

EMBRACING COMPLEXITY IN THE SCIENCE OF SCIENCE

Dakota S. Murray

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Department of Informatics,

Indiana University

August 26, 2021

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Doctoral Committee

Cassidy R. Sugimoto

Cassidy R. Sugimoto, PhD

Yong-Yeol Ahn

Yong-Yeol Ahn, PhD

Staša Milojević

Staša Milojević, PhD

Santo Fortunato

Santo Fortunato, PhD

Guillaume Cabanac

Guillaume Cabanac, PhD

August, 23, 2021

Acknowledgments

First and foremost I want to thank the careful time and patience of my doctoral committee. Dr. Cassidy Sugimoto was my first mentor. She took a chance on me, some kid in a pile who did stuff with computers and bees and NASA, and now I ended up here. Her support throughout this process has been extensive and unwavering, and I will forever be grateful to have learned so much science from her. Dr. Yong-Yeol Ahn came later as my co-advisor, but was no less influential; I thank him for all the tools and techniques and ways of doing science that he taught me, and how he set me up for success wherever I go. Drs. Staša Milojević, Santo Fortunato, and Guillaume Cabanac also all took the time and energy to make my career possible; they gave me their time, their kindness, and their knowledge. I can only hope to one day pay forward the kindness and attention of my committee to students of my own.

I also want to thank all the co-authors appearing on works in this dissertation, including Dr. Kyle Siler, Dr. Vincent Larivière, Dr. Wei Mun Chan, Andrew M. Collings, Dr. Jennifer Raymond, Dr. Kevin Boyack, Wout Lamers, Dr. Ludo Waltman, Dr. Nees Jan van Eck, Dr. Rodrigo Costas, Dr. Woo-Sung Jung, Clara Boothby, Huimeng Zhao, Vanessa Minik, Nicolas Bérubé, Sadamori Kojaku, and Jisung Yoon. Every one of these collaborators made essential contributions to the studies appearing in this dissertation, and some of the words appearing herein belong to them.

I also want to thank the members of both Dr. Sugimoto's and Dr. Ahn's research groups, who have endured many talks about elements of my dissertation, and exhibited patience and appreciation for my work, and whose commetsns contributed to the work you see here. In particular I want to thank Lili Miao, my fellow PhD student and one of my first and longest-lasting friends in Bloomington. Her advice, feedback, and companionship working in the neighboring desk contributed in no small way to my mental health, and the eventual fruition of my work.

I give special thanks for the peer reviewers who took the time to do a mostly thankless job

to assess and suggest improvements for the studies in this manuscript. Most of them have done a fantastic job and made this work better.

I had the good fortune to be supported by several grants throughout my graduate study, among them funding from the National Science Foundation (SciSIP #1561299) and the Air Force Office of Scientific Research under award number FA9550-19-1-0391. Additionally, I thank Indiana University and the Luddy School of Informatics, Computing, and Engineering for providing the space and resources for doing this work, and the tremendous intellectual community of the Department of Informatics. I hope that the sheer disciplinary breadth of my training and community are shown in my writing.

Finally, I thank my wife, Dongeun Shin, who's patience, support, and understanding have been extraordinary, and who I cannot imagine a life without. I also thank my newborn daughter, Elle, for her invaluable contributions to hurrying things along.

For Dongeun & Eleanor, who provide a life of love and a life I love living.

Preface

I've never been good at wrapping gifts, and this dissertation has been a particularly tricky gift to wrap. It should be easy. After all, its only a "stapled" dissertation, the papers are already done! I just need to attach them together with some rudimentary semantic glue. But it wasn't easy. Together, these papers make up a very large and very oddly-shaped gift, that proved difficult to write in a single narrative wrapper. At least not in a wrapping that I would be willing to gift my wife. Individually, each study is small, focused, and its meaning clear. Yet much like the complex systems this dissertation ended up being about, together they form something entirely new, some great n -sided polygon with jutting towers, jagged edges, and sharp corners. How could I possibly wrap such an awful shape and tie a bow on top? More than that, I wanted to add something *of my own* into the mix; all of those studies were written with the help of other co-authors; I did the majority of the work, but that is *our* work. I wanted to add *my* work, to see if I could turn these five years of doctoral education into something productive. So now I not only needed to somehow wrap these studies together; the box they are in better be nice too.

I went through several iterations. It could be about measuring success in science! Or maybe about self-organizing systems! Maybe even science policy could be there too? On the fourth or fifth or whatever-numbered version, I was forced to commit; the baby was nearly here after all, a gift that would be altogether more beautiful, perfectly-wrapped, and infinitely demanding of my attention.

The final decision was simple: make the dissertation about complexity. Namely, how might these individual publications, each an exemplar from the field of Metascience, be viewed through the perspective of Complexity Science? What else might we learn about how science works, that wasn't already revealed within these studies? I didn't pull this narrative out of the void. After all, I attended the Santa Fe Institute's Complex Systems Summer School. And I work in a department

of Complex Systems. And my advisors consist of several complexity scientists. So I thought I would finally put that exposure to use.

And so I started writing. And I kept writing. Past the theories and conceptual framework, past the discussion, and through the introduction. What you see here is the eventual destination of that initial thread, "what is complexity". I'm unhappy with a lot of it. The edges remain rough, the pieces don't stick together quite as well as I would have hoped but there are parts of it that I like, too. I like Complexity Science, I think it reflects reality best. I think it reflects life best. After all, life is complex. The studies in this dissertation were written over a span of years; in that time, happenstance put me and my wife together, my social networks gave me access to wonderful opportunities, and cumulative advantage helped propel me into the makings of an actual academic career. Oh the complexity! Where things could have gone different, what subtle differences might have seen me careening into an entirely different trajectory for my life? And now I have a baby—an agent of chaos and a master of re-writing fates. How would this dissertation be different if this baby never came? Would I have committed to this narrative path, or would I have changed my mind, winding up with an entirely different dissertation, talking to you now about the implications of evaluation metrics or success or whatever? Who's to know. That's the beauty of complexity, and its the lesson I took away from writing this dissertation.

Life is complex, and science, being made up of lives, is tremendously more complex. The field of Metascience should try and embrace that complexity, or at least stop working against it. I think that if a complexity perspective were more widely adopted in Metascience, then we might learn something new about how science works. About how babies, marriages, friendships, languages, geographies, cultures, economies, colleagues, ideas, and so many more all work together to produce the science that we see today. And the most important lesson of all? That everything could have been different. Each individual scientist's life could have been so different. And so science, made up of scientists, could look very different too. Knowing that things could look different gives us hope

for making things better, but also fear for making things worse. I hope that Complexity Science can help us know what shapes science and how, and maybe even give advice on making the entire scientific system better and more fair. It will take a lot of work, mountains of data, and more than a little imagination, but I think we can do it.

So that's where I am. As I write the baby is fussing from the next room, waking up from a nap. Elle is her name. She still hasn't forgiven us for taking her to the doctor to get her shots yesterday, and she lets us know. She's lucky that she is so adorable. Oh how things could have been different, how they could have been worse, how they will keep getting better. So here is my gift. Its ugly. The box is bulging and the wrapper is torn all over, straining to hold whats inside only with liberal use of tape and patched holes. I mostly made it for me. To whoever is reading, I hope you like it too.

Dakota S. Murray

EMBRACING COMPLEXITY IN THE SCIENCE OF SCIENCE

Science, already a massive and global enterprise, continues to grow in both size and complexity. Accompanying this growth is a deluge of data on scientific activity and advancements in computational techniques that open new avenues for its analysis. Leveraging these new data and techniques, the field of *Science of Science* turns the tools of science upon science itself, aiming to understand its composition and behavior. While making significant contributions, the study of science remains constrained by its vastness, heterogeneity, and interconnectedness. Here, I argue that the *Science of Science* benefits from viewing science as a *complex system*, and drawing on the conceptual framework developed to understand such systems in other domains. Specifically, I detail a *complexity perspective* that conceptualizes science as a self-organizing system of interconnected scientists in which bottom-up interactions between individuals give way to emergent global structure and behavior. I demonstrate the value of this perspective by using it to interpret four studies covering diverse topics in *Science of Science*: bias in peer review, prejudice in teaching evaluations of university faculty, the incidence of disagreement in science, and the landscape of global scientific mobility. Viewing my findings through the lens of complexity, I disentangle the forces that contribute to the behavior of individual scientists and illustrate how feedback mechanisms simultaneously entrench social structures while maintaining the potential for revolutionary change. This perspective also reveals how the inherent complexity of science poses challenges for its study, confounding attempts at objective measurement. Finally, I argue that the complexity perspective is uniquely positioned to benefit from future advancements in data availability and methodology and to provide further insights into the fundamental composition and behaviors of science. Embracing complexity offers a promising direction for the *Science of Science*, with deep implications for the understanding and governance of the global scientific enterprise.

Cassidy R. Sugimoto

Cassidy R. Sugimoto, PhD

Yong Yeol Ahn

Yong-Yeol Ahn, PhD

Staša Milojević

Staša Milojević, PhD

Santo Fortunato

Santo Fortunato, PhD

Guillaume Cabanac

Guillaume Cabanac, PhD

Contents

Acknowledgments	iii
Dedication	v
Preface	vi
Abstract	ix
1 Introduction	1
1.1 Science, its complicated	1
1.2 Embracing complexity	5
1.3 Historical context	8
1.4 Structure of dissertation	15
2 Science as a Complex System	18
2.1 What is a complex system?	18
2.2 Science as a complex system	22
2.3 Science as a self-organized system	26
2.4 Individual-level characteristics	29
2.5 Organizing principles of science	36
2.6 What complexity contributes to Metascience	55
3 Study 1: Peer review at <i>eLife</i>	58
3.1 Foreword	58
3.2 Abstract	60
3.3 Introduction	60
3.4 Consultative peer review and <i>eLife</i>	64

3.5	Data and methods	65
3.6	Results	74
3.7	Discussion	85
4	Study 2: Student-teacher evaluations at U.S. Universities	96
4.1	Foreword	96
4.2	Abstract	98
4.3	Introduction	98
4.4	Data and methods	101
4.5	Results	106
4.6	Discussion	113
4.7	Conclusion	122
5	Study 3: Measuring disagreement in science	124
5.1	Foreword	124
5.2	Abstract	126
5.3	Introduction	126
5.4	Literature Review	127
5.5	Materials and Methods	130
5.6	Results	142
5.7	Discussion	145
6	Study 4: The landscape of global scientific mobility	152
6.1	Foreword	152
6.2	Abstract	154
6.3	Introduction	154
6.4	Embeddings provide functional distance between locations	159

6.5 word2vec and the gravity model	160
6.6 Embeddings capture global structure of mobility	163
6.7 Conclusion	171
7 Discussion	185
7.1 Applying the complexity perspective	185
7.2 The self-organization of scientists	186
7.3 Stability, change, and feedback in science	190
7.4 Taking measure of a complex system	199
7.5 Moving forward with complexity in the Science of Science	202
8 Conclusion	206
Bibliography	209
A Study 1: Peer review at <i>eLife</i>	281
A.1 Text	281
A.2 Figures	283
A.3 Tables	293
B Study 2: Teaching evaluations	302
B.1 Text	302
B.2 Figures	307
B.3 Tables	316
C Study 3: Measuring disagreement	326
C.1 Text	326
C.2 Figures	337
C.3 Tables	337

D Study 4: Embedding mobility	352
D.1 Text	352
D.2 Figures	364
D.3 Tables	390

Chapter 1

Introduction

You cannot get through a single day without having an impact on the world around you. What you do makes a difference, and you have to decide what kind of difference you want to make.

Jane Goodall

The history of science will be as complex, chaotic, full of mistakes, and entertaining as the ideas it contains, and these ideas in turn will be as complex, chaotic, full of mistakes, and entertaining as are the minds who invented them.

Paul Feyerabend

1.1 Science, its complicated

Science is growing. Once an activity of a select few scattered individuals, it has developed into "big science" [1, 2], involving nearly nine million scientists (and counting) and millions more students and professionals around the world, supported by an ever-increasing amount of public and private investment [3–5]. As science has grown, it has also become more complex.

No longer constrained to a select few countries or a small number of elite organizations [6, 7], science is now global in scope. Once peripheral nations like China [8–10] and South Korea [11–13] have become major funders and producers of scientific knowledge [14]—examples of a broader

trend of deconcentration of science [15, 16]. The geography [17], politics [18], economics [19], priorities [20], and culture [21, 22] of each of these countries gives way to unique scientific systems. Science happens in more and more-diverse kinds of places than ever before, shaping the careers of scientists in a myriad of ways.

Scientists themselves are also more diverse than ever before, representing a greater variety of backgrounds than in the past. Following centuries of exclusion [23, 24], the United States has seen increased representation of women and marginalized minorities among its scientific workforce [25]. The global deconcentration of scientific production has also allowed more people from different nations and cultures to become a part of the global scientific enterprise [26, 27]. While representation remains uneven [28–32], parity a long way off [33], and many scientists still endure marginalization and discrimination [34–39], global science in 2021 is far more diverse than a century prior. More cultural, ethnic, national, and social backgrounds are represented in science than ever before, each experiences science in different ways and bringing their own experiences and perspectives to their work.

The kind of work scientists do has also broadened, becoming more complex and heterogeneous. Incentivized to pursue novelty [40], scientists mix and match different ideas, constantly expanding the boundaries of their field or filling in knowledge gaps [41–44]. Indeed, the extent of topics researched in disciplines has grown alongside their total publications [45, 46]. Yet the resource, knowledge, and labor inputs necessary for continued progress in many fields has also increased, necessitating that scientists become increasingly specialized both in terms of their knowledge [47] and the kinds of tasks they perform [48, 49]. These specialized scientists work in ever-larger teams [50], in which they can leverage their diverse perspectives and skills [51]. Today, scientists pursue a greater variety of topics more narrowly, and in more different social arrangements, than ever before.

Science is further complicated by scientists themselves, who move across these large geographic

and disciplinary spaces more freely than before. Increased globalization has allowed for extensive international collaboration [52–54] and mobility [55, 56], allowing the circulation of people and ideas around the world. Interdisciplinarity too, is on the rise, as scientists increasingly weave together disparate ideas [57, 58] and collaborate between fields [59, 60], even as scientists on the whole has become more specialized. Just as science has become more heterogeneous, scientists themselves also cross more boundaries, exchanging ideas and expertise across countries and disciplines, and creating an increasingly interconnected science.

The *Science of Science* [61] is a discipline that aims to turn the tools of science upon itself, leveraging massive bibliographic datasets and modern computational techniques in order to better understand the composition, organization, and behavior of global science. As science has grown, the Science of Science and related fields such as *Scientometrics* [**leydesdorff scientometrics'2012**] have provided valuable quantitative analysis of science, producing macroscopic maps of disciplines [62–64], informing models of success and failure [65, 66], and identifying inequalities and biases in need of intervention [67, 68]. However, the increasing complexity of science poses unique challenges to its continued study.

The size and heterogeneity of science puts limits on how far findings can be generalized. Much like Paul Feyerabend's criticism of a universal scientific method [69], an exception can be found somewhere in science, for any generalization of a finding in the Science of Science. For example, a model of success that holds true for Physics may break down when applied to Sociologists or Biologists, due to their distinct cultures [70] and social organizations [40]. Similarly, nations have their own distinct specializations [14, 71] and funding priorities [18], so that the kinds of scientists who can become successful in the United States may be very different than in China. A patchwork of different incentives and policies across nations, institutions, and departments also shape the day-to-day behavior of scientists, and how they choose to navigate their careers. For example, the choice to move internationally is influenced by national policies which often incentive foreign and

return travel [72–74]. There are as many ways of doing science as there are scientists, imposing strict limits on any finding or theory in the Science of Science.

The multi-faceted and interconnected nature of science makes it nearly impossible to identify causal mechanisms or to disentangle the vast array of factors involved in its behavior. Any one aspect of science is the result of a multitude of forces. For example, a scientist’s success stems not only from their performance, but from the biased judgements of their peers [65], which are influenced by everything from socioeconomic class [75], gender [76], existing reputation [77–79], and more. Science is also deeply interconnected, such that the behaviors and status of scientists emerge out of their interaction with others. For example, a scientist’s success might be a product of the relationship with their mentor [80] or with their eminent co-authors [81], whereas a nation’s scientific impact is shaped by their international collaborations [82–85]. The multitude of forces and interactions that give way to the structure and dynamics of science challenge the ability of the Science of Science to locate and disentangle causal relationships [86], further limiting the potential of the Science of Science.

Science is also *chaotic*, its behavior seemingly random and unpredictable. Major scientific breakthroughs and discoveries are inherently unpredictable “black swan”-like events [87], yet they have deep and profound impacts on their discipline. Papers that are initially unsuccessful may eventually become central to a new scientific field, accruing a huge number of citations decades after their publication [88]. In spite of advances in big data and machine learning technologies, the ability to predict the success of funding proposals and individuals remains elusive [89]. Prediction in science is further complicated because science is not isolated, but rather an *open system* subject to further chaos from external geopolitical, cultural, and economic events [90], such as the collapse of the Soviet Union [91, 92], or more recently the COVID-19 pandemic [93]. The limits to prediction confounds the ability of the Science of Science to create, validate, and replicate models or theories, constraining progress in the field.

Science is a massive, and massively-complex global enterprise. The Science of Science is poised to make progress in understanding the structure and dynamics of science by leveraging increasingly-available data on peer review outcomes [94] and publication full-text data [95], improvements to existing bibliographic databases [96, 97], and recent advances in machine learning tools [98–101]. However, just as these same data and tools expand the scope of the field, it also reveals the true complexity of science, and shows it to be increasing further still, posing important conceptual and methodological challenges to its study. Moving forward, the Science of Science will need to recognize this complexity and appreciate and adapt to its challenges, or else the field risk stalled progress and a fragmented, fleeting understanding of how science works.

1.2 Embracing complexity

In this dissertation, I propose that the Science of Science conceive of science as a *complex system*, drawing on the conceptual framework of Complexity Science in order to reason about its structure and dynamics. Specifically, I outline a *complexity perspective* of science, which views science as a vast network of interconnected individuals, and the behavior of science as emerging out of the interactions of these individuals. I argue that by adopting this perspective, the Science of Science will gain access to a wealth of valuable conceptual tools for appreciating and adapting to the complexity of science, providing important insights into its inner workings.

A complex system is, roughly, one that exhibits qualities such as emergent phenomenon, hierarchical organization, chaotic behavior, adaptation, non-linearity, and self-organization [102, 103]. Examples include networks of synapses in the brain [104], ecological food webs [105], and financial markets [106]. The scientific enterprise, too, can be easily conceived of as a massive globe-spanning complex system [107], in which individual scientists are connected through various interactions such as friendship, collaboration, referencing, or via shared community or institutional affiliation. The overall structure and behavior of science, then, is located not in top-down processes or general rules,

but instead emerge out of the bottom-up interactions of these individual scientists. The *complexity perspective* of science, as I define it here, defines sciences as a complex system, and lays out a set of detailed criteria and considerations relevant to the Science of Science (outlined more completely in Chapter 2). Adopting this perspective brings several important implications that I argue will augment the better Science of Science.

First, the complexity perspective shifts the focus of the Science of Science away from universal rules and top-down processes, and instead towards the everyday actions of scientists. For example, the physical location of a scientists office influences who they might meet and collaborate with [108, 109]. Their professional network, in turn, dictates where a scientist chooses to move in the future [110], or who a journal's editor might recruit to conduct a review [111]. A scientist's demographics, too, can shape the topics they choose to pursue [112, 113], influence how others perceive their accomplishments [34, 36, 114], and mediate the impact of other life events, such as childbirth [115]. Where a person applied for permanent employment, and who hires them, is similarly a vital element to spread of ideas [116]. When viewed through the complexity perspective, these local, idiosyncratic, and seemingly incidental factors are critical to the organization of global science, and so the everyday considerations of science become important topics for the Science of Science.

Second, by conceiving science as a complex system, the complexity perspective provides access to a diverse range of conceptual tools for reasoning about and understanding science. Concepts like *non-linearity*, for example, help describe why some scientists achieve such massive success compared to their peers [65], or how the potential for interaction gives cities an edge when it comes to producing research [117]. *Emergence* can also help explain how bottom-up processes give rise to the overall structure, and complex behavior of science [118]. Of particular interest to this dissertation is the concept of *feedback*, and how ubiquitous institutions in science act as feedback mechanisms that drive much of its organization and behavior. These concepts also prove natural

extensions of existing topics in the Science of Science literature, such as *non-linearity* building on existing research into scale-free properties of science [119], and *feedback* extending well-studied phenomenon like the Matthew Effect [79]. By adopting the complexity perspective, the Science of Science can, with little difficulty, integrate a powerful conceptual framework that has already found success in a wide variety of domains, and has the potential to do the same for the study of science.

Third, whereas the Science of Science tends to study scientists as isolated and independent, the complexity perspective instead views each scientist as embedded in a complex environment of people, institutions, and ideas. For example, a person’s career success is partially driven by their innate ability [120], but also by their interactions with their mentors [121, 122], collaborators [123], and their institutional community [124]. The complexity perspective, however, does not view scientists as passive members of this environment; rather, there is a *mutual shaping* between the individual and the system that in which they are embedded. On the one hand, a scientist is influenced by their environment, such as when trained in the dominant intellectual tradition of their field. On the other hand, they also act upon and change their environment, such as by publishing ideas that upend the existing paradigm. By adopting this view, the Science of Science is encouraged to treat phenomena not as inherent to a singular individuals, but instead to the dynamic *interaction* between them and their environment.

Finally, the complexity perspective is an inherently humbling framework that places explicit limits on the extent to which science can be understood. Science is vast, heterogeneous, interconnected, and chaotic system. These qualities complicate any attempt to generalize findings, make predictions, or establish universal theories of science. The Science of Science is further complicated by its reliance on observational data, which while massive, is often systematically biased against certain disciplines and languages [125, 126]. The complexity perspective doesn’t offer a solution to these issues, and makes no claim to predictive power or universal laws of science. Rather, the complexity perspective encourages recognizing the inherent complexity of science, and the limits

to its study. In adopting this perspective, researchers Science of Science should approach their subject with caution and to realize that any finding is contingent, and may have been different for a different discipline, a different country, or different time, or for different people.

By adopting this complexity perspective, the Science of Science can gain a powerful new lens with which to view science, and the conceptual tools to derive new insights into its structure and dynamics. In this dissertation, I illustrate the potential of the complexity perspective by applying it towards the interpretation of four studies in the Science of Science, each of which makes use of novel data or methodology. Specifically, I examine the incidence and nature of demographic bias at a particular life sciences journal *eLife*, I investigate the factors driving how students rate their tenure and tenure-track university faculty, I create a text-based measure to assess the degree of disagreement across millions of scientific articles, and I employ a novel method to capture and explore the many factors contributing to global scientific mobility. Each study is a significant contribution in and of themselves, representing different topical areas in the Science of Science. However, when viewed together with the complexity perspectives, they also shed light into how a multitude of individual-level forces and simple feedback mechanisms drive the behavior of science as a whole. Moreover, the perspective also provides important context to some of the fundamental challenges in the Science of Science, and offers potential steps for their mitigation. The complexity perspective also offers a promising foundation for future development, where it can be pushed forward with new data, techniques, to provide an even greater understanding of how science works.

1.3 Historical context

This dissertation is drawn out of two distinct fields, each with their own histories that provide important context to this work. The first is the *Science of Science*, a field which "...places the practice of science itself under the microscope, leading to a quantitative understanding of the genesis of scientific discovery, creativity, and practice and developing tools and policies aimed at

accelerating scientific progress” [61]. The second is *Complexity Science*, a much more general field that ”...studies the complex systems, which consist of a large number of components that interact with each other to produce nontrivial phenomena that cannot be explained by analyzing the individual constituent elements” [107]. Each of their histories are noisy, following not a single thread of development but rather drawing from many different historical moments. In this section, I provide a brief overview of some of the most important of these moments.

Science of Science

Though the name itself has history stretching back to de Solla Price in the 1960s [2], the *Science of Science* has emerged only recently as a cohesive field [61]. Taxonomically, it can be placed as a sub-category of *Metascience*—a term capturing all the fields that consider science as its object of study. Yet whereas Metascience includes qualitative disciplines, such as those that make use of ethnographic methods, the Science of Science is instead primarily concerned with the *quantitative* study of science. It is thus the history of quantitative science studies that I overview here, a topic that has been taken up by several fields, and gone through many versions, over its more than a century of history.

The earliest versions of quantitative study of science involved simple counts of the number of books or documents contained within an archive. Such tallies go as far back as the ancient Library of Alexandria [127], and persisted into the 1800s in the libraries of Munich, Paris, and across the U.S. [128]. The goal of these counts were often logistical—accounting for what each library owned. However, they also proved compelling for scholars, who found that such counts could act as rough estimates of the quantity of human knowledge [129]. Others would pursue more extensive or specialized counts, creating tabulations that included scholarly articles in the medical sciences [130] or published patents [131]. In other cases, such counts were used to demonstrate national prestige [132, 133]. These counts would become the basis for future *bibliometric* methods

in the Science of Science.

Around the turn of the 20th century, the ethically-wrought Eugenics movement would discover that these bibliometric counts could also be used to study *individual* performance, just as they did for nations. Specifically, the prominent Psychologist and then-editor of *Science*, McKeen Cantell, would study the productivity of so-called "great men" of science using their number of published papers [134, 135], a topic pushed further by later scholars [136, 137]. Although not the first attempt at counting individual-level publications¹, the Eugenics movement was still the first to integrate bibliometric measures into a scientific program aimed at studying the characteristics of scientists.

The difficulty of tabulating publications would constrain research efforts for several decades more, at least until the development of the *Science Citation Index* (SCI), a powerful tool that would prove essential to the Science of Science and Metascience. A citation index is intended to expand beyond simple *bibliometric* measures, like paper counts, and instead incorporate *bibliographic* quantities, such as the number of citations that a document received. The earliest citation indices were developed in the legal profession in the 1700s and 1800s, in which the ability to trace citations between court documents to an original precedent is crucial [138]. The first scientific citation index, however, is the manually-curated index curated by Paul Gross & E.M. Gross in 1927 [139], in which it was argued that librarians could use the number of citations received by Chemistry journals as a metric to make stocking decisions. These early developments would eventually serve as inspiration for Eugene Garfield, who in 1955 published a famous article in *Science* [140], in which he argued for the creation of a universal citation index of scientific publications; such an index, Garfield argued, would allow researchers to easily explore associations between ideas, and to assess the quality of articles and journals. Garfield himself took up the challenge to create his proposed index, founding the Institute of Scientific Information (ISI) in 1956, and would release the first version of the index

¹The earliest known uses of individual-level publication counts were actually in the 1830s [132], which measured the "contributions towards improving natural knowledge" by counting the number of memoirs published in the *Philosophical Transactions* by the Royal Society of London. However, their use did not become a major topic of research or assessment until later in the 1800s and in the early 1900s.

in 1963 [141]. While this initial release was commercially unsuccessful, the index would soon find a user among sociologists of science such as Stephen & Jonathan Cole, who conceived of citations as indicators of the *quality* of a scientist's work [142]. Not long after, the term "*bibliometrics*" would be coined to this powerful new way of studying quality in science [143].

The next significant development by Eugene Garfield and the ISI would be the creation of the Journal Citation Index in the 1970s, alongside the formulation of the Journal Impact Factor (JIF) [144, 145]. The JIF measured the average number of citations that papers published in a given journal receive within a fixed time window. Originally marketed as a tool for librarians to select journals to purchase, the indicator would also find itself as a means for journals to flaunt their prestige, and a tool for researchers to compare journals and individuals [146, 147].

While Eugene Garfield was laying out the infrastructure that would soon underpin the contemporary Science of Science, another scholar, Derek de Solla Price, was detailing a conceptual foundation. Specifically, Price published his famous *Big Science, Little Science*, in which he used bibliometric and other data to plot the historical growth of science, and to speculate on its future trajectory [2]. This work would inspire the new formation of *Scientometrics*, a field concerned with the measurement of science [148].

The new field of scientometrics would advance hand-in-hand with increases in the size and accessibility of bibliographic data. The SCI, eventually purchased and renamed the *Web of Science*, would soon be just one among many databases, with competitors such as *Scopus*, *SciElo*, *Microsoft Academic Graph*, and *Dimensions*, to name just a few. Scientometricians wold use these data to devise new indicators of quality, such as the H-index [149], and applying these databases towards topics ranging from quantifying the career paths of scientists [120] to conducting global demographic assessments [28]. The same data would also find a home in institutional evaluations, where they would be used to assess the performance of scientists [150], and rank the relative research impact of universities [151]. Others have sought to expand the scope of scientometrics, scraping data

from social media and news organizations to construct indicators of social impact referred to as *Altmetrics* [152], and constructing indicators of service using recent tools like Publons [94]. Still today, the field of scientometrics remains a prominent field that approaches the study of science using quantitative methods and often drawing on massive bibliographic databases.

The Science of Science emerged more recently as a field, most clearly following the publication of an article in the journal *Science* [61], and later a book [153] of the same name. Drawing a line between Scientometrics and the Science of Science can prove difficult, each similar in their topic and approach. The most clear demarcation is perhaps historical: whereas Scientometrics has existed for decades, with specialized practitioners and a dedicated journal, the Science of Science is instead composed of scholars from many different disciplines, most often Physics and Computer Science, pulled together under a shared interest for studying the structure and dynamics of science, and a shared appreciation for big data, mathematical modelling, and sophisticated computational techniques. Moreover, Scientometrics has long concerned itself with the evaluation of individuals and institutions in science (or criticism of such evaluation), whereas the Science of Science is not attached to such a use case. While I locate this dissertation in the Science of Science, it is really a part of both traditions, drawing heavily on the foundations of Scientometrics, while also integrating the advancements and directions from this newer field.

Complexity science

The history of Complexity Science is made up of much more complicated and loosely-related movements. While there is a common topic and some core methodological and conceptual tools at the heart of the field, it remains a composition of many different disciplines contributing their own perspectives. Rather than tracing the histories of each of the major fields involved in Complexity Science, I instead choose to focus on only a few historical contributions that are important for contextualizing this dissertation.

Perhaps one of the most foundational observations for complexity science was the discovery of *chaos*—about how minute changes in the system can entail drastic consequences. The notion was first formulated by Henry Poincaré, who in attempting to solve the so-called *many-body problem* in 1887, observed that a system of even only *three* bodies, such as orbiting planets, tiny errors in the measurement of their initial positions would result in completely different outcomes; because every measurement is associated with error, Poincaré argued that it was impossible to predict the behavior of such systems [154]. Nearly a century later, a meteorologist named Edward Lorenz, would stumble on this same idea. Lorenz observed that interrupting a weather simulation, and re-starting it would result in entirely different patterns. Investigation would reveal that key values in the program were rounded to the third decimal point between stopping and re-running—an seemingly minor different, yet enough to significantly alter the behavior of the simulation [155]. Other scientists would expand upon Lorenz’s discovery, kick-starting the new field of *chaos science* [156]. Benoit Mandelbrot would uncover the fractal structure of chaotic systems [157], and Mitchell Feigenbaum would even find universal characteristics in their behavior [158]. The field would even obtain a degree of fame thanks to the popular science book *Chaos: Making a New Science* by James Gleick [159], which brought chaos theory into mainstream attention. So much was its impact, that the 1990 book and 1993 movie *Jurassic Park* would draw heavily on the themes of chaos, even featuring a fictional chaos theorist played by Jeff Goldblum ². Even today, chaos remains an important topic of study in mathematics. However, one of its lasting contributions is serving as a crucial element of the new field of complexity science.

Network Science also sits at the core of contemporary Complexity Science, in which networks are used to represent and make sense of many different kinds of complex systems. The origins of Network Science lie in the history of graph theory, a branch of mathematics concerned with graphs: representations containing pairwise relationships between discrete objects. One of the earliest ap-

²Chaos theory even proved a pivotal idea in Liu Cixin’s popular 2008 book *The Three Body Problem*, in which mysterious events, and the shape of a civilization, emerges out of the difficulty and near impossibility of modelling and predicting the behavior of a tri-solar star system.

pearances of graph theory was in 1741, when Leonhard Euler attempted to solve the "Seven bridges of Köningsburg" problem [160], which asked whether it was possible to walk through the city of Köningsburg crossing each of seven bridges once and only once (no). Centuries later, the mathematicians Paul Erdős and Alfréd Rényi expanded on Euler's ideas, introducing models of *dynamic* networks that were essential for studying many real-world phenomenon [161]. The concept of networks would also find success outside of mathematics, and in the annals of social science. Stanley Milgram, for example, was a Social Psychologist who devised an experiment to investigate the structure of human social relationships [162], finding that a letter given to some person, addressed to some person they didn't know thousands of miles away, could eventually find its way to its destination by being passed along social networks—family, friends and acquaintances³. Milgram discovered that social networks have "small-world" properties, a finding later developed and observed in other networks by the Sociologist Duncan Watts [163]. That this property was so common across so many networks intrigued others to delve deeper into what could be learned about their structure. Albert-László Barabási & Réka Albert discovered another common property of networks, that they were *scale free*, meaning they were made up only a few highly-connected "hub" nodes, and many less well-connected "leaf" nodes, the degree distribution of which could be represented using a power law [119]. Due to their high interconnectivity, complex systems have served as a ripe application for networks. So central are networks to complexity, in fact, that feature prominently in the definition of a complex system [103]. Like chaos, network science has its own history, and future, independent of complexity; however, their importance to complexity necessitates that their history is appreciated, and their contributions to complexity science acknowledged.

The field of Complexity Science itself is unique in that its origins rest not in a single idea or discovery, but rather an organization—the Santa Fe Institute. The institute was founded by a

³Forced to rely on only people the participants already knew, they would pass the letter to a colleague who thought would be closest. For example, a participant in Kansas, knowing the destination is in Boston, might pass the letter to their local Preacher who attended Seminary in Massachusetts; the preacher might then send the letter to their old teacher in Boston, who might send it to a colleague, who might know the sibling of the destination person, who could then complete the chain.

group of researchers, mostly physicists, from Los Alamos National Laboratory in New Mexico, in the United States. Among its founders were several prominent scientists such as Murray Gell-Man and George Cowan [164], whose pedigree likely helped the institute's notoriety. The goal of the institute was to foster a field that would study the chaotic, non-deterministic, non-linear, and high-dimensional *complex systems* that were encountered in more and more fields. Aside from physics, other contributions came from the likes of Claude Shannon and his pioneering work into information and entropy [165], Fredreich Hayek's ideas on local information and distributed economies [166], along with new methodological tools such as agent-based modelling [167], and the developments of fields like Network Science. The institute continues act as a nexus for Complexity Science today, maintaining its own researchers, hosting summer programs for students, and bringing together their many externally-affiliated faculty. Yet the field has also moved beyond the institute, now operating their own research centers and departments around the world, and continuing to develop the theory and methods necessary to make sense of complex systems.

1.4 Structure of dissertation

This dissertation concerns the synthesis of both Science of Science and Complexity Science, specifically the adoption of a *complexity perspective* of science. This combination is by no means new or unique [168], but a thorough treatment of the two, using complexity as a lens to understand new and past literature in the Science of Science, is still a useful contribution. My aim for this work is that it lays a foundation that future scholars can build upon, further exploring how viewing science as a complex system can aid its understanding.

To support this argument, I first clarify what I mean by the *complexity perspective*, by providing a more explicit definition about what it is, and it conceives of science, in the second chapter of this dissertation. I also detail several of the concepts from Complexity Science that are relevant in this dissertation. Afterwards, I map existing concepts of the Science of Science, such as cumulative

advantage, homophily, and social capital, onto the complexity perspective. This chapter closes with a brief discussion of the value of Complexity Science for understanding science and interpreting the findings of the four studies.

After defining of the complexity perspective, I present four distinct studies split across the following four chapters. Each study includes a foreword briefly explaining the context and significance of each paper, as well as the details of its authorship, funding, and publication. After the foreword, each study includes its own self-contained introduction, methodology, results, and discussion sections (or variants thereof) that can be read independently of the narrative of the overall dissertation. The first of these studies (chapter 3) considers the incidence and extent of bias in peer review at *eLife*, a prestigious life science journal that implements a unique form of peer review. The next study (chapter 4) examines the factors contributing to student's ratings of the teaching of tenure- and tenure-track faculty at universities in the United States, finding evidence of bias that might hinder the careers of scientists. The third study (chapter 5) quantifies the extent of disagreement across millions of scientific articles using a novel cue-word based measure, observing heterogeneity across fields based on their characteristics and organization. Finally, the fourth study (chapter 6) applies a novel neural embedding technique *word2vec* to the study of scientific mobility, creating a dense and meaningful representation of global mobility, which I use to disentangle many of the factors contributing to scientist's movements around the world. Each study has either been published, or is under review or revision, and as such I maintain the format and style of their respective venues. Individually, every one of these studies constitutes a significant contribution to the Science of Science, yet when viewed together with the *complexity perspective*, also reveal insights into the structure and dynamics of science.

In the seventh chapter, I present a cohesive discussion section incorporating the findings of all four studies. Specifically, I interpret the results of these studies with the complexity perspective, attempting to explain them in terms of concepts from complexity science, and mapping them

onto broader potential explanations about the inner workings of scientific research. Specifically, I disentangle many of the bottom-up forces that give rise to the findings in each of the studies. I also consider the mechanism of *feedback* in science, and peer review, metrics, disagreement, mobility, and more are potential feedback mechanisms that give rise to the structure of science, maintaining it over time, while also allowing the possibility for revolutionary systemic change. The complexity perspective can also provide much-needed context for common issues for research in the Science of Science, and what paths forward might be. Finally, I close the discussion by laying out a path for future work, including other concepts from Complexity Science that might be useful for the Science of Science.

I provide a brief conclusion to this dissertation in chapter 8, where I summarize the main ideas of the dissertation, and consider some of their implications. Following this conclusion, I also provide a short postscript, in which I outline some of my own, highly opinionated, feelings about the direction of the Science of Science as informed by my work here, and the implications for the governance of science.

Chapter 2

Science as a Complex System

I never knew anybody . . . who found life simple. I think a life or a time looks simple when you leave out the details.

Ursula Le Guin

At a high level, the *complexity perspective* is about viewing science as a complex system, its organization and behavior emerging primarily out of bottom-up forces affecting individuals and their interactions. In this chapter, I provide necessary context for understanding what is a complex system, and define the particular framework that I use to conceptualize science. I outline aspects of complex systems that are important for this dissertation, especially *self-organization*, a concept which I draw on heavily in later chapters. Following this, I provide a non-exhaustive listing of individual characteristics and individual-level forces that are relevant for understanding the structure of behavior of science. This perspective gives us a unique way of viewing science, and opens the door to new interpretations and novel insights.

2.1 What is a complex system?

What exactly is a complex system? This turn out to be a complex question. The field that studies such systems—Complexity Science—originates in Physics yet spans disciplines as diverse as Computer Science, Biology, Anthropology, Design, and more. This diversity of fields provides many angles for understanding complex systems, yet their different perspective also introduce confusion. For example, there is no widely-agreed upon definition of a complex system. Melanie Mitchell, in her 2009 book *Complexity, a Guided Tour* [103], attempts to bridge the various communities with a single consensus definition of a complex system in terms of their common characteristics, stating

that,

"A complex system is one in which **large networks** of components with **no central control** and **simple rules of operation** give rise to complex collective behavior, sophisticated information processing, and adaptation via learning or evolution" (emphasis mine)

While not a formal or clear mathematical definition, this definition does lay out many of the most relevant aspects of complex systems. For one, the structure of complex systems can almost always be conceived of as a *large network* of individual elements (nodes), a representation that highlights the linkages and interactions (edges) between entities. Their structure is also usually *decentralized*, and their behavior lacking any sort of central control. Rather, the behavior of the system *emerges* out of simple rules that operate at the level of individuals. This quality of *emergent* behavior is a critical facet of complex systems, and even features as a defining characteristic of complex systems in Mitchell's second, shorter definition:

"A system that exhibits nontrivial **emergent** and **self-organizing** behaviors".

This notion of *emergence* is at the core of the field of Complexity Science. Emergence refers to novel property of a system that results from the interactions of its individual elements, and which is not necessarily predictable or expected given those simple, lower-level rules [102]. That is, complex systems are inherently *non-linear*, such that "the whole is not the sum of its parts", and that complex behaviors can result from simple rules¹.

Self-organization is similar to emergence, yet focuses on how the *structure* of a system emerges from simple rules affecting its individual elements. Complex systems can manifest tremendously complicated structures and patterns, without any centralized control or direction, solely from

¹There are also different categories of emergence based on what kind of behavior or property a system exhibits. *Weak emergence* is used to refer to cases of emergence in which the path by which individual elements produce the emergent property is intelligible, even if complication. For example, a flock of birds behaves very differently than a single bird; however, the behavior of the flock can still, in principle, traced back to that of the individual birds. In contrast, *strong emergence* refers to the case where this connection is unintelligible, such that there is no clear path how or why a property of the system should emerge from the constituent elements. Nearly every case of emergence is weak emergence, with strong emergence saved to refer to certain physical states, like superfluidity and superconductivity (though with controversy), and the most famous of all: the emergence of consciousness (or rather, mental states, subjective experience, or *qualia*) from non-conscious particles and neurons. The possibility and existence of strong emergence remains a topic of debate in Physics and Philosophy. More discussion of this can be found in David Chalmer's essay *Strong and weak emergence* [169].

the decentralized actions of their individual elements. These structures are often hierarchically-organized [102, 170] or follow fractal patterns [157].

Another way of defining complex systems is not through their qualities, but instead by the kinds of problems they present to scientific research. The physicist Warren Weaver, in 1948, wrote a very prescient piece about the three basic types of scientific problems [171]. The first were problems of *simplicity*, meaning those problems that consist of only a few variables, and which could be studied, and perhaps even solved through relatively simple mathematical techniques. The second were problems of *disorganized complexity*, which consisted of upwards of millions, or even billions of variables, but which did not interact; by aggregating and modelling these *linear* variables, these kinds of problems could be approached using statistical techniques. The third type of problem were those of *organized complexity*, those that consisted of at least hundreds or thousands of variables, but in which these variables were mutually-dependent, interacting in non-linear and complex ways, and which none of the then-established techniques could hope to understand. Problems of organized complexity are those that characterize the study of Complex Systems, and though new methodological techniques and equipment have put them more within reach, addressing these problems remains one of the biggest challenges of contemporary scientific research. Weaver himself believed that science should re-make itself in order to approach problems of organized complexity, requiring that science shift from a system of individuals working alone to teams of researchers working together [50], and towards increased specialization [47]. Needless to say, Weaver's predictions came to pass, a testament to his insight. Problems of organized complexity have proliferated across science, presenting unique challenges that Complex Science aim to understand.

Perhaps the most informative means of defining complex systems is not through their usual characteristics, but rather through illustrative real-world examples. The diversity of disciplines that contribute to Complexity Science reflects the ubiquity of complex systems in almost every

domain of science. One of the most common examples of a complex system is a flock of birds: in spite of a lack of centralized control or communication between birds, simple rules governing each individual's behavior give rise to complex patterns and behaviors [172]. Entire ecosystems, too, are examples of complex systems. The interactions between plant and animal species—through eating, breeding with, fighting, or socializing with one another—create complicated yet resilient structures that evolve alongside and adapt to changing environments [105, 173]. Other complex systems span the entire planet. The Earth's climate, for example, is the product of immeasurable chaotic atomic-scale interactions between molecules in the atmosphere, landscape, and oceans, which eventually give rise to complex weather patterns, transcontinental currents, and micro-climates [174]. Other complex systems are much smaller, but still tremendously-complex. The human brain, for example, is comprised of billions of neurons that, through their interactions, form the basis for cognition, memory, perception, problem solving, and more [104]. And human society too, the product of billions of brains and bodies, is a complex system. Social interactions give rise to entire communities, cultures, and nations.

Science is also a complex system. It consists of millions of scientists, interconnected through collaborations, citations, community membership, and more, and their interactions give rise to collaborations, disciplinary communities, and new knowledge. The Science of Science also faces the same difficulties as Weaver's *problems of organized complexity*, in that science comprises many millions of variables that all interact with one another in a host of complicated ways. Past research has also recognized science as a complex system [107], with comparable size and complexity to other systems. Recognizing the complexity of science is the first step towards defining and applying the complexity perspective.

2.2 Science as a complex system

What does it mean, precisely, to call science a complex system? Although I showed that science fits with their common definitions and qualities, the comparison remains vague and poorly defined, with many open questions. Who is included in my definition of *science*, how do they interact, and what are their relevant characteristics? In this section, I provide necessary precision to my definition of the complexity perspective of science.

The first step is to establish the scale, or the level of analysis that I consider when talking about science. Scientists themselves are massively-complex organisms, with bodies containing multitudes of cells, bacteria, and their behavior and emotions are dictated by many molecular and neurological processes which are themselves comprised of the interactions of atoms and subatomic particles. Like any other complex system, Science is just atoms all the way down! Atoms, however, are not the most sensible level of analysis to use when talking about a system made of people, and at any scale lower than an individual human, the detail and complexity is simply too great to be practical. Yet neither should the scale be made too broad. Science is made up of larger organizational units, ranging in scale from research groups up to institutions, cities, countries, and international organizations. This scale obfuscates important details, aggregate interpersonal relationships and interactions to broad units which, while useful for some studies, is not appropriate for my analyses. Here, when I speak of science, I aim to understand it in terms of individual *human* actors².

Just as the scale needs to be understood, so too do the *boundaries* of the system. Science is one component of the massive and massively-complex social network that is Humanity, to which science is deeply interwoven. All people, scientists are not, are connected to the scientific system through its organization or its products. For example, a government legislator might enact policies with the aim of supporting or directing scientific research, and an university administrator may work to maintain

²I make the distinction of humans here, as complex systems can be composed of a mix of different types of actors, some of them non-human. For example, in Bruno Latour's Actor-Network Theory [175], which is distinct from, though conceivable as describing a complex network—is composed of both human and non-human actors, such as research instruments and natural objects.

and operate the facilities and institutions that scientists work within. Without even knowing it, others may be themselves deeply connected to science: a glassblower, for example, may craft glass containers and vessels used to store chemicals in a lab somewhere, whereas a factory workers may craft the screws and bolts holding together million-dollar scientific instruments located a world away (literally, in the case of extra-terrestrial probes). We are all also consumers of objects and processes produced by science: I am typing this dissertation on an astonishingly-complex machine built atop the foundations of Mathematics and Engineering developed by Ada Lovelace, Alan Turing, and John Von Neumann, the same general-purpose machine that allows a toddler to watch *Paw Patrol* on an iPad and a cashier to record an automated purchase. Scientific knowledge is everywhere, and touches everything. Even the line between *scientist* and *non-scientist* is blurred by distinctions between academic vs. corporate research, and by professional vs. amateur or citizen scientists, and by STEM fields vs. the humanities. So where exactly should I draw the boundary? For my studies, I consider only *scientists*, whom I broadly define as individuals professionally-employed for the goal of contributing to public knowledge³. By public, I mean actively publishing work in venues that, at least in an ideal world, are openly accessible by society at large such as academic journals and conference proceedings. This definition mostly includes university faculty and those working at public or private research institutes, as well as scientists actively publishing their work from industry⁴. I also adopt a broad definition of knowledge production, including the fields of Science and Engineering, as well as the Humanities under the umbrella term "scientist", in spite of the many known differences between their institutions and cultures [176]. Non-scientist actors who still influence science, such as administrators, legislators, technicians, and others are excluded from my definition of the scientific system, and I consider their actions and influence on the system

³While the choice to limit to producers of public knowledge has theoretical benefits—limiting to those who work with others in a wider academic community—it is also a practical choice; research that is done in private, such as in secretive industry and government labs, can not be observed in bibliometric data, or other data on scholarly activity, which forms the basis of my work.

⁴Researchers at Bell Labs, and more recently Microsoft Research, are examples of Industry scientists who would be included due to their intense research and publication activity.

to be an *external force*. Science, as I consider it, is made up of scientists⁵.

These individual scientists are not, however, faceless or nameless robots programmed to churn out knowledge. Rather, each person in scientist is a unique individual, with their own unique mix of knowledge, skills, experiences, preferences, demographics, and circumstances that shape how they interact the world, with one another, and how they approach their work as a scientist. Needless to say, even an individual scientist is complex, so this vast space of possible characteristics needs to be simplified to a more manageable set of variables. Here, I introduce a few of the individual characteristics of scientists that are relevant to the work in this dissertation (to be discussed in more detail later). The first are the individual's demographics, such as their gender, race, and nationality⁶, which all play an important role in structuring their social experiences. Then, there are the individual's capital, which is a term I use to refer to all of the resources available to a researcher and which they can leverage to do scientific work, and to improve their status in their scientific community; this includes their funding, equipment, and facilities (material capital), their technical knowledge and skills (human capital), their social relationships (social capital), their ability to navigate the cultural workings of their field (cultural capital), the prestige derived from their credentials and affiliations (symbolic capital), and their general social status within the scientific profession (reputation).

So how do individuals, with their individual characteristics, actually interact with one another in science? As with the scales and boundaries of science, the simplicity or complexity of interactions in science need to be demarcated. In an overly-complex case, two scientists could be thought of as "interacting" even if they never spoke, or perhaps even modern scientists *interact* with Marie Curie every time they draw on her theories of radioactivity. At the other extreme, we might consider two scientists to be interacting *only* when they are enter a formalized relationship, such as mentor-

⁵In this dissertation, I use the terms "scientist" interchangeably with "researcher" and "scholar"

⁶Nationality, in practice, will be operationalized as the country of the author's affiliation, rather than the country of their birth. Due to the ubiquity of mobility in science [177], these are not always the same, especially for high-immigrant countries such as the U.S. [178], yet still can serve as an effective proxy for many factors related to nationality.

mentee or co-authors. Here, I draw a line somewhere in the middle, consider co-authorship, citation, affiliation, and journal publication as indicating important interpersonal interactions between scientists. Affiliation with an organization can similarly indicate interactions with the researcher, and other affiliates, especially for the purposes of hiring and promotion, wherein current professors of the organization will review a scientist's record and make a decision on their employment. Journal publication is similar—first, when a scientist submits to a journal, it indicates a connection between their work, and that of the disciplinary community that the journal represents; second, a group of fellow scholars will serve as the editor and peer reviewers of the submitted article, deciding on its acceptance and necessary changes. Co-authorship is when two scientists appear together on the byline of a published scientific article, indicating that they collaborated and worked on the work. Once a paper is published, it forms an object that others in the community can interact with, and in a way interact with the authors, through citation. By citation, I mean when an author publishes a paper that references a past study, an important way that scientists indicate a connection between their work, and that of others in their community; this introduces some potentially-confounding cases, such as interactions between living and long-dead scientists (such as when a student, in 2021, cites Auguste Comte), but the ubiquity and importance of citations to science merits their inclusion regardless. One commonality of these interactions considered here is that elements of each are visible in bibliometric databases—databases containing metadata of published scientific papers and their authors—or else through other quantitative sources that I had access to for this dissertation work. This necessarily restricts the kind of interactions I can consider; a professional mentorship, for example, may result in no publications or no visible interactions, even though the relationship is an important aspect of both of their careers.

To complete this model of science as a complex system, I'll introduce the idea of a *network*. Networks are vital for understanding complexity, they even appear in Melanie Mitchell's definition: "A complex system is one in which large ***networks*** of components..." [103]. Network models are

incredibly useful for modelling how discrete entities, represented as "nodes", are connected with some number of other entities, where these relationships are called "edges" or "links". These models serve as a fantastic representation of complex and high-dimensional systems, so much so that they are central to our understanding of some industries. For example, networks are the dominant way of representing the connections between friends or followers social media; so central is this definition, that the 2010 movie about the rise of Facebook was called *The Social Network*. Similarly, networks can be used to model the flow of information and viruses through populations, ecological food webs, and metabolic networks within the the human body. Networks are also ubiquitous for conceptualizing and studying the global system of science, whether it be networks of collaborators [179], citations between papers [180], or mobility between places [177]. For this dissertation, I'm primarily interested in thinking about science in terms of individuals (nodes), each associated with their relevant characteristics (demographics, capital, etc.). Then, these nodes are connected with one another if they have interacted with one another (through co-authorship, citations, affiliation, journal, etc.). Since this is primarily a theoretical and metaphorical model, rather than a methodological one, I don't distinguish between different types of interaction links. It's also important to note that the network of scientists, like most other real-world networks, is dynamic, and constantly evolving from moment to moment ad researchers enter or fall out of science, and as they interact with and form links with one another in their day-to-day work. This complex network model of science—of scientists connected through interpersonal interactions—serves as an important vehicle for thinking about and making sense of the structure of global knowledge production.

2.3 Science as a self-organized system

One of the most important properties of complex systems, and specifically of the complex network of science is that it is *self-organizing*, a property expressed in Melanie Mitchell's second definition:

”a system that exhibits nontrivial emergent and self-organizing behaviors” [103]. Self-organization loosely refers to the ability of the system to spontaneously form structures without central control [102]. For example, complex and beautiful snowflake forms out of molecular-level physical forces that cause crystallization, and ants build colonies, search for food, and fight wars based on individual-level actions, distributed agency, and pheromone trails. Within science, there are some aspects of centralized control—politicians and administrators attempt to structure and guide their national scientific systems—but largely the social structures of science organize themselves based on the agency of millions individual scientists, each making their own decisions based on local information.

To draw a more precise description of the self-organizing properties of science, I turn to the definition given by the International Encyclopedia of Social and Behavioural Sciences [181], which outlines four main qualities of self-organized systems: pattern formation, autonomy, robustness and resilience, and dynamics. The complex social network of science reflects all of these characteristics.

