this end, universities must provide an attractive research environment that caters to the needs of researchers focused on unconventional avenues. Moreover, it is crucial that public universities have researchers who can address ethical, social, and other potential societal issues related to AI. During a panel discussion held in February 2023, Aidan Gomez, the CEO of Cohere and co-author of the seminal paper that established most state-of-the-art AI technologies [Vaswani et al., 2017], expressed his disappointment that the transformer architecture proposed in the paper may not be surpassed. He stated,

"I hope that the transformer will die one day", and expressed his desire for a more elegant and effective technology or mathematical approach to further advance AI research. While this pivotal paper emerged from research conducted at Google, it is not guaranteed that industrial AI labs will continue to produce and publish groundbreaking foundational research. The recent trend of limited release of LLMs [Solaiman, 2023] may indicate a direction towards greater control and restriction in AI research.

To advance the development of efficient, ethically aligned, and robust AI architectures, policymakers and companies should collaborate to establish conditions that facilitate fundamental and accountable AI research. It is crucial to incentivize AI researchers to work in both the public and private sectors and ensure that suitable conditions are in place to support their work.

4.6 Appendix

4.6.1 Results only AI focal papers

Table 4.4: Descriptive statistics after filtering, moving average,lag and before percent rank on only AI papers

	mean	sd	min	q1	median	q3	max	obs	NAs
Nb. Citations	18.776	138.432	0	1	5	14	28,782	759,983	194,616
Network Centrality (Academia)	0.707	0.965	0	0.167	0.458	0.917	27.840	759,983	183,338
Network Centrality (Industry)	0.072	0.838	0	0	0	0	80.905	759,983	183,338
Disruptiveness	1.414	2.236	0	0	0.333	2.167	76	759,983	0
DL Experience	0.225	0.418	0	0	0	0	1	759,983	0
Novelty	-0.073	0.583	-5.388	-0.412	-0.107	0.107	4.944	759,983	183,338
Nb. Papers	1.635	2.151	0	0.500	1	2	93	759,983	183,338
Academic Age	12.211	9.699	1	5	10	17	62	759,983	0
Top Institution	0.434	0.496	0	0	0	1	1	759,983	91,187

Table 4.5: Survival analysis only AI papers

	(1)	(2)	(3)	(4)	(5)	(6)
Academic Age	-0.191^{***}	-0.191^{***}	-0.196^{***}	-0.196^{***}	-0.191^{***}	-0.192^{***}
	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.008)
DL Experience		0.843^{***}		0.765^{***}	0.329^{***}	0.655^{***}
		(0.064)		(0.071)	(0.061)	(0.061)
Top Institution				1.474^{***}	1.211^{***}	1.321^{***}
				(0.073)	(0.072)	(0.072)
Network Centrality (Academia)			0.005	-0.074^{**}	-0.319^{***}	-0.171^{***}
			(0.029)	(0.030)	(0.024)	(0.017)
Network Centrality (Industry)			0.137^{***}	0.141^{***}	0.170^{***}	0.162^{***}
			(0.013)	(0.012)	(0.008)	(0.006)
Novelty					-0.275^{**}	0.133
					(0.138)	(0.137)
Disruptiveness					3.050^{***}	
					(0.113)	
Nb. Citations						1.704^{***}
						(0.117)
Observations	95,383	95,383	95,209	95,209	95,163	95,163
R^2	0.034	0.036	0.033	0.039	0.045	0.041

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4.6: Difference-In-Differences analysis only AI papers

	Dependent variable:					
	Nb. Citations		Disruptiveness		Novelty	
	(1)	(2)	(3)	(4)	(5)	(6)
Switcher	0.046^{***}	0.046^{***}	0.008	0.008	-0.013^{**}	-0.013^{**}
	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)
Year of transition	0.054^{***}	0.080^{***}	0.017^{**}	0.034^{***}	0.008	0.013^{*}
	(0.008)	(0.009)	(0.007)	(0.008)	(0.006)	(0.007)
Years since transition		-0.006^{***}		-0.004^{***}		-0.001
		(0.001)		(0.001)		(0.001)
Interaction switcher/Year of transition	0.007	-0.011	0.011	0.001	0.011	0.010
	(0.010)	(0.012)	(0.009)	(0.011)	(0.008)	(0.010)
Interaction switcher/Years since transition		0.004^{**}		0.002		0.0002
		(0.002)		(0.002)		(0.001)
Observations	11,430	11,430	11,430	11,430	11,430	11,430
R^2	0.316	0.318	0.276	0.277	0.238	0.238

Note: *p<0.1; **p<0.05; ***p<0.01

General Conclusion

This thesis provides insights concerning the dynamics of collaboration and mobility in science and its link with creativity. The first chapter delves into collaboration within the global health science system. Chapters 2 and 3 extend this exploration, examining the health science system and collaboration through the lens of creativity. The final chapter delves into the dynamic relationship between creativity and mobility.

Contributions

Addressing urgent challenges stemming from an exogenous shock requires a country to strategically utilize its existing resources, general or related science capacity, and established partnerships. **Chapter 1** provides insights into this phenomenon. It appears that the capacity for coronavirus-related research (CRR) that countries had built up before the pandemic played a crucial role in their initial response. After the initial shock, capacities unrelated to CRR before the pandemic were the main factor of national and international CRR. This suggests that countries and institutions leveraged their existing partnerships to quickly form collaborative efforts to address the urgent challenges posed by the pandemic. Furthermore, existing grants were redirected from global research toward CRR during the pandemic. This shift in funding priorities highlights the adaptability and resource allocation strategies that countries employed to respond effectively to the crisis. Additionally, nations that relied heavily on international funding for their research experienced minimal disruptions in their CRR during the initial stages of the pandemic. Still, in the subsequent years, dependence on international funding had detrimental effects on their research output, emphasizing the importance of building domestic research capacity and resilience.

Chapter 2 introduces *Novelpy*, an open-source Python package designed to com-

pute novelty and disruption indicators for scientific documents or patents. This tool serves as a centralized resource for the scientometrics community, enabling the analysis and comparison of different measures of novelty and disruptiveness. By addressing a gap in the scientometrics field, the development of Novelpy lays the foundation for future research that aims to investigate the relationship between these indicators systematically. This package was used throughout the chapters and in future research.

Chapter 3 of the thesis focuses on exploring creativity within the Health Science System by examining the relationship between cognitive diversity in scientific teams and their ability to generate innovative ideas and gain scientific recognition. The chapter introduces an author-level metric for measuring cognitive diversity and emphasizes the concept of "potential novelty" as a measure of the likelihood of novel knowledge combinations within teams. We analyze the impact of cognitive diversity on novelty indicators and citation impact, revealing an inverse U-shaped relationship, indicating that a balanced combination of exploratory and exploitative individuals leads to the most disruptive and distant knowledge combinations. This emphasizes the critical role of team composition in scientific creativity.

Chapter 4 of the thesis delves into the relationship between intersectoral mobility among AI researchers and creativity. The study highlights concerns about the migration of AI researchers from academia to industry, particularly in fields like deep learning, driven by collaborations with tech companies like Google, Microsoft, and Facebook. The chapter reveals that this migration can potentially lead to a talent drain from academia and a decline in academic creativity, emphasizing the importance of science policy discussions to balance AI technology growth with the preservation of vibrant exploratory public research in AI.

Limits and future research

However, this work has room for further expansion, particularly in exploring the relationship between collaboration, mobility, and creativity. In this thesis, we have only scratched the surface of the complex relation between them. To conclude this thesis, we discuss potential avenues for future research that could enhance our comprehension of these phenomena and shed more light on the process of the scientific system. This, in turn, can provide policymakers with valuable insights and tools to better support research and researchers.

Firstly, we characterize the transition phenomenon as a brain drain, with re-

searchers leaving academia. We have not yet delved into the transformation of the collaboration network for those in transition. Although the number of individuals returning to academia is minimal, rather than solely perceiving it as a brain drain, we can consider it a "brain circulation," as outlined in Geuna [2015], where these transitions spark new collaborations between universities and industry. Do these transitioned researchers continue to collaborate with their previous partners? Do they remain within the same academic community? Does this community still pursue the same research topics? Is the research now primarily driven by industry or universities? Addressing these collaboration-related questions is important for gaining a more comprehensive understanding of the actual impact of this mobility.

Secondly, it's worth noting that there is a limited understanding of the dynamics of novelty. Our examination has focused on the static level of novelty. We have not explored how a researcher's profile evolves throughout their career. We also have not investigated whether a transition has consequences on researchers' exploratory profiles, and if so, whether all transitions have the same effect. The measure developed in the **third chapter** can help us better understand the different trajectories of researchers. Additionally, we haven't examined whether novelty tends to follow a cyclical pattern by nature or if disruptiveness consistently follows a peak of novelty in a particular field.

Lastly, only the **the first chapter** takes funding into account, even though it is not the chapter's main focus. I firmly believe that funding plays a pivotal role in influencing the field of science. Recently, OpenAlex has made grant metadata accessible within its database. This data could be leveraged in future research endeavors to gain deeper insights into authors' strategies and profiles. One potential avenue for investigation could involve examining the evolution of funding dynamics for a specific field and its subsequent impact on the reorientation of research given authors' profiles.

Conclusion Générale

Cette thèse donne un aperçu de la dynamique de la collaboration et de la mobilité dans le domaine scientifique et leur lien avec la créativité. Le premier chapitre explore la collaboration au sein du système de santé. Les chapitres 2 et 3 approfondissent cette exploration en examinant le système de santé et la collaboration sous l'angle de la créativité. Le dernier chapitre se penche sur la relation dynamique entre la créativité et la mobilité.

Contributions

Pour relever les défis urgents découlant d'un choc exogène, un pays doit utiliser de manière stratégique ses ressources existantes, sa capacité scientifique générale et ses partenariats établis. Le **Chapitre 1** offre des perspectives sur ce phénomène. Il apparaît que la capacité de recherche liée au coronavirus pré pandémie mondiale a joué un rôle crucial dans la réponse initiale. Après le choc initial, les ressources qui n'étaient pas liées au CRR avant la pandémie sont devenues le facteur principal de la CRR nationale et internationale. Cela suggère que les pays et les institutions ont exploité leurs partenariats existants pour former rapidement des efforts de collaboration visant à relever les défis urgents posés par la pandémie. De plus, pendant la pandémie, il y a eu une réaffectation de subventions existantes destinées à la recherche mondiale vers la recherche pour la COVID-19. Ce changement de priorités de financement met en lumière l'adaptabilité et les stratégies d'allocation des ressources que les pays ont employées pour répondre efficacement à la crise. De plus, les nations qui dépendaient fortement du financement international pour leur recherche ont connu peu de perturbations dans leur recherche lié au coronavirus au cours des premières étapes de la pandémie. Cependant, dans les années suivantes, la dépendance au financement international a eu des effets préjudiciables sur leur production de recherche, soulignant l'importance de renforcer la capacité de recherche nationale et la résilience.

Le **Chapitre 2** introduit "Novelpy", un package Python open-source conçu pour calculer les indicateurs de nouveauté et de disruption pour les documents scientifiques ou les brevets. Cet outil sert de ressource centralisée pour la communauté scientométrique, permettant l'analyse et la comparaison de différentes mesures de nouveauté et de disruption. En comblant un manque dans le domaine de la scientométrie, le développement de Novelpy crée les bases de futures recherches visant à enquêter sur la relation entre ces indicateurs. Ce package a été utilisé dans les chapitres suivants et sera utilisé dans mes futures recherches.

Le **Chapitre 3** de la thèse se concentre sur l'exploration de la créativité dans le système de santé en examinant la relation entre la diversité cognitive au sein des équipes scientifiques et leur capacité à générer des idées novatrices et à obtenir une reconnaissance scientifique. Le chapitre introduit une métrique au niveau des auteurs pour mesurer la diversité cognitive et met l'accent sur le concept de "nouveauté potentielle" comme mesure de la probabilité de combinaisons de connaissances nouvelles au sein des équipes. Nous analysons l'impact de la diversité cognitive sur les indicateurs de nouveauté et l'impact des citations, révélant une relation en forme de U inversé, indiquant qu'une combinaison équilibrée d'individus exploratoires et exploitatifs conduit aux combinaisons de connaissances les plus perturbatrices et éloignées. Cela souligne le rôle crucial de la composition des équipes dans la créativité scientifique.

Le **Chapitre 4** de la thèse explore la relation entre la mobilité intersectorielle des chercheurs en intelligence artificielle et la créativité. L'étude met en lumière les préoccupations liées à la migration des chercheurs en IA de l'académie vers l'industrie, notamment dans des domaines tels que l'apprentissage profond, sous l'impulsion de collaborations avec des entreprises technologiques telles que Google, Microsoft et Facebook. Le chapitre révèle que cette migration peut potentiellement entraîner une fuite des talents de l'académie et un déclin de la créativité académique. Cela souligne l'importance des discussions autour des politiques scientifiques pour favoriser la croissance de la technologie de l'IA tout en préservant de la recherche publique exploratoire et vibrante en IA

Limites et futures recherches

Cependant, ce travail ouvre des pistes d'expansion, en particulier pour explorer la relation entre la collaboration, la mobilité et la créativité. Dans cette thèse, nous n'avons fait qu'effleurer la surface de la relation complexe entre elles. C'est pour

cela que pour conclure cette thèse, nous engageons une discussion sur les pistes potentielles de futures recherches qui pourraient améliorer notre compréhension de ces phénomènes et éclairer davantage le fonctionnement du système scientifique. Cela pourrait fournir aux décideurs politiques des informations précieuses et des outils pour mieux soutenir la recherche et les chercheurs.

Tout d'abord, nous caractérisons le phénomène de transition comme une fuite des cerveaux, avec des chercheurs quittant l'académie. Nous n'avons pas encore exploré la transformation du réseau de collaboration pour ceux en transition. Bien que le nombre d'individus retournant dans l'académie soit minimal, plutôt que de le percevoir uniquement comme une fuite des cerveaux, nous pouvons le considérer comme une "circulation des cerveaux", comme décrit dans Geuna [2015]. Ces transitions suscitent de nouvelles collaborations entre les universités et l'industrie. Est-ce que ces chercheurs en transition continuent à collaborer avec leurs anciens partenaires? Restent-ils au sein de la même communauté académique? Cette communauté poursuit-elle toujours les mêmes sujets de recherche? La recherche est-elle désormais principalement dirigée par l'industrie ou les universités? Répondre à ces questions liées à la collaboration est important pour obtenir une compréhension plus complète de l'impact réel de cette mobilité.