Pattern formation is one of the most important aspects of a self-organizing system, and certainly the most interesting (science is interested in the patterns, after all). Within science, spontaneous patterns of social organization are ubiquitous, and occur at nearly all levels of analysis. At the lowest level, individual researchers interested in the same topic may collaborate and exchange ideas and materials, forming an ”invisible college” [71]. Given enough time, attention, and success, these invisible colleges may grow into entire disciplines, which are themselves a kind of pattern of social organization, arranged and bounded based on the communities formed around the study of specific topics, and the interactions (or lack thereof) between these communities [182, 183]. Prestige hierarchies among universities [184] and journals [185] naturally emerge based on the competition for prestigious employment and seemingly-random choices of individuals, hiring committees, and simple serendipity [170, 186].

Autonomy refers to the *self* in self-organized systems, namely that the system is organized

based on individual-level actions distributed across all the actors, and lacking any one centralized control. While centralization exists in science, it is largely a de-centralized endeavor, in which the actions of individuals choose research topics and collaborate with the aim of improving their professional reputation [40] or in response to economic incentives [19], though there are of course differences in the individual autonomy, and the breadth of individuals' choices, between fields [70].

Self-organized systems are also *robust*, or resistant to change, and *resilient*, or able to recover from shocks. Science is certainly robust—in spite of the constant changes of the world both within and around science, disciplines and institutions remain stable over decades, or in some cases even centuries⁷. In terms of epistemology, scientific communities are also conservative, reticent to change their current theoretical paradigm except in rare instances of "revolution" [187]; yet when such revolutions do occur, science also demonstrates itself to be resilient, and quickly re-organizes itself around the new paradigm. Through reflexivity [188], science is able, to an extent, to detect and remedy mistakes and errors, perhaps illustrated best with the recent reckoning in many scientific fields over statistical significance and spurious results [189, 190], a process that leads to consensus, and ideally, valid knowledge [191, 192]. Similarly, scientific communities are exposed to more minor shocks all the time—their members retire or die, taking their knowledge and perspective with them; still, after an initial loss [193], the community keeps going, sometimes even evolving after the death of influential (perhaps too much so) individuals [81]. Science, in some form, has existed for centuries; while most aspects of science have changed, and certainly grown in scale, the system as a whole has persisted, a testament to its robustness and resilience.

Finally, self-organized systems are *dynamic*, meaning that the system is constantly changing over time. Indeed, the system of science is intensely dynamic, changing from day-to-day, or even moment to moment, as researchers go about their work and interact with one another. A scientists might run into a colleague from across campus, and invite them for coffee; while talking together, they

⁷Oxford University, for example, has a history bordering on a millennium, and scientific Astronomy has a history going back hundreds of years

discover a shared research interest and an interesting idea, which they promise to discuss further; one of the scientists might put their doctoral student to work on the idea, eventually sparking a new collaboration, publication, and potentially revolutionizing a field. Scientists carry their knowledge and experiences with them, spreading them like pollen through their social environments, helping to diffuse ideas across communities [116, 194, 195], and with a little luck and a serendipitous encounters with the right person [196–198], can snowball into breakthroughs that power entire scientific disciplines and industries. In parallel to the continuous individual-level churn of people and ideas, science has also grown substantially over the past century, starting with the era of "Big science" [2], and continuing now as new nations develop their scientific capacity, contributing millions of scholars to the global community [9, 16, 199], all the while growing more specialized [47], team-oriented [50], and likely even more unequal [200–202]

Science is a self-organized system, where the complex structures of collaborations, labs, institutions, and disciplines all emerge from the actions of individuals. The entire system continues to evolve over time, in response to internal feedback and fluctuations, as well as in response to outside influence and shocks. From individual scientists, comes science.

2.4 Individual-level characteristics

In some complex systems, individuals can effectively be thought of as identical: particles swirling in the atmosphere may, for many purposes, behave similarly if they were hydrogen or oxygen, just as the individual personalities of birds don't need to be understood when modelling their flocking behavior. In science however, treating every individual as identical obfuscates a great deal, and too much, of the detail and behavior of the system. Each scientist has a unique mix of demographics, experiences, personality, aspirations, expectations, and relations that determine how they enter science, and how they behave once there. Here, I draw a distinction between two types of individual-level traits, which I term *demographic* characteristics, and *capital*.

Demographics

I define demographic traits as those that a person would have whether or not they became a researcher. A women, for example, is a woman whether she becomes a researcher, a deep-sea welder, or an acupuncturist. Here, I consider only three major demographic characteristics: gender, race, and nationality.

One's gender has deep implications for how they enter and navigate science. Historically, science has been inaccessible to women, with figures like Ada Lovelace and Marie Curie were the exceptions, and faced many challenges as a result of their gender⁸. Discrimination, prejudice, and the gendered-structure of society itself precluded women from participating. Still today, women comprise only about 30 percent of scientists worldwide [28], and an even lower figure in fields such as Data Science and Artificial Intelligence [203]. Now, in many fields, more women are becoming scientists, yet few make up a smaller portion of tenure and tenure-track faculty positions [204], and when they get jobs, they suffer a larger prestige penalty than their male counterparts [184]. Progress has been made, but true gender parity in science is likely decades away [205], and is likely to be worsened by the uneven impacts of the COVID-19 pandemic on women scientists [206, 207], which at the time of writing is still ongoing. A mix of structural discrimination and bias fuel these disparities. For example, gendered-stereotypes can prevent women from entering science altogether [34], and once they become scientists, women may be relegated to particular work roles [208], face biases lowering their odds of receiving a grant [35, 209, 210], and getting their work accepted into top-tier journals [211]. Women's work is also held to a higher standard [212] and suffer the consequences of unprofessional peer review [37], and even during presentations, their work is more intensely questioned [213] than their male counterparts. The contributions of women in science have also had a long history of being ignored all-together, or even attributed to men; so common

⁸One impressive figure, Mary Somerville, made enormous contributions to mathematics in the 19th century, but was only able to pursue her career following the untimely death of her husband, whose prejudice had precluded her from doing her work

is this phenomenon that the historian Mary Rossiter gave it its own name, the "Matilda Effect" of science [214]⁹. Women scientists also have more constraints on their time than men, such as disproportionate service burdens [215, 216] and the pressures of parenthood [115, 217], which limit their ability to do scientific work. In so many ways, gender shapes the experiences of scientists, and is a vitally-important characteristics when considering the structure and dynamics of science.

Race, too, has deep consequences on the career of scientists¹⁰. Just as women have historically been excluded from science [23], so too have White individuals held a monopoly over the scientific enterprise [24]. In the U.S., Black and Hispanic people have been under-represented in science, especially in traditionally-STEM fields [218–221]. These groups face many barriers in science, including a lower likelihood of being hired as postdocs [36], to more likely have their grant applications rejected [222], and to receive lower ratings in student-teacher evaluations [223], all important components of hiring and promotion. Black scholars are also more likely to research areas where there is less available funding [113]. This is to say nothing of the day-to-day struggle that scientists in these minoritized groups may face during their work [224]. Race is an important subject in the U.S. in 2021, and changes are occurring in the Society, as in other areas of society; while there is hope for change, one's race remains an important factor in shaping scientists' experiences and careers.

A researcher's nationality—the place in the world they are born, with all if its entailing political and cultural consequences—has massive consequences on the shape of their career. Major and prestigious scientific institutions, such as elite universities, have historically and remained concentrated in Europe and North America, only recently developing in more peripheral nations [15, 16, 225], such as the massive growth of Chinese science in recent decades [9, 199, 226]. The resources to conduct science are also unevenly distributed, with some countries having higher scientific capacity

⁹In a particularly-relevant example for the field of Scientometrics, Rossiter notes in her article that Robert Merton's famous paper on the Matthew Effect in science was inspired by the ideas of his wife, Harriet Zuckerman, who was left un-credited on the paper.

¹⁰Being an American, here I primarily consider Race in the context of the U.S. in 2021, the politics and cultural meaning of which would differ dramatically in another time or place.

than others [227]. Researchers in nations with lower scientific capacity will have less access to funding, facilities, and information to conduct high-impact research. For example, scientists working in the scientific periphery have less access to pay-walled publications [228], and end up relying on sources such as SciHub to access cutting-edge information [229]. Scientists in these countries may also need to rely on collaborators to access funding and vital equipment, leading to a leader and follower dynamic in international collaboration [82], which inevitably leads to a focus on the research agenda of the richer country [20]. Language is also an important component of where one is born. English remains the *lingua franca* of science [230]. Native-english speakers, and those with near-native proficiency in English are at an advantage when it comes to publishing [231], whereas those without sufficient English proficiency can find it more difficult to get their work published [232, 233]. Even after overcoming barriers to resources and language, research published by a team from a low-income country is valued less than that from more scientifically-advanced ones [234]. Where one is born dictates a lot about one's life, about whether someone can become a scientist at all. Even if they become a scientist, the consequence of their birth will continue to shape their entire career in both obvious, and other more subtle ways.

Capital

If demographics are largely the circumstances of one's birth, then *capital* exists only as a consequence of becoming a scientist, and is mutable through their career. Here, I loosely define capital as the resources—both material and abstract—available to a scientist to conduct scientific work, and to improve their professional status, or their "success" [65]. Another way of thinking about capital or success is as a measure of an individual's ability to act upon and influence the system of science through the creation of new interactions and objects of study; for example, using money (material capital), a researcher can employ students and technicians to conduct science, and using their professional networks (social capital), they can form new collaborations, both of which will end up

shaping the system of science. The concept of capital can be expanded in a myriad of different ways, but here I focus on only four main forms: material capital, human capital, social capital, and "reputation", which I use as an umbrella term for various kinds of cultural and symbolic capital in science.

The first, and most obvious form of capital is what I call *material capital*, which refers to the money and property available to a scientist to use towards their research. Material capital is necessary, because Science costs money [19]. For example, a Molecular Biologist may need to pay for lab space, expensive equipment such as centrifuges and chemical fume hoods, and a range of objects such as beakers, pipettes, and safety goggles. High-energy physics instead requires massive funding, often through governments and large, centralized organizations, to build and maintain tremendously-complex particle accelerators. Even for fields that are seemingly cheap, such as History, scholarship may require paying the costs of travel to attend conferences, to send a student to an archive, as well as paying article processing charges to journals for publication. Money buys the labor and materials of science, and so those who have access to material capital will be more productive, produce more impactful work, and have an out-sized impact on the system of science.

Human capital, in contrast to material, is not measured in dollars or objects, but rather in the knowledge, skills, and experience that a researcher can put to work towards research [235, 236]. There are many kinds of human capital, and many ways of conceptualizing it. For example, an Aerospace Engineer may have extensive knowledge of different materials and their properties for the construction of aircraft wings; they likely also have the forms of fluid dynamics equations memorized, as well as knowledge of the technical details of equipment they use. This encyclopedic-like information is *explicit* knowledge, because it can easily be made explicit, and is (relatively) easy to transfer to others, such as through reading a textbook. However, aerospace engineering is not about memorizing functional forms and the tensile strength of materials to the 6th decimal point, but also about understanding how the properties of materials work together, about using the

industrial equipment and tools to design, test, and construct an aircraft components, and about having an intuitive sense of what the plane and what can go wrong. This difficult-to-transfer knowledge is called *tacit knowledge* [237], the kind that can't be easily encoded or learned from reading a textbook or watching a YouTube video, but instead requires practice, experience, and social learning. Another way to envision human capital is using Scott Page's notion of a *cognitive toolbox*: the mix of perspectives, interpretations, heuristics, and causal models that a person can draw on to do work [51]; in this model, each researcher has access to a set of "tools", which are perspectives, interpretations, heuristics, and causal models that they gain throughout their life, and that shape how they approach problems. A vital aspect of the cognitive toolbox model is that each tool is not equally-useful in every domain: both a *Cirque de Soleil* performer and a Mycologist have powerful and essential cognitive tools, but an acrobat would be as uncomfortable studying fungal spores as the mycologist would be on a trapeze. Therefore, the value of a particular cognitive tool, as with the value of any bit of explicit or tacit knowledge, is dependent on the context where the researcher is working. Knowing the value of a particular skill to a domain is difficult, however, as the skill requirements of science are always changing [47], and the value of a particular skill to a problem cannot always be known ahead of time [51, 238]. While this ambiguity makes quantification difficult, human capital remains a useful cognitive tool for thinking about how training and experience can shape one's success in science.

Whereas material and human capital are embodied in a person, a scientist's *social capital* is the resources available to them from their professional and social network. The popular imaginings of major scientific figures—Newton, Einstein, Curie, and Darwin, to name a few—portray their achievements as entirely their own, absent of help or contributions of others. While this may have once been true, science is now by no means a solitary endeavor. Teams not dominate the production of scientific knowledge [50], and the ability to work in teams, and to leverage social connections, is more important than ever. Social capital is meant to capture the resources that a

researcher might derive from the relationships with their colleagues, acquaintances, peers, mentors, and friends, which are often thought of in terms of their social network. The kinds of social connections can dictate what resources they make available. For example, weak ties [239] can introduce novel information and ideas, and notify an individual of professional opportunities and resources. Strong ties [123], in contrast, can foster improved communications and collaboration, allowing for enhanced productivity and impact. Social capital is not only a product of the size or types of a researchers' links, but also *who* they are connected to. For example, having a prestigious mentor can be necessary to "chaperone" one's work into elite journals, sparking a lifetime of elite access [80], whereas collaborating with highly-productive and important authors can similarly help a researcher improve their status [240], or give them access to more professional opportunities, such as mobility [110]. Social capital can also produce disincentives that directly impact one's career, such as through nepotism; authors submitting to journals, for example, tend to have their work rated more highly when reviewed by someone they know [241–243]. Science is a social enterprise; social capital gives one way of making sense of how a scientist's place in the social network of scientists can shape their success, and their influence over the system.

Finally, *reputation* is perhaps one of the most vital forms of capital in science [40]. I define a scientist's reputation as, loosely, how highly a scientist's peers think of them and their achievements. More specifically, I use "reputation" as an umbrella term to refer to various kinds of capital derived from a scientist's credentials and accomplishments. Symbolic capital is one form of reputation, referring to the prestige derived from affiliation with famous individuals or institutions, and which can benefit one's career. For example, a graduate of an elite university like Harvard can leverage the university's prestige to gain access to more jobs than a graduate of a lower-ranked school [184, 244], as well as an advantage in peer review [245, 246]. Similarly, already-famous researchers are more likely to get their papers accepted [247], through virtue of their recognition. Another kind of capital that falls under reputation is *cultural capital*, which loosely refers to the knowledge and

adherence to the culture of one's scientific field [248]. Cultural capital might correspond to one's economic class: someone born in a poor family may lack knowledge of how to behave and present one's self in professional contexts in science, a profession for which most members come from middle or upper-middle class backgrounds. Cultural capital is also specific to individual fields [236]; for example, in male-dominated fields, masculine-typed behavior may be expected, whereas feminine-typed behavior is punished. Similarly, in some Business and Management schools, regularly wearing a suit is expected, whereas in Computer Science, it will likely be met with some derision. Knowledge of and the ability to adhere to these norms can help a researcher build their reputation in their field. Reputation, or how scientists judge each other, is an essential and perhaps even an organizing principal of science, and so it is vitally-important to consider reputation when thinking about science as a complex system.

Capital is a powerful idea for thinking about science and the ability of certain scientists to influence their surroundings. Unlike demographic characteristics, however, a scientist's capital is abstract, and cannot easily be quantified or categorized. Thus, while I spend a great deal of time directly studying demographics in this dissertation, I cannot directly and empirically comment on these other forms of capital. Yet, capital is a powerful theoretical tool for making sense of scientific outcomes and how science is organized. Like the notion of complex systems itself, I use *capital* as a theoretical and metaphorical construct to understand and interpret my research findings.

2.5 Organizing principles of science

Individuals, interacting with one another, structure the systems of science, but what other principles govern the system? Why do two people work together? Why does one science cite another? How does a student just starting out in science end up winning a Nobel prize? Understanding *why* scientists interact the way they do is essential for understanding the complex system of science. Competition, whether for employment opportunities or resources, is at the root of many governing

principles in science. For example, there are many more PhD graduates than open academic positions [249], and the scarcity of funding leads to many hours spent preparing grant applications that could be spent on other work [250], and a selection against novel research ideas [251]. Yet, competition, while important, is not the only principles that drives interactions in the complex system of science. A range of other forces and principles, all interacting with one another, act as a sort of "invisible hand" that guides the structure of science and the production of knowledge. here, I list a few of these principles, though the list is certainly not non-exhaustive, that help to understand how and why scientists work the way they do.

Reputation

Scientists are distinct from other professionals, such as Lawyers and Doctors, in that the status of a scientist is not conferred by clients or external stakeholders (though they still might play a role), but rather through the judgements of other members of their profession. Reputation (discussed previously as a form of capital) is the sum of judgements made about a scientist by their peers, and can be thought of as the *coin of the realm* in Science [40]. A high reputation can confer special access to opportunities, resources, and professional attention Those with high reputations also become scientific leaders, serving as gatekeepers (editors, reviewers, hiring & promotion committee members), and dictating the direction of knowledge in their field. The quest for reputation, then, is perhaps the most central aim of a scientist's career.

Whitley, in discussing the organization of the sciences, defined science as a unique *reputational* system [40]. By this, Whitley means that scientists' self-organize through their search for reputation, and that the different structure of disciplinary communities can be understood through the paths to reputation in those fields. The diversity of stakeholders in a field, for example, determines how and where a scientist can bolster their reputation. In a field like Management Studies, there are many potential venues to publish one's research, and there are also many different stakeholders

interested in the work, ranging from sociologists to CEOs; lacking centralized gatekeepers, the field is decentralized and only loosely hierarchical, lacking singular theories or agreed-upon methodology¹¹. In contrast, a field like Economics is intensely centralized and hierarchical, with broad consensus on methods and approaches stemming from the careful disciplinary-control over credentialing and the more narrow set of stakeholders, all of which shape what kind of research an economist can do to become successful¹².

In centralized and hierarchical fields, such as Economics, competition for reputation can be intense and dictate a scientist's career and research choices. However, a new economist can barely hope to compete with established and famous economists. Indeed, the grip of an eminent researcher over their field can be strong, made evident by Max Planck's principle that "science progresses one funeral at a time", which was later empirically-validated [81]. To avoid direct competition with these imposing incumbents, new scientists will instead opt to circumvent them by choosing to specialize, seeking out topics in and specializing in sub-fields where they can themselves become eminent [40]. For example, a student trained in Economics might specialize in Environmental Economics, Health Economics, Digital Economy, or some other specialty where they hope for less competition. This may explain why, during the continued growth of science [2], that researchers have become more and more specialized [47], each generation competing with an ever-larger population of incumbents. Similarly, this may explain why marginalized groups pursue more niche [113] or novel [112] research topics, as they aim to escape areas of study dominated by prejudicial figures.

Cumulative advantage

Science is highly unequal, and is becoming more so over time [200]. Scale-free, or *power law*, patterns in citation counts and scientific networks [119] speak to the existence of so-called *superstars*,

¹¹Management Studies is what Whitley terms a "Fragmented Adhocracy", which are fields that are not very centralized, and which produce diffuse knowledge on common-sense objects of study, such that the boundaries of the field are permeable. Other examples include political and literary studies.

¹²Economics is what Whitley would call a Partitioned Bureaucracy, which are rule-governed and hierarchical fields, where there is clear delineation between "applied" and "theoretical" areas of study.

with success far and away above that of the broad population of less-acclaimed, more typical individuals. Science is ostensibly a meritocracy in which ideas, and their progenitors, are evaluated and rewarded on their own merits; how then, can one scientists have 1000s of times more citations than another? Are they really 1000s of times better (some may argue yes, but I'll settle on "no")? Evaluating the performance of scientists and their ideas is difficult; oftentimes, even ideas take years to become relevant or to be rediscovered [88], and sensibly and fairly allocating credit to the main contributor of a scientific paper can similarly be a challenge [252], especially as teamwork becomes more central to science [50]¹³. When evaluation is hard, as in science, success stems not necessary from performance, but from other people's *evaluation* of their performance [65]; Namely, success comes from other people. This means that scholars are likely to pay more attention to, and judge more positively, those scientists who already have strong reputations, and who are already successful. This process has many names. Observing the ability of famous scientists to become ever-more successful, Robert Merton coined the term "Matthew Effect" to describe it [79], drawn from the biblical Gospel of Matthew, and which is best summarized as "the rich get richer and the poor get poorer". The effect is also more generally referred to as *cumulative advantage* [201]¹⁴. Driven by the difficulty of evaluation, and the competitiveness of science [255], cumulative advantage are ubiquitous, and can be observed for paper citations [256, 257], journal impact factors [258], research funding [259], institutional prestige [260], national scientific performance [261, 262], mentorship [122], and career success [78]. In each case, these cumulative advantages fuel inequalities between scientists, diverting resources to the already-successful.

Cumulative advantage illuminates two crucial, yet seemingly competing aspect of science as a

¹³Another issue in credit allocation is that references to ideas tend to become invisible over time, following a process termed "obliteration by incorporation" [253], whereby an idea becomes so ingrained in a field that it no longer needs to be referenced. For example, "Newton's Law of Gravity" is so central that it can be written as such, with no explicit reference given.

A similar issue is known as "Stiglar's Law of Eponyms", referring for the tendency of these named ideas, or eponyms, to not actually be attributed to the original discoverer [254]. For example, the Hubble Constant was not actually discovered by Edwin Hubble, but rather the mathematician Georges Lemaître.

¹⁴In network science, the term "preferential attachment" is used, which describes the same effect in the growth of scale free networks [119].

complex system: *chaos*, and *inertia*. Here, chaos means that science has a strong sensitivity upon initial conditions: minor differences in a scientist’s early career can have drastic and far reaching, and perhaps even chaotic consequences on their trajectory. For example, having a famous and successful mentor can allow a young scholar to publish in high-prestige venues, making it easier to access these places in the future [80], just as attending an elite institution can lead to a similarly-elite job [184], and increased access to resources that make them more productive [124]. Over time, these researchers can leverage their early success into more capital, which they then expend to produce more and higher-impact research, and improve their status ever further. In contrast, those who didn’t have a famous mentor or a prestigious affiliation, or who were unlucky in their choice of research topic, may struggle to launch their careers, and may be relegated to low-status positions or pushed out of science entirely. These small differences early in one’s career can send them careening down entirely different career paths.

Conversely, cumulative advantage also gives science inertia: those who succeed will continue to do so, whereas those who don’t succeed will have trouble ever improving their status. This is not to say that scientists cannot become successful later: many scientific papers are ”sleeping beauties” meaning that they becomes famous only many years, upwards of decades after publication [88]. This idea characterizes the life of the biologist Katalin Karikó, who for years struggled to make others value her research, but who after being forced out of the academy, has found widespread notoriety for her work on the mRNA vaccine technology that contributed to the end of the COVID-19 pandemic [263]. However, these stores of late breakout are notable because they are rare exceptions, whereas the norm is that one’s long-term career success is largely a function of their *early* success.

The idea of inertia in science may seem to compete with chaos, but the two are natural outgrowths of cumulative advantage. Small differences can lead to huge differences in success, but the successful will likely continue to be so throughout their career. These two effects can plainly be seen across science. For example, scientific disciplines, as a whole, are conservative, reticent to alter

the dominant paradigm. yet while disciplines tend towards inertia, small anomalies, discoveries, and ideas have the potential to upend the community. Following the accumulation of theoretical anomalies and issues, and maybe after the retirement or death of a few older scholars, disciplines can experience revolution, replacing the old paradigm with a new, and hopefully more powerful one. [187]. Success, and how scientists achieve it, is essential to understanding the dynamics of science, because the most successful individuals with the highest reputation are also the most influential ones, and have an outsized impact on the structure and direction of their field.

Demographic bias

While individual success in science may appear chaotic, it is by no means random. Ideally, science should adhere to Merton's norm of Universalism [264], such that the evaluation of a scientists and their ideas is considered on their merits, and not any other particularistic criteria like gender, race, or nationality. However, empirical studies have shown that science often fails to live up to this idea. Instead, demographic biases in the academic system, whether individual or systematic, have a direct impact on how a scientist is judged by their peers, and their access to capital.

Women, for example, have historically been excluded from science [23]. Those women who do work as scientists have been systematically-disadvantaged through implicit bias, such that others, conscious of unconscious, evaluate women more negatively than men. For example, women tend to receive lower ratings during grant peer review than men [35, 209] and are evaluated less highly in hiring [34, 184], have their work cited less often [265, 266]¹⁵, often have their contributions ignored or attributed to men [214], and their careers are more negatively impacted by the labor requirements of parenthood [115]. Women's work in science also tends to be relegated to lower-status positions. In research groups, women are disproportionately represented in non-leadership roles, doing labor in the lab rather than conceptualizing or leading projects [48, 208], and when women's work is

¹⁵Another interesting citation-based gender disparity is that women are less likely than men to cite themselves [267]. The consequence of this is that women's citation metrics are less inflated than men's which may have negatively impact them during hiring and promotion.

published, it is often to lower-impact venues [211, 268]. Together, these biases and disparities, whether large or small, serve to cumulatively *disadvantage* women scientists [269], burdening their careers [270] or pushing them out of science entirely [68]. As a result, women represent only about 30 percent of scientists in the world, and only about 43 percent of scientists in the U.S. [28], and only about 37 percent of tenure and tenure-track faculty in the U.S. are women [204].

Similar to women, researchers from marginalized racial backgrounds can face biases and prejudice that constrain their careers. Black scholars, for example, have long been excluded from science [24], and those that do work as scientists have less funding than their white counterparts [271], partially as a result of them being pushed towards under-funded fields [113]. Compounding on this, Black & Latin scientists are perceived as less hireable for postdoctoral opportunities [36], non-White scholars are more likely to suffer from unprofessional behavior during journal peer review [37], their published research is cited less often [272], and they are less likely to be recognized in news articles about their work [273]. Together, these biases place a barrier on the career prospects of Black and Latin scientists, and others from marginalized populations, contributing to their severe underrepresentation in science [204].

Pushing out or otherwise excluding and marginalizing populations from science can have far-reaching consequences for human's collective understanding about the world, limiting knowledge to the perspective of the dominant group (White, male, Western). This is because who does science has a direct impact on what kind of science gets done. For example, women are more example to study how clinical effects differ by sex in medical studies [274]; given that women have long been excluded from medicine, this has in turn led to a lack of knowledge about women's response to medical treatments [23]. Similarly, Black and other non-White scientists tend to pursue different, and potentially more novel research topics than their white counterparts [112, 113]. Rich countries also dominate medical research, which naturally focuses attention towards the medical issues of the Western, developed world, rather to the illnesses afflicting much of the developing world [20]; in

turn, scientists in these developing countries collaborating with rich countries might have to adhere to Western research interests, in the hope of maintaining their funding and collaboration [82]. Graduates of elite universities, clustered in rich, developed countries, also have a privileged position to diffuse their ideas to other departmental communities [116]. More subtle cultural dimensions can also find their way into scientific thought, shaping how certain problems are conceptualized and approached. For example, the history of Computer Science has been heavily shaped by the prominence of the game Chess as an indicator of intelligence in the Western world, leading to the prioritization of certain kinds of technical problems and algorithms; a historical trajectory that instead centered Go, a very different game common to East Asia, would have likely required very different choices, and a very different kind of Computer Science [275]. A homogeneous science leads to a narrow understanding of the world; demographic bias, by marginalizing certain populations or forcing them out entirely, restricts who is allowed to do science and thus narrows the production of knowledge.

Homophily

Demographics also shape science in a very different way, dictating not only who becomes successful and a leader of science, but also structuring who works together. Homophily is the idea that people tend to associate with others like themselves, best exemplified in the proverb "birds of a feather flock together". Through homophily, self-organizing systems can spontaneously form segregated structures, simply through the preferences (not even necessarily strong ones) of individual [276, 277]. Scientists, like birds, tend to interact and collaborate with people like themselves, which has deep consequences to the structure of science.

Scientists are likely sort themselves based on their demographic characteristics. For example, men are more likely to collaborate with other men [278], and cite other male researchers [266] than would be expected by chance. Similarly, authors tend to collaborate with individuals of their same

ethnicity [279]. The implications of homophily can shape science in other, more subtle ways. For example, journal editors, who are mostly men, recruit peer reviewers from their, usually gender-homophilous, social networks, leading to male-skewed peer reviews in major journals [111]. However, homophily is not always symmetric. All-male collaborations, for example, are more common in science than all-women ones [280].

Homophily is a power force shaping interactions in science, but its effect can also constrain knowledge production. Homogeneous teams, containing people all from the same demographics, discipline, or background, can struggle to tackle complex problems of science; cognitively-diverse¹⁶ teams, in contrast, are able to draw on their distinct perspectives, heuristics, interpretations, and causal models to generate creative and effective solutions [51]. Perhaps this is why interdisciplinary [281, 282] and ethnically-diverse [279] produce research with higher-impact than their homogeneous counterparts. Homogeneous professional and social networks can also block a researcher off from unique or novel information, entrenching scientists in an echo chamber [239, 283, 284]. Diversity can make science more effective, but it is not without its drawbacks; conflict, communication breakdowns, and adversarial dynamics can stem from the dispatched preferences and expectations of diverse groups [51, 285]. Whether or not its a net benefit, understanding homophily is necessary for understanding how science is organized.

Proximity

Another, often overlooked principle that governs how people interact is their proximity to one another or simply how close they are to one another in geographic terms. For example, a lifelong collaboration or a revolutionary idea may form from the spark of a 5-minute conversation between two researchers in a hallway, who would have never interacted if they hadn't shared

¹⁶In his book, Scott Page draws a distinction between cognitive diversity, and other kinds of diversity. Specifically, he speaks of the benefits (and potential drawbacks) of cognitive diversity, rather than gender or demographic diversity. However, as he notes, demographic diversity often correlates with cognitive diversity, as the distinct backgrounds and experiences of different demographic groups can lead to different ways of thinking about problems.

an office in the same building. In fact, the impact of shared office space was revealed in the case of a science department in a French university; following the discovery of rooftop asbestos, the faculty in the department were moved to other labs, scattered across the campus; following their expulsion, however, it was found that the moved researchers were more likely to collaborate with their new neighbors, even though they might have studied very different topics, and may have never interacted otherwise [108]. In another case, cheaper airline tickets between two cities spurred additional scientific collaboration between them, the lower price effectively increasing the researchers' proximity [286]. These cases demonstrate that proximity can lead to serendipitous interactions, and potentially new and powerful combinations of ideas [198]. The power of serendipity is well known in economics of cities. *Knowledge spillovers* occur when skilled individuals working in close proximity interact, whether formally or informally, and cause a spillover of ideas or knowledge from one to another [287–289]; chance interactions of proximate individuals is a big reason why cities become agglomerative clusters, hosting dense collections of skilled individuals (often within a specific industry) and often building atop the knowledge infrastructure of local universities [290, 291]¹⁷.

Proximity does not necessarily have to refer to *geography*, but instead other more abstract notions. *Disciplinary proximity*, for example, can dictate the probability of interdisciplinary collaboration. Scientists are more likely to interact with someone in their own discipline than another, simply as a matter of being in the same community, attending the same conferences, and sharing the same interests. Interdisciplinary collaboration, in contrast, is partially a function of the proximity between the fields [281]: a mechanical engineer, for example, is more likely to collaborate with a Physicist than a German cultural historian (though I don't discount the latter possibility).

Social proximity also matters, referring to the proximity of a researchers colleagues and peers [109,

¹⁷Perhaps the most famous agglomerative cluster in the United States is Silicon Valley, California, where high skilled individuals in the local computing and semiconductor industries drew talent and expertise from local universities like Stanford and UC Berkeley. Another example is the Research Triangle in North Carolina, which hosts a strong biotech cluster supported by Duke University, North Carolina State University, and UNC Chapel Hill (the three points on the "triangle").

241]; a researcher may be more likely to start a new collaboration with the colleague of a present co-author, due to their social proximity, usually operationalized through collaboration networks¹⁸. There are countless other kinds of proximity which might dictate the likelihood of interaction between two people, including their language, socioeconomic status, their demographics, the visa guidelines between their countries, and more, all of which help shape the structure of science.

Mobility

Proximity isn't static. Rather, researchers are *mobile* throughout their career, moving from one institution to another, between cities, and across continents. With every move, the researchers who they are proximate to, geographically, but also socially, and can broaden their professional networks and gain access to new ideas. Mobility has long been a central phenomenon in academia and scholarship¹⁹, though the flow of scholars around the world has further exploded in recent decades [177]. This contemporary international mobility has been found to drive collaborations between countries [84, 292, 293], and is associated with greater career citation impact [294, 295]. One major benefit of mobility is that it facilitates the diffusion of ideas [116, 296], promoting innovation [194, 297] that can power local industry [298]. When a scientist moves, they are able to transfer their *tacit knowledge* (or human capital) [237, 299] to people at their destination [300, 301]. In some cases, the extensive mobility of scholars from a single institution has cultivated entire intellectual traditions by diffusing their ideas [116, 296]²⁰. The flow of information is not however, one way. A mobile scientist can bridge the social communities of their origin and destination,

¹⁸The most familiar way of thinking about social proximity is probably through friend networks. Assuming I, Dakota, am at the center of my network, then my immediate friends might have a proximity of 1. The friends of my friends, in contrast, are more distant, with a proximity of 2; their friends, in turn, would have a proximity of 3, and so on. This is also well-exemplified in the professional social networking website LinkedIn, which visualizes the "degree of connection" between two people, based on voluntary connections between pairs of users.

¹⁹A tradition of mobility existed in Medieval Europe, when scholars could (relatively) easily gain experience from different institutions. One famous example of this is Erasmus of Rotterdam, a scholar and theologian who had a life spread across England, Italy, and the Holy Roman Empire, learning from and working alongside new scholars at each institution.

²⁰Some of the most famous examples of this include the Chicago School of Sociology [302]. By graduating and placing students trained at the university, the intellectual tradition of its department was able to be spread far and wide, and became dominant in the field.

exposing the communities to each other's work, and easing the flow of ideas and knowledge between them [303–305]. Noticing the potential benefits of mobility, a "global competition for talent" [306] has emerged, such that countries compete with one another to attract and retain high-skilled workers. So far, countries like the United States, United Kingdom, and China have attracted the largest shares of scholars, helping to power their scientific systems [307], with the United States, in particular, enjoying decades serving as a hub for international researchers, brought on by a strong education and research environment [308].

Individual scholars have their own reasons to be mobile, often very different from national interests. Most mobile researchers aim to expand their professional networks [110], to gain access to prestigious institutions and opportunities [309, 310], to obtain entry into high-performing research groups [311], and to obtain resources for research [312, 313] or employment [314] that are rare in their home countries. In many cases, scholars can reap the benefits of mobility²¹ without permanent moves; temporary mobility, usually lasting from a few weeks to months, or the length of a faculty's sabbatical, are common in academia [317], and co-affiliation—the holding of multiple simultaneous affiliations, often splitting time or resources between them—is also ubiquitous [318]. A researcher's choice of destination may be further shaped by language [319, 320], visa and immigration policies [321], and family considerations [311]

Although only a fraction of scientists are mobile [295, 318], the effects of mobility are far-reaching, and often beneficial. Its consequence on science are similarly deep. By altering the landscape of proximity, and facilitating the accumulation of social (and sometimes material) capital, mobility is yet another vital principle that governs who interacts with who, and how, in science.

²¹If should be noted that, especially at the individual level, mobility is not an entirely positive or beneficial activity. For example, women have historically faced more constraints to their ability to be mobile, such as disproportionate family burdens [315]. Policies encouraging scientific mobility discriminate against those with families or dependents, scholars with disabilities, or those who simply prefer not to move [316].

Formal evaluation

So far I've spoken mostly of informal, individual-level factors that shape science, but many scientific communities also govern themselves through more centralized means using formal evaluation, most often through peer review and performance metrics. Peer review in its current form is relatively recent, emerging after World War II in order to help manage the ever-growing scientific ecosystem [2]²². Now, peer review is the most ubiquitous form of evaluation in science²³, used to select candidate to hire, manuscripts to publish, and grant proposals to fund [326]. With the growing-importance of peer review, the role of the peer-reviewer has also grown into a powerful gatekeeping position, able to influence what kind of science gets done, in many cases biasing certain kinds of science (and scientists) over others. For example, journal peer reviewers are more likely to criticize papers that challenge prevalent theoretical perspectives, whereas methodological and applied works can go through review relatively unscathed [327]. By disproportionately criticizing papers that offer novel theoretical perspectives, peer review can constrain the frontier of science to the interests and ideas of a narrow set of elite gatekeepers [328]. Biases in peer review can also be personal, driven by the demographics of the author [329], the personal relationships between the authors and reviewers [241–243], and the competing disciplinary norms and expectations of the reviewers themselves [326]. Peer review is a means for a community to act as gatekeeper, dictating which people and ideas that are acceptable, and thus dictating what kinds of science gets done and controlling the boundaries of the field [330].

Quantitative science performance metrics, appearing as early as the 1830s [132], have proliferated in recent decades as a more seemingly-object way of judging papers, scientists, institutions, and countries. The most common science performance metrics currently in use are the paper cita-

²²The origin of peer review can be traced to various starts, the most widely-accepted being systems of refereeing papers introduced by the Royal Society of Edinburgh in 1731, and a similar system implemented by the Royal Society of London in 1752 [322].

²³Alternative review formats have been recently proposed, such as open peer review at Nature journals [323], collaborative peer review at Frontiers [324], and consultative peer review at eLife [325] (King, 2017) offer potential solutions, though the implications of these consequences are not well understood.

tion count [150], which judge the *impact* of a paper by the number of citations it receives [331]; the H-index, which quantifies a scientist performance as an aggregate of their number of publications, and citations those publications receive [149]; the Journal Impact Factor, which scores a journal based on the average citations received by papers published in it; and university rankings, such as the *Leiden Rankings*, which organize universities and institutions along various dimensions of productivity and impact [151, 332]. The move towards quantification has also lead to the use of performance metrics in other areas of academia [333]. Teaching performance is often quantified through student-evaluations of teachers, which aim to use student feedback to evaluate faculty's instruction [334, 335]. And now, services like Publons aim to credit researchers for their service activity by counting the number of journal peer reviews they conduct [94]. These various performance metrics aim to objectively measure scientific performance and impact, however they can also shape the way science gets done. This is best exemplified by Goodhart's Law, which states that "when a measure becomes a target, it ceases to become a good measure" [336]; in other words, if individuals and institutions are rewarded based on a metric, they may change their practices to optimize the metric itself, rather than what the metric is attempting to measure. For example, some journals have adopted a practice of coercing submitting authors to cite other papers in the venue, in order to inflate their impact factor [337]; similarly, some authors have formed citation carters whereby they cite one another's work, even when unnecessary to do so, in order to increase their H-indexes [338]. Individuals can also "salami-slice" their publications, or publish duplicate or redundant papers, with the goal of inflating their total publications and citations [339]. Scientists may also chase safe topics that will reliably result in moderately-impactful publications, rather than riskier, but perhaps more transformative research [19, 340]. Universities, too, have aimed to grow their rankings, relentlessly chasing individual metrics such as class size, student-to-teacher ratio, and library size, rather than underlying education [341, 342]²⁴. Metrics are powerful tools

²⁴One important metric in university rankings is the average class size at a university, though this is easy to manipulate. A university can (and does) organize classes such that many are relatively small (maybe around 20 students), but with a few incredibly large ones (for example, 200). The result is that the average class size will

with which science can be evaluated, yet they are not neutral; but rewarding optimization over scholarship, metrics can shape how scientists do their work, and what topics they pursue.

Object of study

The topic a scientist chooses to study can also have consequences on how they organize themselves and interact with their community. For example, High-Energy physics requires massive and tremendously-expensive machinery and instruments, necessitating that scientists studying in this area collaborate with one another, pooling resources and labor in order design and carry out experiments [70]. In another area, researchers studying extra-planetary geology on Mars requires that teams of scientists collaborate with engineers in order to negotiate tasks and plan routes for the few extant Martian rovers [343]. Similarly, Glaciology research in Antarctica necessitates careful planning and organization of research teams, and careful coordination with logistical and support staff (especially the U.S. Navy) [344]. Climatologists studying the Earth's atmosphere, in contrast, are more distributed, utilizing data collected from a myriad of observation stations around the world, often with the help of distant collaborators or even volunteers [174]. In biomedical and genetics research, scientists are instead more distributed, yet rely on animal models such as mice, drosophila, axolotls. However, generalizable research using these organisms requires they be identical, at least as nearly as possible; this requires that their genotype be homogenized across the population and genetic integrity maintained, a task requiring careful coordination across labs and institutions [345, 346]. In contrast to all of these, Humanities scholars are more diffuse, their work requiring little if any collaboration between them, leading to an individual, and sometimes antagonistic approach to their field [347]. The practicalities of science are often overlooked when understanding its structure, yet certain objects of study require certain kinds of capital and specific social organizations.

be relatively small, but that the average student will almost certainly have one (or more) of these massive courses, simply because one 200 person class has more students than 9 20 person classes [65].

External forces

So far, I've focused almost entirely on *internal* processes of science—how scientists organize themselves based on principles of reputation, cumulative advantage, homophily, and so on. However, science don't exist in a vacuum. Rather, science is just one institution among many overlapping and interacting systems that constitute global humanity. The consequence of this is that what happen *outside* science also has an effect on whats happening *within*. To demonstrate how, I'll draw on three illustrative examples of how forces external to science itself can have a direct impact on who does science and what kind of science gets done; namely, I explore national immigration, national crises, and corporate interests.

The U.S. has long benefited from immigrants arriving on its shores and contributing not only to its economic welfare [348], but also to the nation's scientific capacity. Historically, many of these immigrants were not so much attracted to the U.S. as they were pushed our of their home countries. In the 1930s, the Nazi Party of Germany removed Jewish scientists from their universities, decimating their own scientific capacity in one cruel decision—one that would precede many, many more. Many of these scientists, now without jobs and living in a country that was growing increasingly-hostile towards them decided to move, scattering across other parts of Europe, and many to the United States. In the U.S., these immigrant (and refugee) scientists contributed immensely to the economic and industrial capacity of the country [349], with some, such as Albert Einstein, becoming intellectual celebrities. Following World War II, the U.S. imported (perhaps against their will) German scientists such as Wernher von Braun, a rocket scientists with a past ranging from complicated to repugnant, but who proved indispensable for America's burgeoning spaceflight program, including the development of the early Redstone rocket and, eventually, the Mercury, Gemini, and Apollo programs. Years later, the U.S.S.R would collapse, and with it, the iron curtain that separated U.S. and Soviet science, leading to the diffusion of knowledge between them; many Soviet mathematicians sought to move to the U.S., and in turn introduced their find-

ings and ideas to the U.S. scientific community [92]. More recently, the Trump administration in the U.S. institutes travel restrictions targeting many majority-Muslim countries, which had the effect of ceasing immigration and the free movement of their scholars, upending scientific relations and collaborations [350]. The COVID-19 pandemic, beginning in China in 2019 and quickly spreading around the world through 2020 broad some of the most widespread travel restrictions ever seen, rapidly altering the structure of global scientific mobility. Mobility, ranging from short-term visits to permanent immigration, is a central phenomenon in science, yet it is also one of the most directly impacted by the goings on of the world outside of science. The cruel decisions of the Nazi Party, the end of World War II, the rise and fall of the U.S.S.R., the political turnover of the U.S. government, and the onset of global pandemic, all had a major and lasting impact on mobility, and thus on the entire history of science.

Nations enact policies for funding and research support that have direct impact on the quantity and quality of science that gets done. However, the choice over how many resources to give to science is not made in a vacuum. There is a widespread, though complicated, belief that funding in science is essential for economic prosperity [351–355], yet funding for science is often cast in opposition to other government spending in welfare, infrastructure, and the military, and so without good and immediate reason to fund science, it will often be de-prioritized, and their national scientific capacity left to decay. However, external threats to a nation can motivate a country to make large commitments to funding research, as well as encouraging scientific and technical careers and supporting their systems of innovation [18]. For example, the dire threat of conflict with between the U.S.S.R. and the U.S. motivated both nations to invest considerable resources into the development of their nation’s research and technological infrastructure, the Space Race being perhaps the most public and prominent example. Other threats are economic; South Korea, for example, emerged out of a painful civil war with a struggling economy, yet facing the prospects of an impoverished future and economic irrelevance, their government instituted policies to expand

their scientific and economic capacity, eventually becoming one of the top per-capita funders of science in the world [11]. Research funding may be essential for developing scientific, and even economic capacity, but it is also an easy item to cut when viewed as unnecessary. In other cases, nations may shift funding priorities in response to changes in geopolitics. During World War 1, for example, the schism between European powers lead to a similar schism between their scientists, producing separate research communities pursuing different ideas without citing or interact with one another [356]. In another case, after the fall of the U.S.S.R., the U.S. shifted funding away from Defense research, such as in Physics and Chemistry, and towards new areas like Computer Science and Oceanography [357]. Crises, threats, geopolitics, Research agendas are also set by current events. The COVID-19 pandemic has seen the intense attention of scientists from around the world [358, 359], often with the support of emergency grants from government agencies. Current events, whether they be the threat of war or economic inferiority, natural crises, or the rise and fall of nations, all have an impact on what kind of science is supported, and how much support it gets, inevitably affecting the makeup and organization of the scientific ecosystem.

Economic stakeholders, such as corporations, can also shape how science gets done by selectively funding science and scientists who's research benefits the company, or otherwise persuade scientists to put professional integrity aside, and conduct misleading or false findings. For example, U.S. Tobacco companies funded numerous studies in the 1950s aiming to cast doubt on the connection between smoking and lung cancer ²⁵, even in some cases espousing the health *benefits* of cigarettes [360]. Similar distortions of science and public opinion was done by U.S. food companies, who funded research detailing the health benefits of grains and sugars—misleading, suspect, and perhaps even fraudulent science that remains the foundation of the "Food Pyramid" which still

²⁵Interestingly, a 1954 book called *How to Lie With Statistics* approached some of the ways that scientists or other motivated actors might mislead a reader with statistical reasoning. This book remains one of the most popular statistics books every written. The author, Darrell Huff, would however become a lobbyist for tobacco companies, himself using spurious statistical methods to try and disprove the smoking-cancer link. The case of Mr. Huff goes to show that contrarianism and self-righteous methodological crusading isn't all that its cracked up to be.

informs nutritional guidelines to this day²⁶. In more recent memory, Oil companies have sought to dispute the connection between greenhouse gas emissions and climate change, with the intention of supplying evidence that could be selectively cited to discourage regulation [360] in spite of the scientific consensus on the subject [361]. Such companies have also been found to encourage EPA regulations with the stated goal of encouraging transparent science by requiring public policy to cite only work that follows open science principles; in practice, however, this was an instance of regulatory capture in science. The regulations were incredibly onerous for the individual scientists who lacked the resources to meet the EPA's guidelines; corporations, on the other hand, had plenty of resources to adhere to these regulations, and so the EPA could use only corporate research to set policy [362]. Science aims towards objectivity, but science doesn't exist in a vacuum. Outside stakeholders can influence the process in countless ways, funding fraudulent or misleading research that ends up distorting the scientific consensus [192].

And many more...

This list hits the highlights (as far as I'm aware) of the main principles governing science, though it is by no means exhaustive. Death, for example, is an exceedingly-important event that has importance not just on the individual doing the dying, but also on the social systems they leave behind. An eminent individual can have prominent positions in their field, stifling competing ideas and young upstarts; that person's death, then, can make room for their competitors to fund, publish, and bring attention to their work, thus changing the field [81]. Underlying social and cultural currents can also shape science, motivating many scholars, independently from one another, to chance hop topics [363] or to attempt to tackle ongoing crises, such as the COVID-19

²⁶The food pyramid is a graphic illustration put out by the U.S. Department of Agriculture attempting to communicate the relative importance of different food groups to a person's nutrition. Grains, such as breads and pastas, sit at the bottom of the pyramid due to their supposed importance to health and well-being. Fats, oils, and sugars, are at the top of the pyramid, meaning that they are supposed to account for only a small part of a person's diet. Nutrition is a sensitive and seemingly-ideological topic for many, and my opinions on the optimal diet is outside the scope of this dissertation. However, most dietary ideologies (and research) will attest to the lack of support for the food pyramid, and will agree that it should be done away with as a tool for dietary recommendations.

pandemic [359]. Communication constraints are also vitally important; science in the 19th century largely relied on handwritten letters mailed between colleagues, making it difficult to facilitate international collaborations; however in the 21st century technology has made communication easy: Zoom, eMail, Slack, and Twitter are tools that have reduced communication friction, opening the door for larger and more seamless international collaborations. Even with these technologies, however, there may be other social forces that constrain the size and composition of scientific teams; Dunbar's number, for instance, refers to the amount of social relationships a person can usually hold (about 150), and its possible that a similar number can structure science, limiting the size of functioning communities and teams to a small and manageable number of relationships. Science is also shaped by the availability of funding, whether from the government or elsewhere; even college (American) football, when profitable for the university, can lead to funding windfalls for its faculty that allow them to pursue more and higher quality work [364]. While I can continue this sort of speculation indefinitely, what is clear from the list in this section is that the workings of science cannot simplified to a single overarching mechanism. Rather, there are many mechanisms, overlapping and interacting, that all together govern how scientists work and work together, giving shape to the complex system of science.

2.6 What complexity contributes to Metascience

After having spent so much time talking about the ins and outs of science as a complex system, its worth pausing and asking why exactly this kind of model is useful. Surely, terming something as "complex" doesn't seem to invite many followup questions; if science is so complex, then why should we spend so much time trying to understand it? Its too complex! Surely all such efforts will lead nowhere. While complexity does dash the hope of certain kinds of research—long-range prediction and forecasting being the most affected—it's far from useless.

Complexity introduces a powerful way of thinking about science. One where the chaos and

disorder of the system is not a problem to be solved or a confounding factor to be waved away, but instead a key component underlying the system. It shifts the perspective of Metascience away from nations, cities, and institutions, and instead individual scientists, making their own decisions in an uncertain world based on limited information of the world around them, being shaped by while also shaping countless overlapping social structures that they may barely be aware of, but in which they are deeply enmeshed.

This dissertation adopts this view of science as a complex system, and one in which individual-level forces are the starting point for the larger structures and behaviors that emerge from the system. This bottom-up view is supported by new datasets offering information on individuals, even in areas of scientific activity that were long invisible to quantitative Metascience. Peer review, the study of which has long been sequestered into qualitative studies on one hand, and analysis of journal-level metrics on the other, can now be studied at the level of individual reviewers and authors thanks to changes in open science. Similarly, new public websites like *RateMyProfessors.com* can provide data-driven valuable insights into the challenges faced by individual faculty, in their own particular circumstances, by subjective measures of their teaching performance—the metrics for which have long been siloed by universities, restricting large-scale analysis. The increasing-availability of the full-text of scientific publications combined with advancements in machine learning makes it possible to study at scale how scientists interact with one another through the text of their publications, and in turn how these local interactions fit within the cultural norms of entire disciplines. Through author-name disambiguation, bibliographic databases like the Web of Science make it possible to study the career mobility of millions of individual around the world, considering how their trajectories are shaped by the complicated cultural, linguistic, symbolic, and geographic factors that underpin human mobility.

While complexity is a powerful framework, it can also be constraining. Metascience not only aims to understand how science works, but also aims to improve the scientific process, whether it

be through new tools, correcting of biases, or the crafting of effective science policy. However, as research in this dissertation will show, the factors that shape science are numerous, subtle, and interact with one another. Predicting the future directions of science, or the actual outcomes of a policy, can prove fruitless. Attempts to improve the system can just as easily backfire and cause harm. What is the point of Metascience then? If science is a complex system, then is it even possible to put our understanding of it (however limited) to use to make meaningful improvements? I think yes. Though only in certain situations, and only to a certain extent. I hope that the following chapters, which go through each of the four studies that make up this dissertation, will provide insight into what I mean.

Chapter 3

Study 1: Peer review at *eLife*

3.1 Foreword

I see no reason to address the—in any case erroneous—comments of your anonymous expert. On the basis of this incident I prefer to publish the paper elsewhere.

Albert Einstein

Peer review is the gold standard for evaluation science. Versions of peer review are implemented to inform faculty hiring decisions, editorial decisions are journals, and the allocation of resources from granting agencies. Although universal, the system has long had its detractors. Albert Einstein offered an early criticism, being offended by the decision of the journal *Physical Review* to call on some “*uninformed*” person to judge his work [365]. More recently, however, peer review has been accused of taking too long [366], inconsistency [367], nepotism [241, 243], and bias against marginalized populations [368].

In this chapter, I present a study investigating the extent of bias in peer review at the journal *eLife*, a recent yet prestigious journal in the life sciences that implements a very unique form of *quantitative* review. Making use of unique data provided by *eLife*, I observe disparities in the acceptance rates between men and women authors, and between U.S. and non-U.S. authors. Yet interestingly, I also find that these disparities reduce depending on the demographics of the reviewer team. Mixed-gender or mixed-nationality teams tend to produce more equitable review outcomes. Not only does this finding identify potential bias in peer review, it also lays out a path forward: make the reviewers more diverse.