Deuxièmement, il convient de noter que notre compréhension de la dynamique de la nouveauté est limitée. Notre analyse s'est concentré sur le niveau statique de la nouveauté. Nous n'avons pas exploré comment le profil d'un chercheur évolue tout au long de sa carrière. Nous n'avons pas non plus étudié si une transition a un impact sur les profils exploratoires des chercheurs, et le cas échéant, si toutes les transitions ont le même effet. La mesure développée dans le **troisième chapitre** peut nous aider à mieux comprendre les différentes trajectoires des chercheurs. De plus, nous n'avons pas examiné si la nouveauté tend à suivre un modèle cyclique par nature ou si la disruption arrive après un pic de nouveauté dans un domaine particulier.

Enfin, seul **le premier chapitre** prend en compte le financement, même s'il n'est pas le principal sujet du chapitre. Je suis convaincu que le financement joue un rôle essentiel dans la structure de la science. Récemment, OpenAlex a rendu les métadonnées des subventions accessibles dans sa base de données. Dans de futures recherches, ces données pourraient être exploitées pour obtenir des informations plus approfondies sur les stratégies et les profils des auteurs. Une piste potentielle d'investigation pourrait consister à examiner l'évolution de la dynamique du financement pour un domaine spécifique et son impact ultérieur sur la réorientation de la

recherche en fonction des profils des auteurs.

Bibliography

The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies (New Edition). Princeton University Press, 2007. ISBN 9780691138541. URL http://www.jstor.org/stable/j.ctt7sp9c.

M. Abdalla and M. Abdalla. The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 287–297, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462563. URL https://doi.org/10.1145/3461702.3462563.

W. L. Abdi, H. "jackknife". *Encyclopedia of Research Design*, 497:655–660, 2013.

G. Abramo, C. A. D'Angelo, and L. Zhang. A comparison of two approaches for measuring interdisciplinary research output: The disciplinary diversity of authors vs the disciplinary diversity of the reference list. *Journal of Informetrics*, 12(4): 1182–1193, 2018.

H. A. Abt and E. Garfield. Is the relationship between numbers of references and paper lengths the same for all sciences? *Journal of the American Society for Information Science and Technology*, 53(13):1106–1112, 2002.

D. Acemoglu and P. Restrepo. The wrong kind of ai? artificial intelligence and the future of labor demand. Technical report, National Bureau of Economic Research, 2019.

J. Adams. The rise of research networks. *Nature*, 490(7420):335–336, 2012.

J. Adams. The fourth age of research. *Nature*, 497(7451):557–560, 2013.

N. Ahmed, M. Wahed, and N. C. Thompson. The growing influence of industry in ai research. *Science*, 379(6635):884–886, 2023.

C. Akiki, G. Pistilli, M. Mieskes, M. Gallé, T. Wolf, S. Ilić, and Y. Jernite. Bigscience: A case study in the social construction of a multilingual large language model. *arXiv preprint arXiv:2212.04960*, 2022.

B. Alberts, M. W. Kirschner, S. Tilghman, and H. Varmus. Rescuing us biomedical research from its systemic flaws. *Proceedings of the National Academy of Sciences*, 111(16):5773–5777, 2014.

J. Allik, K. Lauk, and A. Realo. Factors predicting the scientific wealth of nations. *Cross-Cultural Research*, 54(4):364–397, 2020.

B. Allyn. Google employees call black scientist's ouster 'unprecedented research censorship', Dec 2020. URL `https://www.npr.org/2020/12/03/942417780/google-employees-say-scientists-ouster-was-unprecedented-research-censorship`.

M. Amendola, J.-L. Gaffard, et al. Novelty, hysteresis, and growth. Technical report, Observatoire Francais des Conjonctures Economiques (OFCE), 2014.

S. Ankrah and A.-T. Omar. Universities–industry collaboration: A systematic review. *Scandinavian Journal of Management*, 31(3):387–408, 2015.

D. Archibugi and A. Filippetti. The retreat of public research and its adverse consequences on innovation. *Technological Forecasting and Social Change*, 127:97–111, 2018.

L. Argote, S. L. Beckman, and D. Epple. The persistence and transfer of learning in industrial settings. *Management science*, 36(2):140–154, 1990.

J. C. Arnott, R. J. Neuenfeldt, and M. C. Lemos. Co-producing science for sustainability: can funding change knowledge use? *Global Environmental Change*, 60: 101979, 2020.

A. Arora, S. Belenzon, A. Patacconi, and J. Suh. The changing structure of american innovation: Some cautionary remarks for economic growth. *Innovation Policy and the Economy*, 20(1):39–93, 2020.

S. Arts, J. Hou, and J. C. Gomez. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, 50(2):104144, 2021.

S. Aviv-Reuven and A. Rosenfeld. Publication patterns' changes due to the covid-19 pandemic: a longitudinal and short-term scientometric analysis. *Scientometrics*, 126(8):6761–6784, 2021.

C. Ayoubi, M. Pezzoni, and F. Visentin. At the origins of learning: Absorbing knowledge flows from within the team. *Journal of Economic Behavior & Organization*, 134:374–387, 2017.

C. Ayoubi, M. Pezzoni, and F. Visentin. Does it pay to do novel science? the selectivity patterns in science funding. *Science and Public Policy*, 48(5):635–648, 2021.

P. Azoulay. Small research teams 'disrupt'science more radically than large ones, 2019.

D. Beaver and R. Rosen. Studies in scientific collaboration: Part i. the professional origins of scientific co-authorship. *Scientometrics*, 1(1):65–84, 1978.

D. Beaver and R. Rosen. Studies in scientific collaboration part iii. professionalization and the natural history of modern scientific co-authorship. *Scientometrics*, 1(3): 231–245, 1979.

T. R. Behrens and D. O. Gray. Unintended consequences of cooperative research: impact of industry sponsorship on climate for academic freedom and other graduate student outcome. *Research policy*, 30(2):179–199, 2001.

J. Ben-David. The scientist's role in society: A comparative study. 1971.

E. Bender, T. Gebru, A. McMillan-Major, and S. Schmitchell. On the dangers of stochastic parrots: Can language models be too big? 2021. URL `http://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf`.

S. Bianchini, M. Müller, and P. Pelletier. Artificial intelligence in science: An emerging general method of invention. *Research Policy*, 51(10):104604, 2022.

M. Bikard, F. Murray, and J. S. Gans. Exploring trade-offs in the organization of scientific work: Collaboration and scientific reward. *Management science*, 61(7): 1473–1495, 2015.

M. Bikard, K. Vakili, and F. Teodoridis. When collaboration bridges institutions: The impact of university–industry collaboration on academic productivity. *Organization Science*, 30(2):426–445, 2019.

N. Bloom, C. I. Jones, J. Van Reenen, and M. Webb. Are ideas getting harder to find? *American Economic Review*, 110(4):1104–44, 2020.

M. Boden. Creativity and knowledge. *Creativity in education*, pages 95–102, 2001.

L. Bornmann and H.-D. Daniel. What do citation counts measure? a review of studies on citing behavior. *Journal of documentation*, 2008.

L. Bornmann, S. Devarakonda, A. Tekles, and G. Chacko. Do disruption index indicators measure what they propose to measure? the comparison of several indicator variants with assessments by peers. retrieved december 6, 2019, 2019a.

L. Bornmann, A. Tekles, H. H. Zhang, and Y. Y. Fred. Do we measure novelty when we analyze unusual combinations of cited references? a validation study of bibliometric novelty indicators based on f1000prime data. *Journal of Informetrics*, 13(4):100979, 2019b.

L. Bornmann, S. Devarakonda, A. Tekles, and G. Chacko. Disruptive papers published in scientometrics: meaningful results by using an improved variant of the disruption index originally proposed by wu, wang, and evans (2019). *Scientometrics*, pages 1–7, 2020.

R. Boschma. Proximity and innovation: a critical assessment. *Regional studies*, 39 (1):61–74, 2005.

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

T. Braun and W. Glänzel. International collaboration: Will it be keeping alive east european research? *Scientometrics*, 36(2):247–254, 1996.

Y. Bu, L. Waltman, and Y. Huang. A multi-dimensional framework for characterizing the citation impact of scientific publications. *arXiv preprint arXiv:1901.09663*, 2019.

G. Buenstorf and D. P. Heinisch. When do firms get ideas from hiring phds? *Research Policy*, 49(3):103913, 2020.

R. S. Burt. Structural holes and good ideas. *American journal of sociology*, 110(2): 349–399, 2004.

X. Cai, C. V. Fry, and C. S. Wagner. International collaboration during the covid-19 crisis: autumn 2020 developments. *Scientometrics*, 126(4):3683–3692, 2021.

T. Caliari and T. Chiarini. Knowledge production and economic development: Empirical evidences. *Journal of the Knowledge Economy*, 12(2):1–22, 2021.

A. C. Cameron and P. K. Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005.

U. Cantner and B. Rake. International research networks in pharmaceuticals: Structure and dynamics. *Research Policy*, 43(2):333–348, 2014.

N. Carayol and J.-M. Dalle. Sequential problem choice and the reward system in open science. *Structural Change and Economic Dynamics*, 18(2):167–191, 2007.

N. Carayol, L. Agenor, and L. Oscar. The right job and the job right: Novelty, impact and journal stratification in science. *Impact and Journal Stratification in Science (March 5, 2019)*, 2019.

M. Chahrour, S. Assi, M. Bejjani, A. A. Nasrallah, H. Salhab, M. Fares, H. H. Khachfe, H. A. Salhab, and M. Y. Fares. A bibliometric analysis of covid-19 research activity: a call for increased output. *Cureus*, 12(3), 2020.

C. M. Christensen and C. M. Christensen. *The innovator's dilemma: The revolutionary book that will change the way you do business*. HarperBusiness Essentials New York, NY, 2003.

J. Clark and G. K. Hadfield. Regulatory markets for ai safety. *arXiv preprint arXiv:2001.00078*, 2019.

L. Compagnucci and F. Spigarelli. The third mission of the university: A systematic literature review on potentials and constraints. *Technological Forecasting and Social Change*, 161:120284, 2020.

R. Costas and T. Franssen. Reflections around 'the cautionary use'of the h-index: response to teixeira da silva and dobránszki. *Scientometrics*, 115(2):1125–1130, 2018.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

K. B. Dahlin and D. M. Behrens. When is an invention really radical?: Defining and measuring technological radicalness. *research policy*, 34(5):717–737, 2005.

P. A. David. 8 innovation and europe's academic institutions–second thoughts about embracing the bayh–dole regime. *This page intentionally left blank*, page 251, 2003.

P. A. David and B. H. Hall. Property and the pursuit of knowledge: Ipr issues affecting scientific research. *Research Policy*, 35(6), 2006.

J. Davidson Frame and M. P. Carpenter. International research collaboration. *Social studies of science*, 9(4):481–497, 1979.

J. Davidson Frame, F. Narin, and M. P. Carpenter. The distribution of world science. *Social studies of science*, 7(4):501–516, 1977.

D. Dekker, D. Krackhardt, and T. A. Snijders. Sensitivity of mrqap tests to collinearity and autocorrelation conditions. *Psychometrika*, 72:563–581, 2007.

E. Di Maria, V. De Marchi, and K. Spraul. Who benefits from university–industry collaboration for environmental sustainability? *International Journal of Sustainability in Higher Education*, 20(6):1022–1041, 2019.

E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

M. Dosso, L. Cassi, and W. Mescheba. Towards regional scientific integration in africa? evidence from co-publications. *Research Policy*, 52(1):104630, 2023.

P. D'Este and P. Patel. University–industry linkages in the uk: What are the factors underlying the variety of interactions with industry? *Research policy*, 36(9):1295–1313, 2007.

P. D'este, O. Llopis, F. Rentocchini, and A. Yegros. The relationship between interdisciplinarity and distinct modes of university-industry interaction. *Research Policy*, 48(9):103799, 2019.

M. Eidsaa and E. Almaas. S-core network decomposition: A generalization of k-core analysis to weighted networks. *Physical Review E*, 88(6):062819, 2013.

E. J. Emond and D. W. Mason. A new rank correlation coefficient with application to the consensus ranking problem. *Journal of Multi-Criteria Decision Analysis*, 11(1):17–28, 2002.

T. C. Erren, D. M. Shaw, and P. Lewis. Small groups, open doors: Fostering individual and group creativity within research communities, 2017.

H. J. Falk-Krzesinski, N. Contractor, S. M. Fiore, K. L. Hall, C. Kane, J. Keyton, J. T. Klein, B. Spring, D. Stokols, and W. Trochim. Mapping a research agenda for the science of team science. *Research Evaluation*, 20(2):145–158, 2011.

R. Fasoulis, K. Bougiatiotis, F. Aisopos, A. Nentidis, and G. Paliouras. Error detection in knowledge graphs: Path ranking, embeddings or both. *arXiv preprint arXiv:2002.08762*, 2020.

R. C. Feenstra, R. Inklaar, and M. P. Timmer. The next generation of the penn world table. *American economic review*, 105(10):3150–3182, 2015.

A. Fernandez-Zubieta, A. Geuna, and C. Lawson. What do we know of the mobility of research scientists and impact on scientific production. In *Global mobility of research scientists*, pages 1–33. Elsevier, 2015.

J. Fitzgerald, S. Ojanperä, and N. O'Clery. Is academia becoming more localised? the growth of regional knowledge networks within international research collaboration. *Applied Network Science*, 6(1):1–27, 2021.

L. Fleck. *Genesis and development of a scientific fact*. University of Chicago Press, 2012.

L. Fleming. Recombinant uncertainty in technological search. *Management Science*, 47(1):117–132, 2001.

A. Flexner. The usefulness of useless knowledge. In *The Usefulness of Useless Knowledge*. Princeton University Press, 2017.

M. Fontana, M. Iori, F. Montobbio, and R. Sinatra. New and atypical combinations: An assessment of novelty and interdisciplinarity. *Research Policy*, 49(7):104063, 2020.

D. Foray and F. Lissoni. University research and public–private interaction. In *Handbook of the Economics of Innovation*, volume 1, pages 275–314. Elsevier, 2010.