As it is, this study provides valuable and actionable insights into peer review, yet they take

new significance when viewed through the complexity perspective. Interpersonal forces, such as demographic bias and homophilic preferences, clearly play a role in shaping review decisions. Yet more generally, I argue that peer review is an example of a *feedback mechanism* in science, in that those who succeed in peer review will have access to more resources that allow their further success. Because of this feedback, even small biases in the process can rapidly accumulate into long-term career consequences [270], and sit at the heart of the underrepresentation of women and other marginalized groups in science. The complexity view reveals that peer review is a mechanism that can perpetuate inequalities in science, making it all the more important that biases are identified and addressed.

The work shown in this chapter was posted to the preprint server *bioRxiv* in 2019 under the title *Author-reviewer homophily in peer review* as a finished public manuscript. The paper was written in collaboration with Kyle Siler, Vincent Larivière, Wei Mun Chan, Andrew M. Collings, Jennifer Raymond, and Cassidy R. Sugimoto. I, and my collaborators, are grateful for the editing and feedback provided by Susanna Richmond (Senior Manager at eLife), Mark Patterson (Executive Director at eLife), Eve Marder, Anna Akhmanova, and Detlef Weigel (Deputy Editors at eLife). We are also grateful for the work of James Gilbert (Production Editor at eLife) for extracting the data used in this analysis. This work was partially supported by a grant from the National Science Foundation (SciSIP #1561299). This work also wouldn't be possible without all the reviewers working at *eLife* and whose contributions make up this data; they are not merely subjects of analysis for meta-analyses, but rather individuals making essential, yet often thankless contributions to science. This analysis is not meant to impugn their work, but rather to find a path forward that makes the process better for both them, the authors they review, and science as a whole.

3.2 Abstract

The fairness of scholarly peer review has been challenged by evidence of disparities in publication outcomes based on author demographic characteristics. To assess this, we conducted an exploratory analysis of peer review outcomes of 23,876 initial submissions and 7,192 full submissions that were submitted to the biosciences journal *eLife* between 2012 and 2017. Women and authors from nations outside of North America and Europe were underrepresented both as gatekeepers (editors and peer reviewers) and authors. We found evidence of a homophilic relationship between the demographics of the gatekeepers and authors and the outcome of peer review; that is, there were higher rates of acceptance in the case of gender and country homophily. The acceptance rate for manuscripts with male last authors was seven percent, or 3.5 percentage points, greater than for female last authors (95% CI = [0.5, 6.4]); this gender inequity was greatest, at nine percent or about 4.8 percentage points (95% CI = [0.3, 9.1]), when the team of reviewers was all male; this difference was smaller and not significantly different for mixed-gender reviewer teams. Homogeneity between countries of the gatekeeper and the corresponding author was also associated with higher acceptance rates for many countries. To test for the persistence of these effects after controlling for potentially confounding variables, we conducted a logistic regression including document and author metadata. Disparities in acceptance rates associated with gender and country of affiliation and the homophilic associations remained. We conclude with a discussion of mechanisms that could contribute to this effect, directions for future research, and policy implications.

3.3 Introduction

Peer review is foundational to the development, gatekeeping, and dissemination of research, while also underpinning the professional hierarchies of academia. Normatively, peer review is expected to follow the ideal of “universalism” [79], whereby scholarship is judged solely on its intellectual merit. However, confidence in the extent to which peer review promotes the best scholarship

has been eroded by questions about whether social biases [369], based on or correlated with the demographic characteristics of the scholar, could also influence outcomes of peer review [368, 370, 371]. This challenge to the integrity of peer review has prompted increased interest in assessment of the disparities and potential influence of bias in their peer review processes.

Several terms are often conflated in the discussion of bias in peer review. We use the term *disparities* to refer to unequal composition between groups, *inequities* to characterize unequal outcomes, and *bias* to refer to the degree of impartiality in judgment or lack thereof. Disparities and inequities have been widely studied in scientific publishing, most notably with regard to gender and country of affiliation. Globally, women account for only about 30 percent of scientific authorships [28] and are underrepresented even when compared to their numbers in the scientific workforce [372, 373]. Underrepresentation of articles authored by women is most pronounced in the most prestigious and high-profile scientific journals [211, 268, 374–377]. Similar disparities are observed across countries, for which developed countries dominate the production of highly-cited publications [26, 378].

The under-representation of authors from certain groups may reflect differences in submission rates, or it may reflect differences in success rates during peer review (percent of submissions accepted), or both. Analyses of success rates have yielded mixed results in terms of the presence and magnitude of such inequities. Some analyses have found lower success rates for female-authored papers [242, 379] and grant applications [35, 209], while other studies have found no gender differences in review outcomes (for examples, see [380–384]). Inequities in journal success rates based on authors' nationalities or country of affiliation have also been documented, with reports that authors from English-speaking and scientifically-advanced countries have higher success rates [39, 385]; however, other studies found no evidence that the language or country of affiliation of an author influences peer review outcomes [39, 234, 386]. These inconsistencies could be explained by several factors, such as the contextual characteristics of the studies (e.g., country, discipline) and

variations in research design and sample size. Another possible explanation is that these gender and national disparities emerge from *bias* in peer review.

The possibility that bias contributes to inequities in scientific publishing and the nature of any such bias is highly controversial. Implicit bias—the macro-level social and cultural stereotypes that can subtly influence everyday interpersonal judgments and thereby produce and perpetuate status inequalities and hierarchies [387, 388]—has been suggested as a possible mechanism to explain differences in peer review outcomes based on socio-demographic and professional characteristics [368]. When faced with uncertainty—which is quite common in peer review—people often weight the social status and other ascriptive characteristics of others to help make decisions [389]. Hence, scholars are more likely to consider particularistic characteristics (e.g., gender, institutional prestige) of an author under conditions of uncertainty [390, 391], such as at the frontier of new scientific knowledge [392]. However, given the demographic stratification of scholars within institutions and across countries, it can be difficult to pinpoint the nature of a potential bias. For example, women are underrepresented in prestigious educational institutions [184, 393, 394], which conflates gender and prestige biases. These institutional differences can be compounded by gendered differences in age, professional seniority, research topic, and access to top mentors [29]. Another potential source of bias is what is dubbed cognitive particularism [395], whereby scholars harbor preferences for work and ideas similar to their own [241]. Evidence of this process has been reported in peer review in the reciprocity (i.e., correspondences between patterns of recommendations received by authors and patterns of recommendations given by reviewers in the same social group) between authors and reviewers of the same race and gender [396] (see also [38, 111]). Reciprocity can exacerbate or mitigate existing inequalities in science. If the work and ideas favored by gatekeepers are unevenly distributed across author demographics, this could be conducive to Matthew Effects [79], whereby scholars accrue accumulative advantages via *a priori* status privileges. Consistent with this, inclusion of more female reviewers was reported to attenuate biases that favored men in the awarding of

RO1 grants at the National Institute of Health [379]. However, an inverse relationship was found in the evaluation of candidates for professorships [397] when female evaluators were present, male evaluators became less favorable toward female candidates. Thus the nature and potential impact of cognitive biases during peer review are multiple and complex.

Another challenge is to disentangle the contribution of bias during peer review from factors external to the review process that could influence success rates. For example, there are gendered differences in access to funding, domestic responsibilities, and cultural expectations of career preferences and ability [398, 399] that may adversely impact manuscript preparation and submission. On the other hand, women have been found to hold themselves to higher standards [212] and be less likely to compete [400], hence they may self-select a higher quality of work for submission to prestigious journals. At the country level, disparities in peer review outcomes could reflect structural factors related to a nation’s scientific investment [378, 401], publication incentives [147, 402], local challenges [403], and research culture [404], all of which could influence the actual and perceived quality of submissions from different nations. There are also several intersectional issues: there are, for example, differences in socio-demographic characteristics of the scientific workforce across countries—e.g., more women from some countries and disproportionately less professionally-senior women in others [28]. Because multiple factors external to the peer review process can influence peer review outcomes, unequal success rates for authors with particular characteristics do not necessarily reflect bias in the peer review process itself; conversely, equal success rates do not necessarily reflect a lack of bias.

Here, we assessed the extent to which bias contributes to gender and country disparities in peer review outcomes by analyzing the extent to which the magnitude of these disparities vary across different gender and country compositions of gatekeeper teams. In particular, we focused on the notion of homophily between the reviewers and authors. This analysis examined the outcomes of peer review at *eLife*—an open-access journal in the life and biomedical sciences. Peer review

at *eLife* differs from other traditional forms of peer review used in the life sciences in that it is done through deliberation between reviewers (usually three in total) on an online platform. Previous studies have shown that deliberative scientific evaluation is influenced by social dynamics between evaluators [326, 405]. We examine how such social dynamics manifest in *eLife*'s deliberative peer review by assessing the extent to which the composition of reviewer teams correlates with peer review outcomes. Using all research papers (Research Articles, Short Reports, and Tools and Resources) submitted between 2012 and 2017 ($n=23,876$), we investigate the extent to which a relationship emerges between the gender and country of affiliation of authors (first, last, and corresponding) and gatekeepers (editors and invited peer reviewers), extending the approach used in previous work [369].

3.4 Consultative peer review and *eLife*

Founded in 2012 by the Howard Hughes Medical Institute (United States), the Max Planck Society (Germany), and the Wellcome Trust (United Kingdom), *eLife* is an open-access journal that publishes research in the life and biomedical sciences. Manuscripts submitted to *eLife* progress through several stages. In the first stage, the manuscript is assessed by a Senior Editor, who may confer with one or more Reviewing Editors and decide whether to reject the manuscript or encourage the authors to provide a full submission. When a full manuscript is submitted, the Reviewing Editor recruits a small number of peer reviewers (typically two or three) to write reports on the manuscript. The Reviewing Editor is encouraged to serve as one of the peer reviewers. When all individual reports have been submitted, both the Reviewing Editor and peer reviewers discuss the manuscript and their reports using a private online discussion system hosted by *eLife*. At this stage the identities of the Reviewing Editor and peer reviewers are known to one another. If the consensus of this group is to reject the manuscript, all the reports are usually sent to the authors. If the consensus is that the manuscript requires revision, the Reviewing Editor and additional peer

reviewers agree on the essential points that need to be addressed before the paper can be accepted. In this case, a decision letter outlining these points is sent to the authors (the original reports are not usually released in their entirety to the authors). When a manuscript is accepted, the decision letter and the authors' response are published along with the manuscript. The name of the Reviewing Editor is also published. Peer reviewers can also choose to have their name published. This process has been referred to as consultative peer review (see [325, 406] for a more in-depth description of the *eLife* peer-review process).

3.5 Data and methods

Ethics statement

This research underwent expedited review by the Institutional Review Board at Indiana University Bloomington and was determined to be exempt (Protocol #: 1707327848).

Data

Metadata for research papers submitted to *eLife* between its inception in 2012 and mid-September, 2017 (n=23,876) were provided to us by *eLife* for analysis. As such, these data were considered a convenience sample. Submissions fell into three main categories: 20,948 Research Articles (87.7 percent), 2,186 Short Reports (9.2 percent), and 742 Tools and Resources (3.1 percent). Not included in this total were six Scientific Correspondence articles, which were excluded because they followed a distinct and separate review process. Each record potentially listed four submissions—an initial submission, full submission, and up to two revision submissions (though in some cases manuscripts remained in revision even after two revised submissions). Fig 3.1 depicts the flow of all 23,876 manuscripts through each review stage. The majority, 70.0 percent, of initial submissions for which a decision was made were rejected. Only 7,111 manuscripts were encouraged to submit a full submission. A total of 7,192 manuscripts were submitted as a full submission; this number was

slightly larger than encouraged initial submissions due to appeals of initial decisions and other special circumstances. Most full submissions, 52.4 percent ($n = 3,767$), received a decision of revise, while 43.9 percent ($n = 3,154$) were rejected. A small number of full submissions ($n = 54$) were accepted without any revisions. On the date that data were collected (mid-September, 2017), a portion of initial submission ($n = 147$) and full submissions ($n = 602$) remained in various stages of processing and deliberation (without final decisions). On average, full submissions that were ultimately accepted underwent 1.23 revisions and, within our dataset, 3,426 full submissions were eventually accepted to be published. A breakdown of the number of revisions requested before a final decision was made, by gender and country of affiliation of the last author, is provided in Fig. A.1. A portion of initial and full submissions ($n = 619$) appealed their decision, causing some movement from decisions of “Reject” to decisions of “Accept” or “Revise”; counts of appeals by the gender of author and gatekeepers is shown in Fig. A.2.

The review process at *eLife* is highly selective, and became more selective over time. While only garnering 307 submissions in 2012, *eLife* accrued 8,061 submissions in 2016. Fig 3.2 shows that while the total count of manuscripts submitted to *eLife* has rapidly increased since the journal’s inception, the count of encouraged initial submissions and accepted full submissions has grown more slowly. The encourage rate (percentage of initial submissions encouraged to submit full manuscripts) was 44.6 percent in 2012, and dropped to 26.6 percent in 2016. The acceptance rate (the percentage of accepted full submissions) was 62.4 percent in 2012 and decreased to 53.0 percent in 2016. The overall acceptance rate (percentage of initial submissions eventually accepted) began at 27.0 percent in 2012 and decreased to 14.0 percent by 2016.

In addition to authorship data, we obtained information about the gatekeepers involved in the processing of each submission. We defined gatekeepers as any Senior Editor or Reviewing Editor at *eLife* or invited peer reviewer involved in the review of at least one initial or full submission between 2012 and mid-September 2017. Gatekeepers at *eLife* often served in multiple roles; for

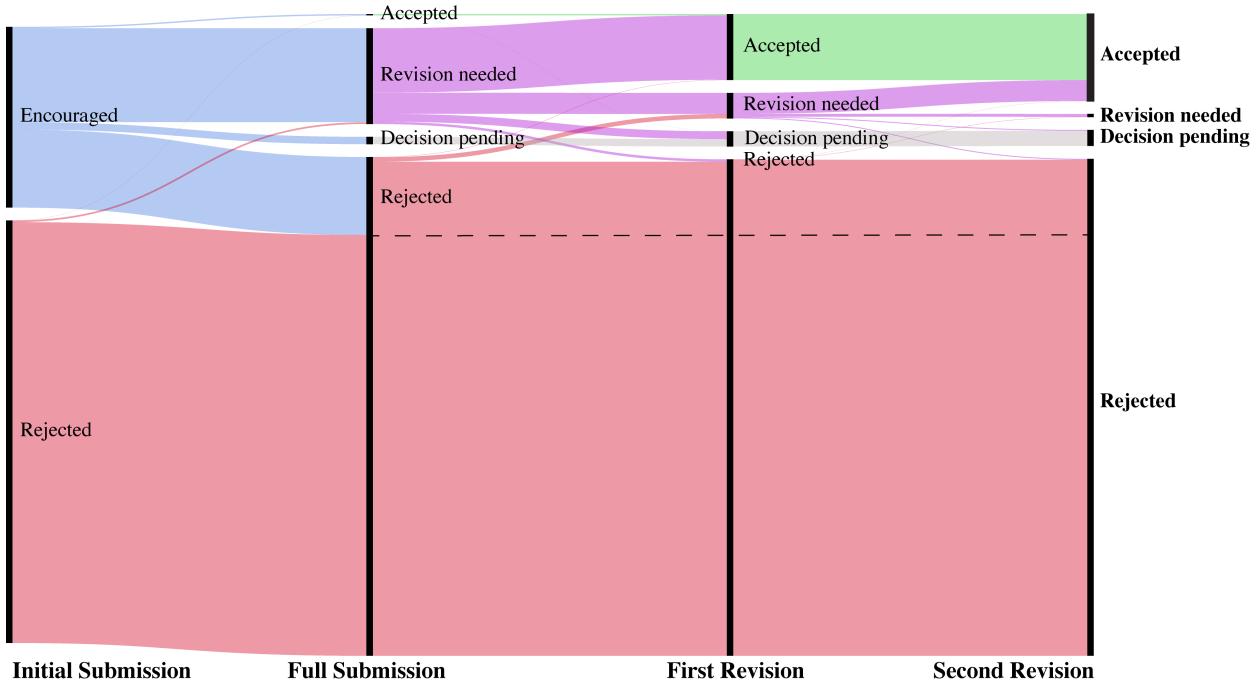


Figure 3.1: Flow of all papers through the *eLife* review process.

Starting from the left, an initial submission is first given an initial decision of encourage or reject, and if encouraged, continues through the first full review and subsequent rounds of revision. “Encouraged”, “Accepted”, “Rejected” and “Revision needed” represent the decisions made by *eLife* editors and reviewers at each submission stage. A portion of manuscripts remained in various stages of processing at the time of data collection—these manuscripts were labeled as “Decision pending”. The status of manuscripts after the second revision is the final status that we consider in the present data. The dashed line delineates full submissions from rejected initial submissions.

example, acting as both a Reviewing Editor and peer reviewer on a given manuscript, or serving as a Senior Editor on one manuscript, but an invited peer review on another. In our sample, the Reviewing Editor was listed as a peer reviewer for 58.9 percent of full submissions. For initial submissions, we had data on only the corresponding author of the manuscript and the Senior Editor tasked with making the decision. For full submissions we had data on the corresponding author, first author, last author, Senior Editor, Reviewing Editor, and members of the team of invited peer reviewers. Data for each individual included their stated name, institutional affiliation, and country of affiliation. A small number of submissions were removed, such as those that had a first but no last author (reflecting compromised data record—even a single-authored manuscript should have duplicate authors across all roles) and those that did not have a valid submission type. Country

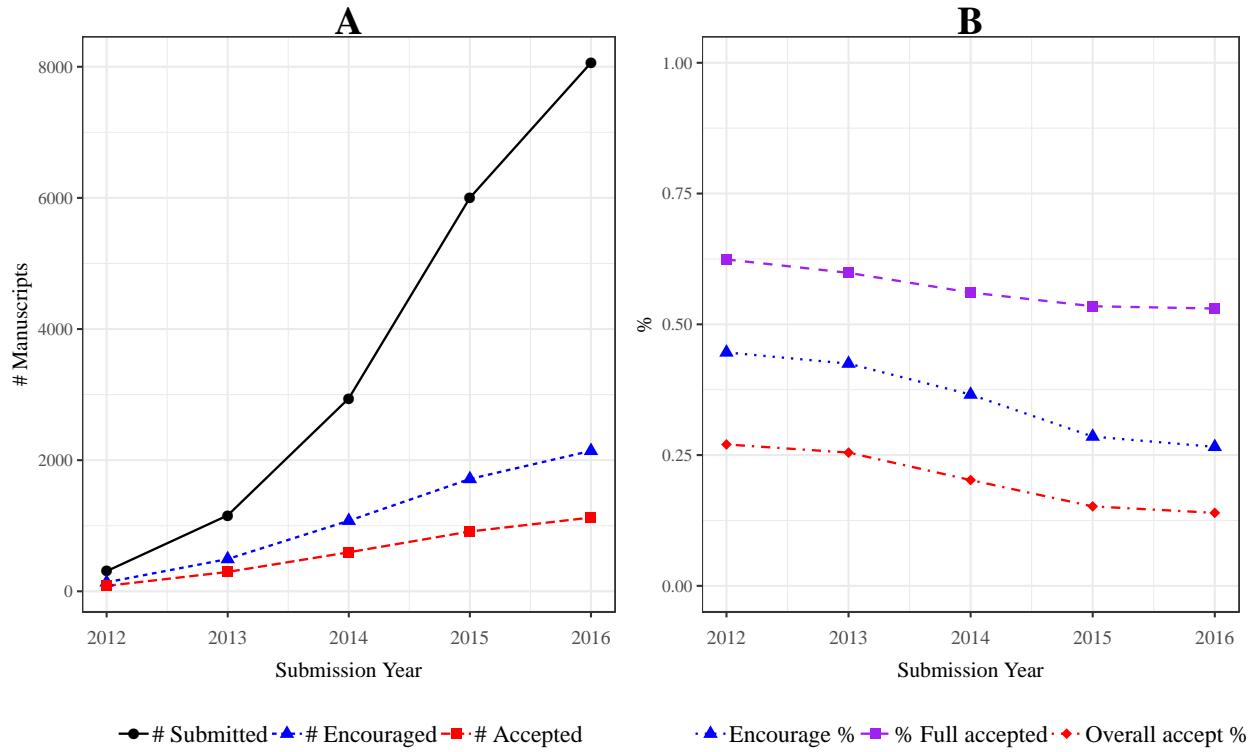


Figure 3.2: Submissions and selectivity of *eLife* over time.

A: Yearly count of initial submissions, encouraged initial submissions, and accepted full submissions to *eLife* between 2012 and 2016; **B:** Rate of initial submissions encouraged (Encourage %), rate of full submissions accepted (% Full accepted) and rate of initial submissions accepted (Overall accept %) between 2012 and 2016. Submissions during the year of 2017 were excluded because we did not have sufficient data for full life-cycle of these manuscripts. Code to reproduce this figure can be found on the linked Github repository at the path figures/selectivity_over_time/selectivity_over_time.rmd.

names were manually disambiguated (for example, by normalizing names such as “USA” to “United States” and “Viet Nam” to “Vietnam”). To simplify continent-level comparisons, we also excluded one submission for which the corresponding author listed their affiliation as Antarctica.

Full submissions included 6,669 distinct gatekeepers, 5,694 distinct corresponding authors, 6,691 distinct first authors, and 5,581 distinct last authors. Authors were also likely to appear on multiple manuscripts and may have held a different authorship role in each: whereas our data included 17,966 distinct combinations of author name and role, this number comprised only 12,059 distinct authors. For 26.5 percent of full submissions the corresponding author was also the first author, whereas

for 71.2 percent of submissions the corresponding author was the last author. We did not have access to the full authorship list that included middle authors. Note that in the biosciences, the last author is typically the most senior researcher involved [407] and responsible for more conceptual work, whereas the first author is typically less senior and performs more of the scientific labor (such as lab work, analysis, etc.) to produce the study [208, 408, 409].

Gender assignment

Gender variables for authors and gatekeepers were coded using an updated version of the algorithm developed in [28]. This algorithm used a combination of the first name and country of affiliation to assign each author's gender on the basis of several universal and country-specific name-gender lists (e.g., United States Census). This list of names was complemented with an algorithm that searched Wikipedia for pronouns associated with names.

We validated this new list by applying it to a dataset of names with known gender. We used data collected from *RateMyProfessor.com*, a website containing anonymous student-submitted ratings and comments for professors, lecturers, and teachers for professors at universities in the United States, United Kingdom, and Canada. We limited the dataset to only individuals with at least five comments, and counted the total number of gendered pronouns that appeared in their text; if the total of one gendered-pronoun type was at least the square of the other, then we assigned the gender of the majority pronoun to the individual. To compare with pronoun-based assignment, we assigned gender using the previously detailed first-name based algorithm. In total, there were 384,127 profiles on *RateMyProfessor.com* that had at least five comments and for whom pronouns indicated a gender. Our first name-based algorithm assigned a gender of male or female to 91.26 percent of these profiles. The raw match-rate between these two assignments was 88.6 percent. Of those that were assigned a gender, our first name-based assignment matched the pronoun assignment in 97.1 percent of cases, and 90.3 percent of distinct first names. While *RateMyProfessor.com* and the

authors submitting to *eLife* represent different populations (*RateMyProfessor.com* being biased towards teachers in the United States, United Kingdom, and Canada), the results of this validation provide some credibility to the first-name based gender assignment used here.

We also attempted to manually identify gender for all Senior Editors, Reviewing Editors, invited peer reviewers, and last authors for whom our algorithm did not assign a gender. We used Google to search for their name and institutional affiliation, and inspected the resulting photos and text in order to make a subjective judgment as to whether they were presenting as male or female.

Through the combination of manual efforts and our first-name based gender-assignment algorithm, we assigned a gender of male or female to 95.5 percent ($n = 35,511$) of the 37,198 name/role combinations that appeared in our dataset. 26.7 percent ($n = 9,910$) were assigned a gender of female, 68.8 percent ($n = 25,601$) were assigned a gender of male, while a gender assignment could be not assigned for the remaining 4.5 percent ($n = 1,687$). This gender distribution roughly matches the gender distribution observed globally across scientific publications [28]. A breakdown of these gender demographics by role can be found in Tables A.1 and A.2.

Gender composition of reviewers

To assess the relationship between author-gatekeeper gender homogeneity and review outcomes, we analyzed the gender composition of the gatekeepers and authors of full submissions. Each manuscript was assigned a reviewer composition category of *all-male*, *all-female*, *mixed*, or *uncertain*. Reviewer teams labeled *all-male* and *all-female* were teams for which we could identify a gender for every member, and for which all genders were identified as either male or female, respectively. Teams labeled as *mixed* were those teams for which we could identify a gender for at least two members, and which had at least one male and at least one female peer reviewer. Teams labeled as *uncertain* were those teams for which we could not assign a gender to every member and which were not mixed. A full submission was typically reviewed by two to three peer reviewers, which may or may

not explicitly include the Reviewing Editor. However, the Reviewing Editor was always involved in the review process of a manuscript, and so we always considered the Reviewing Editor as a member of the reviewing team. Of 7,912 full submissions, a final decision of accept or reject was given for 6,590 during the dates analyzed; of these, 47.7 percent ($n = 3,144$) were reviewed by all-male teams, 1.4 percent ($n = 93$) by all-female teams, and 50.8 percent ($n = 3,347$) by mixed-gender teams; the remaining six manuscripts had reviewer teams classified as uncertain and were excluded from further analysis.

Institutional Prestige

Institutional names for each author were added manually by *eLife* authors and were thus highly idiosyncratic. Many institutions appeared with multiple name variants (e.g., “UCLA”, “University of California, Los Angeles”, and “UC at Los Angeles”). In total, there were nearly 8,000 unique strings in the affiliation field. We performed several pre-processing steps on these names, including converting characters to lower case, removing stop words, removing punctuation, and reducing common words to abbreviated alternatives (e.g., “university” to “univ”). We used fuzzy-string matching with the Jaro-Winkler distance measure [410] to match institutional affiliations from *eLife* to institutional rankings in the 2016 *Times Higher Education World Rankings*. A match was established for 15,641 corresponding authors of initial submission (around 66 percent). Matches for last authors were higher: 5,118 (79 percent) were matched.

Institutions were classed into two levels of prestige: “top” institutions were those within the top 50 universities as ranked by the global *Times Higher Education* rankings. Institutions which ranked below the top 50, or which were otherwise unranked or which were not matched to a Times Higher Education ranking were labeled as “non-top”. One limitation of the Times Higher Education ranking as a proxy for institutional prestige is that these rankings cover only universities, excluding many prestigious research institutes. To mitigate this limitation, we mapped a small number of

well-known and prestigious biomedical research institutes to the “top” category, including: The Max Plank Institutes, the National Institutes of Health, the UT Southwestern Medical Center, the Memorial Sloan Cancer Medical Center, the Ragon institutes, and the Broad Institute.

Geographic distance

Latitude and longitude of country centroids were taken from Harvard WorldMap [411]; country names in the *eLife* and Harvard WorldMap dataset were manually disambiguated and then mapped to the country of affiliation listed for each author from *eLife* (for example, ”Czech Republic” from the *eLife* data was mapped to ”Czech Rep.” in the Harvard WorldMap data). For each initial submission, we calculated the geographic distance between the centroids of the countries of the corresponding author and Senior Editor; we call this the *corresponding author-editor geographic distance*. For each full submission, we calculated the sum of the geographic distances between the centroid of the last author’s country and the country of each of the reviewers. All distances were calculated in thousands of kilometers; we call this the *last author-reviewers geographic distance*.

Analysis

We conducted a series of χ^2 tests of equal proportion as well as several logistic regression models in order to assess the likelihood that an initial submission is encouraged and that a full submission is accepted, as a function of author and gatekeeper characteristics. We supply p-values and confidence intervals as a tool for interpretation; we generally maintain the convention of 0.05 as the threshold for statistical significance, though we also report and interpret values just outside of this range. When visualizing proportions, 95% confidence intervals are calculated using the definition $p \pm 1.96\sqrt{p(1-p)/n}$, where p is the proportion and n is the number of observations in the group. When conducting χ^2 tests comparing groups based on gender, we excluded submissions for which no gender could be identified. When conducting tests for gender and country homogeneity, we report 95% interval confidence intervals of their difference in proportion—we do not report confidence intervals

for tests involving more than two groups. Odds ratios and associated 95% confidence intervals are reported for logistic regression models. Data processing, statistical testing, and visualization was performed using R version 3.4.2 and RStudio version 1.1.383.

Having conducted an exploratory analysis of gender and country inequities in peer review with this univariate approach, we built a series of logistic regression models to investigate whether these differences could be explained by other factors. In each model, we used the submission's outcome as the response variable, whether that be encouragement (for initial submissions) or acceptance (for full submissions). For both initial and full submissions, we added control variables for the year of submission (measured from 0 to 5, representing 2012 to 2017, accordingly), the type of the submission (Research Article, Short Report, or Tools and Resources), and the institutional prestige of the author (top vs non-top). For full submissions, we also controlled for the gender of the first author. Mirroring the univariate analysis, we constructed two sets of models. The first set of models investigates the extent of peer review inequities based on author characteristics. We considered predictor variables for the gender and continent of affiliation of the corresponding author (for initial submissions), and the last author (for full submissions). For the second set of models, we investigated whether these inequities differed based on gender or country homogeneity between the author and the reviewer or editor. In addition to variables from the first model, we considered several approaches to capture the effect of gender-homogeneity between the author and reviewers on peer review inequity (see below). We also included variables for the corresponding author-editor geographic distance (for initial submissions), and last author-reviewers geographic distance (for full submission), and a dummy variable indicating whether this distance was zero; these variables serve as proxies for the degree of country homogeneity between the author and the editor or reviewers. There were a small number of Senior Editors in our data—in order to protect their identity we did not include their gender or specific continent of affiliation in any models; we maintained a variable for corresponding author-editor geographic distance.

Several approaches were considered for modeling the relationship between equity in peer review and the composition of the reviewer team using logistic regression. Approaches such as modelling equity using simple interaction terms or with a two-model approach were also considered but were ultimately excluded due to methodological and interpretive constraints (see Appendix A for more discussion of these models and their results). A third approach modelled equity across groups as a categorical variable consisting of all six combinations of last author gender (male, female) and reviewer team composition (all-male, all-female, mixed); This approach provides a more interpretable means of testing the extent to which gender equity in success rates was related to the interaction between author and reviewer team demographics, and was the focus of our analysis.

3.6 Results

Gatekeeper representation

We first analyzed whether the gender and countries of affiliation of the population of gatekeepers at *eLife* was similar to that of the authors of initial and full submissions. The population of gatekeepers comprised primarily of invited peer reviewers, as there were far fewer Senior and Reviewing Editors. A gender and country breakdown by gatekeeper type has been provided in Tables A.2, and A.3.

Fig 3.3 illustrates the gender and country demographics of authors and gatekeepers. The population of gatekeepers at *eLife* was largely male. Only 21.6 percent ($n = 1,440$) of gatekeepers were identified as female, compared with 26.6 percent ($n = 4,857$) of corresponding authors (includes authors of initial submissions), 33.9 percent ($n = 2,272$) of first authors, and 24.0 percent ($n = 1,341$) of last authors. For initial submissions, we observed a strong difference between the gender composition of gatekeepers and corresponding authors, $\chi^2(df= 1, n = 17,119) = 453.9, p \leq 0.00001$. The same held for full submissions, with a strong difference for first authorship, $\chi^2(df= 1, n = 6,153) = 844.4, p \leq 0.0001$; corresponding authorship, $\chi^2(df= 1, n = 6,647) = 330.04, p \leq 0.0001$; and last authorship, $\chi^2(df= 1, n = 5,292) = 17.7, p \leq 0.00003$. Thus, the gender proportions of

gatekeepers at *eLife* was male-skewed in comparison to the authorship profile.

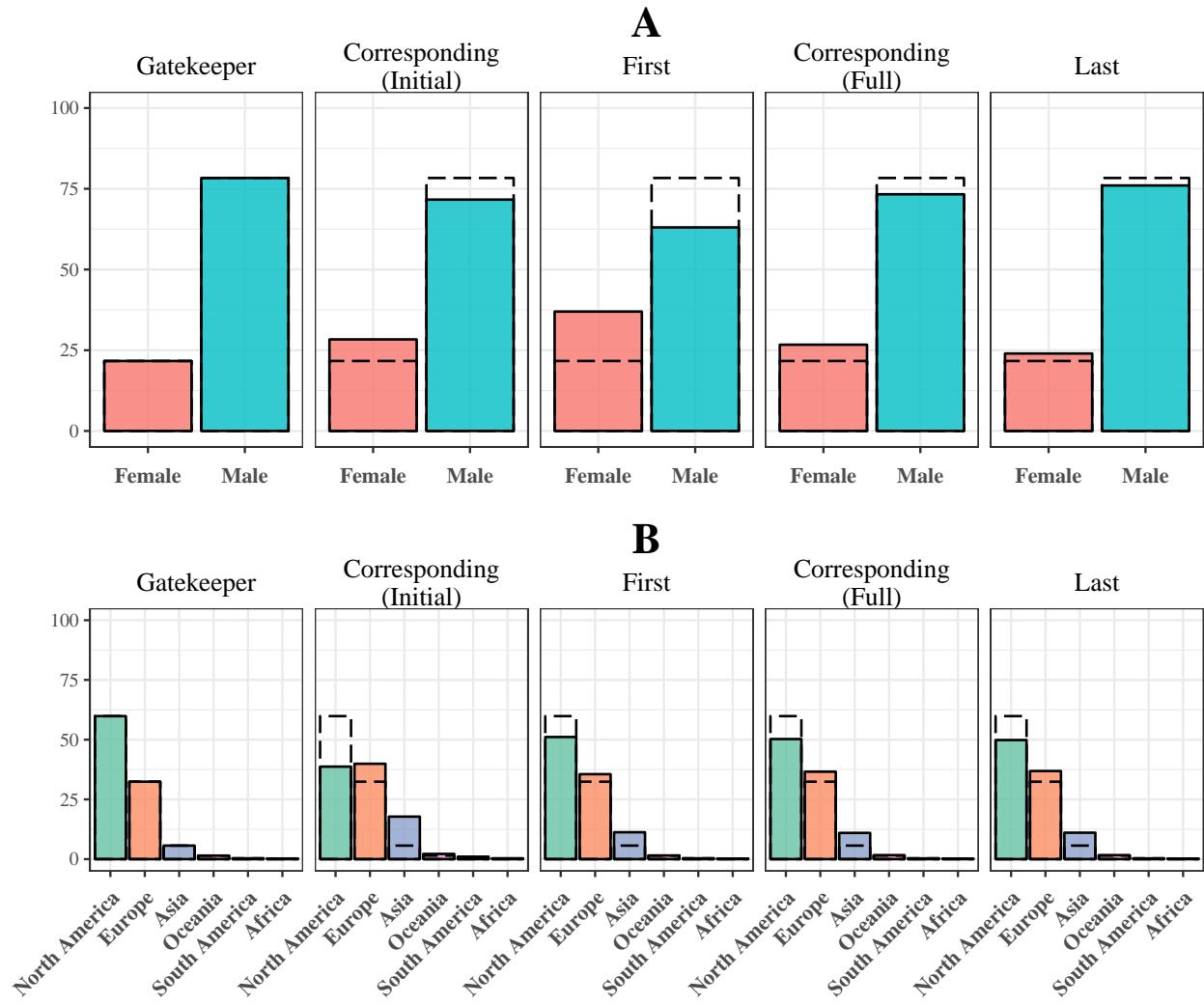


Figure 3.3: **Gender and country of affiliation demographics of authors and gatekeepers at *eLife*.**

A: proportion of identified men and women in the populations of distinct gatekeepers (Senior Editors, Reviewing Editors, and peer reviewers) and of the populations of distinct corresponding authors on initial submissions, and first, corresponding and last authors on full submissions; percentages exclude those for whom no gender was identified. **B:** proportion of people with countries of affiliation within each of six continents in the population of distinct gatekeepers, and for the population of distinct corresponding, first, and last authors. Black dashed lines overlaid on authorship graphs indicate the proportion of gatekeepers within that gendered or continental category. Values used in this graph can be found in Tables A.1 and A.4. Code to reproduce this figure can be found on the linked Github repository at the path figures/gatekeeper_representation/gatekeeper_representation.rmd.

The population of gatekeepers at *eLife* was heavily dominated by those from North America,

who constituted 59.9 percent ($n = 3,992$) of the total. Gatekeepers from Europe were the next most represented, constituting 32.4 percent ($n = 2,162$), followed by Asia with 5.7 percent ($n = 378$). Individuals from South America, Africa, and Oceania each made up less than two percent of the population of gatekeepers. As with gender, we observed differences between the country composition of gatekeepers and that of the authors. Gatekeepers from North America were over-represented whereas gatekeepers from Asia and Europe were under-represented for all authorship roles. For initial submissions, there was a significant difference in the distribution of corresponding authors compared to gatekeepers $\chi^2(df= 5, n = 18,195) = 6738.5, p \leq 0.00001$. The same held for full submissions, with a significant difference for first authors, $\chi^2(df= 5, n = 6,674) = 473.3, p \leq 0.00001$, corresponding authors, $\chi^2(df= 5, n = 6,669) = 330.04, p \leq 0.00001$, and last authors $\chi^2(df= 5, n = 5,595) = 417.2, p \leq 0.0001$. The international representation of gatekeepers was most similar to first and last authorship (full submissions), and least similar to corresponding authorship (initial submissions) due to country-level differences in acceptance rates (see Fig 3.4). We also note that the geographic composition of submissions to *eLife* has changed over time, attracting more submissions from authors in Asia in later years of analysis (see Fig. A.4).

Peer review success rates by author gender, country of affiliation

Male authorship dominated *eLife* submissions: men accounted for 76.9 percent ($n = 5,529$) of gender-identified last authorships and 70.7 percent ($n = 5,083$) of gender-identified corresponding authorships of full submissions (see Fig. A.3). First authorship of full submissions was closest to gender parity, although still skewed towards male authorship at 58.1 percent ($n = 4,179$).

We observed a gender inequity favoring men in the outcomes of each stage of the review processes. The percentage of initial submissions encouraged was 2.1 percentage points higher for male corresponding authors—30.83 to 28.75 percent, $\chi^2(df= 1, n = 22,319) = 8.95, 95\% \text{ CI} = [0.7, 3.4], p = 0.0028$ (see Fig. A.3). Likewise, the percentage of full submissions accepted was

higher for male corresponding authors—53.7 to 50.8 percent $\chi^2(df= 1, n = 6,188) = 3.95$, 95% CI = [0.03, 5.8], $p = 0.047$. The gender inequity at each stage of the review process yielded higher overall acceptance rates (the percentage of initial submissions eventually accepted) for male corresponding authors (15.6 percent) compared with female corresponding authors (13.8 percent), $\chi^2(df= 1, n = 21,670) = 10.96$, 95% CI = [0.8, 2.9], $p = 0.0009$ for a male:female success ratio of 1.13 to 1.

Gender disparity was only apparent in the senior authorship roles. Fig 3.4.A shows the gendered acceptance rates of full submissions for corresponding, first and last authors. There was little to no relationship between the gender of the first author and the percentage of full submissions accepted, $\chi^2(df= 1, n = 5,971) = 0.34$, 95% CI = [-1.8, 3.5], $p = 0.56$. There was, however, a significant gender inequity in full submission outcomes for last authors, as also observed for corresponding authors—the acceptance rate of full submissions was 3.5 percentage points higher for male as compared to female last authors—53.5 to 50.0 percent, $\chi^2(df= 1, n = 6,505) = 5.55$, 95% CI = [0.5, 6.4], $p = 0.018$.

Fig 3.4.B shows the proportion of manuscripts submitted, encouraged, and accepted to *eLife* from corresponding authors originating from the eight most prolific countries (in terms of initial submissions). Manuscripts with corresponding authors from these eight countries accounted for a total of 73.9 percent of all initial submissions, 81.2 percent of all full submissions, and 86.5 percent of all accepted publications. Many countries were underrepresented in full and accepted manuscripts compared to their submissions. For example, whereas papers with Chinese corresponding authors accounted for 6.9 percent of initial submissions, they comprised only 3.0 percent of full and 2.4 percent of accepted submissions. The only countries that were over-represented—making up a greater portion of full and accepted submissions than expected given their initial submissions—were the United States, United Kingdom, and Germany. In particular, corresponding authors from the United States made up 35.8 percent of initial submissions, yet constituted 48.5 percent of full

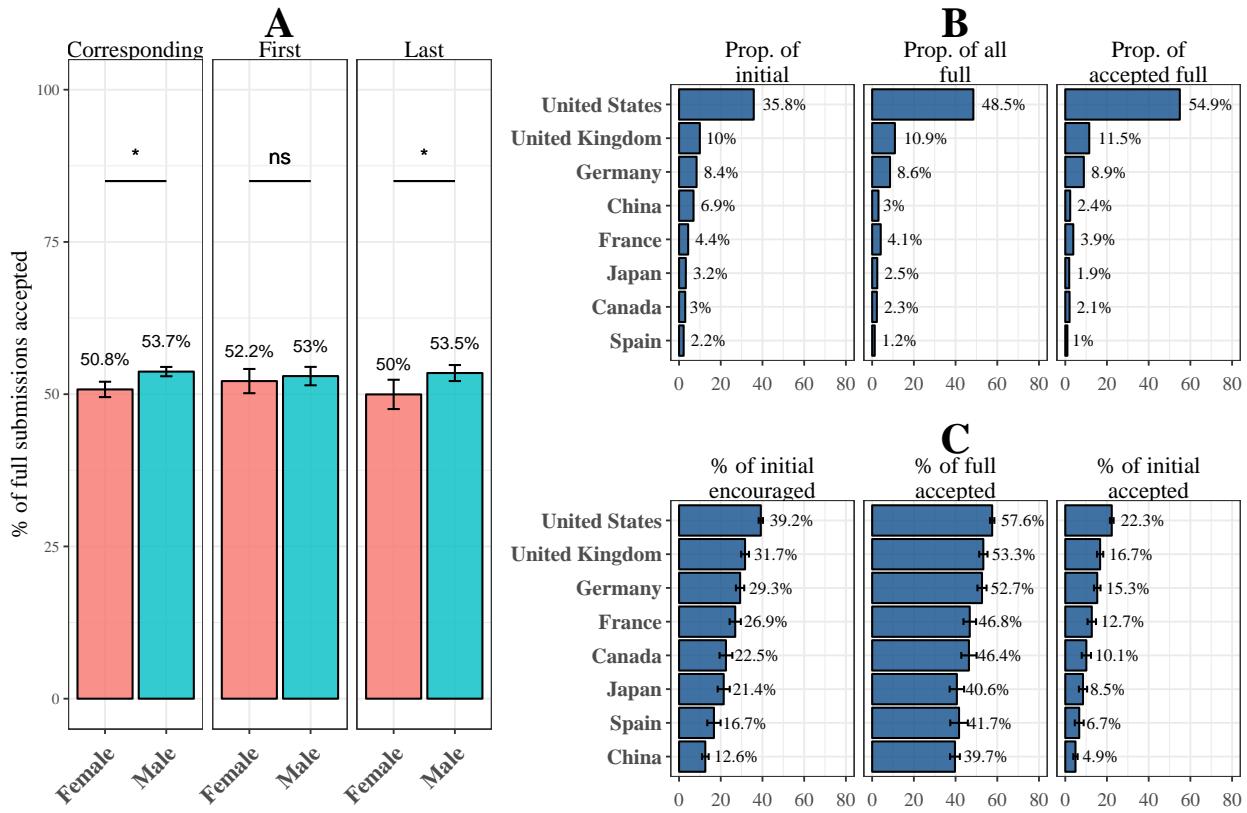


Figure 3.4: Peer review success rates by gender and country of authors.

A: Percentage of full submissions that were accepted, shown by the gender of the corresponding author, first author, and last author. Authors whose gender was unknown were excluded from analysis. See Fig A.3 for an extension of this figure including submission rates, encourage rates, and overall acceptance races. Error bars indicate 95% confidence intervals of the proportion of accepted full submissions. Asterisks indicate significance level of χ^2 tests of independence of frequency of acceptance by gender; "****" = $p < 0.001$; "***" = $p < 0.01$; ** = $p < 0.05$; "-" = $p < 0.1$; "ns" = $p \geq 0.1$. **B:** Proportion of all initial submissions, full submissions, and accepted full submissions by the country of affiliation of the corresponding author for the top eight most prolific countries in terms of initial submissions. **C:** Encourage rate of initial submissions, acceptance rate of full submissions, and acceptance rate of initial submissions by country of affiliation of the corresponding author for the top eight more prolific countries in terms of initial submissions. Error bars indicate 95% confidence intervals for each percentage. This same graph with the top 16 most prolific nations can be found in Fig A.7. Code to reproduce this figure can be found on the linked Github repository at the path figures/author_outcomes/submission_outcome_by_gender.rmd.

submissions and the majority (54.9 percent) of accepted submissions.

Each stage of review contributed to the disparity of country representation between initial, full, and accepted submissions, with manuscripts from the United States, United Kingdom, and Germany more often encouraged as initial submissions and accepted as full submissions. Fig 3.4.C

shows that initial submissions with a corresponding author from the United States were the most likely to be encouraged (39.2 percent), followed by the United Kingdom (31.7 percent) and Germany (29.3 percent). By contrast, manuscripts with corresponding authors from Japan, Spain, and China were comparatively less likely to be encouraged (21.4, 16.7, and 12.6 percent, respectively). These differences narrowed somewhat for full submissions: the acceptance rate for full submissions with corresponding authors from the U.S. was the highest (57.6 percent), though more similar to other countries, such as the United Kingdom, Germany, and France than for encourage rates.

There were gendered differences in submissions by country of affiliation (Fig. A.5), but there were insufficient data to test whether gender and country of affiliation interacted to affect the probability of acceptance.

The gender and country inequities evident in the univariate analyses were subsequently affirmed using a logistic regression model that controlled for a number of potential confounds (see Fig 3.5. We modeled peer review outcomes based on not only the gender and affiliated continent of the corresponding author (initial submissions) and last author (full submissions), but also the prestige of the author's institution, the year in which the manuscript was submitted, and the submission type (Research Article, Short Report, or Tools and Resources). For full submissions, we also controlled for the gender of the first author. The results of this regression for initial and full submissions are shown in Fig 3.5.

For both initial and full submissions, the prestige of the author's institution was the strongest predictor of a positive peer review outcome (initial: $\beta = 1.726$, 95% CI = [1.663, 1.789], $p \leq 0.0001$; full: $\beta = 1.379$, 95% CI = [1.272, 1.486], $p \leq 0.0001$). A more recent year of submission was associated with a lower odds of acceptance, (initial: $\beta = 0.918$, 95% CI = [0.894, 0.942], $p \leq 0.0001$; full: $\beta = 0.888$, 95% CI = [0.847, 0.929], $p \leq 0.0001$), reflecting the increasing selectivity of *eLife* (see Fig 3.2). Compared to Research Articles, both Short Reports, ($\beta = 0.742$, 95% CI = [0.638, 0.847], $p \leq 0.0001$), and Tools and Resources ($\beta = 0.740$, 95% CI = [0.567, 0.913], $p \leq$

0.0001) were less likely to have a positive review outcome at the initial submission stage.

Even when controlling for these variables, there were still inequities by the gender and country of affiliation of the author, affirming trends illustrated in Fig 3.4. Initial submissions with a male corresponding author were associated with a 1.12 times increased odds of being encouraged (95% CI = [1.048, 1.182], $p = 0.0014$), and full submissions with a male last author with a 1.14 times increased odds of being accepted (95% CI = [1.03, 1.26], $p = 0.025$), compared to submissions with female corresponding or last authors. In contrast, for the first author position, there was no significant difference in outcomes by gender. The logistic regression also provided evidence for the presence of geographic inequities, with lower odds of success for submissions with authors outside of North America. Compared to submissions with a corresponding author from North America, an initial submission with a corresponding author from Europe was 0.68 times as likely to be encouraged (95% CI = [0.3236, 0.783], $p \leq 0.0001$), and a corresponding author from Oceania was 0.56 times as likely to be encouraged (95% CI = [0.34, 0.78], $p \leq 0.0001$), followed by corresponding authors from Africa ($\beta = 0.53$, 95% CI = [-0.18, 1.088], $p = 0.027$), Asia ($\beta = 0.40$, 95% CI = [0.30, 0.49], $p \leq 0.0001$), and South America ($\beta = 0.21$, 95% CI = [-0.269, 0.679], $p \leq 0.0001$). Geographic disparities were also present, although less pronounced for full submissions, with significantly lower odds of acceptance for submissions with a last author from Europe ($\beta = 0.86$, 95% CI = [0.75, 0.97], $p = 0.008$) or Asia ($\beta = 0.59$, 95% CI = [0.41, 0.76], $p \leq 0.0001$) compared with North America.

Peer Review Outcomes by Author-Gatekeeper Homogeneity

The higher acceptance rates for male authors manifested largely from instances when the reviewer team was all male (Fig 3.6). When all reviewers were male, the acceptance rate of full submissions was about 4.7 percentage points higher for male compared to female last authors ($\chi^2 = 4.48$ (df=1, $n = 3,110$), 95% CI = [0.3, 9.1], $p = 0.034$) and about 4.4 points higher for male compared to

female corresponding authors (Fig. A.6; $\chi^2(df=1, n=2,974) = 3.97$, 95% CI = [0.1, 8.7] $p = 0.046$). For mixed-gender reviewer teams, the disparity in author success rates by gender was smaller and not statistically-significant. All-female reviewer teams were too rare to draw firm conclusions (only 81 of 6,509 processed full submissions), but in the few cases of all-female reviewer teams, there was a higher acceptance rate for female last, corresponding, and first authors that did not reach statistical significance. There was no significant relationship between first authorship gender and acceptance rates, regardless of the gender composition of the reviewer team. In sum, greater parity in outcomes was observed when gatekeeper teams contained both men and women. Notably, the acceptance rate for female authors was not lower for all-male reviewer teams compared with mixed reviewer teams, rather the gender disparity arose from a higher acceptance rate for submissions from male authors when they were reviewed by a team of all-male reviewers. We refer to this favoring by reviewers of authors sharing their same gender as *homophily*.

Homophily was also evident in the relationship between peer review outcomes and the presence of country homogeneity between the last author and reviewer. We defined last author-reviewer country homogeneity as a condition for which at least one member of the reviewer team (Reviewing Editor and peer reviewers) listed the same country of affiliation as the last author. We only considered the country of affiliation of the last author, since it was the same as that of the first and corresponding author for 98.4 and 94.9 percent of full submissions, respectively. Outside of the United States, the presence of country homogeneity during review was rare. Whereas 88.4 percent of full submissions with last authors from the U.S. were reviewed by at least one gatekeeper from their country, country homogeneity was present for only 29.3 percent of full submissions with last authors from the United Kingdom and 26.2 percent of those with a last author from Germany. The incidence of reviewer homogeneity fell sharply for Japan and China which had geographic homogeneity for only 10.3 and 9.9 percent of full submissions, respectively. More extensive details on the rate of author/reviewer homogeneity for each country can be found in Fig. A.5.

Last author-reviewer country homogeny tended to result in the favoring of submissions from authors of the same country as the reviewer. We first pooled together all authors from all countries ($n = 6,508$ for which there was a full submission and a final decision), and found that the presence of homogeny during review was associated with a 10.0 percentage point higher acceptance rate, (Fig 3.6.B; $\chi^2(1, n = 6,508) = 65.07$, 95% CI = [7.58, 12.47], $p \leq 0.00001$). However, most cases of homogeny occurred for authors from the United States, so this result could potentially reflect the higher acceptance rate for these authors (see Fig 3.4), rather than homophily overall. Therefore we repeated the test, excluding all full submissions with last authors from the United States, and we again found a significant, though statistically less confident homophilic effect, $\chi^2(df= 1, n = 3,236) = 4.74$, 95% CI = [0.52, 10.1], $p = 0.029$. We repeated this procedure again, excluding authors from both the United States and United Kingdom, (the two nations with the highest acceptance rates, see 3.4), and we identified no homophilic effect, $\chi^2(df= 1, n = 1,920) = 0.016$, 95% CI = [-4.6, 7.7] $p = 0.65$. Thus, the effects of last-author reviewer country-homophily were largely driven by the United States and United Kingdom.

For authors from outside the United States, not only was the presence of author-reviewer country homogeny rare, but the tendency for a homophilic effect on peer review outcome appeared to vary, depending on the country. Fig 3.6.C shows acceptance rates for last authors affiliated within the eight most prolific nations submitting to *eLife*. For the United States, presence of homogeny was associated with a 6.9 percentage point higher likelihood of acceptance compared to no homogeny $\chi^2(df= 1, n = 3,270) = 6.25$, 95% CI = [1.4, 12.4], $p = 0.0124$. Similarly, papers from the United kingdom were 8.0 percentage points more likely to be accepted if there was last author-reviewer homogeny $\chi^2(df= 1, n = 739) = 3.65$, 95% CI = [-0.1, 16.2], $p = 0.056$. In contrast, submissions with last authors from France were 23 percentage points *less* likely to be accepted if there was country homogeny $\chi^2(df= 1, n = 204) = 4.34$, 95% CI = [-42.8, -3.4], $p = 0.037$. There was a similar, though non-significant effect for Canada and Switzerland (also French-speaking countries).

Due to the rarity of country homogeneity outside of the U.S., more data are needed to draw firm conclusions on a per-country basis.