S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, et al. Science of science. *Science*, 359(6379), 2018.

J. G. Foster, A. Rzhetsky, and J. A. Evans. Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5):875–908, 2015.

J. G. Foster, F. Shi, and J. Evans. Surprise! measuring novelty as expectation violation. 2021.

M. R. Frank, D. Wang, M. Cebrian, and I. Rahwan. The evolution of citation graphs in artificial intelligence research. *Nature Machine Intelligence*, 1(2):79–85, 2019.

C. Franzoni, P. Stephan, and R. Veugelers. Funding risky research. *Entrepreneurship and Innovation Policy and the Economy*, 1(1):103–133, 2022.

K. Frenken, S. Hardeman, and J. Hoekman. Spatial scientometrics: Towards a cumulative research program. *Journal of informetrics*, 3(3):222–232, 2009.

J. Friedmann. A general theory of polarized development. 1967.

C. V. Fry, X. Cai, Y. Zhang, and C. S. Wagner. Consolidation in a crisis: Patterns of international collaboration in early covid-19 research. *PloS one*, 15(7):e0236307, 2020.

R. J. Funk and J. Owen-Smith. A dynamic network measure of technological change. *Management science*, 63(3):791–817, 2017.

N. Gai, K. Aoyama, D. Faraoni, N. M. Goldenberg, D. N. Levin, J. T. Maynes, M. J. McVey, F. Munshey, A. Siddiqui, T. Switzer, et al. General medical publications during covid-19 show increased dissemination despite lower validation. *Plos one*, 16(2):e0246427, 2021.

R. Garcia, V. Araújo, S. Mascarini, E. Santos, and A. Costa. How long-term university-industry collaboration shapes the academic productivity of research groups. *Innovation*, 22(1):56–70, 2020.

A. Gazni, C. R. Sugimoto, and F. Didegah. Mapping world scientific collaboration: Authors, institutions, and countries. *Journal of the American Society for Information Science and Technology*, 63(2):323–335, 2012.

A. Geuna. *The economics of knowledge production: funding and the structure of university research*. Edward Elgar Publishing, 1999.

A. Geuna. *Global mobility of research scientists: The economics of who goes where and why*. Academic Press, 2015.

W. Glänzel. National characteristics in international scientific co-authorship relations. *Scientometrics*, 51(1):69–115, 2001.

M. Gofman and Z. Jin. Artificial intelligence, human capital, and innovation. *Human Capital, and Innovation (August 20, 2019)*, 2019.

M. Gofman and Z. Jin. Artificial intelligence, education, and entrepreneurship. *Journal of Finance, Forthcoming*, 2022.

B. González-Pereira, V. Guerrero-Bote, and F. Moya-Anegon. The sjr indicator: A new indicator of journals' scientific prestige. *arXiv preprint arXiv:0912.4141*, 2009.

C. A. E. Goodhart. *Monetary theory and practice: The UK experience*. Springer, 1984.

Q. Gui, C. Liu, and D. Du. International knowledge flows and the role of proximity. *Growth and Change*, 49(3):532–547, 2018.

F. H. Guleid, R. Oyando, E. Kabia, A. Mumbi, S. Akech, and E. Barasa. A bibliometric analysis of covid-19 research in africa. *BMJ Global Health*, 6(5):e005690, 2021.

R. Gustafsson and E. Autio. A failure trichotomy in knowledge exploration and exploitation. *Research Policy*, 40(6):819–831, 2011.

M. Haghani and M. C. Bliemer. Covid-19 pandemic and the unprecedented mobilisation of scholarly efforts prompted by a health crisis: Scientometric comparisons across sars, mers and 2019-ncov literature. *Scientometrics*, 125:2695–2726, 2020.

D. Hain, R. Jurowetzki, T. Buchmann, and P. Wolf. Text-based technological signatures and similarities: how to create them and what to do with them. *arXiv preprint arXiv:2003.12303*, 2020.

D. S. Hain and R. Jurowetzki. Incremental by design? on the role of incumbents in technology niches. In *Foundations of Economic Change*, pages 299–332. Springer, 2017.

S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 2125–2126, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2945386. URL `http://doi.acm.org/10.1145/2939672.2945386`.

T. Hale, N. Angrist, R. Goldszmidt, B. Kira, A. Petherick, T. Phillips, S. Webster, E. Cameron-Blake, L. Hallas, S. Majumdar, et al. A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nature human behaviour*, 5(4):529–538, 2021.

M.-G. Hâncean, M. Perc, and J. Lerner. The coauthorship networks of the most productive european researchers. *Scientometrics*, 126(1):201–224, 2021.

K. Hao. "I started crying": Inside Timnit Gebru's last days at Google, 2020. URL `https://www.technologyreview.com/2020/12/16/1014634/google-ai-ethics-lead-timnit-gebru-tells-story/`.

S. Hemlin, C. M. Allwood, B. R. Martin, and M. D. Mumford. Why is leadership important for creativity in science, technology, and innovation? introduction. In *Creativity and leadership in science, technology, and innovation*, pages 1–26. Routledge, 2013.

R. M. Henderson and K. B. Clark. Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative science quarterly*, pages 9–30, 1990.

S. J. Herstad, T. Sandven, and B. Ebersberger. Recruitment, knowledge integration and modes of innovation. *Research Policy*, 44(1):138–153, 2015.

J. Hoekman, K. Frenken, and F. Van Oort. The geography of collaborative knowledge production in europe. *The annals of regional science*, 43:721–738, 2009.

B. Hofstra, V. V. Kulkarni, S. Munoz-Najar Galvez, B. He, D. Jurafsky, and D. A. McFarland. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17):9284–9291, 2020.

J. Homolak, I. Kodvanj, and D. Virag. Preliminary analysis of covid-19 academic information patterns: a call for open science in the times of closed borders. *Scientometrics*, 124:2687–2701, 2020.

E. Horlings and P. Van den Besselaar. Convergence in science: Growth and structure of worldwide scientific output, 1993–2008. In *2011 Atlanta conference on science and innovation policy*, pages 1–19. IEEE, 2011.

F. Jacob. Evolution and tinkering. *Science*, 196(4295):1161–1166, 1977.

L. B. Jeppesen and K. R. Lakhani. Marginality and problem-solving effectiveness in broadcast search. *Organization science*, 21(5):1016–1033, 2010.

Z. Jiang, M. Yang, M. Tsirlin, R. Tang, Y. Dai, and J. Lin. "low-resource" text classification: A parameter-free classification method with compressors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6810–6828, 2023.

B. F. Jones. The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies*, 76(1):283–317, 2009.

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

K. Jonkers and R. Tijssen. Chinese researchers returning home: Impacts of international mobility on research collaboration and scientific productivity. *Scientometrics*, 77:309–333, 2008.

M. Kato and A. Ando. National ties of international scientific collaboration and researcher mobility found in nature and science. *Scientometrics*, 110:673–694, 2017.

J. S. Katz. Scale-independent indicators and research evaluation. *Science and Public Policy*, 27(1):23–36, 2000.

J. S. Katz and B. R. Martin. What is research collaboration? *Research policy*, 26 (1):1–18, 1997.

Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 112(24): 7426–7431, 2015.