To further assess the contribution of author-reviewer homogeneity to inequity in peer review outcomes, we extended the logistic regression approach shown in Fig 3.5. For full submissions, we compared two logistic regression models, one that considered author-reviewer geographic homogeneity but only main effects of reviewer team gender composition (Fig 3.7.A) and one that included terms to model the effects of author-reviewer geographic and gender homogeneity (Fig 3.7.B). To model the extent to which gender equity differed based on the gender composition of the reviewer team, we modelled interactions using a variable combining factor levels for last author gender and reviewer team composition (Fig 3.7.B). To model the degree of country homogeneity between the author and the author and the reviewers, we included in the model the last author-reviewers geographic distance, defined as the sum of the geographic distance between the centroids of the last author's country, and the country of all of the peer reviewers. All distances were calculated in thousands of kilometers; for example, the geographic distance between the United States and Denmark is 7.53 thousands of kilometers. We included a dummy variable indicating whether the distance was zero. A similar analysis was performed to assess the effect of author-editor homogeneity on the outcomes of initial submissions (Table A.8); this excludes any analysis of homophily between the author and Senior Editor in order to protect the identity of the small number of Senior Editors.

Fig 3.7.A shows that there were similar main effects of author gender and country, in terms of direction and magnitude, as in Fig 3.5.B. Even after controlling for reviewer team composition, a full submission with a male last author was 1.14 times more likely to be accepted than a submission with a female last author (95% CI = [1.020, 1.256], $p = 0.032$). In addition, there were inequities based on author continent of affiliation, although smaller than in Fig 3.5.B. Affiliation within Asia was associated with a 0.779 times reduced odds of acceptance compared to North America (95% CI = [0.565, 0.992], $p = 0.022$)—a smaller effect size than the 0.585 times reduced odds observed

in Fig 3.5.B. Submissions with a last author from Oceania were associated with a 1.494 times increased odds of acceptance compared to North America, though with wide confidence intervals (95% CI = [1.020, 1.968], $p = 0.097$); this diverges from the non-significant negative effect observed in Fig 3.5. The effect of control variables—submission year, submission type, author institutional prestige, and first author gender—were also similar to those in Fig 3.5.

The extended model in Fig 3.7.A revealed a main effect of reviewer team gender composition. Compared to mixed-gender reviewer teams, submissions reviewed by all-male reviewers were 1.15 times more likely to be accepted (95% CI = [1.051, 1.252], $p = 0.0059$); there was no significant difference between all-female and mixed-gender teams. In addition, this model revealed an influence of author-reviewer geographic homogeneity. Every 1000km of last author-reviewer distance was associated with a 0.988 times lower odds of acceptance (95% CI = [0.982, 0.994], $p \leq 0.0001$). This negative effect of last author-reviewers geographic distance provides additional evidence for the observations from Fig 3.6—that homogeneity between the author and reviewers was associated with a greater odds of acceptance, even when controlling for the continent of affiliation of the author and other characteristics of the author and submission. A last author-reviewers geographic distance of zero (indicating that all reviewers were from the same country as the corresponding author) was not associated with a strong effect beyond that predicted by distance.

Finally, we modelled interactions between last author gender and reviewer-team composition by combining them into a single categorical variable containing all six combinations of factor levels (Fig 3.7.B). Full submissions with a male last author and which were reviewed by a team of all-male reviewers was associated with a 1.22 times higher odds of being accepted than a full submission with a female last author that was reviewed by an all male team (95% CI = [1.044, 1.40], $p = 0.027$). No significant differences were observed for other combinations of author gender and reviewer gender composition. The absolute difference in parameter estimates between male and female authors among mixed-gender teams (0.084) was less than half that of all-male reviewer teams (0.198), sug-

gesting greater equity among submissions reviewed by mixed-gender teams than by all-male teams. Taken together, these findings suggest that gender inequity in peer review outcomes tended to be smaller for mixed-gender reviewer teams, even controlling for many potentially confounding factors. These results provide evidence affirming observations from the univariate analysis in Fig 3.6.

3.7 Discussion

We identified inequities in peer review outcomes at *eLife*, based on the gender and country of affiliation of the senior (last and corresponding) authors. Acceptance rates were higher for male than female last authors. In addition, submissions from developed countries with high scientific capacities tended to have higher success rates than others. These inequities in peer review outcomes could be attributed, at least in part, to a favorable interaction between gatekeeper and author demographics under the conditions of gender or country homogeneity; we describe this favoring as *homophily*, a preference based on shared characteristics. Gatekeepers were more likely to recommend a manuscript for acceptance if they shared demographic characteristics with the authors, demonstrating homophily. In particular, manuscripts with male (last or corresponding) authors had a significantly higher chance of acceptance than female (last or corresponding) authors when reviewed by an all male review team. Similarly, manuscripts tended to be accepted more often when at least one of the reviewers was from the same country as the corresponding author (for initial submissions) or the last author (for full submissions), though there may be exceptions on a per-country basis (such as France and Canada). We followed our univariate analysis with a regression analysis, and observed evidence that these inequities persisted even when controlling for potentially confounding variables. The differential outcomes on the basis of author-reviewer homogeneity is consistent with the notion that peer review at *eLife* is influenced by some form of bias—be it implicit bias [242, 368], geographic or linguistic bias [245, 385, 412], or cognitive particularism [395]. Specifically, homophilic interaction suggests that peer review outcomes may sometimes be associated with factors

other than the intrinsic quality of a manuscript, such as the composition of the review team.

The opportunity for homophilous interactions is determined by the demographics of the gatekeeper pool, and the demographics of the gatekeepers differed significantly from those of the authors, even for last authors, who tend to be more senior [208, 407–409]. The underrepresentation of women at *eLife* mirrors global trends—women comprise a minority of total authorships, yet constitute an even smaller proportion of gatekeepers across many domains [31, 377, 413–419]. Similarly, gatekeepers at *eLife* were less geographically diverse than their authorship, reflecting the general underrepresentation of the “global south” in leadership positions of international journals [30].

The demographics of the reviewer pool made certain authors more likely to benefit from homophily in the review process than others. Male lead authors had a nearly 50 percent chance of being reviewed by a homophilous (all-male), rather than a mixed-gender team. In contrast, because all-female reviewer panels were so rare (accounting for only 81 of 6,509 full submission decisions), female authors were highly unlikely to benefit from homophily in the review process. Similarly, U.S. authors were much more likely than not (see Table A.5) to be reviewed by a panel with at least one reviewer from their country. However, the opposite was true for authors from other countries. Fewer opportunities for such homophily may result in a disadvantage for scientists from smaller and less scientifically prolific countries.

Increasing representation of women and scientists from a more diverse set of nations among *eLife*’s editor may lead to more diverse pool of peer reviewers and reviewing editors and a more equitable peer review process. Editors often invite peer reviewers from their own professional networks, networks that likely reflect the characteristics of the editor [420–422]; this can lead to editors, who tend to be men [31, 377, 413–419] and from scientifically advanced countries [30] to invite peer reviewers who are demographically similar to themselves [111, 423, 424], inadvertently excluding certain groups from the gatekeeping process. Accordingly, we found that male Reviewing Editors at *eLife* were less likely to create mixed-gender teams of gatekeepers than female Reviewing

Editors (see Fig A.8). We observed a similar effect based on the country of affiliation of the Reviewing Editor and invited peer reviewers (see Fig A.9). Moreover, in Table A.12 we conducted a regression analysis considering only the gender of the Reviewing Editor, rather than the composition of the reviewer team; we found similar homophilous relationships as in Fig 3.7, suggesting the importance of the reviewing editor to the peer review process at *eLife*.

The size of disparities we observed in peer review outcomes may seem modest; however these small disparities accumulate through each stage of the review process (initial submission, full submission, revisions). These cumulative effects yield an overall acceptance rate (the rate at which initial submissions were eventually accepted) for male and female corresponding authors of 15.6 and 13.8 percent respectively; in other words, manuscripts submitted to *eLife* with male corresponding were published at a rate 1.13 times the rate of those with female corresponding authors. Similarly, initial submissions by corresponding authors from China were accepted at only 22.0 percent the rate of manuscripts submitted by corresponding authors from the United States (with overall acceptance rates of 4.9 and 22.3 percent, respectively). Success in peer review is vital for a researcher's career because successful publication strengthens their professional reputation and makes it easier to attract funding, students, postdocs, and hence further publications. Even small advantages can compound over time and result in pronounced inequalities in science [67, 259, 270, 425].

Our finding that the gender of the last authors was associated with a significant difference in the rate at which full submissions were accepted at *eLife* stands in contrast with a number of previous studies of journal peer review that reported no significant difference in outcomes of papers submitted by male and female authors [247, 426, 427], or differences in reviewer's evaluations based on the author's apparent gender [428]. This discrepancy may be explained in part by *eLife*'s unique context, policies, or the relative selectivity of *eLife* compared to journals where previous studies found gender equity. In addition, our results point to a key feature of study design that may account for some of the differences across studies: the consideration of multiple authorship

roles. This is especially important for the life sciences, for which authorship order is strongly associated with contribution [408, 409, 429]. Whereas our study examined the gender of the first, last, and corresponding authors, most previous studies have focused on the gender of the first author (e.g., [369, 430]) or of the corresponding author (e.g., [381, 431]). Consistent with previous studies, we observed no strong relationship between first author gender and review outcomes at *eLife*. Only when considering lead authorship roles—last authorship, and to a lesser extent, corresponding author, did we observe such an effect. Our results may be better compared with studies of grant peer review, where leadership roles are more explicitly defined, and many studies have identified significant disparities in outcomes favoring men [210, 379, 432–434], although many other studies have found no evidence of gender disparity [380, 382, 383, 435–437]. Given that science has grown increasingly collaborative and that average authorship per paper has expanded [50, 438], future studies of disparities would benefit from explicitly accounting for multiple authorship roles and signaling among various leadership positions on the byline [48, 407].

The relationship we found between the gender and country of affiliation of gatekeepers and peer review outcomes also stands in contrast to the findings from a number of previous studies. Studies of gatekeeper country of affiliation have found no difference in peer review outcomes based on the country of affiliation or country of affiliation of the reviewer [439, 440], though there is little research on the correspondence between author and reviewer gender. One study identified a homophilous relationship between female reviewers and female authors, [441]. However, most previous analyses found only procedural differences based on the gender of the gatekeeper [381, 427, 428, 442] and identified no difference in outcomes based on the interaction of author and gatekeeper gender in journal submissions [427, 439, 443] or grant review [382]. One past study examined the interaction between U.S. and non-U.S. authors and gatekeepers, but found an effect opposite to what we observed, such that U.S. reviewers tended to rate submissions of U.S. authors more harshly than those of non-U.S. authors [38]. Our results also contrast with the study most

similar to our own, which found no evidence of bias related to gender, and only modest evidence of bias related to geographic region [369]. These discrepancies may result from our analysis of multiple author roles rather than considering only the characteristics of the first author. Alternatively, they may result from the unique nature of *eLife*'s consultative peer review; the direct communication between peer reviewers compared to traditional peer review may render the social characteristics of reviewers more influential.

Limitations

There are limitations of our methodology that must be considered. First, we have no objective measure of the intrinsic quality of manuscripts. Therefore, it is not clear which review condition (homophilic or non-homophilic) more closely approximates the ideal of merit-based peer review outcomes. Second, measuring the relationship between reviewer and author demographics on peer review outcomes cannot readily detect biases that are shared by all reviewers/gatekeepers (e.g., if all reviewers, regardless of gender, favored manuscripts from male authors); hence, our approach could underestimate the influence of bias. Third, our analysis is observational, so we cannot establish causal relationships between success rates and authors or gatekeeper demographics—there remain potential confounding factors that we were unable to control for in the present analysis, such as the gender distribution of submission by country (see **Proportion of women corresponding authors by country**). Proportion of female corresponding authors on initial submissions for each country having more than 200 initial submissions during the period of study. Code to reproduce this figure can be found on the linked Github repository at the path figures/general_infromation/supp_gender_prop_by_country.rmd). Along these lines, the reliance on statistical tests with arbitrary significance thresholds may provide misleading results (see [444]), or obfuscate statistically weak but potentially important relationships. Fourth, our gender-assignment algorithm is only a proxy for author gender and varies in reliability by continent.

Further studies will be required to determine the extent to which the effects we observed generalize to other peer review contexts. Specific policies at *eLife*, such as their consultative peer review process, may contribute to the effects we observed. Other characteristics of *eLife* may also be relevant, including its level of prestige [211], and its disciplinary specialization in the biological sciences, whose culture may differ from other scientific and academic disciplines. It is necessary to determine the extent to which the findings here are particularistic or generalizable; it may also be useful in identifying explanatory models. Future work is necessary to confirm and expand upon our findings, assess the extent to which they can be generalized, establish causal relationships, and mitigate the effects of these methodological limitations. To aid in this effort, we have made as much as possible of the data and analysis publicly available at: <https://github.com/murrayds/elife-analysis>.

Conclusion and recommendations

Many factors can contribute to gender, country, and other inequities in scientific publishing[399, 401, 445–448], which can affect the quantity and perceived quality of submitted manuscripts. However, these structural factors do not readily account for the observed effect of gatekeeper-author demographic homogeneity associated with peer review outcomes at *eLife*; rather, relationships between the personal characteristics of the authors and gatekeepers are likely to play some role in peer review outcomes.

Our results suggest that it is not only the form of peer review that matters, but also the composition of reviewers. Homophilous preferences in evaluation are a potential mechanism underpinning the Matthew Effect [79] in academia. This effect entrenches privileged groups while potentially limiting diversity, which could hinder scientific advances, since diversity may lead to better working groups [449] and promote high-quality science [25, 450]. Increasing gender and international representation among scientific gatekeepers may improve fairness and equity in peer review outcomes and accelerate scientific progress. However, this must be carefully balanced to avoid overburdening

scholars from minority groups with disproportionate service obligations.

Although some journals and publishers, such as *eLife* and Frontiers Media, have begun providing peer review data to researchers (see [111, 451]), data on equity in peer review outcomes is currently available only for a small fraction of journals and funders. While many journals collect these data internally, they are not usually standardized or shared publicly. One group, *PEERE*, authored a protocol for open sharing of peer review data [452, 453], though this protocol is recent, and the extent to which it will be adopted remains uncertain. Watchdog groups, such as BiasWatchNeuro, are now tracking and posting the representation of women authors in some journals. To both provide better benchmarks and to incentivize better practices, journals should make analyses on author and reviewer demographics publicly available. These data include, but would not be limited to, characteristics such as gender, race, sexual orientation, seniority, and institution and country of affiliation. It is likely that privacy concerns and issues relating to confidentiality will limit the full availability of the data; but analyses that are sensitive to the vulnerabilities of smaller populations should be conducted and made available as benchmarking data. As these data become increasingly available, systematic reviews can be useful in identifying general patterns across disciplines and countries.

Some high-profile journals have experimented with implementing double-blind peer review as a potential solution to inequities in publishing, including *Nature* [454] and *eNeuro* [376], though in some cases with low uptake [455]. Our findings of homophilic effects may suggest that single-blind review is not the optimal form of peer review; however, our study did not directly test whether homophily persists in the case of double blind review. If homophily is removed in double-blind review, it would reinforce the interpretation of bias; if it is maintained, it would suggest other underlying attributes of the manuscript that may be contributing to homophilic effects. Double-blind peer review is viewed positively by the scientific community [456, 457], and some studies have found evidence that double-blind review mitigates inequities that favor famous authors, elite

institutions [246, 247, 458], and those from high-income and English-speaking nations [234].

There may be a tension, however, in attempting to further double blind peer review while other aspects of the scientific system become more open. More than 20 percent of *eLife* papers that go out for review, for example, are already available as preprints, which complicates the possibility of truly blind review. To a lesser extent, several statements required for the responsible conduct of research—e.g., conflicts of interest, funding statements, and other ethical declarations—would require altered administrative treatment to implement double blind review. Other options involve making peer review more open—one recent study showed evidence that more open peer review did not compromise the integrity or logistics of the process, so long as reviewers could maintain anonymity [459].

Other alternatives to traditional peer review have also been proposed, including study pre-registration, consultative peer review, and hybrid processes (eg: [324, 325, 460–463]), as well as alternative forms of dissemination, such as preprint servers (e.g., arXiv, bioRxiv) which have in recent years grown increasingly popular [464]. Currently, there is little empirical evidence to determine whether these formats constitute more equitable alternatives [368]. In addition, some journals are analyzing the demographics of their published authorship and editorial staff in order to identify key problem areas, focus initiatives, and track progress in achieving diversity goals [377, 423, 426]. More work should be done to study and understand the issues facing peer review and scientific gatekeeping in all its forms and to promote fair, efficient, and meritocratic scientific cultures and practices. Editorial bodies should craft policies and implement practices to mitigate disparities in peer review; they should also continue to be innovative and reflective about their practices to ensure that papers are accepted on scientific merit, rather than particularistic characteristics of the authors.

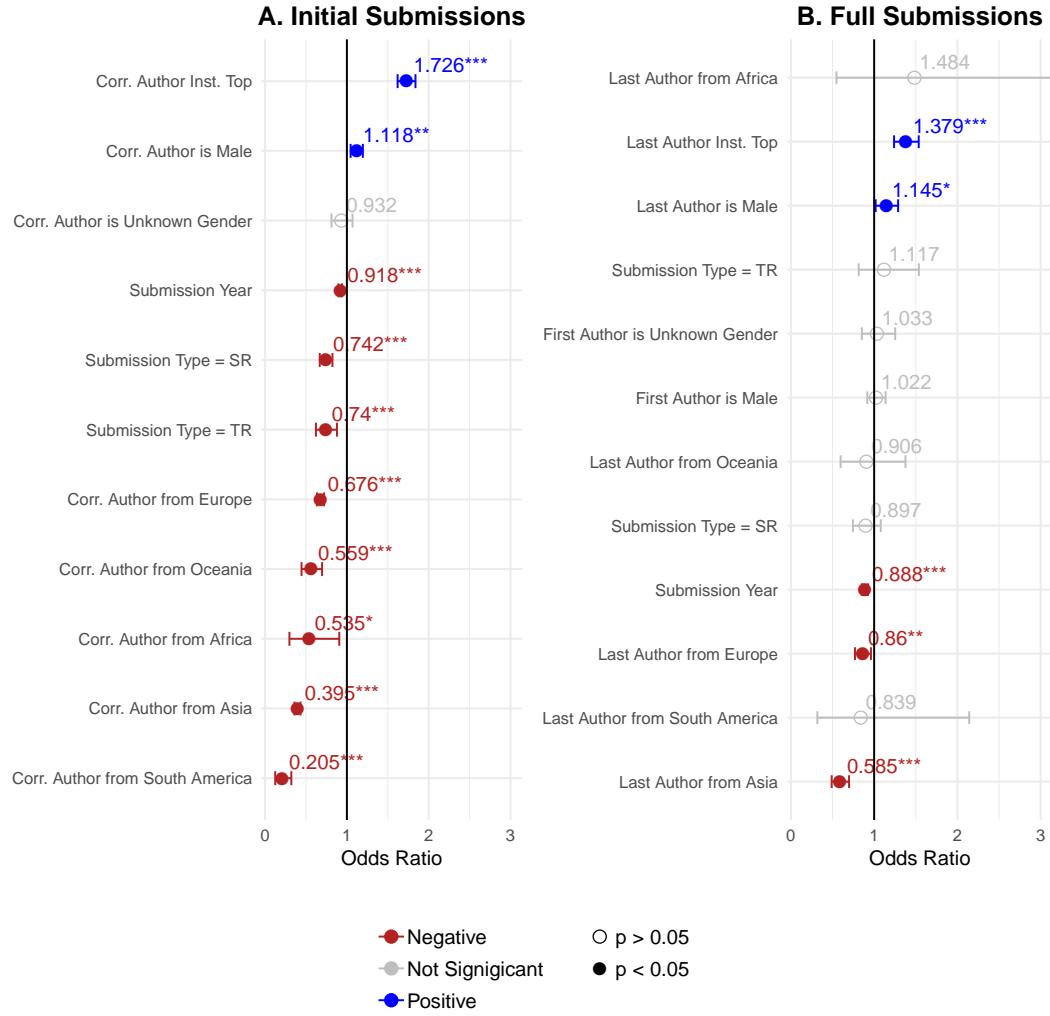


Figure 3.5: Modelling success rates of initial and full submissions based on author characteristics.

A: Estimates of a logistic regression model of initial submissions using whether the submission was encouraged as the response variable, and available information on the corresponding author as predictors. **B:** Estimates of a logistic regression model of full submissions using whether the submission was accepted as the response variable, and available information about the first and last authors as predictors. For both initial and full submissions, control variables included author's institutional prestige, the year of submission, and the submission type. For full submissions, there is also a control variable for the gender of the first author. For continent of affiliation, we held "North America" as the reference level. For submission type, "RA" (research article) was used as the reference level; the submission type "SR" means "Short Reports", and "TR" means "Tools and Resources". Blue, red, and grey points indicate positive, negative, and non-significant effects, respectively. The numbers above each point label the size of the effect, as an odds ratio. Bars extending from either side of each point indicate 95% confidence intervals. Asterisks next to each label indicate significance level: "***" = $p \leq 0.001$; "**" = $p \leq 0.01$; "*" = $p \leq 0.05$; otherwise, $p > 0.05$. Some confidence intervals are cropped; a table detailing full effects are included in Tables A.6, and A.7. Code to reproduce this figure can be found on the linked Github repository at the path figures/regression_analysis/regression_analysis_simple.rmd.

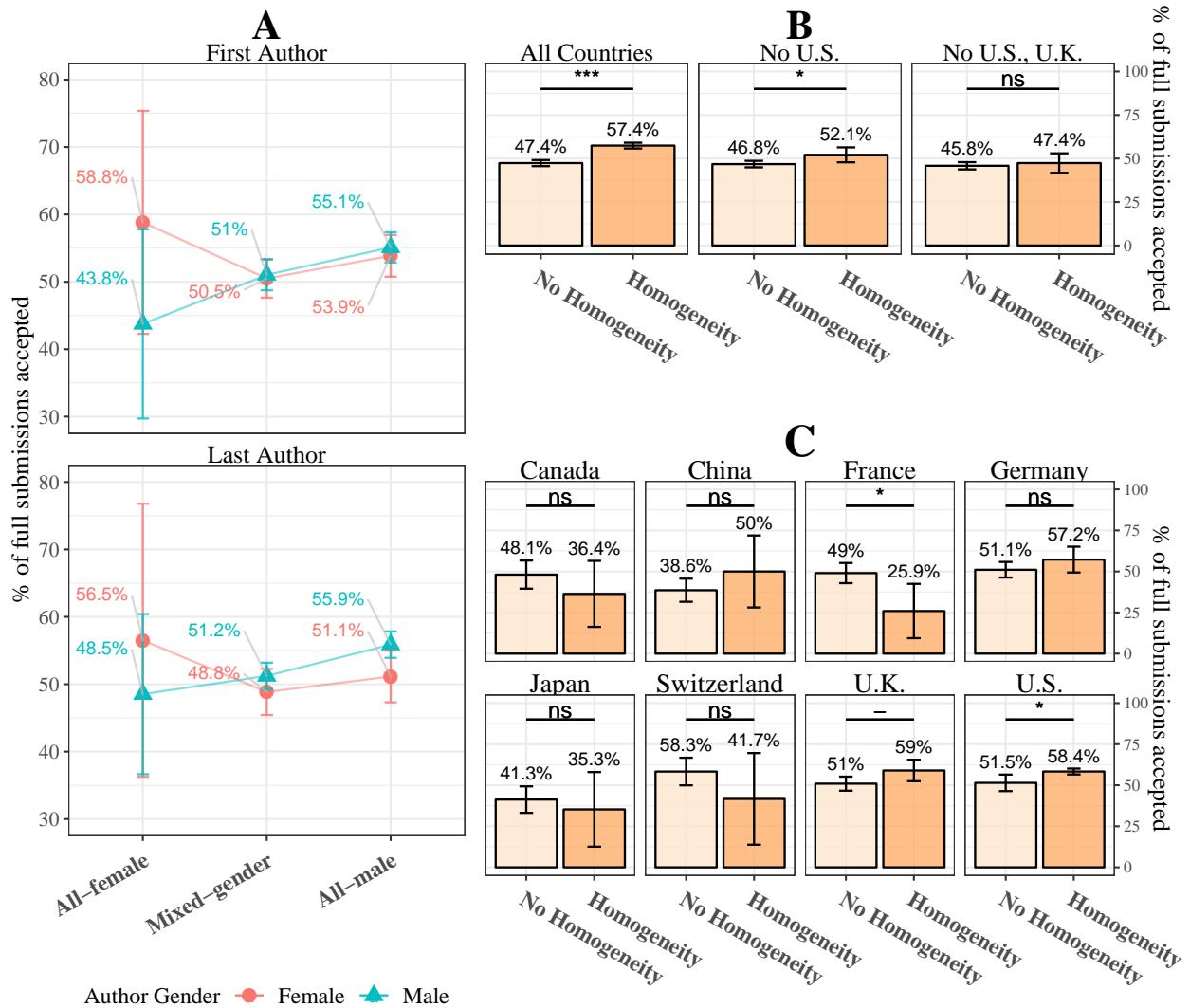


Figure 3.6: Relationship between author-reviewer homogeneity and peer review outcomes.

A: Percentage of full submissions that were accepted by gender of the first author (top) and last author (bottom), partitioned by the gender composition of the peer reviewers. The y-axis has been cropped between 30 percent and 80 percent in order to draw attention to the relevant effect. See Fig. A.6 for more information. **B:** Peer review outcome by presence of country homogeneity (last author from the same country as at least one reviewer) for all submissions (left), excluding submissions from the United States (middle) and excluding submissions from the United States and the United Kingdom, the two countries with the highest acceptance rates (right). **C:** Acceptance rate of full submissions by country homogeneity, shown for individual countries. Shown are the top eight most prolific countries in terms of number of initial submissions. For all panels: vertical error bars indicate 95% percentile confidence intervals. Values at the base of each bar indicate the number of observations within each group. Asterisks indicate significance level of χ^2 tests of independence comparing frequency of accepted full submissions between presence and absence of homogeneity and within each country. “***” = $p < 0.001$; “**” = $p < 0.01$; “*” = $p < 0.05$; “-” = $p < 0.1$; “ns” = $p \geq 0.1$. Code to reproduce this figure can be found on the linked Github repository at the path `figures/gatekeeper_author_outcomes/gatekeeper_author_outcomes.rmd`.

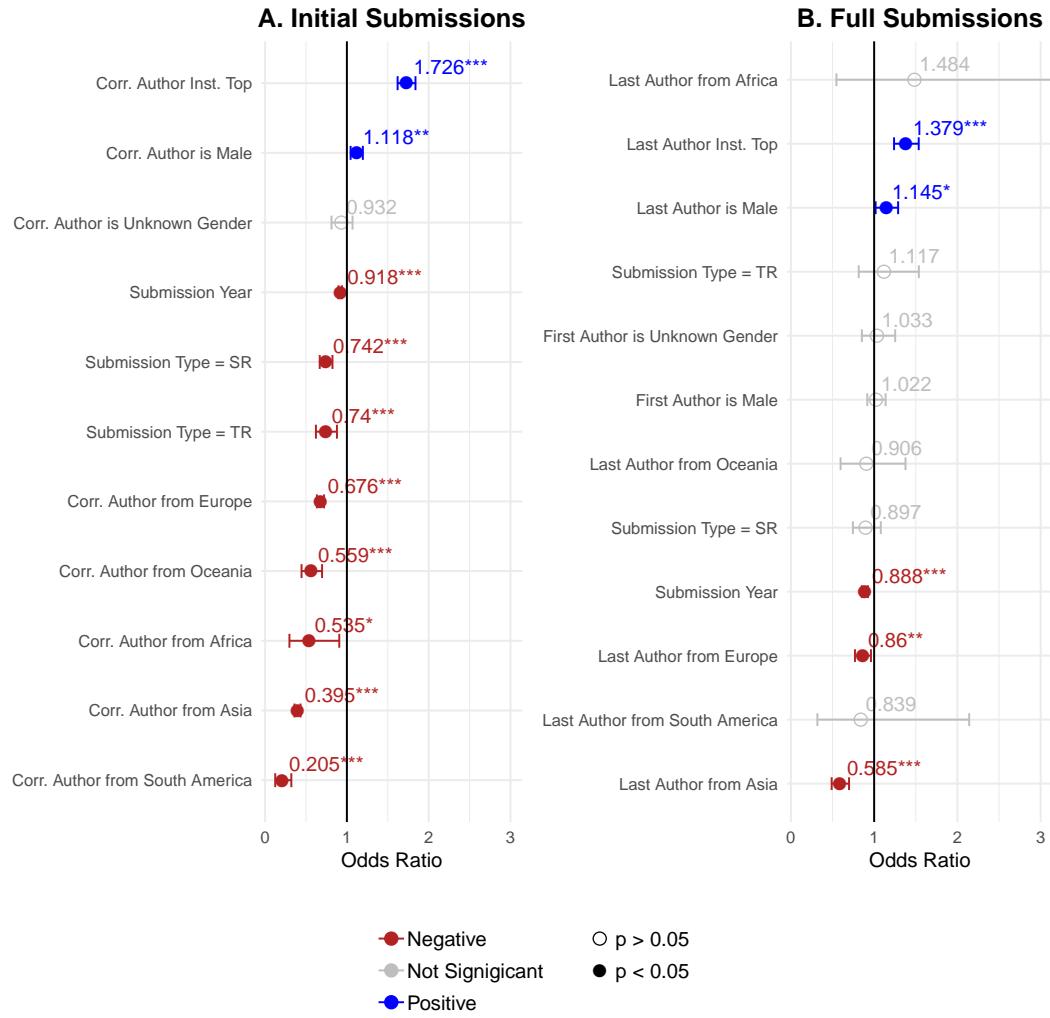


Figure 3.7: Modelling success of full submissions with author-reviewer homogeneity.
Estimates of logistic regression models of full submissions using whether the submission was accepted as the response variable. **A:** Includes as predictors the demographic and geographic characteristics of last author and gatekeepers, along with an indicator or the level of last author-reviewer geographic homogeneity. **B:** Includes all predictors as in **A** but with the last author gender and reviewer gender composition combined into a single, six-level categorical variable. Control variables for both panels include author's institutional prestige, year of submission, submission type, and gender of the first author. For continent of affiliation, "North America" was used as the reference level. For submission type, "RA" (research article) was used as the reference level; the submission type "SR" means "Short Reports", and "TR" means "Tools and Resources". For the combination variable of last author gender and reviewer team gender composition, we held "last author female—all rev. male" as the reference level. Blue and red points indicate positive and negative effects, respectively. The numbers above each point are the size of the effect as an odds ratio. Bars extending from either side of each point indicate 95% confidence intervals. Asterisks above each label indicate significance level: "***" = $p < 0.001$; "**" = $p < 0.01$; *" = $p < 0.05$; otherwise, $p > 0.05$. Some confidence intervals are cropped; a table detailing full effects is included in Table A.9. Code to reproduce this figure can be found on the linked Github repository at the path figures/regression_analysis/regression_analysis_interaction.rmd.

Chapter 4

Study 2: Student-teacher evaluations at U.S. Universities

4.1 Foreword

Your worst enemy is your best teacher.

The Buddha

Performance metrics have exploded in popularity alongside the rising availability of quantitative data about scholarly performance. Bibliometric indicators like the H-index [149], the journal impact factor [146], and university rankings [151] are widely used to evaluate and allocate resources towards individuals, journals, and institutions. Ideally, these metrics should provide a more objective measure of performance. Yet like peer review, these metrics have been challenged as being invalid [150], biased [266, 465], and for encouraging bad evaluative and scientific practices [258, 337, 338].

In this chapter, I explore the factors driving a metric that is so-far understudied in the Science of Science, but which plays a role in scientists' careers: student evaluations of teaching. Specifically, I draw on public data from *RateMyProfessors.com*, enriching it and identifying a subset of tenure and tenure-track faculty using a new dataset from the company *Academic Analytics*. With this merged data I observe how non-pedagogical demographic and aesthetic characteristics, such as an instructor's gender, ethnicity, accent, and attractiveness contribute to the ratings they receive from students, affirming previous accusations levied at student-teacher evaluations [466–469]. Moreover, the ratings appear driven more by the difficulty and work requirements of the course, than any measure of teaching ability. These findings demonstrate the issues of student-teacher evaluations at a scale and diversity larger than any previous analysis, illustrating why they should be re-imagined or abandoned, lest they continue to harm the careers of scientists and teachers.

Like the case of peer review, performance metrics can also be understood through the *complexity perspective* as a kind of feedback mechanism that perpetuates inequalities in science by hindering the careers of marginalized groups. Not only do metrics shape the composition of science, but they also affect the behavior of scientists. By creating an incentive structure, metrics like student-teacher evaluations or citation metrics can encourage *gaming*, whereby individuals optimize for the metric rather than what it is trying to measure [336]. Evaluation has consequences, and they can not always be predicted in advance; by recognizing the complexity of science, strategies can still be adopted to identify and mitigate the most deleterious effects of evaluation.

This study appears in the form that it was published to the journal *PLoS One* in 2020, under the title *Exploring the personal and professional factors associated with student evaluations of tenure-track faculty*. It was written in collaboration with Clara Boothby, Huimeng Zhao, Vanessa Minik, Nicolas Bérubé, Dr. Vincent Larivière, and Dr. Cassidy R. Sugimoto. This work was funded by the National Science Foundation award #1561299 (EAGER: Illuminating the role of science funding on disparities in science) awarded to Dr. Sugimoto. I also thank the company *Academic Analytics*, who provided us with data on Tenure and Tenure-Track faculty in the U.S., making this study possible.

4.2 Abstract

Tenure-track faculty members in the United States are evaluated on their performance in both research and teaching. In spite of accusations of bias and invalidity, student evaluations of teaching have dominated teaching evaluation at U.S. universities. However, studies on the topic have tended to be limited to particular institutional and disciplinary contexts. Moreover, in spite of the idealistic assumption that research and teaching are mutually beneficial, few studies have examined the link between research performance and student evaluations of teaching. In this study, we conduct a large scale exploratory analysis of the factors associated with student evaluations of teachers, controlling for heterogeneous institutional and disciplinary contexts. We source public student evaluations of teaching from *RateMyProfessor.com* and information regarding career and contemporary research performance indicators from the company *Academic Analytics*. The factors most associated with higher student ratings were the attractiveness of the faculty and the student's interest in the class; the factors most associated with lower student ratings were course difficulty and whether student comments mentioned an accent or a teaching assistant. Moreover, faculty tended to be rated more highly when they were young, male, White, in the Humanities, and held a rank of full professor. We observed little to no evidence of any relationship, positive or negative, between student evaluations of teaching and research performance. These results shed light on what factors relate to student evaluations of teaching across diverse contexts and contribute to the continuing discussion teaching evaluation and faculty assessment.

4.3 Introduction

Performance indicators have come to dominate faculty evaluations of teaching and research at universities in the United States, raising concerns over their consequences [333]. One of the most prominent indicators for teaching are student evaluations of teaching (SETs), in which students anonymously score and comment on their course instructors for the purpose of evaluation and

improvement. However, SETs alone are not sufficient for evaluation of tenure and tenure-track faculty for whom teaching constitutes only a portion of their professional responsibilities. Contemporary research universities are built on the premise that faculty balance research, service to the academic community, and teaching (see Boyer's model of scholarship [470]). Holistic faculty evaluation requires assessments along each of these dimensions and the faculty's ability to balance their commitments. However, quantitative studies of SETs typically have not examined teaching ratings in relation to faculty performance in other professional activities. Studies of SETs are also limited by the difficulty of aggregating data across institutional contexts, which has resulted in a poor understanding of the extent to which SETs depend on institutional and disciplinary factors. There is a need for a large-scale analysis of SETs to provide a more complete understanding of the extent to which these evaluations relate to personal or professional characteristics of teachers, institutional context, and research performance.

Questions of bias in SETs have prompted intense scrutiny and numerous studies on their validity. For example, past research on traditional SETs has identified biases based on gender [468, 469, 471–474], race [223, 469], attractiveness [475], and age [467, 474, 476]. Many have also criticized traditional SETs as invalid measures of teaching quality and student learning [466, 467, 471, 474, 477–480] and warned university administrators against using them for hiring and promotion decisions [481]. In light of these issues, there have been intensifying claims that SETs harm both students and faculty [482] and public calls to stop relying on them for evaluating teaching [483, 484]. In spite of this controversy, SETs have remained one of the most common metrics of teaching performance across a variety of U.S. universities [335]. Given their continued use for hiring and promotion, there remains a need to study the factors contributing to outcomes on SETs.

The *research-teaching nexus* refers to the relationship between time spent doing research, and time spent teaching. The Humboldtian ideal of a university is built on the premise that these tasks are mutually beneficial [485], and many have followed this tradition, positing a strong relationship

between research and teaching [486–489]. However, there is a lack of consensus surrounding the presence, extent, and nature of the nexus. While some studies have found evidence of *positive* research-teaching nexus—a mutually-beneficial relationship [490–492], other studies have instead observed a *negative* research-teaching nexus, suggesting that faculty incentive structures encourage research at the expense of teaching quality [493–495]. Conflicting with both the positive and negative nexus hypotheses, a landmark meta-analysis instead suggested a *neutral* research-teaching nexus, observing no evidence of a relationship between research and teaching [496]. Taken together, these studies offer no clear understanding of the research-teaching nexus; moreover, these studies have tended to be small and limited to particular institutional contexts. There remains a pressing need to understand the research-teaching nexus at scale and across institutional contexts.

In this study, we conduct a large-scale exploratory investigation of the extent to which demographic characteristics and research performance relate to SETs for tenured and tenure-track faculty in the United States. We leverage public teaching evaluations from *RateMyProfessor.com*, a public data source of public SETs which, despite criticism [497, 498], has been found to correlate with traditional evaluations [499–502]. We match these teaching evaluations with records from *Academic Analytics*, a research analytics company which provided us with a list of active tenured and tenure-track faculty in the United States, along with indicators of their number of publications, citations, grants, and professional awards. In performing this analysis, we hope to gain a more complete understanding of how individual, classroom, university characteristics, and research performance correlate with university faculty’s teaching evaluations. We also aim to shed light on the research-teaching nexus, the relationship between research and teaching.

4.4 Data and methods

Academic Analytics

Academic Analytics is a U.S. based company that sells access to their proprietary dataset of individual-level bibliometric indicators for use by university administrators in the United States and the United Kingdom to assess their departments. This data is derived from a mix of direct cooperation with research institutions and collection from publicly available sources such as institutional websites, CrossRef, and Federal agencies. We maintain a contract with Academic Analytics, through which we are granted a copy of their 2017 data release (AA2017).

The version of AA2017 used in this study contained demographic and bibliometric data for 165,666 tenure and tenure-track faculty at 399 universities and research institutions in the United States. AA2017 contains full names, departmental and institutional affiliations, year of doctoral attainment, and disciplinary classification. The dataset also included bibliometric indicators of recent scholarly performance: indexed publications produced in the previous five years; citations to those publications; grants held in the previous five years; lifetime professional awards won; and books published within the past ten years. Details and definitions of the relevant variables from AA2017 can be found in Table B.1.

RateMyProfessor.com

RateMyProfessor.com is a website offering students at institutions of higher education the opportunity to review their teachers and to read reviews by other students. Founded in 1999, the most recent version of *RateMyProfessor.com* allows students to anonymously review teachers along dimensions of overall quality, level of difficulty, and until recently, "hotness"—a binary rating implicitly associated with physical attractiveness (see Appendix B for discussion of rating types removed from the website). Ratings on *RateMyProfessor.com* have been found to correlate with

traditional student-evaluations of teachers (see Appendix B). Students are also encouraged to post comments to elaborate on their experience, and to select from a list of pre-defined “tags” that describe the common characteristics of the teacher and the course. Teachers, courses, and schools are all added to *RateMyProfessor.com* by users, and so the presence of any individual depends on the effort of students. Although the website has passed through many iterations, these core features have remained roughly consistent over time. *RateMyProfessor.com* remains one of the only and most popular large-scale, publicly available source of students’ evaluations of teachers, boasting “... more than 19 million ratings, 1.7 million professors and over 7,500 schools” [503]. We collected these data in January of 2018. Details and definitions of relevant variables from this data can be found in Tables B.2 and B.3.

Disciplinary aggregation

The AA2017 dataset used a hierarchical three-tiered disciplinary taxonomy, with the most granular tier consisting of 171 distinct classifications that were applied based on each individual’s departmental affiliation. When an individual held multiple affiliations or when a program was classified as more than one discipline, *Academic Analytics* duplicated their entire record, changing only their disciplinary classification. Thus, while there were 165,666 unique tenure and tenure track faculty represented in AA2017, 42,500 of these individuals had at least one duplicate record, which resulted in 225,877 total records.

To streamline the large variety of AA disciplinary classifications, we manually mapped each of the AA2017 171 detailed classifications to one of the five NSF classifications of research discipline: “*Natural Sciences*”, “*Medical Sciences*”, “*Social Sciences*”, “*Humanities*”, and “*Engineering*”. After we applied these broad disciplinary classifications, 16,254 individuals had duplicate records with distinct NSF classifications, compared to the 42,500 with distinct *Academic Analytics* classifications.

Processing research indicators

We added a new research indicator for each individual, *Publication Count*, which we defined as the sum of their indexed conference proceedings, book publications, and article publications; this combined indicator simplifies analysis, and captures the range of publications types that have distinct disciplinary distributions [332] (see distributions in Fig. B.1). The final indicators included the number of recent publications (5 years for articles and conferences, 10 for books), the number of citations to those recent publications, the number of grant dollars currently held, and the number of lifetime professional awards held by the individual. We field-normalized each AA2017 research indicator by the mean across the 171 granular disciplinary categories. This was performed for each record, normalizing by the mean of that record’s associated granular discipline. For example, if an individual published ten times within the past five years, and had two records, one for discipline A, with a field-mean of 5 publications, and one for discipline B with field-mean of 15 publications, then that individual’s records would have field-normalized scores of 2.0 and 0.667, respectively.

We also created discretized versions of each continuous field-normalized indicator of research performance. We binned each research indicator into an ordered factor containing a value of “None”, “Moderate”, or “High”. A classification of “None” meant that a count of zero is reported for that indicator. “Moderate” meant that the reported count is between the 1st and 90th percentile (inclusive) for that research indicator, calculated on the population of individuals who have a count greater than one. “High” meant that the reported count was above the 90th percentile of those with a count of at least one for that indicator. We performed this discretization because each field-normalized indicator is strongly zero-inflated and right-skewed (see the log-log distribution of indicators in Fig. B.1); these categories mitigated the impact of outliers and allowed for a clearer comparison between those with and without recent research activity.

Record matching

After the above pre-processing steps, we attempted to match records between the AA2017 and RMP2018 datasets. For each individual in AA2017, we attempted to find a likely match within RMP2018. After extensive experimentation and parameter tuning we settled on using Jaro-Winkler string distance [504–506] as the measure of distance between records. This measure offers flexibility to handle minor variation in instructor and department names. Distance between two strings is based on the number of character matches that occur in similar indexes in both strings, and includes a penalty factor that penalizes strings that have a mismatch within the first few characters. Given that this measure prioritizes matches early in the string, we format match strings for records in AA2017 and RMP2018 as follows,

[LAST NAME] [MIDDLE INITIAL] [FIRST NAME] [PROGRAM AFFILIATION]

where [PROGRAM AFFILIATION] is the “Program Name” variable in AA2017 and the “Department” variable in RMP (see Table B.2 and Table B.3 for descriptions of these variables). Using this format, *Jaro-Winkler* distance will tend to enforce strict similarity between last names while allowing for some increased variation in first names and department names. This is especially useful for faculty who use informal nicknames while teaching; for example, an individual in AA2017 with the match string “*Smith Robert Applied Mathematics*” results in a relatively high similarity score with an individual from RMP2018 with the name “*Smith Bob Applied Mathematics*”.

We calculated pairwise *Jaro-Winkler* string distances between the match strings for each individual in AA2017 and each profile from RMP2018. If the largest similarity metric between a record from AA2017 and any profile on RMP was lower than 0.1, then we excluded that individual from the final dataset. If at least one RMP profile has a similarity score above the threshold, then the most similar profile was selected as a match. This process resulted in 47,509 matches between individuals in AA2017 and RMP, representing 34.5 percent of AA2017 records, and 3.0 percent

of all RMP2018 records; this small population of matched RMP2018 records is expected because *RateMyProfessor.com* included non-tenured/non-tenure track faculty, faculty who are no longer active, and faculty from countries not represented in our version of AA2017.

A discussion of the representativeness and potential biases in our matching process can be found in Appendix B.

Gender assignment

We assigned a gender to each record in the matched dataset by comparing the number of masculine and feminine pronouns that appeared in text reviews left on faculty's profiles on *RateMyProfessor.com*. If the reviews of a profile contained more of one type of gendered pronouns than the square of the other, then we assigned their gender using the gender of the majority pronoun. For example, if one profile's reviews contained a total of ten masculine pronouns (e.g.: "he", "him", "himself"), but only three feminine pronouns (e.g.: "she", "her", "herself"), that profile would be assigned a gender of male ($10 > 3^2$); however if a profile contained four masculine and three feminine pronouns, then no gender was assigned ($3^2 > 4$). Using this method, we assigned a gender of male or female to 99.7 percent of tenure and tenure-track professors in the final matched dataset.

Race assignment

We infer a race for each individual in our dataset from their surname. We retrieved the dataset of surnames from the US Census, which contains, for each surname, the percentage of individuals having that name that are White, Black, Asian, Hispanic, Native American or Pacific Islander, and two or more races, as determined by the census. We adopt a conservative and coarse-grained approach to inferring race from these information; An individual in our dataset is assigned as *Likely White* when at least 70 percent of those having the same surname are White. Otherwise, an individual is assigned *Likely Non-White*. When an individual's surname does not appear in the Census dataset, then they are assigned a race of *Unknown*.

Final dataset

For those individuals in AA2017 who had duplicate records due to multiple affiliations, we selected one record at random and excluded others. We also removed records that were not assigned a value for their Scientific Age in AA2017 for which no gender could be assigned, and which had fewer than three reviews on *RateMyProfessor.com*. We excluded faculty who had fewer than five reviews in order to mitigate noise. The final matched dataset contained 18,946 records. Finally, we enriched these data with university characteristics from the 2018 Carnegie Classification of Higher Education Institutions. Analysis was conducted on a set of relevant variables extracted from the matched and enriched dataset. Descriptions of these final variables, identified following an extensive literature review of factors relevant to teaching performance, can be found in Table 1. These variables reflect a range of individual, classroom, university, and professional characteristics of the faculty and their teaching. These data, and the code for processing it, can be found at https://github.com/murrayds/aa_rmp.

4.5 Results

We fit a linear regression model with the overall teaching quality as the response, and all other variables from Table 4.1 as predictors. The resulting model had a R^2 of 0.514. Fig. 4.1.A visualizes the estimates of this regression (also shown in Table B.4). Because this is an exploratory analysis, we do not report p-values or significance levels for parameter estimates.

Table 4.1: **Description of final variables.** Extracted from *RateMyProfessor.com* (RMP2018), the 2017 version of *Academic Analytics* (AA2017), and the Carnegie Classification of Higher Education Institutions (Carnegie) for matched profiles

Variable	Source	Description
Overall Quality	RMP2018	The average of all 1-5 point reviews of overall quality left for a professor on <i>RateMyProfessor.com</i> between 2012 and 2017. Ratings are aggregated across all courses
Difficulty	RMP2018	The average of all 1-5 point reviews of difficulty left for a professor on <i>RateMyProfessor.com</i> between 2012 and 2017. Ratings are aggregated across all courses
Interest	RMP2018	The average of all 1-5 point reviews of student interest left for a professor on <i>RateMyProfessor.com</i> between 2012 and 2017. Original levels marked by an order set of five qualitative levels. These levels were mapped to values between 1 and 5 to accommodate numeric calculations. Ratings are aggregated across all courses
Number of reviews	RMP2018	The number of reviews left for the professor between 2012 and 2017. We use this as a control variable
Mentions Accent	RMP2018	True if the word "accent" appears at least once in the text of reviews for an individual
Mentions TA	RMP2018	True if the word "TA" or "Teaching Assistant" appears at least once in the text of reviews for an individual
Has Chili Pepper	RMP2018	True if the individual is given a "chili pepper" symbol, implicitly a rating of physical attractiveness
Gender	Mixed	Gender assigned to each individual of the dataset. Assigned using pronouns included in comments from RMP2018 data
Inferred Race	Mixed	Inferred race assigned to each individual in the dataset based on their family name.
Discretized: Citedness; Output; Awards Won; Grants Held	AA2017	Four variables: Citedness, scholarly output, awards won, and grants held. Each variable represents a count of recent field-normalized research items, categorized into three discrete groups. More detail on how each of these research items is counted by AA is included in supplementary information. Assigned category of "None" if no research item. Assigned "Moderate" if not None, and if between the 1st and 90th percentile (inclusive) of those with at least one of that research item; assigned "High" if greater than 90 th percentile
Scientific Age	AA2017	Number of years, in decades, since the individual obtained their terminal degree
Discipline	AA2017	High-level discipline of individual. One of Natural Sciences, Medical Sciences, Social Sciences, Engineering, or Humanities. In case a user was assigned to multiple disciplines, one was randomly selected
Rank	AA2017	The professional rank of the individual, coded as Associate, Assistant, or Full
Uni. Type	Carnegie	The classification of the research activity of the institution: R1 or Not R1
Uni. Control	Carnegie	The classification of the "control" of the institution that the individual is affiliated with: Public or Private

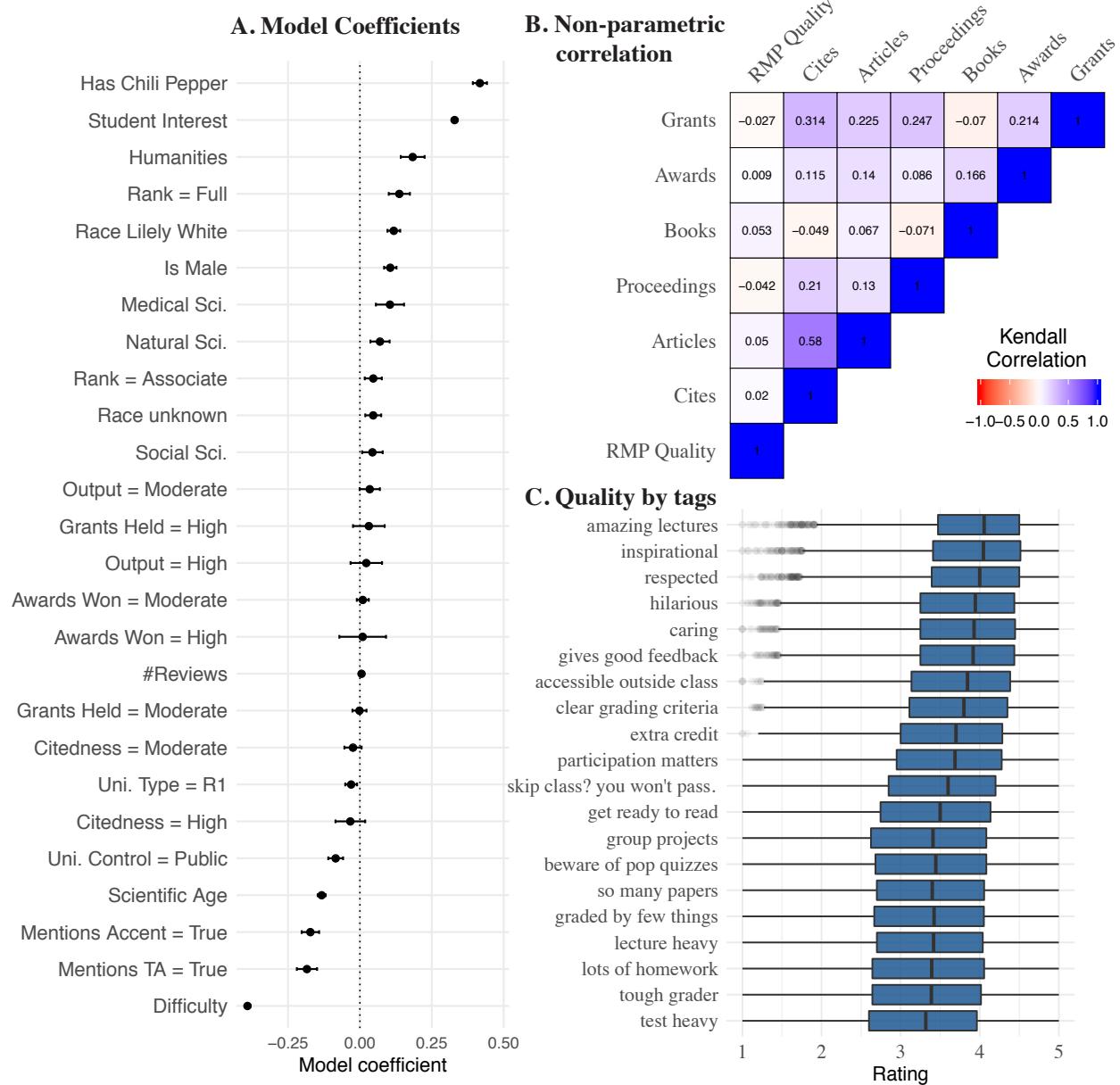


Figure 4.1: Individual, Classroom, University, and Research characteristics associated with overall teaching quality. **A.** Estimates of linear regression model using the overall teaching quality (continuous, 1-5) as the response and all variables from table 1 as the predictor variables. The x-axis corresponds to the estimate for each covariate, which are listed along the y-axis. For binary variables, “false” is always used as the reference level. For Gender, “female” is used as the reference. For race, “Non-White” is used as the reference. For “Rank”, “Assistant” is used as the reference. For Discipline, “Engineering” is set as the reference. For Uni. Control, “Private” is used as the reference. For Uni. Type, “Not R1” is used as the reference. For all research indicators, “Low” is used as the reference. Error bars surrounding each point correspond to the 95th percentile confidence intervals. Results are also shown in Table B.4. **B.** The non-parametric Kendall Rank Tau between research indicators and overall teaching quality. Values map to the correlation between 1 (correlated) and -1 (inversely correlated). Raw values for this test can be found in Table B.3. **C.** The distribution of overall teaching quality ratings for faculty possessing each of the pre-defined “tags” listed on their *RateMyProfessor.com* profile.

Several personal characteristics of faculty were associated with ratings of overall teaching quality in RMP2018. Presence of the “chili pepper” in RMP2018, which implies attractiveness, was associated with 0.41 point higher overall teaching quality ($\beta = 0.42$, 95% CI = [0.39, 0.44]); this was the largest positive estimate from the model. Compared to female faculty, male faculty were associated with 0.11 points greater overall teaching quality ($\beta = 0.11$, 95% CI = [0.08, 0.13]). Faculty having a commonly White surname were associated with 0.118 points greater overall teaching quality ($\beta = 0.12$, 95% CI = [0.10, 0.14]), whereas those with unknown race were associated with slightly higher ratings ($\beta = 0.05$, 95% CI = [0.019, 0.074]). Faculty who were mentioned as having an accent in a comment left on their RMP2018 profile were associated with 0.17 point lower ratings of overall quality than those for whom an accent was not mentioned ($\beta = -0.17$, 95% CI = [-0.20, -0.14]). Scientific age was negatively correlated with overall teaching quality such that each additional decade was associated with 0.13 point lower rating ($\beta = -0.13$, 95% CI = [-0.15, -0.12]). Professional rank had some association with ratings of overall teaching quality. Compared to assistant professors, full professors were associated with 0.14 point higher ratings of overall teaching quality ($\beta = 0.14$, 95% CI = [0.1, 0.17]); associate professors were associated with only 0.05 point higher ratings ($\beta = 0.047$, 95% CI = [0.017, 0.076]). Personal characteristics may also interact; for example, we observe evidence that White male faculty are higher than their Non-White, female counterparts ($\beta = 0.051$, 95% CI = [0.002, 0.10]), among other weaker interaction effects (Table B.5).

Characteristics of the class itself were also associated with ratings of overall teaching quality. The rated difficulty of the course was largest negative estimate from the model; each additional point of difficulty was associated with 0.39 lower points of overall teaching quality ($\beta = -0.39$, 95% CI = [-0.40, -0.38]). The student interest ratings of a faculty was the second largest positive estimate; each additional point in interest was associated with 0.33 points higher overall teaching quality ($\beta = 0.33$, 95% CI = [0.32, 0.34]). Faculty for whom a comment on RMP2018 mentioned a teaching assistant were associated with 0.18 point lower ratings of overall quality ($\beta = -0.18$, 95% CI =

$[-0.22, -0.15]$).

Associations between university characteristics and ratings of overall teaching quality were found to be weaker than for individual and class characteristics. Compared to all others, R1 universities—doctoral universities with very high research activity (as identified by the Carnegie Classification of Higher Education Institutions)—were associated with 0.03 point lower ratings of overall quality ($\beta = 0.03$, 95% CI = $[-0.05, -0.01]$). Compared to those in private universities, faculty affiliated with public universities were associated with 0.08 point lower teaching evaluations ($\beta = -0.08$, 95% CI = $[-0.11, -0.06]$).

There were notable differences in ratings of overall teaching quality between disciplines. All other disciplines were rated as having higher overall quality than Engineering, the reference level; Engineering was chosen as the reference because it had the lowest ratings of overall teaching quality. Compared to Engineering, faculty in the Humanities were associated with 0.18 point higher overall quality ratings ($\beta = 0.18$, 95% CI = $[0.14, 0.23]$). After the Humanities, faculty in Medical Science were associated with 0.11 points higher ratings than those in Engineering ($\beta = 0.11$, 95% CI = $[0.056, 0.153]$), followed by faculty in the Natural Sciences ($\beta = 0.07$, 95% CI = $[0.037, 0.10]$) and finally faculty in the Social Sciences ($\beta = 0.044$, 95% CI = $[0.008, 0.079]$). Distinct disciplinary contexts can also interact with other variables. While a complete cross-disciplinary analysis is out of the scope of the present study, we conduct a preliminary analysis of how gender interacts with discipline. We found that while male faculty get higher ratings, and Humanities and Natural Science faculty are rated more highly than those in Engineering, this total disparity fell when considering the Male/Humanities and Male/Natural Sciences combinations of factors (Table B.5).

Research indicators were only weakly or trivially associated with ratings of overall research quality. During analysis, we designated three levels of field-normalized research productivity over the past 5 years: no publications, moderate (at least one publication, less than or equal to the 90th percentile), and high (above the 90th percentile); this was repeated for all research indicators. Com-

pared to faculty with no publications in the past five years, faculty with moderate publication were associated with 0.034 point higher ratings ($\beta = 0.034$, 95% CI = [-0.001, 0.069])—this was the only estimate for which confidence intervals only barely crossed zero. Those with a high level of publications were associated with 0.022 point higher ratings ($\beta = 0.022$, 95% CI = [-0.033, 0.077]). Faculty with a moderate and high level of citations were associated with 0.024 point ($\beta = -0.024$, 95% CI = [-0.054, 0.006]) and 0.033 point ($\beta = -0.033$, 95% CI = [-0.085, 0.018]) lower teaching evaluations, respectively. Faculty with a moderate amount of grants were associated with only 0.002 point lower ratings ($\beta = -0.002$, 95% CI = [-0.026, 0.023]) whereas those with a high amount of grants were associated with 0.031 point higher evaluations than those with no grants ($\beta = 0.031$, 95% CI = [-0.024, 0.086]). Finally, compared to those with no awards, those with a moderate amount of awards were associated with 0.01 point higher ratings of overall teaching quality ($\beta = 0.01$, 95% CI = [-0.011, 0.031]), and those with a high amount of awards were associated with a similar 0.01 point higher ratings ($\beta = 0.01$, 95% CI = [-0.071, 0.091]).

One limitation of this regression analysis was that research indicators, due to their zero-inflated and heavily-skewed distributions, were binned into one of three categorical values; this made them more amenable for analysis but could mask linear relationships. We sought to further assess the presence of the research-teaching nexus by repeating the regression analysis with continuous, rather than categorical variables for research performance indicators (results provided in Table B.3). However, this analysis provided no new evidence for the research-teaching nexus, presenting at most a trivial positive relationship between the field-normalized count of awards and the overall teaching quality ($\beta = 0.008$, 95% CI = [0.001, 0.014]). We computed an ANOVA test to compare the two approaches but observed no significant difference in the variance explained by the models ($p = 0.47$). To mitigate the potential impact of multicollinearity, we also performed a regression model using the principal component of the continuous research indicators but still observed no evidence of a relationship between this variable and ratings of overall teaching quality. Additionally, we observed

no strong evidence of multicollinearity from the adjusted generalized variance inflation factors of both the model with discrete indicators (Table B.3), and the model with continuous indicators (Table B.3). We also sought to assess the impact of omitted variable bias to see how the absence of research indicator could impact other estimates, but observe only trivial differences, with an ANOVA between the basic model and the model with discrete indicators (Fig. 4.1.A) showing no evidence of a difference ($p = 0.49$).

We also investigated the extent to which continuous research performance indicators were correlated with ratings of overall teaching quality using the non-parametric Kendall Rank Tau test (Fig. 4.1.B). Non-parametric approaches may be better suited to understanding these zero inflated and skewed data. We calculated the correlations for all combinations of research indicators and separated total publication count into three variables corresponding to the count of articles, count of conference proceedings, and count of books indexed in AA2017 (these variables are described in Table B.1). However, we observed only trivial correlations between research indicators and ratings of overall teaching quality, the strongest having a value of 0.046 for the number of articles, followed by the number of books. For faculty with positive research indicators, we investigated the distribution of overall teaching quality by decile rank (Fig. B.2) which revealed some evidence of a positive linear relationship between overall teaching quality and citations and publications. However, as these results did not bear out when partitioning by discipline (Fig. B.3), when the linear trends all but disappeared; these small correlations may be confounded by disciplinary differences in publishing patterns and teaching quality. We note that the research indicators collected by *Academic Analytics* include only recent performance (5 years for publications and citations, 10 years for books) and do not represent faculty's full career, which may have proven more predictive of ratings of teaching quality.

Having observed the large estimates of individual and class characteristics from our regression analysis, we further investigated which characteristics of teaching were associated with the *Rate-*

MyProfessor.com overall teaching quality rating. The website allows users posting a review to select from a 20 pre-defined tags that denote common characteristics of university faculty and classes. Fig. 4.1.C shows the distribution of overall quality scores for faculty having each of these tags. The tags associated with the highest ratings of overall teaching quality tend to be personal characteristics of the instructor such as “amazing lectures”, “inspirational”, “respected”, “hilarious”, and “caring”. The tags associated with the lowest ratings instead tend to refer to course characteristics, such as “graded by a few things”, “lecture heavy”, “lots of homework”, “tough grader”, and “lots of tests”. The results from these tags confirm the relationship between difficulty and ratings observed in the regression model.

4.6 Discussion

Ideally, faculty evaluation would be an unbiased performance assessment, uninfluenced by gender, ethnicity, age, attractiveness, or other personal characteristics. However, empirical analyses of student evaluations of teaching (SETs) have demonstrated that they often fall short of this ideal [223, 467–469, 471–476]. Moreover as the ideal of the university posits a mutually beneficial research-teaching nexus, faculty evaluation should be holistic, considering performance across all professional responsibilities; however, assessments of the so-called research-teaching nexus have not produced a clear consensus of its presence, nature, or extent [490–496]. By constructing a large and heterogeneous dataset of tenure and tenure track faculty in the U.S., this exploratory study provides additional evidence of bias in SETs while also demonstrating little to no relationship between common indicators of teaching and research.

Individual characteristics

The strongest correlate with teaching evaluations was whether or not the faculty had a “chili pepper” rating on *RateMyProfessor.com*. The precise implication of the chili pepper is unclear, as

it was never explicitly defined and so its meaning will vary between users. We conceive the "chili pepper" as a rating of the physical attractiveness of the instructor, following past research [507] and widespread cultural understandings [484]. Following controversy, this rating was removed in 2018 (see Appendix B) however it remained in use at the time of data collection. Our finding is consistent with studies of student evaluations in traditional evaluative settings [474, 508], studies of faculty's online self-presentation [509], and past studies of *RateMyProfessor.com* [475, 507]. In unbiased evaluation, a faculty's physical attractiveness should not factor into the quality of their teaching or pedagogical skill. The relationship we observed could result from student's implicit bias favoring physically attractive faculty. It can also be interpreted as a "halo effect" [510], whereby student's positive impressions of one aspect of their professor (e.g.: their attractiveness) influences other aspects of their evaluation. Student's perceptions of physical attractiveness are also likely to differ with the perceived age, race, and gender of both the instructor and the students [511], resulting in different manifestations of this trend across different contexts. For example, younger faculty were more likely to be assigned a chili pepper, demonstrated by the negative trend between scientific age and probability of having a chili apparent in Fig. B.4). While we control for some of these characteristics (e.g., age, gender), we cannot effectively control for others such as ethnicity and student demographics.

We observed a small trend such that male faculty tended to receive higher ratings (of 0.10 points) of overall teaching quality than female faculty. Past studies of traditional SETs have noted gender biases favoring men in experimental settings [468] and in large-scale observational studies[473]. Studies leveraging *RateMyProfessor.com* have observed gendered differences in language used to describe faculty [512] but findings of bias in evaluation scores have been mixed with reports of small or no significant gender bias depending on context [492, 507]. We observed no evidence of gender difference in the distribution of overall ratings based on aggregate data (Fig. B.5), but did observe a relationship when controlling for other variables such as scientific age, disciplines, and university

context (Fig. 4.1.A); This discrepancy and the lack of consensus among studies suggests that gender bias in SETs is contingent on contextual factors of the university, discipline, and student body [471].

Faculty with commonly-White family names tended to be rated more highly than others. This finding affirms past studies that identified racial bias in SETs such that persons of color, particularly black faculty, were rated lower than their White counterparts [223, 469, 513]. However, those with names absent from the U.S. Census data also tended to have higher ratings than Non-White faculty; we cannot speak to the precise demographics of these names, however these names were more common in fields such as Engineering (Fig. B.6), which also tended to have the most Non-White associated family names (Fig. B.7). It is likely that the "unknown" category is therefore a mixture of White and Non-White faculty, the precise demographics of which require further investigation. We found evidence that race and gender interact, such that White Male faculty tended to be rated more highly than others, mirroring inter-sectional narratives. Related to race, faculty for whom an "accent" was mentioned in their evaluations tended to be rated lower than those for whom no accent was mentioned. *RateMyProfessor.com* and *Academic Analysis* offer no means of reliably inferring country of origin of faculty; here, we consider the mention of an accent as a proxy indicating non-native English speaker who may encounter bias and stereotyping in SETs. Whereas students often claim that instructor's accent is less important than their knowledge of the source material [514], accented faculty have been found receive lower evaluations, especially for comprehension [515]. On *RateMyProfessor.com*, a population of Asian-born professors (who may or may not have noticeable accents) were found to receive lower ratings than their U.S. born counterparts [516]. Non-White and foreign-born faculty face additional challenges when teaching such as stereotyping and prejudice. These approaches, while limited, demonstrate how biases can manifest in student's evaluations of faculty, which can hinder their career and produce additional inequality.

More senior faculty, in terms of the number of years since obtaining their Ph.D., tended to

receive lower ratings; each additional decade of scientific age was associated with 0.13 point lower score. Most past research studying the relationship between age and SETs has studied actual age, a value which is likely correlated with the scientific age we study here. One study of data from *RateMyProfessor.com* found evidence that older instructors were rated lower, but that this effect disappears after controlling for other factors, such as their physical appearance and the difficulty of their courses [476]. However, even after controlling for many of the same factors, our findings contribute to the consensus of studies finding that older faculty receive lower evaluations [474, 475].

Related to scientific age is also professional rank; we observed that full professors tended to get higher ratings than both assistant and associate professors, contrary to what we would expect given that younger faculty receive higher ratings. Assistant professors tended to be scientifically younger, whereas full professors tended to be older (shown in Fig. B.8). This suggests the relationship between seniority and SET ratings are not necessarily linear, and that those faculty with experience, though perhaps not too much seniority, tend to do best. One past study compared teaching from non-tenured instructors and tenure/tenure track faculty found that non-tenured instructors had stronger evaluations [517]. However, there is little research examining SETs across tenure ranks (assistant, associate, full). Common wisdom suggests that teaching benefits from experience but evidence suggests that past a baseline level of experience, students generally rate younger professors more highly over more senior and experienced faculty. However, younger professors may more readily relate to students or employ more recent pedagogical techniques. Moreover, the requirements, demands, and roles of faculty change over the course of their career, and teaching may be de-emphasized during certain career stages.

Classroom characteristics

The strongest negative relationship we observed was between overall teaching quality and ratings of class difficulty. Every point increase in difficulty rating (where five is most difficult, and one

is easiest) was associated with a drop of nearly half a point in overall quality. This finding is consistent with past studies identifying a negative relationship between difficulty and quality ratings in traditional SETs [518] and on *RateMyProfessor.com* [507, 518, 519]. One interpretation of this finding is that *RateMyProfessor.com* is a site used by students to complain about difficult courses and low grades, but overall teaching quality scores are actually somewhat skewed towards higher ratings, with median ratings of 3.6 for the matched dataset. Others have suggested that students have varying definitions of “difficulty”. For example, in some studies of SETs, difficulty was associated with perceptions of “fairness” in the course [518, 520]; similar effects were observed on *RateMyProfessor.com* [499]. Other scholars have found that clarity of course material and expectations are also important factors of student’s ratings of difficulty when posting reviews [498, 519]. The form for posting a review on *RateMyProfessor.com* is vague, and so there are boundless interpretations of the difficulty scale, which we cannot directly examine. However, tags associated with low teaching quality (Fig. 4.1.C) tended to relate to quantity and type of course material and grading (“tough grader”, “lecture heavy”, “lots of homework”, “test heavy”).

Ratings of prior interest almost mirrored those of difficulty, and were the second largest positive correlate with overall teaching ratings; each additional point in student interest was associated with 0.36 point higher ratings. Past studies found similar results when investigating SETs [521] and *RateMyProfessor.com* [507], though generally little research has been conducted examining the effect of student’s prior interest. Under the U.S. liberal arts model of higher education, many instructors will teach courses containing a mixture of students with radically different interest levels in the curriculum, from majors in the subject field to students fulfilling general education credits. This dynamic may similarly affect SETs. Indeed, there is some, if limited, evidence that elective courses (which are freely chosen by the student) often receive better student ratings than required courses [522]. Faculty who teach required or general-education courses may be at a systematic disadvantage during performance evaluations if they are subject to the prior interests of

their students. However, there are also difficulties with interpreting the rating of “prior interest” because it assumes that the student is aware of their true interest in a course at the time of posting their review, and that this measure is somehow indicative of their intrinsic interest in the subject. As with the “chili pepper”, ratings of interest may instead reflect a halo effect, such that a student’s rating of “interest” (or other teaching-related categories) is more closely related to their opinion of the professor than the course material.

We observed that faculty whose reviews mentioned a teaching assistant (TA) received lower ratings than those where no TA was mentioned. The presence of a TA is our best (though highly flawed) proxy for whether an instructor teaches large courses as TAs are typically employed for larger classes (though not in all cases, and with variations by discipline and university context). Our finding is however consistent with past studies that observed a small but significant negative effect between class size and SET ratings [472, 523, 524]. However, it is difficult to disentangle the extent that the TA in *RateMyProfessor.com* reviews indicates of the course size, or whether students only mentioned TAs when they were a negative aspect of the course.

University characteristics

Affiliation with public universities was related with lower ratings than affiliation with private universities, by about 0.08 points of overall quality. One reason for this small difference might be that faculty at private universities have been found to give, on average, higher grades to their students [525] and this higher expected grade may positively influence subsequent evaluations [526]. However, the difference we observed might also emerge from the distribution of contextual factors across public and private universities. For example, the sample of private colleges may include many smaller or liberal-arts colleges hosting more faculty in the Humanities and Social Sciences.

We also examined the research classification of universities, but we observed only trivial differences between R1 and non-R1 universities. There is little research examining differences in

SETs across different university types, whether between public and private universities or between research-focused and teaching-focused. Part of this may be because aggregating SETs across institutions is difficult due to their sensitivity. The formats of SETs are also likely to vary between institutions making comparisons between universities difficult. Here we find little difference in ratings of teaching quality based on university types, but more work is needed to understand the role of institutional context in teaching.

Discipline

We observed distinct trends in student's ratings of teaching by discipline; faculty in the Humanities tended to be the highest rated, whereas faculty in Engineering and Social Science tended to have the lowest ratings. These findings are consistent with past studies of discipline and teaching evaluation. For example, faculty teaching traditionally quantitative disciplines were found to receive lower ratings, an effect that was observed for traditional evaluations [466, 477] and on *RateMyProfessor.com* [507]. However, whereas Social Science is not typically associated with quantitative courses, we observed that teaching ratings for faculty in Social Science tended to only be trivially higher than faculty in Engineering. One reason for this discrepancy may be that the high-level classifications used in this study mask the true heterogeneity of disciplines and courses and don't easily allow for "quantitative" / "not quantitative" distinctions. However, we also observed that regression estimates for disciplinary effects differed from the simple average of ratings by discipline, for which Natural Sciences actually have the lowest median rating (Fig B.5); this suggests that some of the disciplinary differences might be explained by contextual factors such as the distribution of faculty demographics, classroom, and university characteristics across disciplines. For example, in our preliminary analysis of the interaction between gender and discipline, we observe differences across fields. More thorough work is necessary to understand discipline and course topic relates to teaching and SETs; in particular, a more comprehensive and thorough statistical

analysis according to discipline, combined with a more fine-grained disciplinary classification could provide additional insight into the relationship between discipline and SETs.

Research-teaching nexus

Applying several different techniques, we observed little to no relationship between indicators of research performance and ratings of overall teaching quality on *RateMyProfessor.com*. In other words, we observed evidence consistent with a neutral research-teaching nexus, as observed in several past studies [496, 527–530]. In the study most similar to our own, a weak correlation was observed between ratings on *RateMyProfessor.com* and journal publication count [492], however the study examined only faculty affiliated within Marketing departments. Other studies have observed positive research-teaching nexus between SETs and research productivity under certain circumstances [531, 532], but generally, empirical evidence is lacking [496]. The results from our analysis contribute to the consensus of a neutral relationship between research and teaching.

The research-teaching nexus is complicated, and difficult to assess. Evidence for a null model tend to use SETs or an equivalent indicator to measure teaching performance. Studies also tend to use output-based bibliometric indicators to measure research performance; our study also only examines recent research output, whereas longer time-scales of output may correlate more strongly with teaching. Such indicators have been called into question as being improper or inadequate tools that don't measure true teaching or research performance [332, 533, 534].

The research-teaching nexus, if it exists, may be intangible or may not manifest in performance measures. Rather than further attempting to empirically verify the existence of the research-teaching nexus using quantitative tools, qualitative methodology may prove more useful to explore perceptions of the nexus [486–489, 535]. Such approaches could reveal the extent to which faculty believe the nexus exists, what they believe about the nature of the nexus, and how the nexus has evolved with increasing faculty time constraints [493, 494, 536]. Moreover, if the relationship

between research and teaching is held as a value of academia, then researchers and administrators should explore ways of actively promoting the research-teaching nexus [527].

Limitations

Our study is subject to several limitations. First, we note that we conducted a preliminary and exploratory study using observational data and as such our methods were not pre-registered and our analysis is subject to issues of multiple comparisons.

Second, our use of *RateMyProfessor.com* as a proxy for SETs is a clear limitation, as reviews on the website suffer from issues of external validity [519] and selection bias, wherein students with extreme opinions are likely to be the ones to post reviews. The website has also endured criticism that reviews not align with effective teaching [498]. While traditional SETs are intended for internal use for faculty evaluation and improvement, the primary purpose of *RateMyProfessor.com* is to help students select courses; the expectations and rating criteria of each likely diverge. Despite these issues, ratings on *RateMyProfessor.com* have been found to correlate with traditional SETs (see Appendix B). Similarly, quantitative measures of teaching and research do not capture quality. The indicators used in this analysis are also limited in that they capture only recent performance—an artifact of *Academic Analytics*—more insights may be gained by examining the historical trends of professor’s research or teaching performance.

Third, we were limited by the evolving nature of our data sources. *RateMyProfessor.com* has undergone many changes since its inception, including changes to the features and indicators provided to raters. While we limit our analysis to relatively recent reviews, during this time certain indicators (such as the separate measures of “Clarity” and “Helpfulness” and “Interest”, as well as the “Chili Pepper”) were removed (see Appendix B).

Fourth, by limiting our analysis to tenure and tenure-track faculty in the United States, our analysis excluded contingent and other non-tenure track faculty who comprise over 70 percent of the

U.S. [537] and more than 50 percent [538] of Canadian faculty appointments, as well as graduate student instructors who may teach a large proportion of courses [539]. These populations face unique challenges [540, 541] that remain unaddressed in the present study. Moreover, these results are limited to faculty within the United States, and so our findings may not generalize to other national contexts.

Finally, our analysis was also limited by the record-matching algorithm which did not capture all relevant faculty. The parameters for record matching favored precision over recall, so the number of matched faculty are a conservative sampling of the population. Additionally, there were many professors who simply did not appear on *RateMyProfessor.com* or in *Academic Analytics*, and so do not appear in the present analysis. Given that there is no known list of all U.S. faculty, it is difficult to assess the extent to which the matched faculty were representative of U.S. tenure and tenure-track faculty as a whole.

4.7 Conclusion

This paper provided an exploratory analysis of the factors relating to online ratings of teaching quality and their relationship to research productivity. We constructed a novel dataset by matching records of known tenure and tenure-track faculty from *Academic Analytics* with individuals listed on *RateMyProfessor.com*. We assessed the effect of the demographics of the teacher, characteristics of the class, of the university, and of the discipline. Faculty tended to receive higher ratings when they were rated as attractive (having the “chili pepper” on *RateMyProfessor.com*), when they were male, when they were young, when they were not mentioned as having an accent, and when they were full and associate professors. Faculty tended to receive lower ratings when the course was difficult, when there was little student interest, or when a teaching assistant was mentioned. We observed some evidence that faculty in private universities were rated slightly higher than those from public universities, but overall university characteristics were weakly related to ratings of

teaching. Faculty from the Humanities tended to be rated most highly, followed by those in the Medical Sciences, Natural Sciences, Social Sciences, and finally Engineering.

In addition to demographic and contextual factors, we also assessed the presence and extent of the so-called *research-teaching nexus*, the relationship between research and teaching. Comparing indicators of recent publications, recent citations, current grant funding, and professional awards, we found evidence consistent with a *neutral* nexus, or no relationship between research and teaching.

These results and data provide a foundation for future large-scale analysis of SETs and of the research-teaching nexus. Future work could delve deeper into this data, comparing patterns of student ratings of teaching across more disciplines, university types, course levels, and even specific departments. *RateMyProfessor.com* also offers a trove of text data from student comments; content analysis and text mining of these data could reveal key insights to the underlying factors of student's ratings, such as gendered language and attitudes [512]. This text data can be leveraged to identify other faculty characteristics, such as their self-disclosed or perceived LGBTQ+ status, allowing study into the unique challenges faced by those faculty of different sexual orientations and gender identities [542–544]. The current dataset could also be enriched with survey data relating to time spent on service-related activities or more detailed bibliometric indicators from the Web of Science or Scopus. It is our hope that the present analysis is the first of many to explore broad trends in the nature of quantitative performance measures across disciplinary, university, and classroom contexts. Despite controversy, student evaluations of teaching dominate faculty evaluation across the United States; given their continued importance, it is important to understand what factors contribute to these scores and how these factors differ between institutional and disciplinary contexts. Our results build on past research that demonstrates the biases, limitations, and deficiencies of SETs. The confluence of research should cause the higher education community to consider whether the student evaluations of teaching should be discounted, rehabilitated, or done away with all together.

Chapter 5

Study 3: Measuring disagreement in science

5.1 Foreword

I don't have to agree with you to like you
or respect you

Anthony Bourdain

Disagreements hold an almost mythical place in science, sitting at the core of its history. The heliocentric view of the solar system, for example, emerged out of intense controversy between Galileo and his contemporaries [69], and even the foundations of modern science were not a matter of obvious consensus, but were the subject of debate between Robert Boyle Bacon and Thomas Hobbes [7]. So central is disagreement, that many theories of science hold it as a requirement for progress [187, 545, 546]. Even in recent years, the importance of disagreement has been re-iterated time and time again [547, 548]. In spite of its importance, empirical studies of disagreement in science are few, limited by the difficulty in identifying instances of disagreement at scale.

In this chapter, I leverage increasingly-available full-text data of scientific publications in order to construct a measure of disagreement, and quantify it across science. Part of this study's contribution is empirical. I observe that disagreement is highest in the social sciences and humanities, and lowest in the physical sciences, closely following the classic "hierarchy of sciences" put forth by Auguste Compte [549]. Yet my findings also speak to the heterogeneity of science, and how individual-level forces can shape the development of a field. For example, fields that can't conduct controlled experiments, such as geology and paleontology, have higher disagreement than other experimental fields. When looked at using the complexity perspective, I argue that various feedback mechanisms transform these local factors and idiosyncratic disagreements into entire social

organizations and disciplinary cultures [40].

This study's contribution is also methodological. This measure of disagreement is based one extensive manual curation and validation, producing a reliable measure of disagreement, the first of its kind to be applied across such a massive quantity of literature. Yet this measure is not perfect. It only returns a fraction of total disagreement, and also retrieves false positives, more so in certain fields that use language or terms that confound our method. The complexity perspective argues that these issues are inherent to a system as massive and heterogeneous as science, and can be mitigated, but never resolved completely. Fortunately, the transparency of this measure makes it easy to identify and address its limitations in a way that other common methods, such as those reliant on machine learning, find difficult.

This manuscript is currently under review at the journal *eLife*, and a pre-print has been posted to arXiv under the name *Measuring disagreement in science*. Data underlying this work has also been made available at <https://doi.org/10.5281/zenodo.5148058>, and code to conduct the analysis is available at github.com/murrayds/sci-text-disagreement. This work was co-authored along with Wout Lamers, Dr. Kevin Boyack, Dr. Vincent Larivière, Dr. Cassidy R. Sugimoto, Dr. Nees Jan van Eck, and Dr. Ludo Waltman. My contributions to this work were supported by the Air Force Office of Scientific Research under award number FA9550-19-1-039. Vincent Larivière acknowledges funding from the Canada Research Chairs program. I would also like to thank Yong-Yeol Ahn, Staša Milojević, Alessandro Flammini, Filippo Menczer, Dashun Wang, Lili Miao, and participants of the A scientometric analysis of disagreement in science seminar held at CWTS at Leiden University for their helpful comments.

5.2 Abstract

Disagreement is essential to scientific progress. However, the extent of disagreement in science, its evolution over time, and the fields in which it happens, remains largely unknown. Leveraging a massive collection of scientific texts, we develop a cue-phrase based approach to identify instances of disagreement citations across more than three million scientific articles. Using this method, we construct an indicator of disagreement across scientific fields over the 2000-2015 period. In contrast with black-box text classification methods, our framework is transparent and easily interpretable. We reveal a disciplinary spectrum of disagreement, with higher disagreement in the social sciences and lower disagreement in physics and mathematics. However, detailed disciplinary analysis demonstrates heterogeneity across sub-fields, revealing the importance of local disciplinary cultures and epistemic characteristics of disagreement. Paper-level analysis reveals notable episodes of disagreement in science, and illustrates how methodological artefacts can confound analyses of scientific texts. These findings contribute to a broader understanding of disagreement and establish a foundation for future research to understand key processes underlying scientific progress.

5.3 Introduction

Disagreement is a common phenomenon in science. The history of science is ripe with histories of famous discoveries, which are often embroiled in debates, controversies, and disputes. Dialectic discourse emerged in ancient Greece, whereby the truth was thought to emerge from the arguments and counterarguments of scholars engaged in dialogue. The modern scientific method arose from a similar dialogue 350 years ago, as two individuals—Robert Boyle and Thomas Hobbes—debated over the meaning of experimental results obtained with newly-invented air pump [7]. Disagreement anchors much of the lore surrounding major scientific discoveries. For example, Alfred Wegener’s theory of plate tectonics was initially rejected by the scientific community, and physics endured a decades-long dispute over the existence of gravitational waves [550] and the value of the Hubble

constant [551]. Other conflicts are influenced by forces external to science, such as the controversies on the link between cigarette and lung cancer or between greenhouse gas and climate change [360]. Disagreement features prominently in key theories of science, such as Popper's falsifiability [546], Kuhn's paradigm shifts (Kuhn, 1996), and Kitcher's scientific division of labor [545].

Despite its importance to science, however, there is little empirical evidence of how much disagreement exists, where it is most common, and the consequences of disagreement. Quantitative measures can be valuable tools to better understand the role and extent of disagreement across fields of science. Previous research has focused on consensus as evidenced by citation networks [191, 552, 553]; on concepts tangential to disagreement in scientific texts such as negative citations, disputing citations, and uncertainty [554–556]; and on word-count based approaches [557]. Studying disagreement is challenging, given the lack of a widely accepted theoretical framework for conceptualizing disagreement combined with major challenges in its operationalization, for instance, the limited availability of large-scale collections of scientific texts.

This paper proposes an operationalization of disagreement in scientific articles that captures direct disagreement between two papers, as well as statements of disagreement within the community. We describe a methodological approach to generate and manually-validate combinations of cue-terms that reliably match to citation sentences (citations) to represent valid instances of disagreement. We then use this approach to quantify the extent of disagreement across more than four million publications in the Elsevier ScienceDirect database, and investigate the rate of disagreement across fields of science.

5.4 Literature Review

It is widely acknowledged that disagreement plays a fundamental role in scientific progress [547, 548, 558]. However, few studies have tried to quantify the level of disagreement in the scientific literature. Part of this may be explained by the fact that disagreement is difficult to both define and

measure. There have been, however, attempts to assess consensus or uncertainty in the literature. Much of the early work on consensus attempted at characterizing differences between so-called hard and soft sciences. Cole [392] described a series of experiments done in several fields, finding no evidence of differences in cognitive consensus along the hierarchy of sciences [549]. Hargens [559] claimed that fields having journals with higher rejection rates had lower consensus. This claim was contested by Cole and colleagues [560], who argued that other variables accounted for the differences, and that reviewer’s assessments would be a better measure of consensus than rejection rates. Fanelli [561] found that positive results—support for the paper’s hypotheses—was far higher in the social sciences than the physical sciences. Although this would suggest higher consensus in the social sciences, it actually reflects the higher level of ambiguity—and thus lower level of consensus—with respect to field-level norms that are more rigorous in the physical sciences.

Recent studies on scientific consensus have made use of citations and text. Through a series of case studies, Shwed and Bearman [191], used network modularity to show that divisions in the citation network declined over time, corresponding to increased consensus. Nicolaisen and Frandsen [562] used a Gini index calculated over bibliographic coupling count distributions to approximate consensus, and found that physics papers showed more consensus on average than psychology papers. Using a corpus of nearly 168,000 papers, Evans et al. [563] calculated the Shannon entropy of language in a set of 8 fields, and found that hard sciences had higher consensus than the social sciences.

Other studies have developed methods to identify uncertainty in scientific texts. For example, Szarvas [564] interpreted uncertainty as a ‘lack of information’ and created an uncertainty detection model based on three datasets (BioScope, WikiWeasel and Fact Bank) previously annotated for uncertainty cue words. Their results suggest some domain dependence of cues but also show that reasonable performance can be obtained for new domains. Yang et al. [565] developed a classifier based on manually annotated uncertainty cues and conditional random fields, and conducted a

series of experiments to assess the performance of their method. Chen, Song and Heo [555] extended previous studies of uncertainty by a) introducing a conceptual framework to study uncertainty that incorporates epistemic status and perturbation strength, b) measuring uncertainty in 24 high-level scientific fields, and c) creating an expanded set of uncertainty cues. They reported the fraction of items from each subject area containing one of those words, with a high of 32 percent for psychology and a low of 4 percent for chemistry. Social sciences have the highest rates of uncertainty, followed by medical sciences, environmental sciences, and engineering. Physical sciences and math had the lowest rates of uncertainty. Note that these rates included all types of uncertainty, whether they be theoretical, conceptual, or experimental, and within or between studies.

Many of the cues used as a starting point by Chen et al. [555] are hedging terms, which are commonly used in scientific writing to express possibility rather than certainty [566]. In addition to being field-dependent, hedging rates have also been found to depend on whether a paper is primarily methodological. Recent work by Small and colleagues [567, 568] showed that citing sentences (i.e. citances) with the word ‘may’ occur much more frequently when citing method papers than non-method papers. More recently, Bornmann, Wray and Haunschild [569] used a similar method to investigate uncertainty associated with specific concepts in the context of highly cited works. While some might equate uncertainty or hedging with disagreement, they are not the same. As mentioned by Small, when citing another work, “hedging does not assert that the paper is wrong, but only suggests that uncertainty surrounds some aspect of the ideas put forward” [568]. In contrast, we attempt to explicitly identify and measure scientific disagreement by using a large set of citances across all fields and by developing a set of cues validated by expert assessment.

Other studies of disagreement have been performed in the context of classification schemes of citation function. In an early attempt to categorize types of citations, disagreement was captured as “juxtapositional” and “negational” citations [570]. However, this scheme was manually developed using a limited sample of papers and citations, and so the robustness and validity of the categories

cannot be easily assessed. More recently, scholars have used larger datasets and machine learning techniques to scale up citation classifications, often including categories of citations similar or inclusive of disagreement. For example, Teufel et al., [571, 572] developed a four-category scheme in which disagreement might be captured under the “weakness” or “contrast” citation type. Bertin and colleagues [573] used n-grams to study location of negative and positive citations, and showed that the word “disagree*” was much less likely to occur than the word “agree*”, irrespective of papers’ sections. In another study that aimed to identify meaningful citations, Valenzuela et al., [574] captured disagreement under the “comparison” citation type. Others have sought more coarse categories: Catalini et al., [554] classified over 750,000 references made by papers published in the Journal of Immunology as either positive or negative, finding that negative references comprised about two percent of all references made. However, while these machine learning approaches are useful for analyzing large text data, they are also black boxes which can obfuscate issues and limit interpretation of their results.

Building on these studies, we develop a definition of disagreement in the scientific literature and a methodological framework for identifying valid instances such disagreement based on a manually validated set of cue-terms. This approach allows us to easily scale our analysis to millions of scientific articles, while also being transparent and reproducible.

5.5 Materials and Methods

Data

We sourced data from the Elsevier ScienceDirect database hosted at the Centre for Science and Technology Studies (CWTS) at Leiden University. This data contains the full-text information of nearly five million English-language research articles, short communications, and review articles published in Elsevier journals between 1980 and 2016. The Elsevier ScienceDirect corpus comprises articles from nearly 3,000 Elsevier journals. Given that Elsevier is the largest publisher in the world,

this corpus is one of the largest multidisciplinary sources of full-text scientific articles currently available, with coverage of both natural sciences, medical sciences, as well as the social sciences and humanities.

Sentences containing in-text citations (citations) were extracted from the full-text of these articles following the procedure outlined in previous work [95]. Articles published by Elsevier were identified using the Crossref REST API and used to download full-text in XML format from the Elsevier ScienceDirect API (Article Retrieval API). Each XML full-text record was parsed to identify major sections and paragraphs (using XML tags), and sentences (using a sentence-splitting algorithm). In-text citations in the main text were identified by parsing the main text (excluding those in footnotes and figure and table captions). XML records without in-text citations were discarded, and publications from before 1998 were omitted from analysis due to poor availability of full-text records before that year. The resulting dataset consisted of 4,776,340 publications containing a total of 145,351,937 citations, spanning from 1998 to 2016.

To facilitate analysis at the level of scientific fields, articles in Elsevier ScienceDirect as well as the references cited in these articles were matched with records in the Web of Science database based on their DOI (where available) and a combination of publication year, volume number, and first page number. We used an existing classification of research articles and review articles in the Web of Science created at CWTS. In this classification, each article published between 2000 and 2015 and indexed in the Web of Science was algorithmically assigned to one of 817 meso-level fields. The classification also includes micro-level fields, but these were not used in order to ensure sufficient coverage per field. The classification was created algorithmically based on direct citation links between articles, using the methodology introduced by Waltman and Van Eck [575] and Traag et al. [96]. The meso-level fields were grouped into five broad fields: Biomedical and Health Sciences, Life and Earth Sciences, Mathematics and Computer Science, Physical Sciences and Engineering, and Social Sciences and Humanities. Linking our dataset to this classification system resulted in a

subset of 3,883,563 papers containing 118,012,368 citances, spanning 2000 to 2015. A visualization of the meso-level classification was created using the VOSviewer software [576].

Operationalizing disagreement

Researchers can disagree for many reasons, sometimes over data and methodologies, but more often because of differences in interpretation [577]. Some of these disagreements are explicitly hostile and adversarial, whereas others are more subtle, such as contrasting findings with past results and theories. We adopt an inclusive definition of disagreement that captures explicit textual instances of disagreement, controversy, dissonance, or lack of consensus between scientific publications, including cases where citing authors are not taking an explicit stance themselves. We define two kinds of disagreement which capture the diversity of obvious and subtle disagreement in the scientific literature: *paper-level disagreement* and *community-level disagreement*.

The first, *paper-level disagreement*, occurs when one publication offers a finding or perspective that is (at least partly) incompatible with the perspective of another (even though there may be no explicit contradiction). Consider the following example of a citation sentence explicitly disagreeing with the conclusion of a past study:

We find that coffee does not cause cancer, contrary to the finding of [ref] that coffee does cause cancer.

Paper-level disagreement can also be more subtle. For example, in the following two disagreement sentences, although they do not resolutely contradict one another, the citing and cited publications use models that are based on incompatible assumptions (first sentence), or observe different effects from different data (second sentence):

Assuming that coffee increases the probability of cancer by 50%, the predicted life expectancy for the Dutch population is 80 years, in contrast to the 85 years proposed by models that assumed coffee does not increase the risk of cancer [ref].

Contrary to previous studies that did not observe evidence to support the hypothesis that coffee causes cancer [ref], our data suggests that drinking coffee increases the probability of cancer by 50%.

Community-level disagreement refers to the situation in which a citing publication, without explicitly disagreeing with a cited publication, instead draws attention to a controversy or lack of consensus in the larger body of literature. For example, the following disagreement sentence notes the disagreement between the referenced studies.

There remains controversy in the scientific literature over whether or not coffee is associated with an increased risk of cancer [ref]

Here, we do not differentiate between paper-level or community-level disagreement, including both under our operationalization of disagreement.

Signal and filter terms

A set of preliminary *signal terms* was derived through an intensive iterative process of manually identifying, classifying, validating, and deliberating on strategies for identifying instances of disagreement. This took place over several meetings, utilizing multiple strategies: e.g., exploring previously used cue words, ranking words by frequency and extracting those associated with disagreement, and identifying synonyms for disagreement. The inductive process included several rounds of inter-rater reliability to generate a robust list. This list of signal terms is intended to have high validity, but is not considered comprehensive.

We queried the database for citances containing each of these signal terms (case insensitive), using wildcards to provide for possible variants of terms (e.g., “challenge”, “challenged”, and “challenges”), excluding generic negation phrases (“no”, “not”, “cannot”, “nor” and “neither” to exclude phrases such as “no conflict”), and for some signal terms excluding citances containing words associated with disciplinary jargon or methods, such as for the signal term “disagreement”, which often appears with Likert-scale descriptions (e.g., “scale”, “agreement”, or “kappa”) for survey-heavy fields. The modifications for the signal terms were derived after several rounds of review and validation. In total, citances returned by signal phrase queries comprise 3.10 percent ($n = 145,351,937$) of the database, though their relative occurrence varied dramatically, with the most coming from

the “*differ**” signal term, and the least from “*disprove**” (see Table 5.1).

Table 5.1: . Specific terms comprising each of the thirteen signal term sets and specific exceptions. The “*” symbol (wildcard) captures possible variants.

Signal term	Variants	Specific negations	Results
challenge*			405,613
conflict*			212,246
contradict*			115,375
contrary			171,711
contrast*			1,257,866
controvers*			154,608
debat*		“parliament* debat*”, “congress* debat*”, “senate* debat*”, “polic* debat*”, “politic* debat*”, “public* debat*”, “societ* debat*”, “different*”, “range”, “scale”, “kappa”, “likert”, “agree*” and/or “disagree” within a ten-word range of each other.	150,617
differ*		“not agree*”, “no agreement”	2,003,677
disagree*		“prove*” and “disprove*” within a ten-word range	52,615
disprov*		“consensus sequence”, “consensus site”	2,938
no consensus		“lack of consensus”	16,632
questionable		“refutab*”	24,244
refut*			10,322
total			4,578,464

In order to more precisely capture valid instances of disagreement and to understand their function within the literature, we also queried for citations containing both the signal terms along with at least one of four sets of filter terms within a four-word window of the signal. As with signal terms, filter terms were derived from iterative manual efforts of the authors to identify terms most associated with valid instances of disagreement. Four distinct sets of terms were identified, corresponding to explicit mentions of terms relating to past studies, ideas, methods, and results (see Table 5.2). As with signal phrases alone, the relative incidence of signal and filter phrase combinations varies widely (Table C.1). Queries were constructed for each combination of signal and filter term.

Table 5.2: Specific terms comprising each of the four filter term sets

studies	studies; study; previous work; earlier work; literature; analysis; analyses; report; reports
ideas	idea*; theory; theories; assumption*; hypothesis; hypotheses
methods	model*, method*, approach*, technique*
results	result*; finding*, outcome*; evidence; data; conclusion*; observation*

Query validation

For each signal/filter term combination, we randomly sampled 50 citations (only 40 citations existed for *no consensus +ideas*), resulting in over 3,000 queried sentences. These citations were manually annotated by two independent coders (selected randomly from the seven authors on this paper) according to whether it constituted a valid instance of disagreement. The label was chosen based only on the text in the citation sentence, without knowledge on the citing paper's title, authors, field of study, or the surrounding text.

Consider the following four example sentences listed below: the first is invalid because the signal term, “conflict”, refers to an object of study, and not a scientific dispute; the second sentence is also invalid because the term “conflicting” refers to results within a single study, not between studies; the third sentence is invalid because “challenge” appears while quoting the cited study; the fourth and fifth sentence are both examples of sentences that would be marked as valid. Similar patterns can be observed for other signal terms, such as *challenge** (Table C.2).

1. **Invalid:** “To facilitate conflict management and analysis in Mr (...) , the Graph Model for Conflict Resolution (GMCR) (...) was used.”
2. **Invalid:** “The 4-year extension study provided ambiguous [...] and conflicting post hoc [...] results.”
3. **Invalid:** “Past studies (...) reviews the theoretical literature and concludes that future empirical research should ‘challenge the assumptions and analysis of the theory’”
4. **Valid:** “These observations are rather in contradiction with Smith et al.’s [...].”
5. **Valid:** “Although there is substantial evidence supporting this idea, there are also recent conflicting reports (...).”

We assessed the labels for each signal/filter term combination with two measures: percent agreement (% agree) and percent valid (% valid). Percent agreement is the proportion of annotated citations in which both coders agreed on the same label of valid or invalid; this measure provides

a simple measure of coder’s consensus. Here, percent agreement is justified over more complicated measures such as Cohen’s kappa due to the small sample of data per signal/filter term combination, and that there are only two categories and coders.

The overall percentage agreement between coders was high, at 85.5 percent (and Cohen’s kappa of 0.66). Given the difficulty of interpreting academic texts, this high percentage agreement demonstrates the robustness of our operationalization of disagreement. Most signal/filter term combinations had high agreement (Figure 5.1a). The signal term with the highest average agreement was *no consensus* (95.8 percent). There were only a few combinations with very low percentage agreement, mostly regarding the signal term *questionable*, which had an average lowest average percent agreement (64 percent); the nature of sentences returned from the *questionable* keyword tended to constitute marginal cases of disagreement. There was virtually no variance between the average percent agreement aggregated across filter terms. However, certain combinations of signal and filter terms were notable in resulting in higher or lower performance. For example, the difference between the highest agreement, *differ* _standalone_* (100 percent), and *differ* +methods* (74 percent) is 26 points—the addition of filter terms can dramatically impact the kinds of citances returned by the query.

We calculate the percent valid as the percentage of citances annotated as valid by both coders; this provides an intuitive measure of the validity and reliability of a query. Signal/filter term combinations that best capture disagreement should have both high percent agreement and high percent validity. Not all signal/filter term combinations were found to be sufficiently valid (Figure 5.1b). Overall, 61.6 percent of all citances were coded as valid, with large variance between the most valid (100 percent), and the least valid (0 percent) combinations. The signal term with the highest average validity regardless of filter term was *no consensus* (94.9 percent), followed by *controversy** (88.8 percent) and *debat** (82.4 percent). Unlike with percent agreement, average validity differs drastically between filter terms, with all having higher average validity than *_standalone_*. The

combinations with highest validity are no consensus + studies (98 percent), no consensus +methods (98 percent), and *no consensus _standalone_* (94 percent) For specific signal terms, the presence of a filter term can have a drastic impact of coded validity; for example, the validity of *contrast* +ideas* (80 percent) is four times greater than of *contrast* _standalone_* and *contrast +methods* (20 percent).

The queries that best capture instances of disagreement are those with the highest validity. We choose a validity threshold of 80 percent and exclude queries with lower validity from subsequent analysis. We also consider several adjustments to the threshold to assess the robustness of our empirical findings. 23 queries sit above this the 80 percent threshold (Figure 5.1c), including all five no consensus and *controvers** queries, four *debat** queries, two *disagree** and *contradict** queries, and one query each for *contrary**, *contrast**, *conflict**, *disprove**, and *questionable*. Because we prioritized precision, these 23 queries comprise only a fraction of total citances: 455,625, representing 0.31 percent of all citances in our dataset. We note that citances returned by queries are not exclusive; for example, a citance containing both *controvers** and no *consensus** would count towards both signal phrases. Similarly, a citance returned with the query *controvers* +methods* would also be returned by the *controvers**. Naturally, more general queries, such as *differ** and *contrast** returned a much greater number of citances. Among queries above the 80 percent threshold, the *controvers** and *debat** produce the highest number of citances (154,608 and 150,617 respectively, Figure 1d).

As a confirmation of overall validity, we measure the rate of disagreement by instances of self-citation and non-self-citation. We expect that authors will be less likely to cite their own work within the context of disagreement. Indeed, we find that the rate of disagreement for non-self-citations is 2.4 times greater than for self-citations (Figure C.3), demonstrating that our indicator of disagreement affirms expectations.

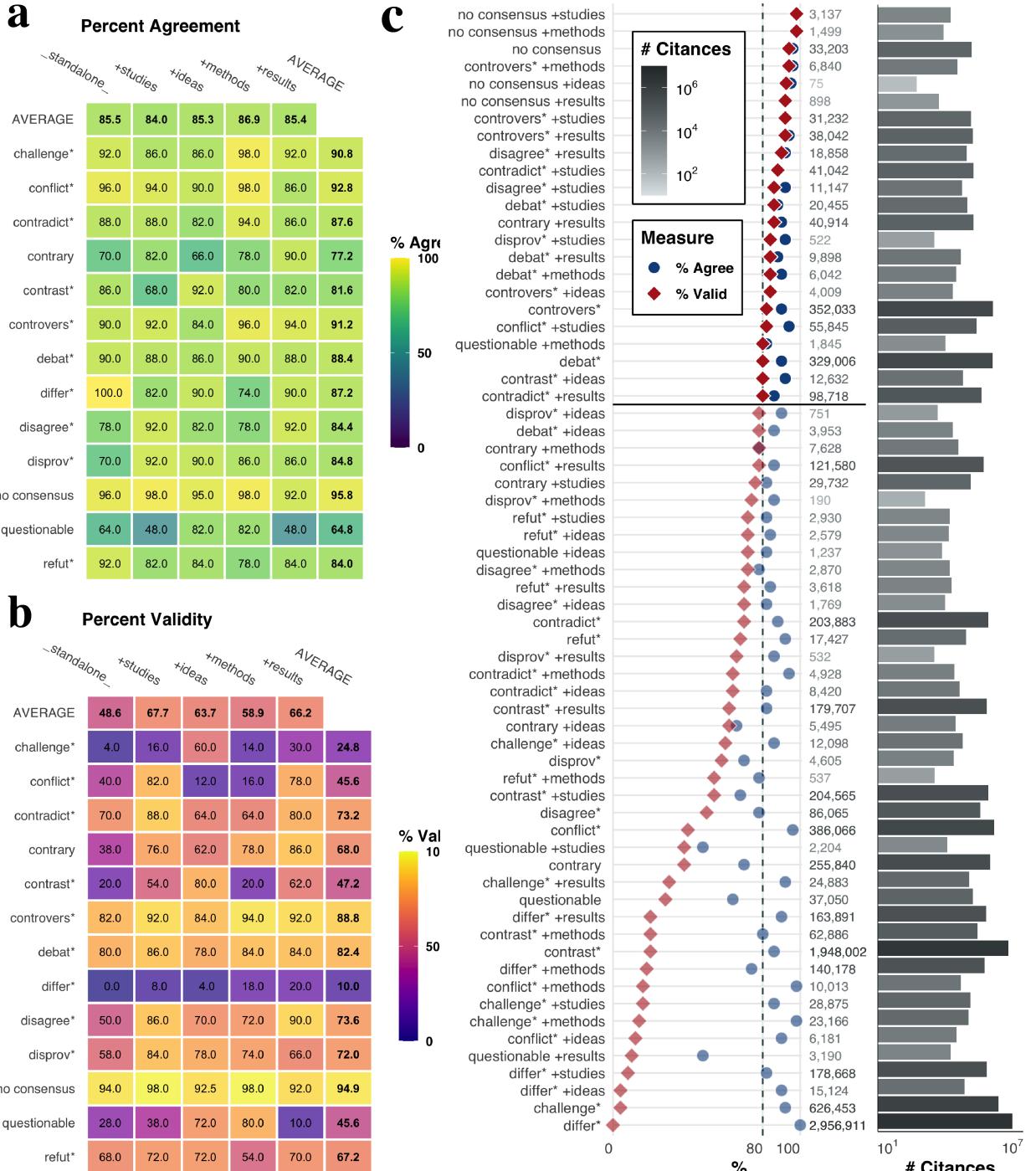


Figure 5.1: **Agreement and validity of signal and filter term combination.** Measures calculated from 50 randomly-sampled citances for each combination annotated as valid or invalid instances of disagreement by two independent coders. **a.** Percentage agreement, or the proportion of citances for which coders independently agreed on the label. **b.** Percentage validity, or the proportion of citances which both coders labeled as valid. Averages are shown by row and column. **c.** Percentage agreement and validity of each signal/filter term combination, ordered from highest percent validity (top) to lowest percent validity (bottom). Numbers on the right are the total number of citances returned by querying using the signal/filter term combination, and are colored according to their log-transformed value. ¹⁴¹ **d.** Log-transformed count of citances returned by each query combination, colored by the (log-transformed) number of citances. Citance counts are non-exclusive, meaning that citances of the form *debat* +studies* will also be counted towards *debat*_standalone_*.