S. Kelchtermans, D. Neicu, and R. Veugelers. Off the beaten path: What drives scientists' entry into new fields? *FEB Research Report MSI_2008*, pages 1–42, 2020.

J. Kim and J. Diesner. Over-time measurement of triadic closure in coauthorship networks. *Social Network Analysis and Mining*, 7:1–12, 2017.

D. A. King. The scientific impact of nations. *Nature*, 430(6997):311–316, 2004.

C. M. Kinsella, P. D. Santos, I. Postigo-Hidalgo, A. Folgueiras-Gonzalez, T. C. Passchier, K. P. Szillat, J. O. Akello, B. Alvarez-Rodriguez, and J. Marti-Carreras. Preparedness needs research: How fundamental science and international collaboration accelerated the response to covid-19. *PLoS pathogens*, 16(10):e1008902, 2020.

J. Klinger, J. Mateos-Garcia, and K. Stathoulopoulos. A narrowing of ai research? *arXiv preprint arXiv:2009.10385*, 2020.

M. D. König, C. J. Tessone, and Y. Zenou. Nestedness in networks: A theoretical model and some applications. *Theoretical Economics*, 9(3):695–752, 2014.

T. Koopmann, M. Stubbemann, M. Kapa, M. Paris, G. Buenstorf, T. Hanika, A. Hotho, R. Jäschke, and G. Stumme. Proximity dimensions and the emergence of collaboration: a hyptrails study on german ai research. *Scientometrics*, pages 1–22, 2021.

T. S. Kuhn. *The structure of scientific revolutions*. University of Chicago press, 1962.

L. Kuld and J. O'Hagan. Rise of multi-authored papers in economics: Demise of the 'lone star'and why? *Scientometrics*, 114(3):1207–1225, 2018.

S. Kumar. Co-authorship networks: a review of the literature. *Aslib Journal of Information Management*, 2015.

V. Kumar, S. Sendhilkumar, and G. Mahalakshmi. Author similarity identification using citation context and proximity. In *2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM)*, pages 217–221. IEEE, 2017.

S. Kuznets. Modern economic growth: findings and reflections. *The American economic review*, 63(3):247–258, 1973.

K. R. Lakhani, K. J. Boudreau, P.-R. Loh, L. Backstrom, C. Baldwin, E. Lonstein, M. Lydon, A. MacCormack, R. A. Arnaout, and E. C. Guinan. Prize-based contests can provide solutions to computational biology problems. *Nature biotechnology*, 31(2):108–111, 2013.

M. T. Larsen. The implications of academic enterprise for public science: An overview of the empirical evidence. *Research Policy*, 40(1):6–19, 2011.

H. Laurençon, L. Saulnier, T. Wang, C. Akiki, A. V. del Moral, T. L. Scao, L. V. Werra, C. Mou, E. G. Ponferrada, H. Nguyen, J. Frohberg, M. Šaško, Q. Lhoest, A. McMillan-Major, G. Dupont, S. Biderman, A. Rogers, L. B. allal, F. D. Toni, G. Pistilli, O. Nguyen, S. Nikpoor, M. Masoud, P. Colombo, J. de la Rosa, P. Villegas, T. Thrush, S. Longpre, S. Nagel, L. Weber, M. R. Muñoz, J. Zhu, D. V. Strien, Z. Alyafeai, K. Almubarak, V. M. Chien, I. Gonzalez-Dios, A. Soroa, K. Lo, M. Dey, P. O. Suarez, A. Gokaslan, S. Bose, D. I. Adelani, L. Phan, H. Tran, I. Yu, S. Pai, J. Chim, V. Lepercq, S. Ilic, M. Mitchell, S. Luccioni, and Y. Jernite. The bigscience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL `https://openreview.net/forum?id=UoEw6KigkUn`.

E. Leahey. From sole investigator to team scientist: Trends in the practice and study of research collaboration. *Annual review of sociology*, 2016.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

D. Lee, Y. Heo, and K. Kim. A strategy for international cooperation in the covid-19 pandemic era: Focusing on national scientific funding data. In *Healthcare*, volume 8, page 204. MDPI, 2020.

J. J. Lee and J. P. Haupt. Scientific globalism during a global crisis: research collaboration and open access publications on covid-19. *Higher Education*, 81:949–966, 2021.

Y.-N. Lee, J. P. Walsh, and J. Wang. Creativity in scientific teams: Unpacking novelty and impact. *Research policy*, 44(3):684–697, 2015.

C. H. Liao. How to improve research quality? examining the impacts of collaboration intensity and member diversity in collaboration networks. *Scientometrics*, 86(3): 747–761, 2011.

Y. Lin. Where do new ideas come from: New directions in science emerge from disconnection and discord. Technical report, University of Chicago, 2021.

Y. Lin, J. Evans, and L. Wu. Novelty, disruption, and the evolution of scientific impact. 2021.

M. Liu and X. Hu. Movers' advantages: The effect of mobility on scientists' productivity and collaboration. *Journal of Informetrics*, 16(3):101311, 2022.

M. Liu, Y. Bu, C. Chen, J. Xu, D. Li, Y. Leng, R. B. Freeman, E. Meyer, W. Yoon, M. Sung, et al. Can pandemics transform scientific novelty? evidence from covid-19. 2021.

R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The lancet*, 395(10224):565–574, 2020.

T. Luukkonen, O. Persson, and G. Sivertsen. Understanding patterns of international scientific collaboration. *Science, Technology, & Human Values*, 17(1):101–126, 1992.

J. Mairesse, M. Pezzoni, et al. The impact of novelty in scientific articles: The case of french physicists. *Revue d'économie industrielle*, 174(2e), 2021.

J. G. March. Exploration and exploitation in organizational learning. *Organization science*, 2(1):71–87, 1991.

G. Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

N. Maslej, L. Fattorini, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, H. Ngo, J. C. Niebles, V. Parli, Y. Shoham, R. Wald, J. Clark, and R. Perrault. *The AI Index 2023 Annual Report*. AI Index Steering Committee,

Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2023, 2023.

K. Matsumoto, S. Shibayama, B. Kang, and M. Igami. Introducing a novelty indicator for scientific research: validating the knowledge-based combinatorial approach. *Scientometrics*, pages 1–25, 2021.

R. M. May. The scientific wealth of nations. *Science*, 275(5301):793–796, 1997.

M. Mazzucato. Mission-oriented innovation policies: challenges and opportunities. *Industrial and corporate change*, 27(5):803–815, 2018.

M. A. McFadyen and A. A. Cannella Jr. Social capital and knowledge creation: Diminishing returns of the number and strength of exchange relationships. *Academy of management Journal*, 47(5):735–746, 2004.

D. L. Medin and C. D. Lee. Diversity makes better science. *APS Observer*, 25, 2012.

J. Melkers and A. Kiopa. The social capital of global ties in science: The added value of international collaboration. *Review of Policy Research*, 27(4):389–414, 2010.

T. Mendes and L. Carvalho. Shifting geographies of knowledge production: The coronavirus effect. *Tijdschrift voor economische en sociale geografie*, 111(3):205–210, 2020.

R. K. Merton. Priorities in scientific discovery: a chapter in the sociology of science. *American sociological review*, 22(6):635–659, 1957.