5.6 Results

Instances of disagreement, operationalized using the 23 validated queries, accounted for approximately 0.31 percent of all citation sentences (citances) extracted from indexed papers published between 2000 and 2015 (Figure 2a). Disagreement was highest in the Social Sciences and Humanities (Soc & Hum, 0.61 percent), followed by Biomedical and Health Sciences (Bio & Health, 0.41 percent), Life and Earth Sciences (Life & Earth, 0.29 percent), Physical Sciences and Engineering (Phys & Engr, 0.15 percent), and Mathematics and Computer Science (Math & Comp, 0.06 percent).

Disagreement has been relatively constant over time (Figure 2b), decreasing at an average rate of about 0.0005 percentage points per year. This is driven by falling disagreement in Phys & Engr (-0.0045 points per year), Soc & Hum (-0.0033 points per year), and Math & Comp (-0.0019 points per year). Phys & Engr stands out not only for its stable decrease each year, but also for its relative size; given a starting rate of one disagreement signal per 529 citances in 2000, by 2015 the rate of disagreement in Physics fell to one disagreement per 809 citances, a 35 percent decrease, compared to a 24 percent decrease for Math & Comp and only a 5 percent decrease in Soc & Hum. In contrast, disagreement has tended to increase somewhat in Bio & Health (+0.0017 points per year) and Life & Earth (+0.0018 points per year). These trends are likely not the result of uses of individual queries; for example, *disagree** queries are over-represented in Phys & Engr (Figure C.2), yet the incidence of these terms is falling or remaining stable (Figure C.1). Similarly, *debat** was over-represented in Soc & Hum and has increased in usage despite slight falling disagreement in the field. That these changes are not confined to any single query suggests that field-level differences represent changes in the level of disagreement within a field rather than linguistic or methodological artifacts. These findings are also consistent at a lower, 70 percent validity threshold for disagreement queries, which includes 13 new queries that bring the total number of disagreement citances to over 650,000 (see Supporting Information).

The more fine-grained meso-fields reveal heterogeneity within the larger fields (Figure 5.3). Overall, meso-field disagreement followed the same pattern as Figure 5.2, with higher scores in Soc & Hum and lower in Math & Comp. However, some meso-fields stand out. For example, some of the highest rates of disagreement found in the Bio & Health meso-field was in more social journals such as *Quality of Life Research*, *Value in Health*, and *Pharmacoeconomics*. Similarly, in Math & Comp, the meso-field with the most disagreement contained journals relating to transportation science, a technical field which draws on management studies and other social science literature. This pattern held in Life & Earth, in which a meso-field with a relatively high share of disagreement contained papers in journals such as *Marine Policy*, *Ecology & Society*, and *Forest policy & Economics*. The high disagreement in these meso-fields lends support to the hypothesis that no matter their high-level field, more socially-oriented topics generate a higher level of disagreement. Also of interest is that, in Life & Earth, several large fields with relatively high disagreement study the distant geological past or other inaccessible objects of studies, comprised of papers in journals such as the *Journal of Vertebrate Paleontology*, *Cretaceous Research*, and *Sedimentary Research*. A similar observation can be made in Phys & Engr, where astronomy-related fields featuring journals such as *Planetary and Space Science* and *Theoretical Biology* also exhibit above-average rates of disagreement, along with fields pertaining to research into superconductivity.

In addition to these quantitative results, we also perform a qualitative investigation of the individual papers that received the most disagreement citations, and papers which themselves included the most disagreement citations. First, we examine the citing paper perspective, that is those papers that issued the most citation sentences (Table C.6). These top papers demonstrate how methodological artefacts can contribute to these more extreme examples. For example, one of these papers considers the pedagogical and evaluative potential of debates in the classroom (Doody & Condon, 2012, Table C.6); the *debat** signal term incorrectly classifies several citations included in this paper as evidence of scientific disagreement. However, other papers offer interesting instances of

disagreement, and exemplify lessons that should be considered when quantifying disagreement. For instance, one such paper concerns meteorite impact structures (French & Koeberl, 2010, Table C.6) and includes discussion on the controversies in the field. Another is a review article arguing for multi-target agents for treating depressive states (Millan, 2006, Table C.6), and catalogs the controversies around the topic. Yet another is a book on *Neurotoxicology and Teratology*, misclassified as a research article, and illustrates how the length of an article can contribute to its likelihood of issuing a disagreement citation (Kalter, 2003, Table C.6).

Considering the cited paper perspective—those papers that received the most disagreement citations—reveals clear instances of disagreement in the literature. Many of the most disagreed with studies (Table C.5) relate to a single longstanding scientific controversy in the Earth sciences concerning the formation of the North Chinese Cranton, a tectonic structure spanning Northern China, Inner Mongolia, the Yellow Sea, and North Korea. This list of most-disagreed-with papers also includes a literature review that is cited as an exemplar of controversy in the field (Munro, 2003, Table C.5), and a paper on fMRI research that is heralded as a methodological improvement in the field, and is often cited to draw a contrast with other methods (Murphy, 2003, Table C.5). A more thorough discussion of papers that issue and receive the most disagreement can be found in the Supporting Information.

We also investigated the extent to which other factors related to disagreement. Younger papers (relative to the citing paper) are more likely to be disagreed with than older ones (Figure C.4). This seems to indicate that older cited papers are more likely to contain established claims, while younger papers are at the forefront of research and therefore more likely to be controversial. Author demographics do not appear to play a strong role; here, we observe little difference in disagreement based on the gender of the citing author (Figure C.7). We also explore whether disagreement relates to citation impact—our preliminary analysis reveals that contrary to the positive effect between conflict and citations observed previously [578], being cited in the context of disagreement has little

impact on citations in the subsequent year (see Supporting Information). Papers that themselves introduce language of disagreement, however, tend to receive more citations over their lifespan.

5.7 Discussion

When it comes to defining scientific disagreement, scholars disagree. Rather than staking out a specific definition, we adopt a broad operationalization of disagreement that incorporates elements of Kuhn’s accumulation of anomalies [187], Latour’s controversies [579], and more recent notions of uncertainty [555] and negative citations [554]. By bringing these past theories, we quantify the rate of disagreement across science. Roughly 0.31 percent of all citances in our dataset are instances of disagreement, a share that has remained relatively stable over time. However, this number is much smaller than in past studies—such as the 2.4 percent for so-called “negative” references [554], and the estimated 0.8 percent for “disputing” citations [556]. This is explained by our operationalization of disagreement, which although conceptually broader than negative or disputing citations, is narrowed to only 23 queries to prioritize precision. It does, however, reinforce that disagreement is an important aspect of citing, but one that is relatively rare, and that can take multiple and subtle forms. Moreover, studies differ in corpus used—one journal covering one field vs. a multidisciplinary corpus. The strength of our analysis is not the absolute incidence of disagreement, but its relative differences across disciplinary and social contexts.

Disagreement across fields can be interpreted using several theoretical frameworks. Differences in disagreement might stem from the epistemic characteristics of fields and their topics of study. For example, Auguste Comte’s hierarchy of sciences model [549] proposed that fields are organized based on the inherent complexity of their subject matter. We reinforce this model, finding that disagreement is highest in high-ambiguity fields like the social sciences and humanities, and lowest in low-ambiguity fields like physics and mathematics. While the hierarchy of sciences model is well-grounded theoretically [392] and bibliometrically [561, 580], other frameworks may be equally

useful in understanding disagreement across fields. For example, the structural characteristics of fields may explain their differences in disagreement. One such characteristic is how reliant the field is on Kuhnian paradigms [187]; so-called “hard” sciences, such as physics, may have strong theoretical paradigms and greater consensus (less disagreement) than “soft” sciences such as those in the social sciences and humanities [581]. The same framework can also help explain the changing rate of disagreement observed in physics and engineering: evidence suggests that the field is increasingly becoming consensual, and less prone to revolutions [582]. Social sciences and humanities have other characteristics that might be associated with more common or more intense conflicts, including low centralization of resources and control over research agendas, high diversity in their audiences and stakeholders, and limited standardization of methods and theories [40]. A field’s cultural characteristics also play a role in its norms of disagreement. Fields have different norms when it comes to consensus formation and the settling of disputes [70], and some fields even value disagreement as an important element of scholarship. For instance, a cultural norm of “agonism”, or ritualized adversarialism, is common in many humanities fields, wherein one’s arguments are framed in direct opposition to past arguments [347]. Fields also have distinct cultures of evaluation, which shapes how they judge each other’s work and impacts whether they are likely to reach consensus [326]. Of course, epistemic, structural, and cultural characteristics of fields are all inter-related—cultural practices emerge in part from structural characteristics of a field, such as access to expensive instruments, which in turn are related to the epistemic aspects of the object of study. Our data does not allow us to disentangle these relationships or argue which is most appropriate, but each offers a useful lens for understanding why disagreement might differ between fields.

Expanding our analysis into a more fine-grained classification of fields reveals greater detail into where disagreement happens in science. We observed that socially-oriented meso-level fields tended to have a higher rate of disagreement, no matter their main field. For example, meso-fields

concerning healthcare policy had higher rates of disagreement than others in Biomedical & Health Sciences, whereas the meso-field concerning transportation science had a higher rate of disagreement than all others in Math & Computer science. Though these fields draw on the expertise of traditionally hard-science fields, they do so in order to study social processes and address social questions. In Life & Earth Sciences, disagreement was especially high in meso-fields that study the earth’s geological and paleontological history. In these fields, much like in the social sciences, researchers cannot easily design experiments, and so progress instead comes from debate over competing theories using limited evidence and reconstructed historical records. This is exemplified by paleontology, in which a 2017 paper sparked controversy and forced a re-interpretation of the fossil record and a 130-year-old theory of dinosaur evolution [583, 584]. Similarly, our approach identified a major controversy in the earth sciences—the formation of the North Chinese Cranton—again illustrating how reliance on historical records might exacerbate disagreement. These cases illustrate the heterogeneity of disagreement in science, and illustrate that existing theoretical frameworks, such as the hierarchy of science, can oversimplify the diversity of cultural norms and epistemic characteristics that manifest at more fine-grained levels of analysis.

Our approach comes with limitations. First, our method prioritizes precision at the cost of recall; as a result, our measure is non-exhaustive and only captures a fraction of disagreement in science. Explicit disagreement could be observed by expanding our set of signal terms, or by using the context around each citance to improve identification. Explicit disagreement may also be more subtle than our approach can detect; scientific writers can use vague words or technical jargon to hide their disagreement, which are not included in our signal terms. Because we capture only a fraction of explicit disagreement, what we do capture may be biased, over-representing certain fields that use our signal terms, and under-representing others. Second, in spite of its overall precision, our approach returns many false positives, especially in particular disciplinary contexts. For example, the signal term *conflict** matches to topics of study and theories in the

fields of sociology and international relations (e.g., “ethnic conflicts”, “Conflict theory”). In other instances, a signal term can even match an author’s name (as in the surname “Debat”). We also find that these artefacts are over-represented among the papers with the most disagreement citations, and those that received the most disagreement citations (see Supporting Information). However, given our extensive validation, these artefacts remain a small minority of all disagreement sentences identified, though they should be considered when interpreting disagreement in small sub-fields. Finally, our inclusive definition of disagreement homogenizes disagreement into a single category, whereas there are many kinds of disagreement in science. For example, the ability to differentiate between paper-level and community-level disagreement could lend insight into how conflict and controversy manifest in different fields. This definition could also be developed to differentiate further between types of disagreement: for example, past citation classification schemes have differentiated between “juxtaposition” and “negational” citations [570], or between “weakness” and “contrast” citations [571, 572] which could also be used here.

Despite these limitations, our framework and study have several advantages. First, in contrast to keyword-based analyses, our approach provides a nuanced view of disagreement in science, revealing the differences in disagreement not only between signal terms, but also based on filter terms. This drives the second advantage of our approach—that its inherent transparency allows us to easily identify confounding artefacts such as when a signal term is an object of study (i.e., “international conflicts”, “public debate”), when it relates to disciplinary jargon (i.e., “disproving theorems” in Mathematics, or “strongly disagree” in survey studies that use Likert scales) or when the keyword is part of a proper name (i.e., “work by Debatin et al.”). These issues are a concern for any dictionary-based or automated analysis of scientific texts—the usage and meaning of words varies across fields. In contrast to black-box style machine learning approaches, ours is transparent and can easily be validated, interpreted, and replicated. Finally, by being open and transparent, our approach is easily adjustable to different contexts. Our initial identification of keywords was the result of an

iterative process of exploration and validation, which eventually resulted in a set of signal and filter terms, and then a final set of validated queries. Any step of this process can be tuned, extended, and improved to facilitate further studies of scientific disagreement—new signals or filters can be introduced, queries can be modified to be even more precise, and the threshold of validity changed; here, for example, we assessed our results by setting a more inclusive threshold for which queries constitute disagreement, and find the results remain robust (see Supporting Information). To assist in further efforts to validate and extend our work, we have made annotated sentences and code that can reproduce this analysis publicly available at github.com/murrayds/sci-text-disagreement.

Our approach is also robust (see Supporting Information), and can be built upon. Whereas black-box machine learning approach have many strengths [585], ours is transparent and intuitive. Its transparency allows to easily identify terms that have field-specific meanings, which may be obfuscated in black-box approaches. Our method is also reproducible and can be refined and extended with additional signal and filter terms. The general method of identifying and manually validating signal and filter terms can also be applied to other scientific phenomena, such as identifying uncertainty [555], negativity [554], and discovery [586].

Future research could refine existing queries and link them to different conceptual perspectives on disagreement in science. Using a refined set of queries, such work could build on our analyses of the factors that may affect disagreement, such as gender, paper age, and citation impact. Disagreement is an essential aspect of knowledge production, and understanding its social, cultural, and epistemic characteristics will reveal fundamental insights into science in the making.

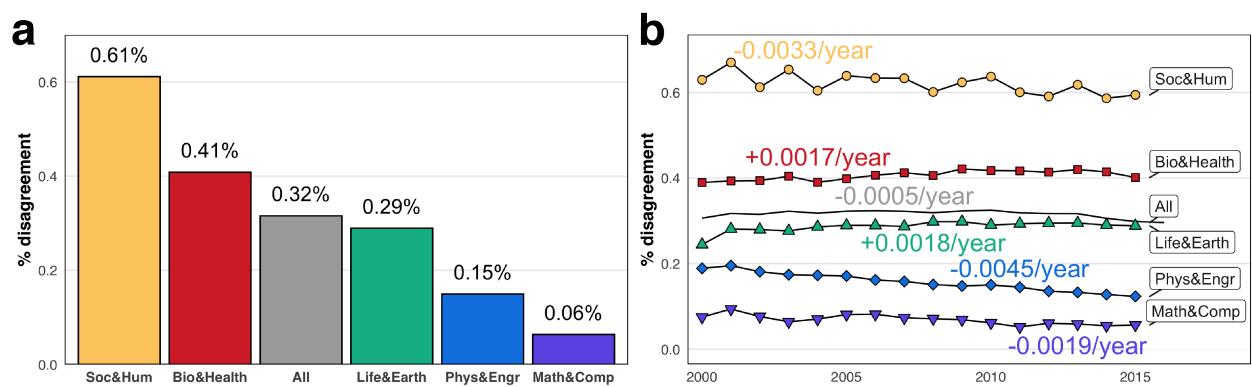


Figure 5.2: **Disagreement reflects a hierarchy of fields.** **a.** Percent of all citances in each field that contain signals of disagreement, meaning they were returned by one of the 23 queries with validity of 80 percent or higher. Fields marked by lower consensus, such as in Soc & Hum, had a greater proportion of disagreement. **b.** Percent of disagreement by field and over time, showing little change overall, but some changes by field. Text indicates the average percentage-point change per-year by field.

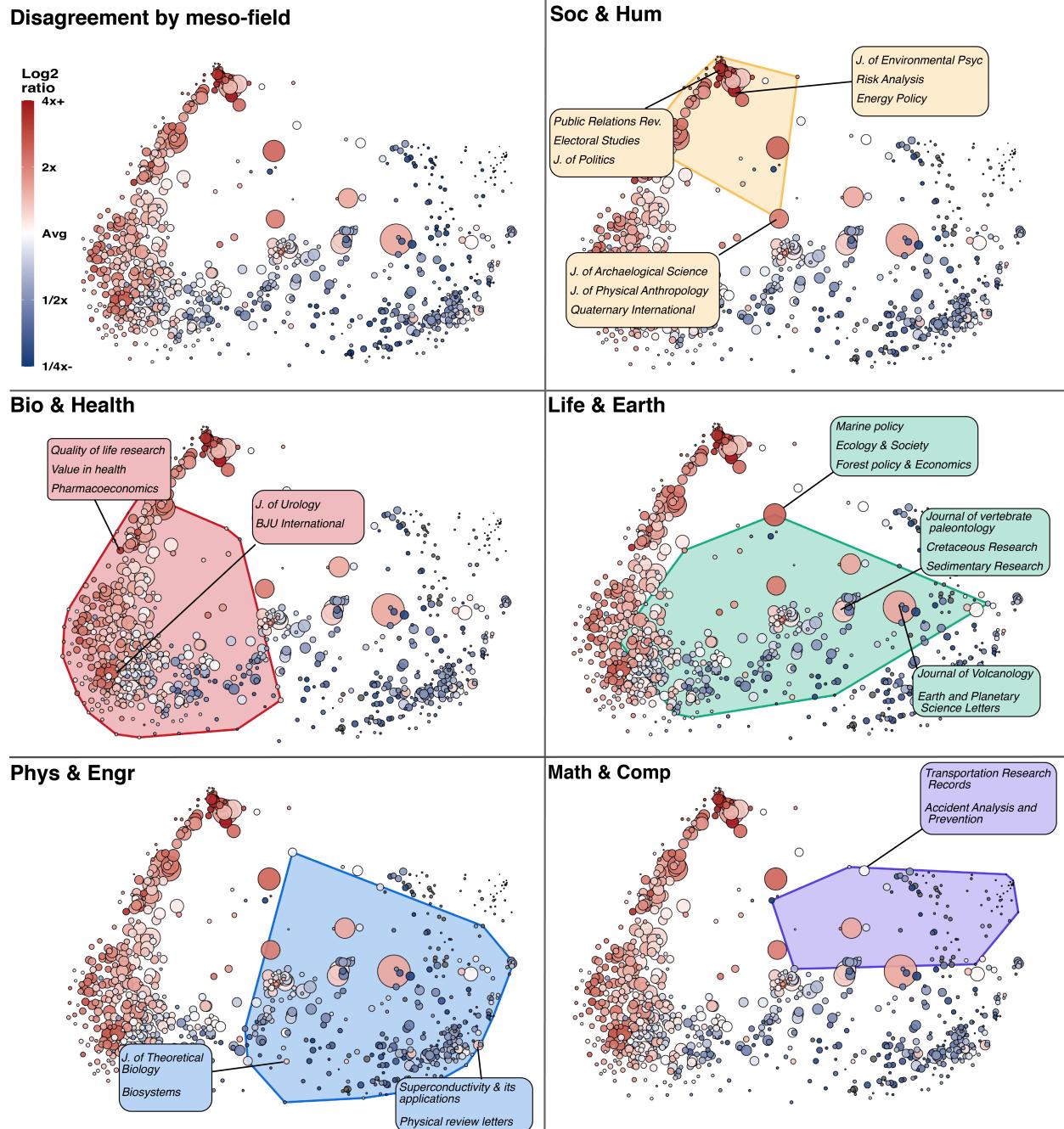


Figure 5.3: **Heterogeneity in disagreement across meso-fields.** Fine-grained view across 817 meso-level fields, each a cluster of publications grouped and positioned based on their citation links derived from the Web of Science database (see Materials & Methods), 2000-2015. The area of each point is proportional to the number of disagreement citances in that field. Color maps to the log ratio of the share of disagreement citances given the mean share across all fields, truncated at 4x greater and 4x lower than the mean. Soc & Hum tends to have a greater proportion of disagreement citances, and Math & Comp the least. Other panels show the same data, but highlight the meso-fields in each high-level field. Meso-fields of interest are highlighted, and labels show a selection of journals in which papers in each field are published. Journals listed in labels are representative of each meso-field in the Web of Science, and is not limited to those represented in the Elsevier ScienceDirect data. An interactive version of this visualization is available online at <https://tinyurl.com/disagreement-meso-fields>.

Chapter 6

Study 4: The landscape of global scientific mobility

6.1 Foreword

Migration is a force of nature, rooted in
human biology and history, along with
that of the scores of other wild species
with whom we share this changing planet.

Over the long history of life on earth, its
benefits have outweighed its costs

Sonia Shah

Science has become more global, with more people travelling between more countries than ever before [55, 56]. This mobility drives scientific impact [294, 295], innovation [297, 587], and the diffusion of ideas [116, 194]. However, the complexity of mobility combined with the difficulty of matching bibliographic records to individuals has so far made studying it, at scale, difficult.

In this chapter, I present a study that leverages a massive dataset of millions of mobile scientists from around the world, and learn a representation of global mobility using recent advancements in *neural embedding* [98]. I demonstrate that this representation better captures real-world mobility than alternatives such as geographic distance or other embedding methods, and examine the structure of mobility by applying techniques unique to neural embeddings, revealing its complicated and multi-faceted structure. These findings point to the complexity of mobility, and the need to understand the many factors that contribute to it, and how they vary at different geographic scales.

When viewed under the complexity perspective, this study highlights how a multiplicity of subtle, individual-level forces can contribute to a single phenomenon in a complex system. Mobility is not a decision made at random, but instead is influenced by a person's career stage, their family,

and the language, culture, and economies of their origin and destination countries. I also find that institutional prestige plays an important role in structuring mobility, evidence of how mobility can act as a feedback loop that both creates and perpetuates status hierarchies among both individuals and institutions.

This manuscript is currently under review and revisions at Nature Human Behavior. A pre-print has been made publicly available on arXiv under the title "Unsupervised embedding of trajectories captures the latent structure of mobility". I was co-first author of this work alongside Jisung Yoon, who together with Sadamori Kojaku made invaluable contributions to developing the mathematical foundations of this project alongside empirical analyses. In addition, this work was co-authored by Rodrigo Costas, Woo-Sung Jung, Staša Milojević, and Yong-Yeol Ahn. I would like to thank the Center for Science and Technology Studies at Leiden University for managing and making available the dataset of scientific mobility, as well as the Goodchoice Company LTD. for making available the dataset of Korean accommodation reservation data. For their comments, I thank Guillaume Cabanac, Cassidy R. Sugimoto, Vincent Larivière, Alessandro Flammini, Filippo Menczer, Lili Miao, Xiaoran Yan, Inho Hong, and Esteban Moro Egido. This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-19-1-0391. Rodrigo Costas is partially funded by the South African DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP). The U.S. airline itinerary dataset can be found at <https://www.transtats.bts.gov/DataIndex.asp>. The raw Korean accommodation reservation dataset, due to privacy concerns, cannot be shared publicly. Due to its proprietary nature, the global scientific mobility dataset, sourced from the Web of Science, cannot be provided; however, metadata and trained neural embeddings have been published at <https://doi.org/10.6084/m9.figshare.13072790.v1>

6.2 Abstract

Human mobility drives major societal phenomena including epidemics, economies, and innovation. Historically, mobility was constrained by geographic distance, however, in the globalizing world, language, culture, and history are increasingly important. We propose using the neural embedding model *word2vec* for studying mobility and capturing its complexity. *Word2ec* is shown to be mathematically equivalent to the gravity model of mobility, and using three human trajectory datasets, we demonstrate that it encodes nuanced relationships between locations into a vector-space, providing a measure of effective distance that outperforms baselines. Focusing on the case of scientific mobility, we show that embeddings uncover cultural, linguistic, and hierarchical relationships at multiple levels of granularity. Connecting neural embeddings to the gravity model opens up new avenues for the study of mobility.

6.3 Introduction

How far apart are two places? The question is surprisingly hard to answer when it involves human mobility. Although geographic distance has historically constrained human movements, it is becoming less relevant in a world connected by rapid transit and global airline networks. For instance, a person living in Australia is more likely to migrate to the United Kingdom, a far-away country with similar language and culture, than to a much closer country such as Indonesia [588]. Similarly, a student in South Korea is more likely to attend a university in Canada than one in North Korea [4]. Although geographic distance has been used as the most prominent basis for models of mobility, such as the Gravity [589] and Radiation [590] models, there have been attempts to define alternative notions of distance, or functional distances [289, 591, 592], from real-world data or *a priori* relationships between geographic entities.

Yet, functional distances are often low-resolution, computed at the level of countries rather than regions, cities, or organizations, and have focused on only a single facet of mobility at a time,

whereas real-world mobility is multi-faceted, influenced simultaneously by geography, language, culture, history, and economic opportunity. Low dimensional distance alone cannot represent the multitude of inter-related factors that drive mobility. Networks offer a solution to representing many dimensions of mobility, yet edges only encode simple relationships between connected entities. Capturing the complexity of mobility requires moving beyond simple functional distances and networks, to learning high-dimensional landscapes of mobility that incorporate the many facets of mobility into a single fine-grained and continuous representation.

Here, we apply a neural embedding framework to real-world mobility trajectories and demonstrate that it can encode the complex landscape of human mobility into a dense and continuous vector-space representation, from which we can not only derive a meaningful functional distance between locations but also probe relationships based on culture, language, and even prestige along with the geographic relationship. We embed trajectories from three massive datasets: U.S. passenger flight itinerary records, Korean accommodation reservations, and a dataset of scientists' career mobility between organizations captured in bibliometric records (Detailed descriptions are available in the Methods).

The flight itinerary data, from the Airline Origin and Destination Survey, consists of records of more than 300 million itineraries between 1993 and 2020 documenting domestic flights between 828 airports in the United States. A trajectory is constructed for each passenger flight itinerary, forming an ordered sequence of unique identifiers of the origin and destination airports. The Korean accommodation reservations consist of customer reservation histories across 2018 and 2020 for 1,038 unique accommodation locations in Seoul, South Korea. A trajectory is constructed for each customer, containing the ordered sequences of accommodations they reserved over time. Finally, we use scientific mobility data that captures the affiliation trajectories of nearly 3 million scientists across ten years. We focus in more detail on scientific mobility due to its richness and importance. Scientific mobility—which is a central driver of the globalized scientific enterprise [177, 306] and

strongly related to innovation [194, 297], impact [294, 295], collaboration [84], and the diffusion of knowledge [116, 194]—is not only an important topic in the Science of Science but also ideal for our study thanks to its well-known structural properties such as the centrality of scientifically advanced countries and the strong prestige hierarchy [184, 593]. In spite of its importance, understandings of scientific mobility have been limited by the sheer scope and complexity of the phenomenon [318, 593], being further confounded by the diminishing role of geography in shaping the landscape of scientific mobility.

Trajectories of scientific mobility are constructed using more than three million name-disambiguated authors who were *mobile*—having more than one affiliation—between 2008 and 2019, as evidenced by their publications indexed in the Web of Science database (see Methods). As a scientist’s career progresses, they move between organizations or pick up additional (simultaneous) affiliations forming *affiliation trajectories* (Fig. 6.1). Thus, the trajectories encode both migration and co-affiliation—the holding of multiple simultaneous affiliations involving the sharing of time and capital between locations—that is typical of scientific mobility [84, 295](see Appendix D).

A vector-space embedding of locations (airports, accommodations, and organizations) is learned by using trajectories as input to the standard with skip-gram negative sampling, or *word2vec* neural-network architecture (see Methods). This neural embedding model, originally designed for learning language models [98], has been making breakthroughs by revealing novel insights into texts [594–599], networks [600, 601] and trajectories [602–607]. It works under the notion that a good representation should facilitate prediction, learning a mapping between words that can predict a target word based on its context (surrounding words). The model is also computationally efficient, robust to noise, and can encode relations between entities as geometric relationships in the vector space [596, 597, 608, 609]. As a result, each location is encoded into a single vector representation, and vectors relate to one another based on the likelihood of locations appearing adjacent to one another in the same trajectory.

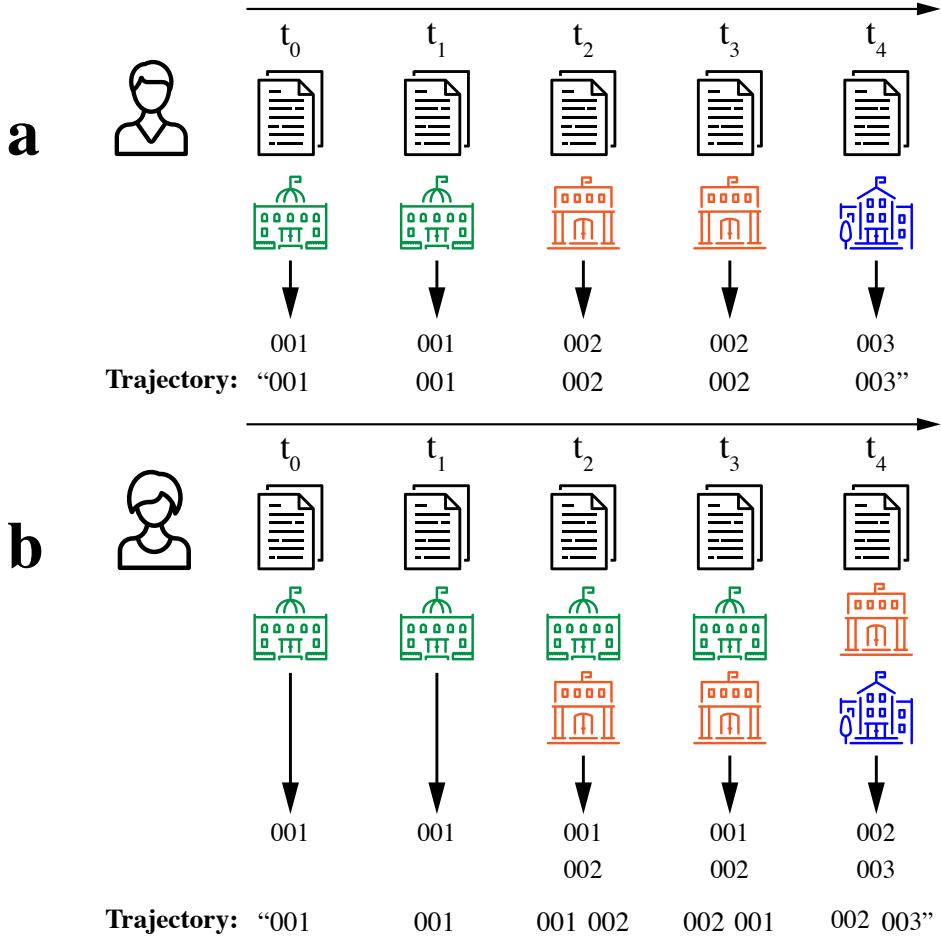


Figure 6.1: **Construction of affiliation trajectories from publication records.** **a.** An author published five papers across five time periods, with only one affiliation listed in the byline of each paper. A unique identifier is assigned to each organization and they are assembled into an affiliation trajectory ordered by year of publication. **b.** If an author lists multiple organization affiliations within the same year, then organization IDs within that year are placed in random order in each training iteration of the *word2vec* model (for more detail, see Appendix D).

To validate our approach, we evaluate the quality of vector representations based on their performance in predicting real-world mobility flows using the gravity model framework [589]. The gravity model is a widely used mobility model [610–613] that connects the *expected flux*, \hat{T}_{ij} , between locations based on their populations and distance:

$$\hat{T}_{ij} = C m_i m_j f(r_{ij}), \quad (6.1)$$

where m_i is the population of location i , $f(r_{ij})$ is a decay function with respect to distance between

locations, and C is a constant estimated from data (see Methods). For the flight itinerary data, we use population m_i as the total number of unique passengers who passed through each airport, for the Korean accommodation reservation data, we use the total number of unique customers who booked with each accommodation, and for scientific mobility, we use the mean annual number of unique mobile and non-mobile authors who were affiliated with each organization. \hat{T}_{ij} , which is often referred to as “expected flux” [590], is the expected frequency of the co-occurrence of location i and j in the trajectory in the gravity model.

The gravity model dictates that the expected flow, \hat{T}_{ij} , ($\hat{T}_{ij} = \hat{T}_{ji}$), is proportional to the locations’ population, $\hat{T}_{ij} \propto m_i m_j$, and decays as a function of their distance, $f(r_{ij})$. We define the distance function in terms of either the geographic distance between locations or their functional distance in the vector space, which is calculated as the cosine distance between their vectors, termed the *embedding distance*. The decay function $f(r_{ij})$ defines the effect of distance, and different decay functions can model fundamentally different mechanisms [614] such as the cost functions for a given distance and the spatial granularity of the observation. For geographic distance, we define $f(r_{ij})$ as the standard power-law function, and for the embedding distance, we use the exponential function, selected as the best performing for each case (Fig. D.7 and Fig. D.8). Though there is no qualitative difference, the embedding distance outperforms geographic distance regardless of the type of gravity model used (see Tables D.3, D.4, D.5). The power-law decay function performs best with the geographic distance, likely resulting from the function’s suitability for large, complex, and scale-free spatial systems [615]; performance of the embedding distance is similar between the functional forms, though best for the exponential form, which stems from an underlying connection between the function and *word2vec* that we discuss later.

6.4 Embeddings provide functional distance between locations

We show that the embedding distance better predicts actual mobility flows than the geographic distance across three disparate datasets. In the case of flight itineraries, the embedding distance explains more than twice the expected flux between airports ($R^2 = 0.51$, Fig. 6.2a) than does geographic distance ($R^2 = 0.22$). Also, the embedding distance produces better predictions of actual flux between airports than does the geographic distance (Fig. 6.2b). In the case of Korean accommodation reservations, embedding distance better explains the expected flux ($R^2 = 0.57$, Fig. 6.2c) than does geographic distance ($R^2 = 0.25$), and predictions made using the embedding distance outperform those made with geographic distance (Fig. 6.2d). This performance is consistent in the case of scientific mobility: the embedding distance explains more than twice the expected flux ($R^2 = 0.48$, Fig. 6.2e) than does the geographic distance ($R^2 = 0.22$), and predictions made using the embedding distance outperform those using the geographic distance (Fig. 6.2f). These patterns hold for the subsets of only domestic (within-country organization pairs, Fig. D.7 and Fig. D.9c) and only international mobility flows (across-country organization pairs, Fig. D.9d). The embedding distance also out-performs alternative diffusion-based network distance measures including the personalized-page rank scores calculated from the underlying mobility network (Fig. D.5, Fig. D.11, Fig. D.12). The embedding distance derived from neural embedding also explains more of the flux and better predicts mobility flows than simpler embedding baselines, such as distances derived from a singular-value decomposition and a Laplacian Eigenmap embedding [616] of the underlying location co-occurrence matrix, Levy's symmetric *word2vec*[608], and even direct optimization of the gravity model (Fig.D.5 and Tables D.3, D.4, D.5). In sum, our results demonstrate that, consistently and efficiently, the embedding distance better captures patterns of actual mobility than does the geographic distance.

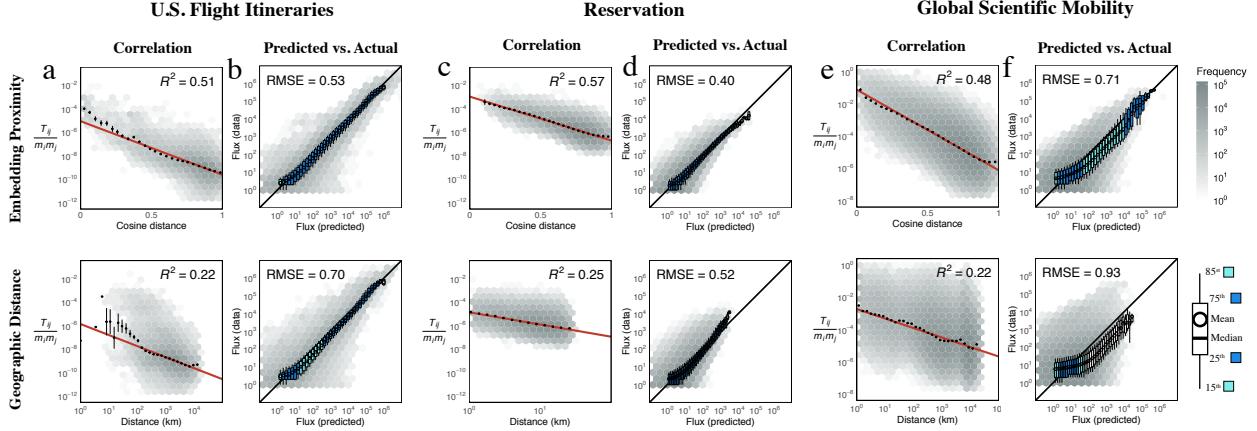


Figure 6.2: Embedding distance encodes functional distance and better predicts mobility in flights, accommodation reservations, and global scientific mobility. **a.** Embedding distance (cosine distance between organization vectors, top) better explains the expected flux of passengers between U.S. airports than does geographic distance (bottom). The red line is the line of the best fit. Black dots are mean flux across binned distances. 99% confidence intervals are plotted for the mean flux in each bin based on a normal distribution. Correlation is calculated on the data in the log-log scale ($p < 0.0001$ across all fits). The lightness of each hex bin indicates the frequency of organization pairs within it. **b.** Predictions of flux between airport pairs made using embedding distance (top) outperform those made using geographic distance (bottom). Box-plots show the distribution of actual flux for binned values of predicted flux. Box color corresponds to the degree to which the distribution overlaps with $y = x$; a perfect prediction yields all points on the black line. “RMSE” is the root-mean-squared error between the actual and predicted values. Results are consistent in the case of scientific mobility. For Korean accommodation reservations, embedding distance better explains the expected flux than does geographic distance (**c**), and produces better predictions (**d**). Similarly, in the case of global scientific mobility, embedding distance explains the expected flux between organizations (**e**) and allows for better predictions (**f**) than geographic distance.

6.5 word2vec and the gravity model

Why does *word2vec* with the skip-gram with negative sampling model (SGNS) work so well to model mobility? The reason for that is the mathematical equivalence between the SGNS model and the gravity model which we demonstrate below.

The *word2vec* model takes a location trajectory, denoted by (a_1, a_2, \dots, a_T) , as input. A target location $a_t = i$ is considered to have a context location $a_{t'} = j$ that appears in the previous or subsequent w locations in the trajectory, i.e., $j \in [a_{t-w}, \dots, a_{t-1}, a_{t+1}, \dots, a_{t+w}]$. *word2vec* learns

an embedding by estimating the probability that location i has context j :

$$P(j | i) := \frac{\exp(\mathbf{u}_j \cdot \mathbf{v}_i)}{Z_i}, \quad (6.2)$$

where the denominator $Z_i = \sum_{j' \in \mathcal{A}} \exp(\mathbf{u}_{j'} \cdot \mathbf{v}_i)$ is a normalization constant, and \mathcal{A} is the set of all locations. Although *word2vec* generates two embedding vectors \mathbf{v}_i and \mathbf{u}_i —referred to as the in-vector and out-vector, respectively—we follow convention to use the in-vector \mathbf{v}_i as an embedding of location i .

Our experiments show that *word2vec* best explains real-world mobility flow when $w = 1$ (Fig. D.4), with the flow predicted by

$$\hat{T}_{ij} \propto P(i)P(j | i) = \frac{P(i)\exp(\mathbf{u}_j \cdot \mathbf{v}_i)}{Z_i}, \quad (6.3)$$

where $P(j)$ is the fraction of j in the data. In general, calculating Z_i is computationally expensive and there are two common approximations: hierarchical softmax [617] and negative sampling [98]. Due to its simplicity and performance, negative sampling is the most widely used strategy, which we also adopt in our study.

Although negative sampling is the most common approximation, it is a biased estimator [618, 619] and fits a different probability model. When taking into account this bias, *word2vec* with skip-gram and negative sampling fits a probability model given by

$$P(j | i) := \frac{P^\gamma(j)\exp(\mathbf{u}_j \cdot \mathbf{v}_i)}{Z'_i}, \quad (6.4)$$

where we redefine the normalization constant as $Z'_i = \sum_{j' \in \mathcal{A}} P^\gamma(j')\exp(\mathbf{u}_{j'} \cdot \mathbf{v}_i)$. (See the Methods and Supporting Information for the full derivation).

Parameter $\gamma = 1$ is a special choice that ensures that, when the embedding dimension is suffi-

ciently large, there exists optimal in-vectors and out-vectors such that $\mathbf{v}_i = \mathbf{u}_i$ [608]. Setting $\gamma = 1$ and substituting $\mathbf{v}_i = \mathbf{u}_i$ into Eq. 6.4, the flow predicted by *word2vec* is given by

$$\widehat{T}_{ij} \propto N(i)P(j | i) = \frac{NP(i)P(j) \exp(\mathbf{v}_j \cdot \mathbf{v}_i)}{Z'_i}. \quad (6.5)$$

where $N(i) = NP(i)$ is the frequency of location i in the data, and $N = \sum_{i \in \mathcal{A}} N(i)$ is the sum of frequencies of all locations.

The flow \widehat{T}_{ij} is symmetric (i.e., $\widehat{T}_{ij} = \widehat{T}_{ji}$) because the skip-gram model neglects whether the context j appears before or after the target i in the trajectory. If we swap i and j in Eq. 6.5, the numerator remains the same but the denominator can be different. Therefore, to ensure $\widehat{T}_{ij} = \widehat{T}_{ji}$, the denominator Z_i should be a constant.

Taken together, the *word2vec* model with the negative sampling predicts a flow in the same form as in the gravity model:

$$\widehat{T}_{ij} = CP(i)P(j) \exp(\mathbf{v}_j \cdot \mathbf{v}_i). \quad (6.6)$$

In other words, *word2vec with the skip-gram negative sampling is equivalent to the gravity model*, with the mass given by the location's frequency $P(i)$, and the distance measured by their dot similarities. While the gravity model predicts mobility flows from the given mass and locations, *word2vec* finds locations that best explain the given mobility flow. Therefore, the embedding distance is inherently tied to the mobility flow, and hence, has greater predictive power than the geographic distance and other baselines.

In practice, because we only have limited amounts of noisy data and the optimization may not find the true optimum, the mathematical result may only approximately hold. Indeed, we find that the in- and out-vectors tend to be different and that the cosine similarity tends to better capture real-world mobility than the dot similarity. This result echos other applications of word

embedding, such as word analogy testing [620], in which cosine distance also outperformed dot similarity. Nevertheless, a model with dot similarity has the second-best performance after cosine similarity (Tables D.3, D.4, D.5), and the embedding distance still outperforms all alternatives we considered.

6.6 Embeddings capture global structure of mobility

In the remainder of the paper, we focus on scientific mobility and interrogate the geometric space generated by the neural embedding to shed light on the multi-faceted relationships between organizations. To explore the topological structure of the embedding, we use a topology-based dimensionality reduction method (UMAP [621]) to obtain a two-dimensional representation of the embedding space (Fig. 6.3a). By showing relationships between individual organizations, rather than aggregates such as nations or cities, this projection constitutes the largest and highest resolution “map” of scientific mobility to date.

Globally, the geographic constraints are conspicuous; organizations tend to form clusters based on their national affiliations and national clusters tend to be near their geographic neighbors. At the same time, the embedding space also reflects a mix of geographic, historic, cultural, and linguistic relationships between regions much more clearly than alternative network representations (Fig. D.13) that have been common in studies of scientific mobility [177, 292].

The embedding space also allows us to *zoom in* on subsets and re-project them to reveal local relationships. For example, re-projecting organizations located in Western, Southern, and Southeastern Asia with UMAP (Fig. 6.3b) reveals a gradient of countries between Egypt and the Philippines that largely corresponds to geography, but with some exceptions seemingly stemming from cultural and religious similarity; Malaysia, with its official religion of Islam, is nearer to Middle Eastern countries in the embedding space than to many geographically-closer South Asian countries. We validate this finding quantitatively with the cosine distance between nations (the centroids of orga-

nizations vectors belonging to that country). Malaysia is nearer to many Islamic countries such as Iraq ($d = 0.27$), Pakistan ($d = 0.32$), and Saudi Arabia ($d = 0.41$) than neighboring but Buddhist Thailand ($d = 0.43$) and neighboring Singapore ($d = 0.48$).

Linguistic and historical ties also affect scientific mobility. We observe that Spanish-speaking Latin American nations are positioned near Spain (Fig. 6.3c), rather than Portuguese-speaking Brazil ($d = 0.35$ vs. $d = 0.54$ for Mexico and $d = 0.39$ vs. $d = 0.49$ for Chile) reflecting linguistic and cultural ties. Similarly, North-African countries that were once under French rule such as Morocco are closer to France ($d = 0.32$) than to similarly geographically-distant European countries such as Spain ($d = 0.39$), Portugal ($d = 0.52$), and Italy ($d = 0.52$). Comparable patterns exist even within a single country. For example, organizations within Quebec in Canada are located nearer France ($d = 0.37$) than the United States ($d = 0.51$).

Mirroring the global pattern, organizations in the United States are largely arranged according to geography (Fig. 6.3d). Re-projecting organizations located in Massachusetts (Fig. 6.3e) reveals structure based on urban centers (Boston vs. Worcester), organization type (e.g., hospitals vs. universities), and university systems (University of Massachusetts system vs. Harvard & MIT). For example, even though UMass Boston is located in Boston, it clusters with other universities in the UMass System ($d = 0.29$) rather than the other typically more highly-ranked and research-focused organizations in Boston ($d = 0.39$), implying a relative lack of mobility between the two systems. Similar structures can be observed in other states such as among New York's CUNY and SUNY systems (Fig. D.14), Pennsylvania's state system (Fig. D.15), Texas's Agricultural and Mechanical universities (Fig. D.16), and between the University of California and State University of California systems (Fig. D.17).

Just as the embedding space makes it possible to *zoom in* on subsets of organizations, it is also possible to *zoom out* by aggregating organizational vectors. In doing so, we can examine the country-level structure that governs scientific mobility. We define the representative vector of each

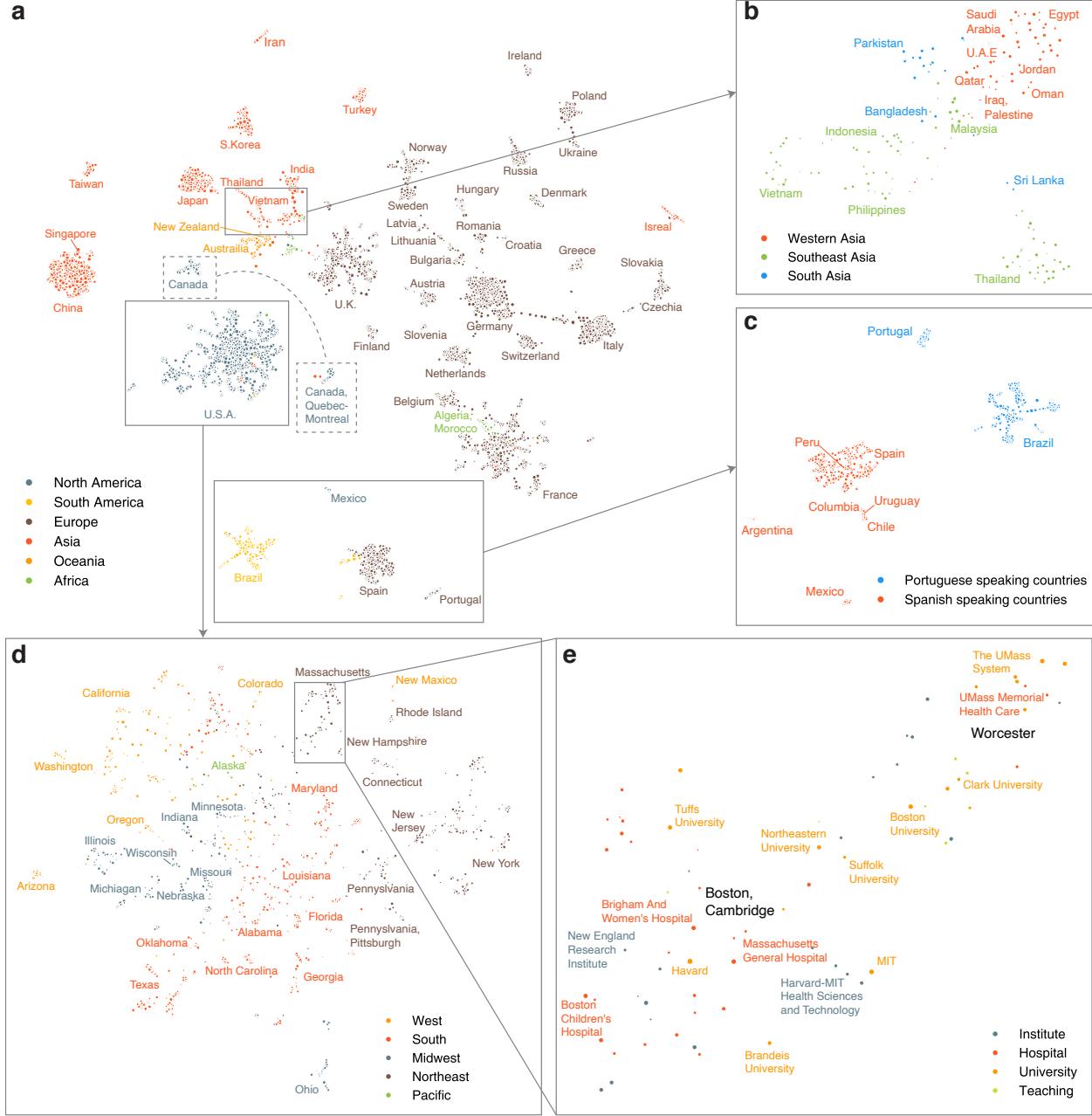


Figure 6.3: Projection of embedding space reveals complex multi-scale structure of organizations. **a.** UMAP projection [621] of the embedding space reveals country-level clustering. Each point corresponds to an organization and its size indicates the average annual number of mobile and non-mobile authors affiliated with that organization from 2008 to 2019. Color indicates the region. The separation of organizations in Quebec and the rest of Canada is highlighted. **b.** Zooming into (re-projecting) the area containing countries in Western, South, and Southeast Asia shows a geographic and cultural gradient of country clusters. **c.** Similarly, zooming into the area containing organizations in Spain, Portugal, South, and Central America shows clustering by most widely-spoken majority language group: Spanish and Portuguese. **d.** Doing the same for organizations in the United States reveals geographic clustering based on state, roughly grouped by Census Bureau-designated regions. **e.** Zooming in further on Massachusetts reveals clustering based on urban center (Boston, Worcester), organizational sector (hospitals vs. universities), and university systems and prestige (UMass system vs. Harvard, MIT, etc.).

country as the average of their organizational vectors and, using their cosine similarities, perform hierarchical clustering of nations that have at least 25 organizations represented in the embedding space (see Fig. 6.4a). The six identified clusters roughly correspond to countries in Asia and North America (orange), Northern Europe (dark blue), the British Commonwealth and Iran (purple), Central and Eastern Europe (light blue), South America and Iberia (dark green), and Western Europe and the Mediterranean (light green). The cluster structure shows that not only geography but also linguistic ties and cultural between countries are related to scientific mobility.

We quantify the relative importance of geography (by region), and language (by the most widely-spoken language of each country) using the element-centric clustering similarity [622], a method that can compare hierarchical clustering and disjoint clustering (geography, language...) at the different level of hierarchy by explicitly adjusting a scaling parameter r , acting like a *zooming lens*. If r is high, the similarity is based on the lower levels of the dendrogram, whereas when r is low, the similarity is based on higher levels. Fig. 6.4b demonstrates that regional relationships play a major role at higher levels of the clustering process (low r), and language (family) explains the clustering more at the lower levels (high r). This suggests that the embedding space captures the hierarchical structure of mobility.

Embeddings capture latent prestige hierarchy

Prestige hierarchy is known to underpin the dynamics of scientific mobility, in which researchers tend to move to similar or less prestigious organizations [184, 593]. Could the embedding space, to which no explicit prestige information is given, encode a prestige hierarchy? This question is tested by exploiting the geometric properties of the embedding space with SemAxis [609]. Here, we use SemAxis to operationalize the abstract notion of academic prestige, defining an axis in the embedding space where poles are defined using known high- and low-ranked universities. As an external proxy of prestige, we use the Times Ranking of World Universities (we also use research

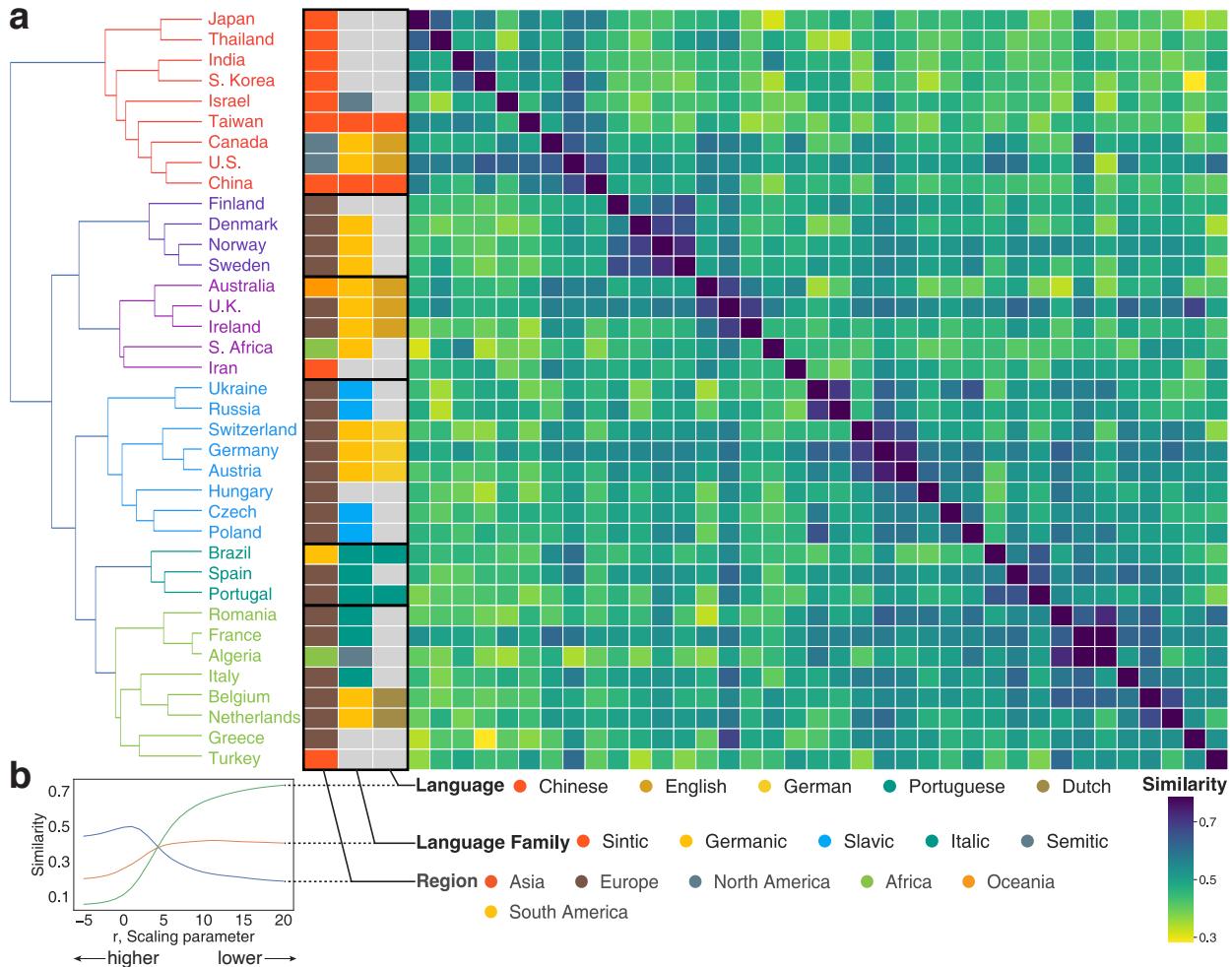


Figure 6.4: Geography, then language, conditions international mobility. **a.** Hierarchically clustered similarity matrix of country vectors aggregated as the mean of all organization vectors within countries with at least 25 organizations. Color of matrix cells corresponds to the cosine similarity between country vectors. Color of country names corresponds to their cluster. Color of three cell columns separated from the matrix corresponds to, from left to right, the region of the country, the language family, taken from the Ethnologue Global Dataset, found at <https://www.ethnologue.com/>, and the dominant language. **b.** Element-centric cluster similarity [622] reveals the factors dictating hierarchical clustering. Region better explains the grouping of country vectors at higher levels of the clustering. Language family, and then the most widely-spoken language, better explain the fine-grained grouping of countries.

impact from the Leiden Ranking [151], see Appendix D); the high-rank pole is defined as the average vector of the top five U.S. universities according to the rankings, whereas the low-rank pole is defined using the five bottom-ranked (geographically-matched by U.S. census region) universities. We derive an embedding-based ranking for universities based on the geometrical spectrum from the high-ranked to low-ranked poles (see Data and Methods).

The embedding space encodes the prestige hierarchy of U.S. universities that are coherent with real-world university rankings. The embedding-based ranking is strongly correlated with the Times ranking (Spearman's $\rho = 0.73$, Fig. 6.5a). For reference, the correlation between the Times ranking and the publication impact scores from the Leiden Ranking [151], a bibliometrically-based university ranking, is 0.87 (Spearman's ρ , Fig. 6.5b). The correlation between the embedding-based ranking and the Times ranking is robust regardless of the number of organizations used to define the axes (Fig. D.18), such that even using only the single top-ranked and bottom-ranked universities produces a ranking that is significantly correlated with the Times ranking (Spearman's $\rho = 0.46$, Fig. D.18a). The correlation is also comparable to more direct measures such as node strength (sum of edge weights, Spearman's $\rho = 0.73$) and eigenvector centrality (Spearman's $\rho = 0.76$, see Appendix D) from the mobility network. The strongest outliers that were ranked more highly in the Times ranking than in the embedding-based ranking tend to be large state universities such as Arizona State University and the University of Florida. Those ranked higher in the embedding-based ranking tend to be relatively-small universities near major urban areas such as the University of San Francisco and the University of Maryland Baltimore County, possibly reflecting exchanges of scholars with nearby high-ranked institutions at these locations. In sum, our results suggest that the embedding space is capable of capturing information about academic prestige, even when the representation is learned using data without explicit information on the direction of mobility (as in other formal models [184]), or prestige.

The axes can be visualized to examine the relative position of organizations along the prestige

axis, and along a geographic axis between California and Massachusetts. Prestigious universities such as Columbia, Stanford, MIT, Harvard, and Rockefeller are positioned towards the top of the axis (Fig. 6.5c). Universities at the bottom of this axis tend to be regional universities with lower national profiles (yet still ranked by Times Higher Education) and with more emphasis on teaching, such as Barry University and California State University at Long Beach. By projecting other types of organizations onto the prestige axis, SemAxis offers a new way of representing a continuous spectrum of organizational prestige for which rankings are often low-resolution, incomplete, or entirely absent, such as for regional and liberal arts universities (Fig. 6.5d), research institutes (Fig. 6.5e), and government organizations (Fig. 6.5f). Their estimated prestige is speculative, though we find that it significantly correlates with their citation impact (Fig. D.22).

We also find that the size (L2 norm) of the organization embedding vectors provides insights into the characteristics of organizations (Fig. 6.6). Up to a point (around 1,000 researchers), the size of U.S. organization's vectors tends to increase proportionally to the number of researchers (both mobile and non-mobile) with published work; these organizations are primarily teaching-focused institutions, agencies, and hospitals that either are not ranked or have a low ranking. However, at around 1,000 researchers, the size of the vector *decreases* as the number of researchers increases. These organizations are primarily research-intensive and prestigious universities with higher rank, research outputs, R&D funding, and doctoral students (Fig. D.23). A similar pattern has been observed in applications of neural embedding to natural language, in which the size of word vectors were found to represent the word's *specificity*, i.e., the word associated with the vector frequently co-appears with particular context words[623]. If the word in question is universal, appearing frequently in many different contexts, it would not have a large norm due to a lack of strong association with a particular context. Likewise, an organization with a small norm, such as Harvard, appears in many contexts alongside many different organizations in affiliation trajectories—it is well-connected. The concavity of the curve emerges in part from the relationship

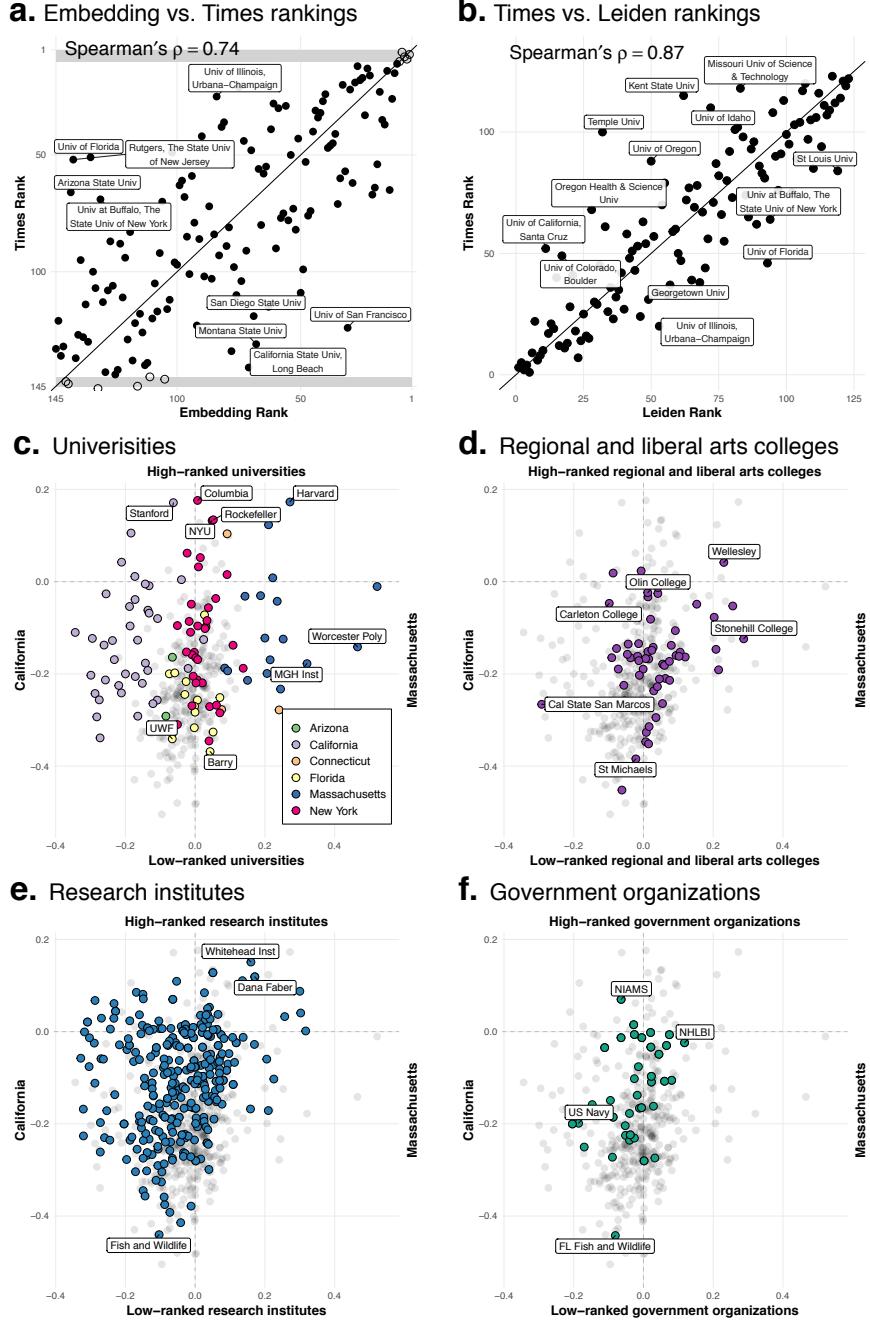


Figure 6.5: Embedding captures latent geography and prestige hierarchy. **a.** Comparison between the ranking of organizations in the Times ranking and the embedding ranking derived using SemAxis. Un-filled points are those top and bottom five universities used to span the axis. Even when considering only a total of ten organization vectors, the estimate of the Spearman's rank correlation between the embedding and Times ranking is $\rho = 0.73$ ($n = 145$, $p < 0.0001$), which increases when more top-and-bottom ranked universities are included (Fig. D.18). **b.** The Times ranking is correlated with Leiden Ranking of U.S. universities with Spearman's $\rho = 0.87$ and $p < 0.0001$. **c-f.** Illustration of SemAxis projection along two axes; the *latent geographic axis*, from California to Massachusetts (left to right) and the *prestige axis*. Shown for U.S. Universities (**c**), Regional and liberal arts colleges (**d**), Research institutes (**e**), and Government organizations (**f**). Full organization names are listed in Table D.1.

between the size of the vector and the expected connectedness of the organization, given its size ($R^2 = 0.17$). Large, prestigious, and well-funded research universities such as Princeton and Harvard have smaller vector norms because they appear in many different contexts compared to more teaching-focused organizations such as NY Medical College, and the University of Michigan at Flint. Some universities, such as the University of Alaska at Fairbanks, have considerably small vectors, which may be a result of their remote locations and unique circumstances.

We report that this curve is almost universal across many countries. For instance, China’s curve closely mirrors that of the United States (Fig. 6.6b). Smaller but scientifically advanced countries such as Australia and other populous countries such as Brazil also exhibit curves similar to the United States (Fig. 6.6b, inset). Other nations exhibit different curves which lack the portions with decreasing norm, probably indicating the lack of internationally-prestigious institutions. Similar patterns can be found across many of the 30 countries with the most total researchers (Fig. D.24; see Appendix D for more discussion).

6.7 Conclusion

Neural embedding approaches offer a novel, data-driven solution for efficiently learning an effective and robust representation of locations based on trajectory data, encoding the complex and multi-faceted nature of mobility. We demonstrated that a functional distance derived from the embedding can be used with the gravity model of mobility to better predict real-world mobility than does geographic distance. Embedding distance outperformed geographic distance across distinct and disparate domains, including U.S. flight itineraries, Korean hotel accommodation reservations, and global scientific mobility. We discover that this performance results from neural embeddings implicitly learning gravity law relationships between locations, making them particularly well suited to representing mobility. Focusing on scientific mobility, we find that the embedding successfully encodes many aspects of scientific mobility into a single representation, including global and re-

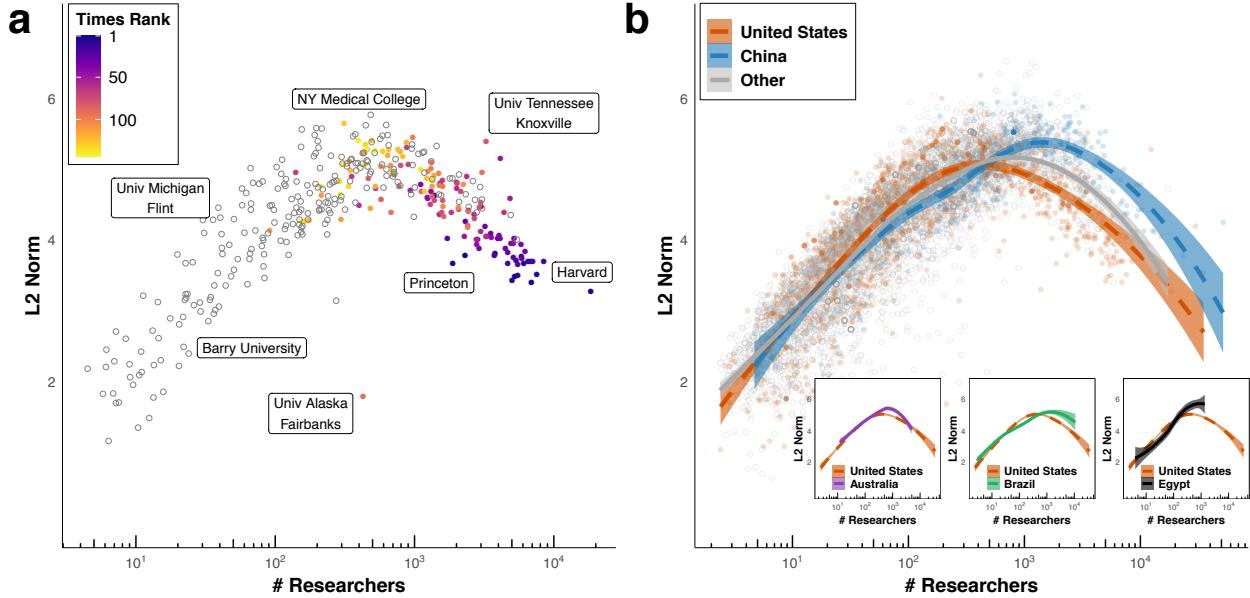


Figure 6.6: Size of organization embedding vectors captures prestige and size of organizations. a. Size (L2 norm) of organization embedding vectors compared to the number of researchers for U.S. universities. Color indicates the rank of the university from the Times ranking, with 1 being the highest ranked university. Uncolored points are universities not listed on the Times ranking. A concave-shape emerges, wherein larger universities tend to be more distant from the origin (large L2 norm); however, the more prestigious universities tend to have smaller L2 norms. b. We find a similar concave-curve pattern across many countries such as the United States, China, Australia, Brazil, and others (inset, and Fig. D.24). Some countries exhibit variants of this pattern, such as Egypt, which is missing the right side of the curve. The loess regression lines are shown for each selected country, and for the aggregate of remaining countries, with ribbons mapping to the 99% confidence intervals based on a normal distribution. Loess lines are also shown for organizations in Australia, Brazil, and Egypt (inset).

gional geography, shared languages, and prestige hierarchies, even without explicit information on these factors.

In revealing the correspondence between neural embeddings and the gravity model, the study of human mobility can move beyond geographic and network-based models of mobility, and instead leverage the high-order structure from individuals' mobility trajectories using these robust and efficient methods. While we focus on three domains of mobility, this approach could be applied to many different domains, such as general human migration, transit-network mobility, and more. Once learned, functional distances between locations, such as countries, cities, or organizations, can be published to support future research. Moreover, this approach can be used to learn a functional

distance even between entities for which no geographic analog exists, such as between occupational categories based on individuals' career trajectories. In addition to providing a functional distance that supports modeling and predicting mobility patterns, the structure of the neural embedding space is amenable to a range of unique applications for studying mobility. As we have shown, the embedding space allows the visualization of the complex structure of scientific mobility at high resolution across multiple scales, providing a large and detailed map of the landscape of global scientific mobility. Embedding also allows us to quantitatively explore abstract notions such as academic prestige, and can potentially be generalized to other abstract axes. Investigation of the structure of the embedding space, such as the vector norm, reveals universal patterns based on the organization's size and their vector norm that could be leveraged in future studies of mobility.

This approach, and our study, also have several limitations. First, the skip-gram *word2vec* model does not leverage directionality, meaning that embedding will be less effective at capturing mobility for which directionality is critical. Future studies may consider bi-directional embeddings, such as BERT [624], to incorporate directionality, as well as their correspondance to asymmetric mobility models, such as the radiation model [590]. Second, the neural embedding approach is most useful in cases of mobility between discrete geographic units such as between countries, cities, and businesses; this approach is less useful in the case of mobility between locations represented using geographic coordinates, such as in the modeling of animal movements. Third, neural embeddings are an inherently stochastic procedure, and so results may change across different iterations. However, in this study we observe all results to be robust to stochasticity, likely emerging from the limited "vocabulary" of scientific mobility, airports, and accommodations (several thousand) and the relatively massive datasets used to learn representations (several million trajectories). Applications of *word2vec* to problem domains where the ratio of the vocabulary to data is smaller, however, should be implemented with caution to ensure that findings are not the result of random fluctuations. Finally, the case of scientific mobility presents domain-specific limitations. Reliance

on bibliometric metadata means that we capture only long-term mobility such as migration, rather than the array of more frequent short-term mobility such as conference travel and temporary visits. The kinds of mobility we do capture—migration and co-affiliation—although conceptually different, are treated identically by our model. Also, our data might further suffer from bias based on publication rates: researchers at prestigious organizations tend to have more publications, leading to these organizations appearing more frequently in affiliation trajectories.

Mobility and migration are at the core of human nature and history, driving societal phenomena as diverse as epidemics[613, 625], and innovation [84, 294, 295, 297]. However, the paradigm of scientific migration may be changing. Traditional hubs of migration have experienced many politically-motivated policy changes that affect scientific mobility, such as travel restrictions in the U.S. and U.K. [350]. At the same time, other nations, such as China, are growing into major scientific powers and attractors of talent [626]. Unprecedented health crises such as the COVID-19 pandemic threaten to bring drastic global changes to migration by tightening borders and halting travel. By revealing the correspondence between neural embedding and the gravity model and revealing their utility and efficacy, our study opens a new avenue in the study of mobility. Mobility is at the core of many global challenges, and the insights brought forth by our tool can be put to use to inform a better understanding of human mobility, and to inform sensible, effective, sustainable, and humane policies.

Methods

U.S. flight itinerary data

We source U.S. airport itinerary data from the Origin and Destination Survey (DB1B), provided by the Bureau of Transportation Statistics at the United States Department of Transportation. DB1B is a sample of 10 percent of domestic airline tickets between 1993 and 2020, comprising 307,760,841 passenger itineraries between 828 U.S. airports. Each itinerary is associated with a trajectory of

airports including the origin, destination, and intermediary stops.

Korean accommodation reservation data

We source Korean accommodation reservation data from collaboration with Goodchoice Company LTD. The data contains customer-level reservation trajectories spanning the period of August 2018 through July 2020 and comprising 1,038 unique accommodation locations in Seoul, South Korea.

Scientific mobility data

We source co-affiliation trajectories of authors from the Web of Science database hosted by the Center for Science and Technology Studies at Leiden University. Trajectories are constructed from author affiliations listed on the byline of publications for an author. Given the limitations of author-name disambiguation, we limit our analyses to papers published after 2008, when the Web of Science began providing full names and institutional affiliations [97] that improved disambiguation (see Appendix D). This yields 33,934,672 author-affiliation combinations representing 12,963,792 authors. Each author-affiliation combination is associated with the publication year and an ID linking it to one of 8,661 disambiguated organizational affiliations (see Appendix D for more detail). Trajectories are represented as the list of author-affiliation combinations, ordered by year of publication, and randomly ordered for combinations within the same year. The most fine-grained geographic unit in this data is the organization, such as a university, research institute, business, or government agency.

Here, authors are classified as mobile when they have at least two distinct organization IDs in their trajectory, meaning that they have published using two or more distinct affiliations between 2008 and 2019. Under this definition, mobile authors constitute 3,007,192 or 23.2% of all authors and 17,700,095 author-affiliation combinations. Mobile authors were associated with 2.5 distinct organizational affiliations on average. Rates of mobility differ across countries. For example, France, Qatar, the USA, Iraq, and Luxembourg had the most mobile authors (Fig. D.2c). However, due

to their size, the USA, accounted for nearly 40 % of all mobile authors worldwide (Fig. D.2a), with 10 countries accounting for 80 % of all mobility (Fig. D.2b). The countries with the highest proportion of mobile scientists are France, Qatar, the United States, and Iraq, whereas those with the lowest are Jamaica, Serbia, Bosnia & Herzegovina, and North Macedonia (Fig. D.2c). In most cases, countries with a high degree of inter-organization mobility also have a high degree of international mobility, indicating that a high proportion of their total mobility is international (Fig. D.2d); However, some countries such as France and the United States seem to have more domestic mobility than international mobility. While the number of publications has increased year-to-year, the mobility and disciplinary makeup of the dataset has not notably changed across the period of study (Fig. D.1).

Embedding

We embed trajectories by treating them analogously to sentences and locations analogously to words. For U.S. airport itinerary, trajectories are formed from the flight itineraries of individual passenger, in which airports correspond to unique identifiers. In the case of Korean accommodation reservations, trajectories comprise a sequence of accommodations reserved over a customer’s history. For scientific mobility, an “affiliation trajectories” is constructed for each mobile author, which is built by concatenating together their ordered list of unique organization identifiers, as demonstrated in Fig. 6.1a. In more complex cases, such as listing multiple affiliations on the same paper or publishing with different affiliations on multiple publications in the same year, the order is randomized within that year, as shown in Fig. 6.1b.

These trajectories are used as input to the standard skip-gram negative sampling word embedding, commonly known as *word2vec* [98]. *word2vec* constructs dense and continuous vector representations of words and phrases, in which distance between words corresponds to a notion of semantic distance. By embedding trajectories, we aim to learn a dense vector for every location,

for which the distance between vectors relates to the tendency for two locations to occur in similar contexts. Suppose a trajectory, denoted by (a_1, a_2, \dots, a_T) , where a_t is the t th location in the trajectory. A location, a_t , is considered to have context locations, $a_{t-w}, \dots, a_{t-1}, a_{t+1}, \dots, a_{t+w}$, that appear in the window surrounding a_t up to a time lag of w , where w is the window size parameter truncated at $t - w \geq 0$ and $t + w \leq T$. Then, the model learns probability $p(a_{t+\tau}|a_t)$, where $-w \leq \tau \leq w$ and $\tau \neq 0$, by maximizing its log likelihood given by

$$\mathcal{J} = \frac{1}{T} \sum_{t=1}^T \sum_{-w \leq \tau \leq w, \tau \neq 0} \log p(a_{t+\tau}|a_t), \quad (6.7)$$

where,

$$p(j | i) = \frac{\exp(\mathbf{u}_j \cdot \mathbf{v}_i)}{Z_i}, \quad (6.8)$$

where \mathbf{v} and \mathbf{u} are the “in-vector” and “out-vector”, respectively, $Z_i = \sum_{j' \in \mathcal{A}} \exp(\mathbf{u}_{j'} \cdot \mathbf{v}_i)$ is a normalization constant, and \mathcal{A} is the set of all locations. We follow the standard practice and only use the in-vector, \mathbf{v} , which is known to be superior to the out-vector in link prediction benchmarks [594–599, 601].

We used the *word2vec* implementation in the python package `gensim`. The skip-gram negative sampling *word2vec* model has several tunable hyper-parameters, including the embedding dimension d , the size of the context window w , the minimum frequency threshold f_{\min} , initial learning rate α , shape of negative sampling distribution γ , the number of the negative samples should be drawn k , and the number of iterations. For main results regarding scientific mobility, we used $d = 300$ and $w = 1$, which were the parameters that best explained the flux between locations, though results were robust across different settings (Fig. D.4). Although the original *word2vec* paper uses $\gamma = 0.75$ [98], here we set $\gamma = 1.0$, though results are only trivially different at different values of γ (Fig. D.5). We used $k = 5$, which is suggested default of *word2vec*. We also use same setting for

U.S. airport itinerary and Korean accommodation reservation data.

To mitigate the effect of less common locations, we set $f_{\min} = 50$, limiting to locations appearing at least 50 times across the training trajectories; 744 unique airport for U.S. airport itinerary, 1004 unique accommodations for Korean accommodation reservation data, and 6,580 unique organizations for scientific mobility appear in the resulting embedding. We set α to its default value of 0.025 and iterate five times over all training trajectories. For scientific mobility, across each training iteration, the order of organizations within a single year is randomized to remove unclear sequential order.

An implicit bias in the negative sampling

Negative sampling trains *word2vec* using a binary classification task as follows. For each target word i , we sample a context word j from the given data and label it as positive, denoted by $Y_j = 1$. Then, we sample k words ℓ from a noise distribution $p_0(\ell)$ and label them as negative, denoted by $Y_\ell = 0$. In *word2vec*, the noise distribution is given by $p_0(\ell) \propto P^\gamma(\ell)$, where $P(j)$ is the fraction of j in the data, and γ is a hyper-parameter. Then, for the sampled words, we fit a logistic regression model

$$P^{\text{NS}}(Y_j = 1; \mathbf{v}_i, \mathbf{u}_j) = \frac{1}{1 + \exp(-\mathbf{u}_j \cdot \mathbf{v}_i)}, \quad (6.9)$$

by maximizing the log-likelihood:

$$\mathcal{J}^{\text{NS}} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{D}} [Y_j \log P^{\text{NS}}(Y_j = 1; \mathbf{v}_i, \mathbf{u}_j) + (1 - Y_j) \log P^{\text{NS}}(Y_j = 0; \mathbf{v}_i, \mathbf{u}_j)], \quad (6.10)$$

where \mathcal{D} is the set of all sampled context words.

This procedure does not guarantee that the embedding optimally converges, even when increasing the training samples and iterations [618, 619]. To make this bias explicit, let us consider the

unbiased variant of negative sampling, i.e., the noise contrastive estimation (NCE). NCE is an unbiased estimator for a probability model P_m of the form:

$$P_m(x) = \frac{f(x)}{\sum_{x' \in \mathcal{X}} f(x')}, \quad (6.11)$$

where f is a non-negative likelihood function of data x , and \mathcal{X} is the set of all data. NCE fits a logistic regression model:

$$P^{\text{NCE}}(Y_j = 1|j) = \frac{1}{1 + \exp[-\ln f(\mathbf{u}_j \cdot \mathbf{v}_i) + \ln p_0(j) + c]}, \quad (6.12)$$

where $c = \ln k + \ln \sum_{x' \in \mathcal{X}} f(x')$ is a constant and maximizes the log-likelihood

$$\mathcal{J}^{\text{NCE}} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{D}} [Y_j \log P^{\text{NCE}}(Y_j = 1|j) + (1 - Y_j) \log P^{\text{NCE}}(Y_j = 0|j)]. \quad (6.13)$$

by calculating the gradients for embedding vectors \mathbf{u}_j , \mathbf{v}_i and iteratively updating them (see Supporting Information for full derivation). Note that NCE is an unbiased estimator that has convergence to the optimal embedding in terms of the original word2vec's objective function, \mathcal{J} if we increase the number of words to sample and the training iterations.

Let us revisit negative sampling from the perspective of NCE. We rewrite the logistic regression model in negative sampling (Eq. 6.9) in form of the posterior probability:

$$P^{NS}(Y_j = 1|j) = \frac{1}{1 + \exp[-(\mathbf{u}_j \cdot \mathbf{v}_i + \ln p_0(j) + c) + \ln p_0(j) + c]} \quad (6.14)$$

$$= \frac{1}{1 + \exp[-\ln f(\mathbf{u}_j \cdot \mathbf{v}_i) + \ln p_0(j) + c]}, \quad (6.15)$$

where we define the likelihood function f by

$$f(\mathbf{u}_j \cdot \mathbf{v}_i) = \exp(\mathbf{u}_j \cdot \mathbf{v}_i + \ln p_0(j) + c), \quad (6.16)$$

which is the unbiased estimator for the probability model

$$P_m^{NS}(\mathbf{u}_j \cdot \mathbf{v}_i) = \frac{f(\mathbf{u}_j \cdot \mathbf{v}_i)}{\sum_{j' \in \mathcal{A}} f(\mathbf{u}_{j'} \cdot \mathbf{v}_i)}, \quad (6.17)$$

$$= \frac{p_0(j) \exp(\mathbf{u}_j \cdot \mathbf{v}_i)}{\sum_{j' \in \mathcal{A}} p_0(j') \exp(\mathbf{u}_{j'} \cdot \mathbf{v}_i)}, \quad (6.18)$$

$$= \frac{P^\gamma(j) \exp(\mathbf{u}_j \cdot \mathbf{v}_i)}{\sum_{j' \in \mathcal{A}} P^\gamma(j') \exp(\mathbf{u}_{j'} \cdot \mathbf{v}_i)} \quad (\because p_0(\ell) \propto P^\gamma(\ell)). \quad (6.19)$$

Taken together, the conditional probability that SGNS *word2vec* actually optimizes is

$$P^{NS}(j \mid i) = P_m^{NS}(\mathbf{u}_j \cdot \mathbf{v}_i) = \frac{P^\gamma(j) \exp(\mathbf{u}_j \cdot \mathbf{v}_i)}{Z'_i}, \quad (6.20)$$

where $Z'_i = \sum_{j' \in \mathcal{A}} P^\gamma(j') \exp(\mathbf{u}_{j'} \cdot \mathbf{v}_i)$.

Distance

We calculate T_{ij} as the total number of co-occurrence between two locations i and j across the data-set. In scientific mobility, $T_{ij} = 10$ indicates that the number of co-occurrence between both organization i and j between 2008 and 2019 is 10, as evidenced from their publications. Here, we treat $T_{ij} = T_{ji}$ for the sake of simplicity and, in the case of scientific mobility, because directionality cannot easily be derived from bibliometric records, or may not be particularly informative (see Appendix D).

We calculate two main forms of distance between locations. The geographic distance, g_{ij} , is the pairwise geographic distance between locations. Geographic distance is calculated as the great circle distance, in kilometers, between pairs of locations. In the case of U.S. flight itinerary and

scientific mobility, we impute distance to 1 km when their distance is less than one kilometer. In the case of Korean accommodation reservation data, because this data is intra-city mobility trajectory at a much smaller scale, we impute distance to 0.01 km when their distance is less than 0.01 km. The embedding distance with the cosine distance, d_{ij} , is calculated as $d_{ij} = 1 - \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|}$, where \mathbf{v}_i and \mathbf{v}_j are the embedding vectors for locations i and j , respectively. Note that d_{ij} is not a formal metric because it does not satisfy the triangle inequality. Nevertheless, cosine distance is often shown to be useful in practice [591, 592, 627]. We compare the performance of this cosine-based embedding distance against those derived using dot product similarity and euclidean distance.

We compare the performance of the embedding distance to many baselines. These include distances derived from simpler embedding approaches, such as Singular Value Decomposition (SVD) and a Laplacian Eigenmap embedding performed on the underlying location co-occurrence matrix. We also use network-based distances, calculating vectors using a Personalized Page Rank approach and measuring the distance between them using cosine distance and Jensen-Shannon Divergence (see Appendix D). Finally, we compare the embedding distance against embeddings calculated through direct matrix factorization, following the approach that *word2vec* implicitly approximates [628].

Gravity Law

We model co-occurrences T_{ij} for locations i and j (referred to as flux), using the gravity law of mobility [589]. The gravity law of mobility, which was inspired by Newton's law of gravity, postulates that attraction between two locations is a function of their population and the distance between them. This formulation and variants have proven useful for modeling and predicting many kinds of mobility [610–612]. In the gravity law of mobility, the *expected flux*, \hat{T}_{ij} between two locations i and j is defined as,

$$\hat{T}_{ij} = C m_i m_j f(r_{ij}), \quad (6.21)$$

where m_i and m_j are the population of locations, defined as the total number of passenger who passed through each airport for U.S. airport itineraries, the total number of customer who booked with each accommodation for Korean accommodation reservations, and the yearly-average count of unique authors, both mobile and non-mobile, affiliated with each organization for scientific mobility. $f(r_{ij})$ is a decay function of distance r_{ij} between locations i and j . Here, we used the most basic gravity model which assumes symmetry of the flow $\hat{T}_{ij} = \hat{T}_{ji}$ and distance $r_{ij} = r_{ji}$, while there are four proposed variants [629]. There are two popular forms for the $f(r_{ij})$: one is a power law function in the form $f(r_{ij}) = r_{ij}^{-\alpha}$ ($\alpha > 0$), and the other is an exponential function in the form $f(r_{ij}) = e^{-\beta r_{ij}}$ ($\beta > 0$) [615]. The parameters for $f(r_{ij})$ and C are fit to given mobility data using a log-linear regression [590, 610–613].

We consider separate variants of $f(r_{ij})$ for the geographic distance, g_{ij} , and the embedding distance, d_{ij} , report the best-fit model of each distance. For the geographic distance, we use the power-law function of the gravity law, $f(g_{ij}) = g_{ij}^{-\alpha}$ (Eq. 6.22). For the embedding distance, we use the exponential function, with $f(d_{ij}) = e^{-\beta d_{ij}}$ (Eq. 6.23).

$$\ln \frac{T_{ij}}{m_i m_j} = \ln C - \alpha \ln g_{ij}, \quad (6.22)$$

$$\ln \frac{T_{ij}}{m_i m_j} = \ln C - \beta d_{ij}, \quad (6.23)$$

where T_{ij} is the actual flow from the data. The gravity law of mobility is sensitive to $T_{ij} = 0$, or zero movement between locations. In our dataset, non-zero flows account for only 4.2 % of all possible pairs of the 6,580 organizations for scientific mobility, while 76.4% of all possible pairs of the 744 airports for U.S. airport Itinerary and 62.5 % of all possible pairs of the 1,004 accommodations for

Korean accommodation reservation data. This value is comparable to other common applications of the gravity law, such as phone calls, commuting, and migration [590]. We follow standard practice and exclude zero flows from our analysis.

SemAxis

SemAxis and similar studies [596, 597, 609] demonstrated that “semantic axes” can be found from an embedding space by defining the “poles” and the latent semantic relationship along the semantic axis can be extracted with simple arithmetic. In the case of natural language, the poles of the axis could be “good” and “bad”, “surprising” and “unsurprising”, or “masculine” and “feminine”. We can use SemAxis to leverage the semantic properties of the embedding vectors to operationalize abstract relationships between organizations.

Let $S^+ = \{\mathbf{v}_1^+, \mathbf{v}_2^+ \dots \mathbf{v}_n^+\}$ and $S^- = \{\mathbf{v}_1^-, \mathbf{v}_2^- \dots \mathbf{v}_n^-\}$ be the set of positive and negative pole organization vectors respectively. Then, the average vectors of each set can be calculated as $\mathbf{V}^+ = \frac{1}{n} \sum_{i=1}^n \mathbf{v}_i^+$ and $\mathbf{V}^- = \frac{1}{n} \sum_{i=1}^n \mathbf{v}_i^-$. From these average vectors of each set of poles, the semantic axis is defined as $\mathbf{V}_{\text{axis}} = \mathbf{V}^+ - \mathbf{V}^-$. Then, a score of organization a is calculated as the cosine similarity of the organization’s vector with the axis,

$$\frac{\mathbf{v}_a \cdot \mathbf{V}_{\text{axis}}}{\|\mathbf{v}_a\| \|\mathbf{V}_{\text{axis}}\|}, \quad (6.24)$$

where a higher score for organization a indicates that a is more closely aligned to V^+ than V^- .

We define two axes to capture geography and academic prestige, respectively. The poles of the geographic axis are defined as the mean vector of all vectors corresponding to organizations in California, and then the mean of all vectors of organizations in Massachusetts. For the prestige axis, we define a subset of top-ranked universities according to either the Times World University Ranking or based on the mean normalized research impact sourced from the Leiden Ranking. The other end of the prestige axis is the geographically-matched (according to census region) set of

universities ranked at the bottom of these rankings. For example, if 20 top-ranked universities are selected and six of them are in the Northeastern U.S., then the bottom twenty will be chosen to also include six from the Northeastern U.S. From the prestige axis, we derive a ranking of universities that we then compare to other formal university rankings using Spearman rank correlation.

Chapter 7

Discussion

7.1 Applying the complexity perspective

Individually, each of the studies presented in this dissertation offer a significant contribution to the Science of Science. Yet when viewed through the complexity perspective, these findings also offer new insights into the inner workings of science. In particular, the complexity perspective seeks explanations for the structure and behavior of science not based on top-down rules or a coherent scientific method, but rather as emergent from local, bottom-up forces. In this chapter, I explore how the findings of this dissertation can be understood through the lens of complexity science. Specifically, I disentangle the various local and individual-level forces that contribute to my observations in peer review, student-teacher ratings, disagreement, and scientific mobility. I then push this framework further, theorizing how the structures that emerge out of foundational *feedback* processes have the effect of existing social structures, while also making possible potential for sudden and radical change. This simple framework that I develop proves a useful and effective lens for making sense of the organization of science and how it changes.

The complexity perspective also provides important context to one of the most fundamental challenges in the Science of Science: measurement. The size and heterogeneity of complex systems complicates any attempt to measure them, and undermines the conceptual and technical foundations of everything from peer review to citation metrics. Recognizing the issue of measurements is not just vital for the Science of Science, but also the use of measures already in use for evaluating scientists. The challenges facing each metric are not isolated, idiosyncratic, or fixable on an *ad hoc* basis, but rather fundamental to any complex system; by appreciating their universality, the complexity perspective offers insights into how these issues should be approached and recognized

in the study of science and the evaluation of scientists.

Viewing science as a complex system and drawing on the framework of complexity science has the potential to offer even more insights into its structure and behavior. Concepts like scaling effects, resilience, and tipping points are valuable for explaining complex behaviors and structures in terms of individual-level forces and simple mechanisms. However, developing the complexity perspective further will be difficult, requiring extending the quantity, quality, and diversity of data about scholarly activity.

Science is massive, massively complex, and growing more-so all the time. As I illustrate in this chapter, a simple change in perspective offers hope for making sense of this size and complexity. Specifically, the Science of Science would benefit by incorporating a *complexity perspective* to inform how scientists approach and understand science. Doing so would not only provide new insights into the structure and dynamics of science, but might also point the way for policies and practices to make science more equitable and effective.

7.2 The self-organization of scientists

Adopting the complexity perspective shifts the attention of Science of Science towards those bottom-up sources of self-organization in science, namely the forces that determine how scientists interact with each other and with their environment. The findings of this dissertation point to the role of several of these forces in affecting the behaviors of scientists, and thus the structure of global scientific system.

Reputation and prestige underlies the behavior of scientists and the structure of science as a whole [40, 78]. Their impact were most visible in my analysis of the landscape of global scientific mobility, in which the centrality of a university in the embedding space revealed a latent hierarchy in scientists' mobility decisions. A university's prestige is characterized, in part, by its ability to attract and employ eminent scholars from other elite schools [170]. Scientists, usually aiming to

maximize their own reputation [40], typically work for the most elite university that they can, their choices constrained by the prestige of their *alma mater* [184]. My analysis of peer review at *eLife* illustrates how prestige can also influence other areas of science, with authors affiliated with high-ranking institutions more likely to have their papers accepted. Past studies have also shown how an author’s institutional prestige and personal reputation can benefit them in peer review decisions [246, 247]. Although not directly observed, prestige may also influence that academic debates quantified in my study of disagreement in science. Specifically, eminent scholars have disproportionate influence over the course of debates in their field [630], and can draw on their reputation as a resource in ongoing controversies [579]. Young scholars hoping to establish their reputation may also benefit from criticizing the elite of their field [347]. Science is driven by the actions of individual scientists; these studies demonstrate how reputation and prestige act as fundamental forces for science, motivating scientists’ behavior while also constraining their options and the course of their careers.

The demographic biases and preferences held by scientists also shapes the behaviors of scientists, disadvantaging marginalized groups and stymieing their careers. I show how peer review at *eLife* may be biased against women lead authors and authors outside of the core set of countries, rejecting their papers at higher rates than their counterparts. Similarly, my investigation of student-teacher evaluations on *RateMyProfessors.com* contributes to the growing consensus [466–469] of the prejudices inherent to such measures, deflating the scores of women and non-White faculty. These findings speak to a broader trend in science whereby evaluative tools—essential to the allocation of resources in science—are biased against certain populations. For example, men are typically favored over women when contending for lab positions [34], women fair worse on the faculty job market than men [184], and funding institutions tend to rate the careers of men more highly, even given equivalent research performance [35]. The consequence of these biases is that they constrain the behavior of the disadvantaged scientists, diminishing their opportunities for employment, funding,

publication, and recognition, and limiting the potential of their careers.

Just as biases can curb a scientists career, homophilous preferences can shape how scientists behave and interact. For example, in the case of *eLife*, I found that papers were more likely to be accepted when there was a high degree of gender or national homophily between the authors and reviewers. Moreover, reviewing editors were likely to recruit reviewers of their same gender or nationality, likely as a result of the homophily of their professional networks [111]. While not directly observed, similar affects might underlie other findings from this dissertation. For instance, in teaching evaluations, students might give higher ratings to faculty with whom they share a demographic identity [631, 632]. As in social networks more generally [633–636], the incidence and extent of disagreements in a field might also relate to the degree of homophily among its scientific community. As with bias, homophilous preferences influence scientists’ careers, in some cases constraining their potential, while in others supporting their success. However, this narrative of homophily in science is complicated by its heterogeneity of its effects. For example, at *eLife*, women authors were less likely to benefit form homophily, due to the underrepresentation of all-woman reviewing teams. Similar heterogeneity can be found in other areas in science. In the case of citations, for instance, men are more likely to cite the work of men than are women to cite other women [266], whereas in the case of collaboration, all-men collaborations are disproportionately common as compared to all-women ones [280]. Together, these findings illustrate the important, though complicated role of homophily in science.

The proximity, or distance, between scientists also dictates how they interact with each other. This is most visible in my analysis of mobility, in which *geographic proximity* was strongest factor in determining where scientists moved in their careers, and proximity also plays a role in sparking collaboration [108], innovation [289], and the diffusion of ideas [637]. However, geographic proximity is not always an appropriate representation of the distance between scientists. For example, cheaper flights increase the degree of collaboration between two cities, all without changing the actual

physical distance between them [286]. By learning an *effective distance* of mobility, I was able to capture these more complicated factors of distance between scientific organizations, revealing also the importance of linguistic, cultural, and political distance between countries. For example, linguistic proximity encourages mobility between Quebec and France, whereas political barriers hinder movement between the United States and Iran. This more complicated notion of distance can also apply to measuring disagreement, in which the linear distance implied by the hierarchy of sciences [392, 549, 561, 580] breaks down when a finer classification of fields is used, their greater heterogeneity not accounted for by a simple metric of "hard" vs "soft" science. Once the notion of distance is broadened beyond geography, it becomes a useful concept for understand the interactions and behaviors of scientists.

Its not just internal social factors that affects scientists' behavior, but also the physical realities of the world that dictate how evidence is gathered, analyzed, and interpreted. Controlled experiments, for example, have historically been the gold standard of empirical evidence [7], their inherent replicability and agreed-upon standards allowing for broad and lasting consensus in fields where experiments are possible. This contributes to the low levels of disagreement observed in physics and other experimental natural sciences observed in my analysis of disagreement. However, many fields cannot conduct controlled experiments, their evidence and analyses laden with assumptions, requiring subjective interpretation that gives rise to disagreements over non-standard and incomplete historical records [638], incommensurate measurements made across locations and with different tools [174], or justifiable differences in analytical approach [639]. These reasons contribute to the high degree of disagreement observed in many of the non-experimental natural sciences, biomedical sciences, and social sciences. The heterogeneity of evidence across science may also confound findings of bias observed at *eLife*, as papers reliant on controversial data or less well-accepted theoretical frameworks may also spark disagreements from reviewers, lowering their chances of success [326, 327]. Just as physical constraints affect how evidence can be interpreted, so too does it

d dictate *where* evidence can come from. Many kinds of science can only be done in certain locations, such as at archaeological sites, within a particular biome, or in an exclusive laboratory [17]. The geographic demands of certain fields contributes to the mobility patterns in this dissertation, such as the centrality of organizations like CERN in Switzerland, which hosts the Large Hadron Collider and other large, expensive, and sophisticated equipment for experimental physics; scientists from fields with fewer geographic constraints, such as molecular biology or the social sciences [640, 641], will be less mobile, limiting their impact on the global landscape of mobility. The physical world sets limits on what scientists can do, and where, giving way to the structures of science observed in this dissertation.

When viewed with the complexity perspective, the studies in this dissertation contribute to disentangling and illustrating the impact of a small number of these forces. Of course, many other factors contribute to scientists' behavior, not the least of which is their own agency and career goals. Yet these few forces already demonstrate how bottom-up processes, affecting how individual scientists interact with one another and with their environments, can give rise to the complex structure of science as a whole. Future research should continue identifying and studying these forces, noting not only the extent of their effect but also how their effects change at different places and times, and how they interact with one another. By understanding the multiplicity of local factors that contribute to scientists' behavior, it may be possible to better grasp the fundamentals of scientific knowledge production.

7.3 Stability, change, and feedback in science

Bottom-up forces shed light on how structures form in science, but they alone do little to explain its dynamics, namely how these structures remain stable in the chaotic system of science, and how despite this stability, revolutionary change still happens. Drawing on the findings from this dissertation and the complexity perspective of science, I argue that *feedback loops*, an essential driver

of the evolution of complex systems, explains both stability and change in science. Specifically, certain social mechanisms in science generate feedback processes that contribute to *cumulative advantages* that entrench structures and hierarchies; at the same time, feedback loops have the potential to *amplify* certain events, under the right conditions, affecting quick and rapid change throughout the system. Of the many mechanisms that give rise to feedback loops, perhaps the most prominent in science are mechanisms for evaluation and selection that are used to allocate resources to scientists. Here, I outline the process by which feedback loops occur by focusing on one illustrative selection mechanism: peer review at the journal *eLife*, which through feedback, has the effect of perpetuating demographic bias and under-representation of marginalized groups in science. Then, I demonstrate how feedback loops can be generalized to other areas of scientific activity, showcasing the utility of this concept as a vital tool for the complexity perspective of science.

In the Science of Science, feedback loops are most commonly recognized under the concept of the "Matthew Effect" [79] or *cumulative advantage* [201, 269]. The core idea is that those who achieve success early gain access to new capital, such as greater reputation or material resources, that they can then leverage into supporting greater productivity and impact, and then into further success. In contrast, those without this early success will find it difficult to access these resources, and struggle to launch their careers. In science, any mechanism that grants access to capital has the potential to spur a feedback loop, entrenching inequalities.

Peer review is one of the most consequential feedback mechanisms in science. It is regarded as the "gold standard" of formal evaluation in science [326], and is ideally concerned only with merit, rather than with any particularistic criteria such as the gender or nationality of an author or applicant [264]. However, in this dissertation I observe how demographic biases can manifest in journal peer review. At first glance, these biases can appear small, accounting for only a few percentage points difference in acceptance rates at *eLife*. However, due to cumulative advantage,

even these small biases can compound into long-lasting career consequences [270]. When an author places their work in a prestigious journal like *eLife*, they establish reputation that improves their chances of winning future funding [259] that supports further high-impact research [642], which again feeds back into more success and continues the loop. However, marginalized groups who find themselves disadvantaged during peer review, such as women, ethnic and racial minorities, and those from countries with lower scientific capacity, will find it difficult to kick-start these feedback loops, a barrier that prevents them from publishing at intensities equivalent to their advantaged peers or pushes them out of science entirely.

By making biased evaluations, peer review dictates which demographics become most successful in science, and later become peer reviewers themselves, reinforcing the same opinions, expectations, and prejudices that aided their own success. At *eLife*, I observed how diverse teams of reviewers, in terms of gender and nationality, produced more equitable review outcomes, demonstrating how the demographics of reviewers might impact their decisions. However, diverse teams accounted for only about half of teams at *eLife*. Reviewers are not drawn from the population of all scientists at random, and are instead often drawn from the personal networks of editors, which tend be homogeneous [111]. Because reviewers are more often men [413, 415, 419, 420], or from a small set of nations [30, 643, 644], the invited peer reviewers will tend to be men, and from these same nations. Indeed, I observe these same homophilous pattern between editor and reviewer demographics at *eLife*. *Who* succeeds in peer review is an input in determining who later gets to become a reviewer; because reviewer demographics relates to outcomes, then *who* becomes a peer reviewer will also contribute to who later becomes successful. This self-reinforcing dynamic—in which reviewer demographics shape who is successful in science, and thus who becomes a peer reviewer or editor in the future—acts as a feedback mechanism that perpetuates demographic biases in science and the underrepresentation of marginalized groups.

Peer review is not the only feedback loop in science. Any mechanism that distributes capital

to scientists based on some selection criteria has the potential to become a feedback loop that perpetuates existing social structures and behaviors.

Performance metrics can perpetuate certain behaviors in science, not just as a natural consequence of cumulative advantage, but also through intentional actions to optimize or "game" the evaluation at the expense of actual performance, a process known as Goodhart's Law [336]. For example, in response to the biases in student-teacher evaluations that I demonstrated using *RateMyProfessors.com*, faculty might design easier courses to elicit higher ratings from their students [645]. This same sort of behavior can be observed in other areas of evaluation in science as well. A journal with a high impact factor can attract more submissions, allowing it to publish the most high-quality work and continue building its success [185, 258]; they may even coerce authors of submitted papers to cite other works in the venue, aiming to inflate their citations [337]. Individual scientists attempting to optimize their H-index may adopt a strategy of quantity over quality, producing many mediocre papers rather than investing time on higher quality works [646]. Authors might also resort to excessive self-citation to inflate their overall citation counts [267, 647] or even form collaborative "citation cartels" to cite each other's work [338, 648]. Individuals who adopt these strategies will tend to achieve further success, and occupy positions allowing them to teach these same practices to their students, who in turn use them to become successful, and perpetuate the cycle. Selection mechanisms, like performance metrics, serve to perpetuate work strategies in science, and given time, responses to selection criteria may become enmeshed in the behaviors of generations of scientists.

The nature and consequence of feedback loops is not uniform across all of science; rather, the social organization of scientific fields can dictate the effects of feedback mechanisms, perpetuating different kinds of structures depending on the field's circumstances. For example, the field of high-energy physics relies on tremendously-expensive equipment and labor demands [19]. As a result, the field is highly-collaborative, consisting of specialized researchers pooling their expertise

and resources in order to construct the necessary experimental apparatuses such as the Large-Hadron Collider [70]. A consequence of this coordination is that the field also becomes heavily centralized, vesting enormous power in a small number of scientists who act as gatekeepers to their field, enforcing a common theoretical paradigm by managing access to funding, publication venues, equipment, and employment [40]. Through this top-down control, the most elite scientists in the field can suppress disagreements and maintain their favored theoretical paradigm; to become successful, scientists must access the resources managed by the scientific elite in their field, as such must submit to their expectations. These successful scientists will have students of their own, who they train under the same cultural tradition which they will eventually spread as they take jobs in other universities [116, 302]. This kind of top-down control in expensive experimental fields likely contributes to their low degree of disagreement observed in this dissertation. In contrast, the higher levels of disagreement in the social sciences and humanities may stem from their relative *disorganization*. Without centralized control, these fields are more likely to split along certain theoretical fault lines, or "schools of thought" [40]. In these fields, success may come not from conformity, but from disagreement itself, whereby young scholars criticize a prominent scholar in order to establish their reputation [347]. In the same way, these successful scientists will perpetuate a culture of disagreement to their students. The central process of feedback, whereby selection criteria impact the future demographics and behaviors of scientists, can manifest differently depending on their particular circumstances, contributing to a diversity of disciplinary traditions while also entrenching each field's culture and social organization.