J. Mervis. Biden, congress roll out big plans to expand national science foundation. *Sciencemag*, Science and Policy, 2021. doi: 10.1126/science.abi8778.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

A. Milewska. Knowledge based economy: Opportunities and challenges. In *Proceedings of the International Scientific Conference" Economic Sciences for Agribusiness and Rural Economy"*, number 2, 2018.

S. Milojević. Quantifying the cognitive extent of science. *Journal of Informetrics*, 9 (4):962–973, 2015.

J. A. Moral Muñoz, E. Herrera Viedma, A. Santisteban Espejo, M. J. Cobo, et al. Software tools for conducting bibliometric analysis in science: An up-to-date review. 2020.

B. Mueller. The building blocks of creativity and new ideas. *RAUSP Management Journal*, 54:242–246, 2019.

J. Nahapiet and S. Ghoshal. Social capital, intellectual capital, and the organizational advantage. *Academy of management review*, 23(2):242–266, 1998.

R. R. Nelson. *An evolutionary theory of economic change*. harvard university press, 1985.

R. R. Nelson and P. M. Romer. Science, economic growth, and public policy. *Challenge*, 39(1):9–21, 1996.

M. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. *Complex networks*, pages 337–370, 2004a.

M. E. Newman. Analysis of weighted networks. *Physical review E*, 70(5):056131, 2004b.

F. Niu and J. Qiu. Network structure, distribution and the growth of chinese international research collaboration. *Scientometrics*, 98:1221–1233, 2014.

I. Nonaka. A dynamic theory of organizational knowledge creation. *Organization science*, 5(1):14–37, 1994.

I. Nonaka, G. Von Krogh, and S. Voelpel. Organizational knowledge creation theory: Evolutionary paths and future advances. *Organization studies*, 27(8):1179–1208, 2006.

B. Nooteboom. *Learning and innovation in organizations and economies*. OUP Oxford, 2000.

B. Nooteboom, W. Van Haverbeke, G. Duysters, V. Gilsing, and A. Van den Oord. Optimal cognitive distance and absorptive capacity. *Research policy*, 36(7):1016–1034, 2007.

E. Noyons, H. Moed, and A. Van Raan. Integrating research performance analysis and science mapping. *Scientometrics*, 46(3):591–604, 1999.

OECD. Effective policies to foster high-risk/high-reward research. (112), 2021. doi: https://doi.org/https://doi.org/10.1787/06913b3b-en. URL `https://www.oecd-ilibrary.org/content/paper/06913b3b-en`.

P. Otlet. *Traite de documentation: le livre sur le livre, theeorie et pratique*. Editiones mundaneum, 1990.

A. Palayew, O. Norgaard, K. Safreed-Harmon, T. H. Andersen, L. N. Rasmussen, and J. V. Lazarus. Pandemic publishing poses a new covid-19 challenge. *Nature Human Behaviour*, 4(7):666–669, 2020.

R. K. Pan, K. Kaski, and S. Fortunato. World citation and collaboration networks: uncovering the role of geography in science. *Scientific reports*, 2(1):1–7, 2012.

A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research, 2020.

P. B. Paulus and B. A. Nijstad. *Group creativity: Innovation through collaboration*. Oxford University Press, 2003.

P. B. Paulus, N. W. Kohn, L. E. Arditti, and R. M. Korde. Understanding the group size effect in electronic brainstorming. *Small Group Research*, 44(3):332–352, 2013.

K. Pavitt. What makes basic research economically useful? *Research policy*, 20(2): 109–119, 1991.

P. Pelletier and K. Wirtz. Novelpy: A python package to measure novelty and disruptiveness of bibliometric and patent data. *arXiv preprint arXiv:2211.10346*, 2022.

A. M. Petersen, M. E. Ahmed, and I. Pavlidis. Grand challenges and emergent modes of convergence science. *Humanities and Social Sciences Communications*, 8(1):1–15, 2021.

H. Poincaré. Mathematical creation. *The Monist*, pages 321–335, 1910.

R. Ponds. The limits to internationalization of scientific research collaboration. *The Journal of Technology Transfer*, 34:76–94, 2009.

W. W. Powell and K. Snellman. The knowledge economy. *Annual review of sociology*, pages 199–220, 2004.

J. Priem, H. Piwowar, and R. Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

A. Pritchard et al. Statistical bibliography or bibliometrics. *Journal of documentation*, 25(4):348–349, 1969.

G. Pruschak and C. Hopp. And the credit goes to. . . -ghost and honorary authorship among social scientists. *Plos One*, 17(5):e0267312, 2022.

A. Purkayastha, E. Palmaro, H. J. Falk-Krzesinski, and J. Baas. Comparison of two article-level, field-independent citation metrics: Field-weighted citation impact (fwci) and relative citation ratio (rcr). *Journal of Informetrics*, 13(2):635–642, 2019.

P. Radanliev, D. De Roure, R. Walton, M. Van Kleek, O. Santos, and L. T. Maddox. What country, university or research institute, performed the best on covid-19? bibliometric analysis of scientific literature. *arXiv preprint arXiv:2005.10082*, 2020.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

C. M. Rawlings and D. A. McFarland. Influence flows in the academy: Using affiliation networks to assess peer effects among researchers. *Social Science Research*, 40(3):1001–1017, 2011.

J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical review E*, 74(1):016110, 2006.

A. Remmel. Us national science foundation set for a funding boom, 2021.

A. Rodríguez-Navarro. Research assessment based on infrequent achievements: A comparison of the u nited s tates and e urope in terms of highly cited papers and n obel p rizes. *Journal of the Association for Information Science and Technology*, 67(3):731–740, 2016.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2021.

M. A. Runco and G. J. Jaeger. The standard definition of creativity. *Creativity research journal*, 24(1):92–96, 2012.

S. Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

I. Sample. 'We can't compete': why universities are losing their best AI scientists, 2017. URL `https://www.theguardian.com/science/2017/nov/01/cant-compete-universities-losing-best-ai-scientists`.

T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

J. Schot and W. E. Steinmueller. Three frames for innovation policy: R&d, systems of innovation and transformative change. *Research policy*, 47(9):1554–1567, 2018.

J. A. Schumpeter et al. *Business cycles*, volume 1. McGraw-Hill New York, 1939.

G. Secundo, S. E. Perez, Ž. Martinaitis, and K. H. Leitner. An intellectual capital framework to measure universities' third mission activities. *Technological Forecasting and Social Change*, 123:229–239, 2017.

F. Shi, J. G. Foster, and J. A. Evans. Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Social Networks*, 43:73–85, 2015.

S. Shibayama, D. Yin, and K. Matsumoto. Measuring novelty in science with word embedding. *PloS one*, 16(7):e0254034, 2021.

S. J. Shin and J. Zhou. When is educational specialization heterogeneity related to creativity in research and development teams? transformational leadership as a moderator. *Journal of applied Psychology*, 92(6):1709, 2007.

I. Solaiman. The gradient of generative ai release: Methods and considerations, 2023.

P. Stephan. *How economics shapes science*. Harvard University Press, 2012.

P. E. Stephan. The economics of science. *Journal of Economic literature*, 34(3): 1199–1235, 1996.

P. E. Stephan. The economics of science-funding for research. *International Centre for Economic Research Working Paper*, (12), 2010.

P. E. Stephan and S. G. Levin. Exceptional contributions to us science by the foreign-born and foreign-educated. *Population research and Policy review*, 20:59–79, 2001.

E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

C. R. Sugimoto and V. Larivière. *Measuring research: what everyone needs to know*. Oxford University Press, 2018.

L. Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10:10–10:29, 2013. ISSN 1542-7730.

I. Tahamtan and L. Bornmann. Creativity in science and the link to cited references: Is the creative potential of papers reflected in their cited references? *Journal of Informetrics*, 12(3):906–930, 2018.

A. Taylor and H. R. Greve. Superman or the fantastic four? knowledge combination and experience in innovative teams. *Academy of management journal*, 49(4):723–740, 2006.

D. Trapido. How novelty in knowledge earns recognition: The role of consistent identities. *Research Policy*, 44(8):1488–1500, 2015.

B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

T. Van Leeuwen, H. Moed, R. Tijssen, M. Visser, and A. Van Raan. Language biases in the coverage of the science citation index and its consequencesfor international comparisons of national research performance. *scientometrics*, 51(1): 335–346, 2001.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

M. Veale and F. Zuiderveen Borgesius. Demystifying the draft eu artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4):97–112, 2021.

R. Veugelers and J. Wang. Scientific novelty and technological impact. *Research Policy*, 48(6):1362–1372, 2019.

W. G. Vincenti. What engineers know and how they know it analytical studies from aeronautical history. 1990.

C. S. Wagner and L. Leydesdorff. Network structure, self-organization, and the growth of international collaboration in science. *Research policy*, 34(10):1608–1618, 2005.

C. S. Wagner, T. A. Whetsell, and L. Leydesdorff. Growth of international collaboration in science: revisiting six specialties. *Scientometrics*, 110:1633–1652, 2017.

C. S. Wagner, T. A. Whetsell, and S. Mukherjee. International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination. *Research Policy*, 48(5):1260–1270, 2019.

M. L. Wallace, V. Larivière, and Y. Gingras. A small world of citations? the influence of collaboration networks on citation practices. *PloS one*, 7(3):e33339, 2012.

L. Waltman. A review of the literature on citation impact indicators. *Journal of informetrics*, 10(2):365–391, 2016.

D. Wang and A.-L. Barabási. *The science of science*. Cambridge University Press, 2021.

J. Wang, R. Veugelers, and P. Stephan. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8):1416–1436, 2017.

L. Wang and X. Wang. Who sets up the bridge? tracking scientific collaborations between china and the european union. *Research Evaluation*, 26(2):124–131, 2017.

E. Wedell, M. Park, D. Korobskiy, T. Warnow, and G. Chacko. Center–periphery structure in research communities. *Quantitative Science Studies*, 3(1):289–314, 2022.

M. L. Weitzman. Recombinant growth. *The Quarterly Journal of Economics*, 113 (2):331–360, 1998.

J. Whitfield. Group theory; what makes a successful team? john whitfield looks at research that uses massive online databases and network analysis to come up with

some rules of thumb for productive collaborations. *Nature*, 455(7214):720–724, 2008.

K. Wiggers. Ai startup cohere launches a nonprofit research lab, June 14 2022. URL `https://techcrunch.com/2022/06/14/ai-startup-cohere-launches-a-nonprofit-research-lab/`.

S. G. Winter and R. R. Nelson. An evolutionary theory of economic change. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*, 1982.

U. Witt. Propositions about novelty. In *Rethinking Economic Evolution*. Edward Elgar Publishing, 2016.

World Bank. *Building knowledge economies: Advanced strategies for development*. The World Bank, 2007.

L. Wu, D. Wang, and J. A. Evans. Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382, 2019.

L. Wu, A. Kittur, H. Youn, S. Milojević, E. Leahey, S. M. Fiore, and Y.-Y. Ahn. Metrics and mechanisms: Measuring the unmeasurable in the science of science. *Journal of Informetrics*, 16(2):101290, 2022.

Q. Wu and Z. Yan. Solo citations, duet citations, and prelude citations: New measures of the disruption of academic papers. *arXiv preprint arXiv:1905.03461*, 2019.

S. Wu and Q. Wu. A confusing definition of disruption, Apr 2019. URL `osf.io/preprints/socarxiv/d3wpk`.

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

Y. Xie and A. A. Killewald. *Is American science in decline?* Harvard University Press, 2012.

F. Xu, L. Wu, and J. Evans. Flat teams drive scientific innovation. *Proceedings of the National Academy of Sciences*, 119(23):e2200927119, 2022.

J. Xu, S. Kim, M. Song, M. Jeong, D. Kim, J. Kang, J. F. Rousseau, X. Li, W. Xu, V. I. Torvik, et al. Building a pubmed knowledge graph. *Scientific data*, 7(1): 1–15, 2020.

A. Yonezawa. The internationalization of japanese higher education: Policy debates and realities. *Higher education in the Asia-Pacific: Strategic responses to globalization*, pages 329–342, 2011.

L. Zhang, W. Zhao, B. Sun, Y. Huang, and W. Glänzel. How scientific research reacts to international public health emergencies: a global analysis of response patterns. *Scientometrics*, 124:747–773, 2020.

M. Zitt, E. Bassecoulard, and Y. Okubo. Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science. *Scientometrics*, 47(3):627–657, 2000.

A. Zomer and P. Benneworth. The rise of the university's third mission. *Reform of higher education in Europe*, pages 81–101, 2011.

J. Zou and L. Schiebinger. Ai can be sexist and racist—it's time to make it fair, 2018.

L. G. Zucker, M. R. Darby, and M. B. Brewer. Intellectual capital and the birth of us biotechnology enterprises, 1994.

List of Figures

List of Tables

Kevin Wirtz

Understanding Scientific Collaboration and Mobility: A Creativity Perspective

RÉSUMÉ

Cette thèse examine la structure et la dynamique du système scientifique, en se concentrant sur la collaboration et la mobilité des chercheurs. Le Chapitre 1 explore la réaction du système de collaboration internationale à un choc exogène, dans le but d'améliorer notre compréhension de son fonctionnement. Les Chapitres 2 et 3 se penchent sur les facteurs qui influencent la créativité en étudiant la dimension cognitive des équipes de recherche, cherchant à révéler ce qui favorise leur capacité à générer des idées innovantes et impactantes. Enfin, le Chapitre 4 se concentre sur les chercheurs spécialisés en intelligence artificielle et étudie leur transition de l'académie à l'industrie. Nous analysons ensuite l'impact de cette transition sur leurs recherche.

Mots clefs: Structure de collaboration; Equipes Scientifiques; Mobilité chercheurs; Nouveauté Combinatoire; Impact Scientifique

RÉSUMÉ EN ANGLAIS

This thesis examines the structure and dynamics of the scientific system, with a focus on researcher collaboration and mobility. Chapter 1 explores the response of the international collaboration system to an exogenous shock, aiming to enhance our understanding of its functioning. Chapter 2 and 3 delves into the factors influencing creativity by studying the cognitive dimension of research teams, seeking to uncover what promotes their ability to generate impactful and innovative ideas. Lastly, Chapter 4 concentrates on researchers specialized in artificial intelligence and investigates their transition from academia to industry. We then analyze the impact of this transition on their research outcomes.

Keywords: Collaboration Structure; Scientific Teams; Researchers' Mobility; Combinatorial Novelty; Scientific Impact