One of the most prominent results of feedback loops are in the creation and perpetuation of hierarchies in science. For example, I find that institutional prestige plays a vital role in the global landscape of scientific mobility, in which the characteristics of organizations structure the mobility decisions of individual scientists. Scientists benefit from mobility by accumulating additional capital, whether in the form of social connections, reputation, access to resources, or new knowl-

edge [110, 235, 236, 306, 649] that they can leverage into more and more high-impact research [294, 295, 313, 650] and additional career opportunities [110, 309]. In this way, mobility functions as another selection mechanism in science, a Matthew Effect that benefits those with the ability and resources to move between institutions. A similar process functions at the level of institutions. Elite universities are more able to place their scholars in other elite positions and attract exceptional talent [184], a process of selection that quickly establishes a rigid hierarchy [170]. These same dynamics result in national hierarchies as well. Countries that attract scientists benefit from the additional human capital, spurring scientific and technological breakthroughs and contributing economic benefits [19, 306, 651, 652], improving the scientific capacity and prestige of the home country, allowing it to attract even more foreign scientists, and further bolster its scientific capacity. With feedback loops, those individuals, institutions, and nations that succeed will find themselves entrenched at the top of a status hierarchy that propels them into sustained future success.

Other feedback loops can be found all across science. For instance, young scientists with eminent mentors can more easily place their work in elite journals, granting them better access to these venues throughout their career [80]. Employment in an elite university grants access to resources and capital that allow for higher research productivity and impact [124]. Winning a grant makes it easier for a scientist to win other grants in the future [259]. Nations able to make investments in science are likely to see economic improvements, leading to further investment and stronger scientific ecosystems [353, 354, 653–655].

All of these various feedback loops establish a kind of conservatism, or *inertia* to science, resulting in stable social structures and hierarchies. In many cases, this inertia is necessary or even beneficial, serving to maintain disciplinary boundaries, support good ideas, and empower effective scientists. Yet as feedback loops maintain social structures, so too do they perpetuate existing biases, inequalities, and disparities [67, 201, 267]. Alone, this view of feedback presents a static and nihilistic view of science: inequalities will not improve, changes cannot be made, and science

will not progress. However, this is clearly an incomplete view. Although inequities remain, science has improved in its representations and support of marginalized groups. Moreover, science has changed, often radically so, in response to advances in theory, technology, politics, and internal social processes. How do feedback loops simultaneously maintain structures, while also allowing for incremental evolution and more thorough revolution?

I argue that feedback loops are not only a mechanism for perpetuating existing social structures, but also *amplifying* certain dynamics, allowing even seemingly-inconsequential events the potential to affect change across all of science. The power of amplification is most visible in the dynamics of success in science. Success is amplified through social processes that ensure that ideas and people that succeed early will tend to gain even greater success. These dynamics have been studied in other complex social domains like business and art. For example products on the crowdfunding website *Kickstarter* that achieve initial success tend to accumulate further success, regardless of their apparent quality [656, 657]. Signals of early success encourage others to give attention to the product, contributing further to its perceived success and attracting even more attention [658]. Eventually, a person may accumulate enough success to become a superstar researcher, whereas an idea might reach a point where it causes a paradigm shift in a field. Yet in order to kick-start amplification, a *spark* of initial success is necessary ¹, though there are many potential sources of this spark.

Ideally, early success stems from the *intrinsic quality* of an idea. In science, quality might take many meanings, such as the robustness of a study, its theoretical implications, or its applicability to real-world problems each of which contribute to the attention it receives and its eventual impact on science. However, quality in science is impossible to objectively measure, even for the peer reviewers tasked with making such assessments [65, 150, 326]. For example, the well-known phenomenon of "sleeping beauties" describes how some scientific papers go ignored for years, perhaps even decades,

¹This initial spark, however, by no means have to come immediately after a product's inception or in the early stages of a person's career. Many famous individuals, including Franz Kafka or Vincent van Gogh, are famous for achieving fame only after their deaths, this "spark" occurring many years after they created their now-famous works.

before eventually being recognized as crucial [88]. The work in this dissertation regarding peer review and teaching evaluations, along with other studies in the Science of Science, demonstrate how intrinsic quality alone cannot account for why some people and ideas achieve early success while others fall flat.

Social forces also play an important role in sparking feedback loops and amplifying success. For example, access to symbolic capital, such as affiliation with an elite university, might attract attention to a person’s work regardless of its merit [260]. Access to capital is, however, moderated by social factors such as demographics and family background [75, 248]. Demographic bias might also inhibit early success for some demographic groups. Women, for example, are disadvantaged during peer review and hiring [29, 34, 35], whereas the names ethnic minorities are often excluded from news articles about their work [273]. The composition of science, where it takes place and who does it, also influences the impact of certain ideas. For example, scientists in wealthy Western countries may ignore findings that are not of immediate national interest, such as research into illnesses common to developing countries [20]. Similarly, a science dominated largely by White men in the U.S. will often undervalue topics pursued by marginalized groups [23, 24, 113]. The merit of an idea or the performance of a scientist is not an objective quality, but rather something determined by social judgements [65], and are subject to all the same biases, subjectivity, and idiosyncrasies as other domains of human perception.

The inherent chaos of science contributes another important force in determining early success: random chance. The importance of chance and luck in science is well-appreciated, often understood as “serendipity” [196, 197]—minor disagreements might lead to unexpected findings that transform a field [547, 548, 558], whereas chance collaborations following an international move may result in a unique and powerful combination of ideas [44, 587, 659]. So too can serendipity trigger early success, launching some ideas to huge and far-reaching impact even while other ideas of comparable or even greater quality are left ignored. Sheer luck may determine which scientists eventually

become superstars [660, 661], or which of a scientist’s papers becomes the most successful [120]. Science also exists within a chaotic world, and so seemingly-random political, technological, and cultural developments can shift which ideas become successful, creating “black swan” events, known for their unpredictability and sudden yet extensive impact [87]. For example, the COVID-19 pandemic saw research into mRNA vaccine technology, largely ignored for decades, suddenly come to prominence [263]. Although intrinsic quality and social forces can help describe the mechanisms of early success, the chaos and apparent randomness of science and the world it exists within ensure that luck, too, plays a vital role.

Amplification is a powerful mechanism explaining how new people and ideas can achieve massive success in science, yet alone it still does not account for how change happens in science. Namely, amplification only explains success, not how old ideas and social structures are cast aside in favor of new ones. Other mechanisms, in addition to feedback, are necessary to understand how the old gives way to the new in science. I briefly speculate on a few possible mechanisms here, and encourage further exploration of these ideas in the future. Turnover among successful people can be explained through retirement or death, in which the resources and capital accumulated in a person’s life are lost or transferred to others [81, 193]. It might also play a role in the turnover of theoretical paradigms, for which the death of the “old guard” means that they can no longer champion the old theories, allowing new ideas to flourish [187]. Yet individuals are also not static in their beliefs, and so death is not the only mechanism for change. Some new ideas may be incommensurate with older theories, such as how Einstein’s work on relativity was at odds with Newton’s classical mechanics—in these cases, theories are in direct competition and so their success is zero-sum. Eventually, scientists will accumulate enough evidence and argue enough points to decide which theory to accept, often involving casting aside of correcting the older theory [662]. However, not all theories are in direct contest with one another, and science as a whole is not necessarily zero-sum. Scientists can escape the debates of one field by exploring topics in sub-

fields [40], forging new areas of study where new ideas are necessary and where new individuals can become successful. The expansion of fields and the development of new theories are also supported by cultural norms of originality in science, often prizing new ideas at the expense of older and more established theories [663]. By supplementing the notion of feedback with these or other social and cultural mechanisms, it becomes possible to form a more complete understanding of how science changes, abandoning or modifying the older ideas and social structures and maintaining a balance between tradition and innovation.

The global scientific enterprise is not stagnant, and neither does it lack a structure. Like all complex systems, science exists at some middle point between order and disorder. Here, I demonstrated how a core mechanism of complex systems, *feedback*, can be used to understand how science entrenches its core structures and hierarchies, yet also how it allows for dynamism by amplifying the impact of new scientists and their ideas. Feedback is not the only relevant mechanism in explaining the structures of science—but it is an essential and powerful one that is able to link demographic biases in peer review and teaching metrics to the underrepresentation of women, local and idiosyncratic disagreements with the formation of disciplinary cultures, and the mobility decisions of individuals to the prestige hierarchies of universities around the world. Further application of this idea may reveal even more insights into the composition of science and how it evolves.

7.4 Taking measure of a complex system

While adopting the framework of complex systems introduces useful tools and concepts to the study of science, it also reveals inherent challenges in the Science of Science, namely the difficulty of measuring the characteristics of the system. Scientists in all disciplines have long struggled with measurement, with famous controversies surrounding the measurement of temperature [664], global weather patterns [174], gravitational waves [550], and psychological phenomenon [665]. Yet the

unique characteristics of complex systems, such as science, bring their own distinctive challenges. The sheer size of science, combined with its immense heterogeneity, makes it difficult to devise measures that accurately describe the system as a whole. Its interconnectedness also confounds the theoretical and conceptual foundations of measurements, making it difficult to know whether the measure is capturing the desired phenomenon, or some combination of many other interacting forces. The work in this dissertation illuminates several of the challenges with the measurement and study of complex systems, and draws attention to the need to recognize these difficulties in future applications of the complexity perspective.

The complexity of science complicates the creation of clear and well-defined measurements, with any measure inevitably failing to account for edge cases or systematic bias within the system. These same issues were encountered in attempting to define a measurement of *disagreement* in science, based only on a set of manually-defined cue phrases and in-text citation sentences. At first glance, the measure performed well across all of science, yet when examined in more detail, the measurement was confounded by the idiosyncrasies of certain sub-fields and contexts. In some cases these idiosyncrasies were random and trivial, and could be safely ignored, such as the author name "*Debatin*" captured in our signal phrase *debat**. However, the language of some certain sub-specialties introduced systematic biases that lead to an overestimation of disagreement. For instance, the signal phrase *conflict** captured irrelevant references to "conflict theory" or "international conflicts", whereas the phrase *controversy** retrieves sentences discussing "public controversies" as an object of study. These issues are not unique to my measure of disagreement, but also plague other measures commonly used in the Science of Science. The journal impact factor, for instance, is intended to capture the impact of a journal, yet the citations a journal receives are skewed towards only a few papers, meaning that the score is not representative of a typical paper in the journal [666, 667]. These kinds of biases might be addressed on an *ad hoc* basis, such as excluding common confounding terms in my measure of disagreement, or making adjustments to citation counts [332,

668, 669]. However, the size of science makes exhaustively fixing every such issue time consuming and perhaps even impossible. Inevitably, measurements of complex systems will fail to capture their size and heterogeneity in a single measure. Researchers in the Science of Science should thus recognize the limits inherent to measures of science, approaching them with caution and a healthy skepticism.

The multiplicity of forces that give rise to the structure and dynamics of science also challenge the conceptual foundations of measurements. This is evident in my study of global scientific mobility, in which I created a measure capturing an abstract notion of *distance* or *affinity* between organizations, learned directly from raw mobility data. What this measure reveals is that data on mobility—such as the number of mobile scientists between countries, regions, or organizations—obscures a variety of underlying factors. Geography plays an important role, but so too do historical connections, linguistic similarity, and academic prestige. Without understanding *what* exactly bibliographic or other data are really capturing, the Science of Science runs the risk of misunderstanding and misinterpreting its measures. For example, interpretations of citation-based metrics are confounded by the many reasons that an author might cite a study, ranging from endorsement to outright condemnation [150, 570, 574, 670, 671]². Connecting the conceptual and theoretical foundations of a phenomenon to the quantity actually being measured remains a tremendously difficult task in the Science of Science, and the study of complex systems more broadly.

These challenges are not only an issue for the measures I created for this dissertation, but also for measures *in the wild*, which are currently used to make real-world editorial and employment decisions. One such measure is peer review, which while intended to assess the merit of a manuscript, funding proposal, or applicant [326], is also shown in this dissertation as susceptible to demographic bias, as well as nepotism [241, 242, 672] and prestige [247]³. I also demonstrate how

²One such controversy existed over the interpretation of a measure of scientific consensus, which its authors justified by arguing that citations were implicit signals of endorsement [191]. However, this measure came under fire, accused of misunderstanding citations and producing a misleading interpretation [552, 553].

³These same biases exist outside of science, too; for example, women are judged harshly in orchestra auditions [114], an effect which vanishes once auditions are blinded, and job-applicants with names perceived as "Black" are less likely

demographic bias pervades student teacher evaluations, and is driven more by the ease of the class or halo effects [474, 510, 674] than anything related to pedagogical ability. When performance is difficult to measure, as in science, it is not performance itself but the *perception* of performance that matters [65]; In each of these real-world cases, measures that aim to assess quality instead end up reflecting a variety of social perceptions, along with the prejudices and biases that accompany them.

The complexity perspective illuminates the challenges of measuring science. Moreover, the perspective re-defines these challenges from being individual and isolated, to instead a fundamental problem in the study of complex systems. This does not imply that measurement is invalid or useless; rather, by approaching the Science of Science with caution, and by recognizing the inherent limits of measurement, it may be possible to gain a humble, yet more accurate understanding of the structure and dynamics of science.

7.5 Moving forward with complexity in the Science of Science

The complexity perspective of science offers insights beyond the studies in this dissertation, with its diverse conceptual framework that may be applied to other areas of Metascience. Here I discuss just some of these concepts which have potential for creating a deeper understanding of how science works.

A property of many complex systems are their patterns of *scaling* [675]. Scale, here, refers to the often non-linear relationships between the size of a complex system and some characteristic of it. For instance, an animal twice the size as another will have *less* than double the metabolic rate (sub-linear scaling). A city twice the size of another, however, will tend to have *more* than double the amount of innovative activity (super-linear scaling). Scaling laws, represented as power law relationships, are found in natural and social system, including the relationship between an
to get a response from employers, even given identical resumes [673].

imals' metabolic rate and size [676], and between a city's science and their degree of in and out migration [610]. Science, too, has already been found to exhibit similar scaling relationships in certain aspects. Networks of scientific collaborators are known to follow scaling laws [119, 677], and scientific knowledge production scales with a city's population [117, 678, 679]. Other areas of science might also be understood through scaling laws. Does a nation's scientific productivity scale linearity with its number of scientists or amount of funding, or are there agglomerative or diminishing returns at larger scales? Similarly, to what degree does a person's citation impact scale with their productivity? What traits of science might be invariant to size? An animal's total lifetime heartbeats, for example, is roughly invariant to their actual lifespan. In science, a similar pattern might emerge in the size of a person's professional network, which might be constrained by the cognitive limits of Dunbar's number ⁴. In the past century science has exploded in size and scope [2], with more people across nations integrating into a global scientific system [9, 15, 16, 177, 199]. By drawing on the notion of scaling laws from Complexity Science, Metascience can gain a better grasp on the role of scale in science, supporting future policies and practices that appreciates the implications of the ever-growing scientific enterprise.

The ability of scientific systems to withstand epistemic revolutions, fraud, retractions, cultural upheavals, war, and more speaks to another aspect of many complex systems: their *resilience* [680]. The concept of resilience has been used to understand how ecological networks survive, and even thrive, amid catastrophes and disruptions [681, 682], and how resilience might be maximized in technological networks [683]. The notion of "resilience" is already commonplace in science, though under a different name. For example, science is colloquially thought to be self-correcting [684], responding to errors and correcting mistakes ⁵, slowly improving the quality and quantity of evidence

⁴Dunbar's number refers to the theorized cognitive limit to the number of stable social relationships that a person can maintain, most often considered to be around 150. This limit may have an effect on the structure of research communities,

⁵In some cases it is not mistake, but outright fraud and deception that risks erroneous knowledge, such as when vested interests aim to distort scientific opinion for material gain [360]. Still, scientific communities may have self-policing mechanisms that make them resistant to fraud [685].

and converging on a theory [192]⁶. Other Metascience findings touch also touch on the notion of resilience; for example, that an eminent scientist's death leads to disruption in their local network, but innovation in the broader scientific community [81, 193], illustrates that science is resilient enough to withstand, and in some cases benefit, from the loss of one of its important members. The concept of resilience in science can be approached more explicitly, taking advantage of the methodological advancements from Complex Systems to better understand how science responds to shocks and errors. Moreover, the related concept of "*antifragility*" [687] might also prove useful for understanding science. Antifragile systems are characterized not only by their resilience, but also by their ability to actually benefit from shocks, such as when human muscles responding to the damage from an intense weightlifting session, or when they gain immunity through exposure to a pathogen. When applied to science, the concept of antifragility might provide insights into how science have withstood centuries of error and misconduct, and how anomalous findings that conflict with establish theory might spur a revolution leads to new knowledge. Applying this concept to science might also contribute to the creation of policies that *encourage* antifragility, creating a scientific system that thrives on internal debates and conflict.

Tipping points may also be a useful idea for understanding the sudden and extensive changes that happen in science. In complex systems, a tipping point is the point at which the accumulation of small changes suddenly results in a cascade of larger changes across the system [688, 689]. This idea already features in theories of science [690]. For example, Kuhn's theory of scientific revolutions expects that, once some critical proportion of scholars adopt a new paradigm, that the rest will quickly follow [187, 662]. Similar tipping points have been observed for opinion spreading in social networks, in which when a critical threshold of people hold a belief, it will rapidly propagate through the entire population [691, 692]. Success may similar exhibit a tipping point, in which a certain level of performance is necessary before feedback mechanisms to have an effect. On the other hand, a critical number of failure may also be necessary before a scientist is likely to find-

⁶The notion of self-correction in science is, however, contentious and open to debate even today [686]

success [66]. The notion of tipping points is essential for understanding the non-linear properties of complex systems, highlighting how discovery, excellence, and success in science do not necessarily rise in proportion to each dollar of financial support, or to how closely a behavior is followed, an important lesson for science policy and governance.

Beyond these, complexity science offers a wealth of other concepts and ideas that can be applied towards understanding science. The notion of cascading failures could be applied to understanding how the epistemic harms of fraud or error reverberate through a field [693, 694]. The memory, or history of science may also allow the system to recover in the face of these failures, or other critical transitions [695, 696]. Complex systems are also *nested*; just as science is nested within a wider social system, so too are disciplines, institutions, and groups within science also complex systems in their own right. Modelling these multi-layered systems could prove useful for making sense of interactions and effects at different levels of analysis.

By thinking about the structure and dynamics of science in terms of complexity, the Science of Science gains access to the extensive and diverse framework of complexity science. The kinds of concepts offered by this framework, which I have outlined in this discussion, provide a powerful lens with which to think about science, and the tools necessary for gaining deeper insights, even as science continues to grow more massive, and more massively complex.

Chapter 8

Conclusion

Science is now bigger, and more complex, than any other time in its history, and growing more so each year. For the Science of Science, challenges accompany this growth. The size, heterogeneity, and inter-connectivity of science confounds existing theoretical frameworks and methodological approaches for studying science. In this dissertation, I argue that *complexity science* holds value for the Science of Science, and can support a better understanding of the structure and dynamics of global science. Specifically, I defining a *complexity perspective* of science, and use it to interpret four studies investigating topics relevant to the science of science. Applying the perspective, I examine how many decentralized, bottom-up forces like demographic bias and homophily give rise to disparities in peer review outcomes and student-teacher evaluations, how the physical requirements of a research topic can shape a field's culture, and how different kinds of proximity shape the mobility decisions of scientists. Pushing the complexity perspective further, I examine how *feedback* organizes science, both perpetuating existing structures and hierarchies, while also maintaining the potential for revolutionary change. Moreover, the complexity perspective also contextualizes measurement issues in the Science of Science, illustrating the inherent difficulty of measuring complex systems and the need to approach their study with caution and humility. In addition to these present contributions, this dissertation also lays out a path for future applications of the complexity perspective to the Science of Science. Broader adoption of the complexity perspective could bring more benefits to the Science of Science. However, adopting this framework also comes with further epistemic and practical implications for the Science of Science, which should be further explored by future scholars.

The epistemic implications of the complexity perspective regard the limits it places on the Science of Science to be able to understand and predict the structure and behavior of science.

The size of science makes collecting and analyzing data on the whole system, rather than smaller components of it, difficult [697]. Its heterogeneity complicates the generalization of findings across subsets of the system and the application of observations of the aggregate to individuals [698]. Science's interconnectedness gives rise to nonlinear effects that confound traditional analytical techniques and constrains models and predictions [89, 687]. Moreover, the events that most impact science—the breakthroughs, revolutionary discoveries, and skyrocketing success—are inherently unpredictable [87]. These challenges do not doom the Science of Science, but instead should be a cause for caution and humility when designing, executing, and interpreting research. They also do not invalidate predictions; rather, much like how weather forecasts incorporate known uncertainty [699], the Science of Science should aim to infer under what conditions, and to what extent, predictions about science can be made and its mechanisms understood [700, 701]. By appreciating the complexity of science and the limits this presents, a stronger Science of Science research program can be established that can produce a more well-rounded and complete understanding of science, uncertainty and all.

The practical implication of the complexity perspective is that it complicates the goal of the Science of Science to use that the findings from the Science of Science can be implemented into effective policies and practices. As I have argued, knowledge about science is deeply uncertain, and predictions about its future inherently fragile. Allocating resources to scientists on the basis of their measured performance is inherently a kind of prediction about who can best apply them. However, these kinds of policies often overlook the contributions of scientists whose value is not immediately apparent¹, and create perverse, unintended incentives that undermine effectiveness [702], and which depend on poorly-defined or inappropriate metrics [150]. Attempts to govern science always have the potential for producing more harm than good. Yet effective governance remains necessary to improve the equity, integrity, openness, and impact of global science. Exist-

¹See, for example, Katalin Karikó, whose work developing mRNA vaccine technology was largely ignored for decades. Other examples include sleeping beauties [88], papers that only become highly cited some time after their publication

ing literature in Management Sciences have explored how complexity can be managed in business and government [703–705]; in adopting the complexity perspective, the Science of Science should examine how these same principles might be applied towards governing and improving the global scientific enterprise.

Assessing and managing the epistemic and practical limits of the Science of Science, as well further developing the complexity perspective of science, will require more and better quality data. Although bibliographic databases continue to grow, their fragmented geographic and disciplinary coverage [125, 126, 706] constraints the generalizability of analysis. Moreover, the complexity perspective encourages analysis of *individual-level* behavior, which remains difficult in spite of advances in author name-disambiguation [97] and adoption of unique author identifiers [228]. Further developing the complexity perspective of science will require greater effort in improving the coverage and quality of these bibliographic databases. Yet the perspective will also benefit from moving beyond bibliometrics, which have been the dominant tool in the Science of science for so long. Fortunately, a wealth of new data has emerged that has can greatly benefit the complexity perspective. Now, the content of scientific publications is growing increasingly available, allowing for new kinds of research into how scientists write and interact. The Open Science movement [707] has also brought transparency to processes like peer review [708], allowing for quantitative investigation into this poorly-understood yet vitally important mechanisms of resource allocation and feedback. The studies in this dissertation are a testament to how new and enriched data sources can empower both the Science of Science and the complexity perspective, and set an example for future work.

Developing a better understanding of science requires diverse expertise from many different fields. In this dissertation, I have argued that Complexity Science offers a useful viewpoint for making sense of science, and that a particular *complexity perspective* can be applied to understand the bottom-up forces of science and mechanisms that give it structure. Choosing to recognize the complexity of science comes with challenges, but with increasingly-available data, new methodolog-

ical techniques, and concerted theoretical effort, there is massive potential payoff for understanding the structure and dynamics of global science. By applying this complexity-focused Science of Science towards real-world policy and governance, a more equitable, effective, and open science might be possible.

Bibliography

1. Weinberg, A. M. Impact of Large-Scale Science on the United States: Big science is here to stay, but we have yet to make the hard financial and educational choices it imposes. *Science* **134**, 161–164 (1961).
2. De Solla Price, D. *Little Science Big Science* 1st (Columbia University Press, 1963).
3. Schneegans, S., Straza, T. & Lewis, J. *UNESCO Science Report: The Race Against Time for Smarter Development* (UNESCO, Paris, 2021).
4. Schneegans, S. *UNESCO Science Report: Towards 2030* 978-92-3-100129-1 (UNESCO, Paris, 2015).
5. Fernández-Cano, A., Torralbo, M. & Vallejo, M. Reconsidering Price's model of scientific growth: An overview. *Scientometrics* **61**, 301–321 (2004).
6. Dolnick, E. *The Clockwork Universe: Isaac Newton, the Royal Society, and the Birth of the Modern World* 1st Edition (Harper, New York, NY, 2011).
7. Shapin, S. & Schaffer, S. *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life* (Princeton University Press, Princeton, N.J, 2011).
8. Glänzel, W., Debackere, K. & Meyer, M. ‘Triad’ or ‘tetrad’? On global changes in a dynamic world. *Scientometrics* **74**, 71–88 (2008).
9. Xie, Y., Zhang, C. & Lai, Q. China’s rise as a major contributor to science and technology. *Proceedings of the National Academy of Sciences* **111**, 9437–9442 (2014).
10. Wu, S. China: How science made a superpower. *Nature* **574**, 25–28 (2019).
11. Noorden, R. V. Science in East Asia — by the numbers. *Nature* **558**, 500–501 (2018).

12. Thelwall, M. & Levitt, J. M. National scientific performance evolution patterns: Retrenchment, successful expansion, or overextension. *Journal of the Association for Information Science & Technology* **69**, 720–727 (2018).
13. Five hubs of Asian science. *Nature* **558**, 499–499 (2018).
14. Miao, L. *et al.* The latent structure of national scientific development. *arXiv:2104.10812 [cs]* (2021).
15. Maisonobe, M., Jégou, L. & Cabanac, G. Peripheral forces. *Nature* **563**, S18 (2018).
16. Maisonobe, M., Grossetti, M., Milard, B., Jégou, L. & Eckert, D. The global geography of scientific visibility: a deconcentration process (1999–2011). *Scientometrics* **113**, 479–493 (2017).
17. Livingstone, D. N. *Putting Science in Its Place: Geographies of Scientific Knowledge* (University of Chicago Press, Chicago, Ill., 2003).
18. Taylor, M. Z. *The Politics of Innovation: Why Some Countries Are Better Than Others at Science and Technology* 1 edition (Oxford University Press, New York, NY, 2016).
19. Stephan, P. *How Economics Shapes Science* 1 edition (Harvard University Press, 2012).
20. Evans, J. A., Shim, J.-M. & Ioannidis, J. P. A. Attention to Local Health Burden and the Global Disparity of Health Research. *PLOS ONE* **9**, e90147 (2014).
21. Klavans, R. & Boyack, K. W. The Research Focus of Nations: Economic vs. Altruistic Motivations. *PLOS ONE* **12**, e0169383 (2017).
22. Wagner, C. S. & Jonkers, K. Open countries have strong science. *Nature News* **550**, 32 (2017).
23. Saini, A. *Inferior: How Science Got Women Wrong-and the New Research That's Rewriting the Story* (Beacon Press, Boston, 2017).
24. Saini, A. *Superior: The Return of Race Science* (Beacon Press, Boston, 2019).

25. Science benefits from diversity. *Nature* **558** (2018).
26. White, K. E., Robbins, C., Khan, B. & Freyman, C. *Science and Engineering Publication Output Trends: 2014 Shows Rise of Developing Country Output while Developed Countries Dominate Highly Cited Publications* NSF 18-300 (National Center for Science and Engineering Statistics, Arlington, VA, 2017).
27. Gonzalez-Brambila, C. N., Reyes-Gonzalez, L., Veloso, F. & Perez-Angón, M. A. The Scientific Impact of Developing Nations. *PLOS ONE* **11**, e0151328 (2016).
28. Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C. R. Bibliometrics: Global gender disparities in science. *Nature News* **504**, 211 (2013).
29. Sheltzer, J. M. & Smith, J. C. Elite male faculty in the life sciences employ fewer women. *Proceedings of the National Academy of Sciences* **111**, 10107–10112 (2014).
30. Espin, J. *et al.* A persistent lack of international representation on editorial boards in environmental biology. *PLOS Biology* **15**, e2002760 (2017).
31. Mauleón, E., Hillán, L., Moreno, L., Gómez, I. & Bordons, M. Assessing gender balance among journal authors and editorial board members. *Scientometrics* **95**, 87–114 (2013).
32. Posselt, J. R. & Grodsky, E. Graduate Education and Social Stratification. *Annual Review of Sociology* **43**, 353–378 (2017).
33. Holman, L., Stuart-Fox, D. & Hauser, C. E. The gender gap in science: How long until women are equally represented? *PLOS Biology* **16**, e2004956 (2018).
34. Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J. & Handelsman, J. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences* **109**, 16474–16479 (2012).

35. Witteman, H. O., Hendricks, M., Straus, S. & Tannenbaum, C. Female grant applicants are equally successful when peer reviewers assess the science, but not when they assess the scientist. *bioRxiv*, 232868 (2017).
36. Eaton, A. A., Saunders, J. F., Jacobson, R. K. & West, K. How Gender and Race Stereotypes Impact the Advancement of Scholars in STEM: Professors' Biased Evaluations of Physics and Biology Post-Doctoral Candidates. *Sex Roles* **82**, 127–141 (2020).
37. Silbiger, N. J. & Stubler, A. D. Unprofessional peer reviews disproportionately harm under-represented groups in STEM. *PeerJ* **7** (2019).
38. Link, A. M. US and non-US submissions: an analysis of reviewer bias. *JAMA* **280**, 246–247 (1998).
39. Primack, R. B. & Marrs, R. Bias in the review process. *Biological Conservation* **141**, 2919–2920 (2008).
40. Whitley, R. *The Intellectual and Social Organization of the Sciences* 2nd edition (Oxford University Press, Oxford England ; New York, 2000).
41. Yan, E., Ding, Y., Milojević, S. & Sugimoto, C. R. Topics in dynamic research communities: An exploratory study for the field of information retrieval. *Journal of Informetrics* **6**, 140–153 (2012).
42. Ramage, D., Manning, C. D. & McFarland, D. A. Mapping three decades of intellectual change in academia, 15.
43. Shi, F., Foster, J. G. & Evans, J. A. Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Social Networks* **43**, 73–85 (2015).
44. Uzzi, B., Mukherjee, S., Stringer, M. & Jones, B. Atypical Combinations and Scientific Impact. *Science* **342**, 468–472 (2013).

45. Milojević, S. Quantifying the cognitive extent of science. *Journal of Informetrics* **9**, 962–973 (2015).
46. Herrera, M., Roberts, D. C. & Gulbahce, N. Mapping the Evolution of Scientific Fields. *PLOS ONE* **5**, e10355 (2010).
47. Jones, B. F. The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder? *The Review of Economic Studies* **76**, 283–317 (2009).
48. Larivière, V. *et al.* Contributorship and division of labor in knowledge production. *Social Studies of Science* **46**, 417–435 (2016).
49. Robinson-Garcia, N., Costas, R., Sugimoto, C. R., Larivière, V. & Nane, G. F. Task specialization across research careers. *eLife* **9** (eds Rodgers, P. & Morgan, A.) e60586 (2020).
50. Wuchty, S., Jones, B. F. & Uzzi, B. The Increasing Dominance of Teams in Production of Knowledge. *Science* **316**, 1036–1039 (2007).
51. Page, S. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies - New Edition* Revised ed. edition (Princeton University Press, Princeton, NJ, 2008).
52. Van Raan, A. F. J. The influence of international collaboration on the impact of research results. *Scientometrics* **42**, 423–428 (1998).
53. Wagner, C. S., Whetsell, T. A. & Leydesdorff, L. Growth of international collaboration in science: revisiting six specialties. *Scientometrics* **110**, 1633–1652 (2017).
54. Wagner, C. S. Six case studies of international collaboration in science. *Scientometrics* **62**, 3–26 (2005).
55. Czaika, M., de Haas, H. & Villares-Varela, M. The Global Evolution of Travel Visa Regimes. *Population and Development Review* **44**, 589–622 (2018).

56. Czaika, M. & Toma, S. International academic mobility across space and time: The case of Indian academics. *Population, Space and Place* **23**, e2069 (2017).
57. Van Noorden, R. Interdisciplinary research by the numbers. *Nature* **525**, 306–307 (2015).
58. Pan, R. K., Sinha, S., Kaski, K. & Saramäki, J. The evolution of interdisciplinarity in physics research. *Scientific Reports* **2**, 1–8 (2012).
59. Bridle, H., Vrieling, A., Cardillo, M., Araya, Y. & Hinojosa, L. Preparing for an interdisciplinary future: A perspective from early-career researchers. *Futures* **53**, 22–32 (2013).
60. Lungceanu, A., Huang, Y. & Contractor, N. S. Understanding the assembly of interdisciplinary teams and its impact on performance. *Journal of informetrics* **8**, 59–70 (2014).
61. Fortunato, S. *et al.* Science of science. *Science* **359**, eaao0185 (2018).
62. Boyack, K. W., Klavans, R. & Börner, K. Mapping the backbone of science. *Scientometrics* **64**, 351–374 (2005).
63. Börner, K. *et al.* Design and Update of a Classification System: The UCSD Map of Science. *PLOS ONE* **7**, e39464 (2012).
64. Bollen, J. *et al.* Clickstream Data Yields High-Resolution Maps of Science. *PLOS ONE* **4**, e4803 (2009).
65. Barabási, A.-L. *The Formula: The Universal Laws of Success* (Little, Brown and Company, New York, 2018).
66. Yin, Y., Wang, Y., Evans, J. A. & Wang, D. Quantifying the dynamics of failure across science, startups and security. *Nature* **575**, 190–194 (2019).
67. Petersen, A. M. & Penner, O. Inequality and cumulative advantage in science careers: a case study of high-impact journals. *EPJ Data Science* **3**, 24 (2014).

68. Huang, J., Gates, A. J., Sinatra, R. & Barabási, A.-L. Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences* **117**, 4609–4616 (2020).
69. Feyerabend, P. *Against Method* (1975).
70. Cetina, K. K. *Epistemic Cultures: How the Sciences Make Knowledge* (Harvard University Press, Cambridge, Mass, 1999).
71. Wagner, C. S. *The New Invisible College: Science for Development* (Brookings Institution Press, Washington, D.C, 2008).
72. Teichler, U. Academic Mobility and Migration: What We Know and What We Do Not Know. *European Review* **23**, S6–S37 (S1 2015).
73. Leung, M. W. H. 'Read ten thousand books, walk ten thousand miles': geographical mobility and capital accumulation among Chinese scholars. *Transactions of the Institute of British Geographers* **38**, 311–324 (2013).
74. Jonkers, K. & Tijssen, R. Chinese researchers returning home: Impacts of international mobility on research collaboration and scientific productivity. *Scientometrics* **77**, 309–333 (2008).
75. Jerrim, J. *Family Background and Access to "High Status" Universities* (Sutton Trust, 2013).
76. Huang, C. Gender differences in academic self-efficacy: a meta-analysis. *European Journal of Psychology of Education* **28**, 1–35 (2012).
77. Fraiberger, S. P., Sinatra, R., Resch, M., Riedl, C. & Barabási, A.-L. Quantifying reputation and success in art. *Science* **362**, 825–829 (2018).
78. Petersen, A. M. *et al.* Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences* **111**, 15316–15321 (2014).
79. Merton, R. K. The Matthew Effect in Science. *Science* **159**, 56–63 (1968).

80. Sekara, V. *et al.* The chaperone effect in scientific publishing. *Proceedings of the National Academy of Sciences* **115**, 12603–12607 (2018).
81. Azoulay, P., Fons-Rosen, C. & Graff Zivin, J. S. Does Science Advance One Funeral at a Time? *American Economic Review* **109**, 2889–2920 (2019).
82. Chinchilla-Rodríguez, Z., Sugimoto, C. R. & Larivière, V. Follow the leader: On the relationship between leadership and scholarly impact in international collaborations. *PLOS ONE* **14**, e0218309 (2019).
83. Jonkers, K. & Cruz-Castro, L. Research upon return: The effect of international mobility on scientific ties, production and impact. *Research Policy* **42**, 1366–1377 (2013).
84. Rodrigues, M. L., Nimrichter, L. & Cordero, R. J. B. The benefits of scientific mobility and international collaboration. *FEMS Microbiology Letters* **363** (2016).
85. Abramo, G., D'Angelo, C. A. & Solazzi, M. The relationship between scientists' research performance and the degree of internationalization of their research. *Scientometrics* **86**, 629–643 (2011).
86. Wagner, A. Causality in Complex Systems. *Biology and Philosophy* **14**, 83–101 (1999).
87. Taleb, N. N. *The Black Swan: Second Edition: The Impact of the Highly Improbable: With a new section: "On Robustness and Fragility"* 2nd ed. edition (Random House Trade Paperbacks, New York, 2010).
88. Ke, Q., Ferrara, E., Radicchi, F. & Flammini, A. Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences* **112**, 7426–7431 (2015).
89. Clauset, A., Larremore, D. B. & Sinatra, R. Data-driven predictions in the science of science. *Science* **355**, 477–480 (2017).
90. San Miguel, M. *et al.* Challenges in complex systems science. *The European Physical Journal Special Topics* **214**, 245–271 (2012).

91. Ferrucci, E. Migration, innovation and technological diversion: German patenting after the collapse of the Soviet Union. *Research Policy*, 104057 (2020).
92. Ganguli, I. Immigration and Ideas: What Did Russian Scientists “Bring” to the United States? *Journal of Labor Economics* **33**, S257–S288 (S1 2015).
93. Myers, K. R. *et al.* Unequal effects of the COVID-19 pandemic on scientists. *Nature Human Behaviour* **4**, 880–883 (2020).
94. Van Noorden, R. The scientists who get credit for peer review. *Nature News* (2014).
95. Boyack, K. W., van Eck, N. J., Colavizza, G. & Waltman, L. Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics* **12**, 59–73 (2018).
96. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, 5233 (2019).
97. Caron, E. & van Eck, N. J. *Large scale author name disambiguation using rule-based scoring and clustering* in *Proceedings of the Science and Technology Indicators Conference 2014* International conference on science and technology indicators (Leiden University, Leiden, Netherlands, 2014), 79–86.
98. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs, stat]* (2013).
99. McKeown, K. *et al.* Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology*, n/a–n/a (2016).
100. Krenn, M. & Zeilinger, A. Predicting research trends with semantic and neural networks with an application in quantum physics. *Proceedings of the National Academy of Sciences* **117**, 1910–1916 (2020).
101. Weis, J. W. & Jacobson, J. M. Learning on knowledge graph dynamics provides an early warning of impactful research. *Nature Biotechnology* (2021).

102. Ladyman, J. & Wiesner, K. *What Is a Complex System?* (2020).
103. Mitchell, M. *Complexity: A Guided Tour* 1st edition (Oxford University Press, Oxford, 2011).
104. Telesford, Q. K., Simpson, S. L., Burdette, J. H., Hayasaka, S. & Laurienti, P. J. The Brain as a Complex System: Using Network Science as a Tool for Understanding the Brain. *Brain Connectivity* **1**, 295–308 (2011).
105. Wilbur, H. M. Experimental Ecology of Food Webs: Complex Systems in Temporary Ponds. *Ecology* **78**, 2279–2302 (1997).
106. Mandelbrot, B. & Hudson, R. L. *The Misbehavior of Markets: A Fractal View of Financial Turbulence* Annotated edition (Basic Books, New York, 2006).
107. Zeng, A. *et al.* The science of science: From the perspective of complex systems. *Physics Reports. The Science of Science: From the Perspective of Complex Systems* **714-715**, 1–73 (2017).
108. Catalini, C. Microgeography and the Direction of Inventive Activity. *Management Science* **64**, 4348–4364 (2017).
109. Kabo, F. W., Cotton-Nessler, N., Hwang, Y., Levenstein, M. C. & Owen-Smith, J. Proximity effects on the dynamics and outcomes of scientific collaborations. *Research Policy* **43**, 1469–1485 (2014).
110. Bauder, H. International Mobility and Social Capital in the Academic Field. *Minerva* **58**, 367–387 (2020).
111. Helmer, M., Schottdorf, M., Neef, A. & Battaglia, D. Gender bias in scholarly peer review. *eLife* **6** (ed Rodgers, P.) e21718 (2017).
112. Hofstra, B. *et al.* The Diversity–Innovation Paradox in Science. *Proceedings of the National Academy of Sciences* (2020).

113. Hoppe, T. A. *et al.* Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. *Science Advances* **5**, eaaw7238 (2019).
114. Goldin, C. & Rouse, C. Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review* **90**, 715–741 (2000).
115. Morgan, A. C. *et al.* The unequal impact of parenthood in academia. *Science Advances* **7**, eabd1996 (2021).
116. Morgan, A. C., Economou, D. J., Way, S. F. & Clauset, A. Prestige drives epistemic inequality in the diffusion of scientific ideas. *EPJ Data Science* **7**, 40 (2018).
117. Nomaler, Ö., Frenken, K. & Heimeriks, G. On Scaling of Scientific Knowledge Production in U.S. Metropolitan Areas. *PLOS ONE* **9**, e110805 (2014).
118. Sawyer, R. K. *Social Emergence: Societies As Complex Systems* (Cambridge University Press, Cambridge ; New York, 2005).
119. Barabási, A.-L. & Albert, R. Emergence of Scaling in Random Networks. *Science* **286**, 509–512 (1999).
120. Sinatra, R., Wang, D., Deville, P., Song, C. & Barabási, A.-L. Quantifying the evolution of individual scientific impact. *Science* **354** (2016).
121. Malmgren, R. D., Ottino, J. M. & Amaral, L. A. N. The role of mentorship in protégé performance. *Nature* **465**, 622–626 (2010).
122. Ma, Y., Mukherjee, S. & Uzzi, B. Mentorship and protégé success in STEM fields. *Proceedings of the National Academy of Sciences* **117**, 14077–14083 (2020).
123. Petersen, A. M. Quantifying the impact of weak, strong, and super ties in scientific careers. *Proceedings of the National Academy of Sciences* **112**, E4671–E4680 (2015).

124. Way, S. F., Morgan, A. C., Larremore, D. B. & Clauset, A. Productivity, prominence, and the effects of academic environment. *Proceedings of the National Academy of Sciences* **116**, 10729–10733 (2019).
125. Mongeon, P. & Paul-Hus, A. The Journal Coverage of Web of Science and Scopus: a Comparative Analysis. *Scientometrics* **106**, 213–228 (2016).
126. Sivertsen, G. & Larsen, B. Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: an empirical analysis of the potential. *Scientometrics* **91**, 567–575 (2012).
127. Parsons, E. A. *The Alexandrian library, glory of the Hellenic world;: Its rise, antiquities, and destruction* (American Elsevier Pub. Co, 1967).
128. *Libraries in American Periodicals Before 1876: Bibliography with Abstracts and an Index* (eds Barr, L. J. & etc) (McFarland & Co Inc, Jefferson, N.C, 1983).
129. Buchner, E. F. Psychological progress. *Psychological Bulletin* **1**, 57–64 (1904).
130. Billings, J. S. An Address on our Medical Literature. *British Medical Journal* **2**, 262–268 (1881).
131. Hulme, E. W. *Statistical Bibliography in Relation to the Growth of Modern Civilization: Two Lectures Delivered in the University of Cambridge in May, 1922* (HardPress Publishing, Place of publication not identified, 1922).
132. Csiszar, A. Peer review: Troubled from the start. *Nature News* **532**, 306 (2016).
133. Fernberger, S. W. On the number of articles of psychological interest published in the different languages; 1936-1945. *The American Journal of Psychology* **59**, 284–290 (1946).
134. Cattell, J. M. A STATISTICAL STUDY OF AMERICAN MEN OF SCIENCE. III. *Science (New York, N.Y.)* **24**, 732–742 (1906).

135. Cattell, J. M. Statistics of American Psychologists. *The American Journal of Psychology* **14**, 310–328 (1903).
136. Franz, S. I. The scientific productivity of American professional psychologists. *Psychological Review* **24**, 197–219 (1917).
137. Nelson, H. The Creative Years. *The American Journal of Psychology* **40**, 303–311 (1928).
138. Shapiro, F. R. Origins of bibliometrics, citation indexing, and citation analysis: The neglected legal literature. *Journal of the American Society for Information Science* **43**, 337–339 (1992).
139. Gross, P. L. & Gross, E. M. COLLEGE LIBRARIES AND CHEMICAL EDUCATION. *Science (New York, N.Y.)* **66**, 385–389 (1927).
140. Garfield, E. Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science* **122**, 108–111 (1955).
141. Garfield, E. in *Science Citation Index* v–xvi (1963).
142. Cole, S. & Cole, J. R. Scientific Output and Recognition: A Study in the Operation of the Reward System in Science. *American Sociological Review* **32**, 377 (1967).
143. Pritchard, A. Statistical Bibliography or Bibliometrics? *Journal of Documentation* **25**, 348–349 (1969).
144. Bensman, S. J. Garfield and the impact factor. *Annual Review of Information Science and Technology* **41**, 93–155 (2007).
145. Archambault, É. & Larivière, V. History of the journal impact factor: Contingencies and consequences. *Scientometrics* **79**, 635–649 (2009).
146. Garfield, E. The History and Meaning of the Journal Impact Factor. *JAMA* **295**, 90–93 (2006).

147. Quan, W., Chen, B. & Shu, F. Publish or impoverish: An investigation of the monetary reward system of science in China (1999-2016). *Aslib Journal of Information Management* **69**, 486–502 (2017).
148. Price, D. d. S. Editorial statements. *Scientometrics* **1**, 3–8 (2005).
149. Hirsch, J. E. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences* **102**, 16569–16572 (2005).
150. Leydesdorff, L., Bornmann, L., Comins, J. A. & Milojević, S. Citations: Indicators of Quality? The Impact Fallacy. *Frontiers in Research Metrics and Analytics* **1** (2016).
151. Waltman, L. *et al.* The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology* **63**, 2419–2432 (2012).
152. Priem, P., Taraborelli, D., Groth, P. & Neylon, C. *altmetrics: a manifesto – altmetrics.org*
153. Wang, D. & Barabási, A.-L. *The Science of Science* 1st edition (Cambridge University Press, Cambridge New York Port Melbourne New Delhi Singapore, 2021).
154. Poincaré, H. *The Three-Body Problem and the Equations of Dynamics: Poincaré's Foundational Work on Dynamical Systems Theory* (Springer International Publishing, 2017).
155. Lorenz, E. N. Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences* **20**, 130–141 (1963).
156. Li, T.-Y. & Yorke, J. A. Period Three Implies Chaos. *The American Mathematical Monthly* **82**, 985–992 (1975).
157. Mandelbrot, B. B. *The Fractal Geometry of Nature* 2nd prt. edition (Times Books, San Francisco, 1982).
158. Feigenbaum, M. J. Quantitative universality for a class of nonlinear transformations. *Journal of Statistical Physics* **19**, 25–52 (1978).

159. Gleick, J. *Chaos: Making a New Science* Anniversary, Reprint edition (Penguin Books, New York, N.Y, 1987).
160. Euler, L. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae* **8**, 128–140 (1741).
161. Erdős, P. & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* **5**, 17–60 (1960).
162. Milgram, S. The small world problem. *Psychology today* **2**, 60–67 (1967).
163. Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998).
164. Cowan, G. A. *Manhattan Project to the Santa Fe Institute: The Memoirs of George A. Cowan* 1st edition (University of New Mexico Press, 2011).
165. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423 (1948).
166. Hayek, F. A. *The Use of Knowledge in Society* SSRN Scholarly Paper ID 1505216 (Social Science Research Network, Rochester, NY, 1945).
167. Schelling, T. C. Dynamic models of segregation. *The Journal of Mathematical Sociology* **1**, 143–186 (1971).
168. Zeng, X. H. T. *et al.* Differences in Collaboration Patterns across Discipline, Career Stage, and Gender. *PLOS Biology* **14**, e1002573 (2016).
169. Clayton, P. & Davies, P. *The Re-Emergence of Emergence* (Oxford University Press, 2008).
170. Kawakatsu, M., Chodrow, P. S., Eikmeier, N. & Larremore, D. B. Emergence of Hierarchy in Networked Endorsement Dynamics. *arXiv:2007.04448 [nlin, physics:physics]* (2020).
171. Weaver, W. Science and Complexity. *American Scientist* **36** (1948).

172. Spector, L., Klein, J., Perry, C. & Feinstein, M. Emergence of Collective Behavior in Evolving Populations of Flying Agents. *Genetic Programming and Evolvable Machines* **6**, 111–125 (2005).
173. Fisher, D. N. & Pruitt, J. N. Insights from the study of complex systems for the ecology and evolution of animal populations. *Current Zoology* **66**, 1–14 (2020).
174. Edwards, P. N. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming* (MIT Press, Cambridge, Massachusetts London, England, 2013).
175. Latour, B. *We Have Never Been Modern* trans. by Porter, C. (Harvard University Press, Cambridge, Mass, 1993).
176. Snow, C. P. *The Two Cultures* (Cambridge University Press, Cambridge, 2012).
177. Czaika, M. & Orazbayev, S. The globalisation of scientific mobility, 1970–2014. *Applied Geography* **96**, 1–10 (2018).
178. Akcigit, U., Grigsby, J. & Nicholas, T. Immigration and the Rise of American Ingenuity. *American Economic Review* **107**, 327–331 (2017).
179. Newman, M. E. J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* **98**, 404–409 (2001).
180. Kuhn, T., Perc, M. & Helbing, D. Inheritance Patterns in Citation Networks Reveal Scientific Memes. *Physical Review X* **4**, 041036 (2014).
181. Gilbert, N. *et al.* in *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)* (ed Wright, J. D.) 529–534 (Elsevier, Oxford, 2015).
182. Börner, K., Chen, C. & Boyack, K. W. Visualizing knowledge domains. *Annual Review of Information Science and Technology* **37**, 179–255 (2003).
183. Boyack, K. W. Mapping knowledge domains: Characterizing PNAS. *Proceedings of the National Academy of Sciences* **101**, 5192–5199 (suppl 1 2004).

184. Clauset, A., Arbesman, S. & Larremore, D. B. Systematic inequality and hierarchy in faculty hiring networks. *Science Advances* **1**, e1400005 (2015).
185. Drivas, K. & Kremmydas, D. The Matthew effect of a journal's ranking. *Research Policy* **49**, 103951 (2020).
186. MacLeod, W. B. & Urquiola, M. *Why Does the U.S. Have the Best Research Universities? Incentives, Resources, and Virtuous Circles* w28279 (National Bureau of Economic Research, 2020).
187. Kuhn, T. S. *The Structure of Scientific Revolutions* 3rd edition (University of Chicago Press, Chicago, IL, 1996).
188. Bourdieu, P. *Science of Science and Reflexivity* 1 edition. Trans. by Nice, R. (University of Chicago Press, Chicago, 2004).
189. Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLOS Medicine* **2**, e124 (2005).
190. Amrhein, V., Greenland, S. & McShane, B. Scientists rise up against statistical significance. *Nature* **567**, 305 (2019).
191. Shwed, U. & Bearman, P. S. The Temporal Structure of Scientific Consensus Formation. *American Sociological Review* **75**, 817–840 (2010).
192. Oreskes, N. *et al.* *Why Trust Science?* (ed Macedo, S.) (Princeton University Press, Princeton, NJ, 2019).
193. Azoulay, P., Graff Zivin, J. S. & Wang, J. Superstar Extinction. *The Quarterly Journal of Economics* **125**, 549–589 (2010).
194. Braunerhjelm, P., Ding, D. & Thulin, P. Labour market mobility, knowledge diffusion and innovation. *European Economic Review* **123**, 103386 (2020).

195. Azoulay, P., Zivin, J. S. G. & Sampat, B. N. *The Diffusion of Scientific Knowledge Across Time and Space: Evidence from Professional Transitions for the Superstars of Medicine* Working Paper 16683 (National Bureau of Economic Research, 2011).
196. Bosenman, M. F. Serendipity and Scientific Discovery. *The Journal of Creative Behavior* **22**, 132–138 (1988).
197. Fink, T. M. A., Reeves, M., Palma, R. & Farr, R. S. Serendipity and strategy in rapid innovation. *Nature Communications* **8**, 1–9 (2017).
198. Johnson, S. *Where Good Ideas Come From: The Natural History of Innovation* Reprint edition (Riverhead Books, New York, 2011).
199. Tollefson, J. China declared world's largest producer of scientific articles. *Nature* **553**, 390–390 (2018).
200. Nielsen, M. W. & Andersen, J. P. Global citation inequality is on the rise. *Proceedings of the National Academy of Sciences* **118** (2021).
201. Allison, P. D., Long, J. S. & Krauze, T. K. Cumulative Advantage and Inequality in Science. *American Sociological Review* **47**, 615–625 (1982).
202. Halfman, W. & Leydesdorff, L. Is Inequality Among Universities Increasing? Gini Coefficients and the Elusive Rise of Elite Universities. *Minerva* **48**, 55–72 (2010).
203. *Artificial Intelligence Index 2019* (AI Index 2019, 2019).
204. Finkelstein, M., Conley, V. M. & Schuster, J. H. *Taking the Measure of Faculty Diversity* (TIAA Institute, 2016).
205. Sugimoto, C. R., Berube, N. & Larivière, V. *On a Trajectory Towards Parity: A Historical Analysis of Gender in Funding from the National Science Foundation* in. 16th International Conference on Scientometrics and Informetrics (Wuhan, China, 2017).

206. Inno, L., Rotundi, A. & Piccialli, A. COVID-19 lockdown effects on gender inequality. *Nature Astronomy* **4**, 1114–1114 (2020).
207. Viglione, G. Are women publishing less during the pandemic? Here's what the data say. *Nature* **581**, 365–366 (2020).
208. Macaluso, B., Larivière, V., Sugimoto, T. & Sugimoto, C. R. Is Science Built on the Shoulders of Women? A Study of Gender Differences in Contributorship. *Academic Medicine: Journal of the Association of American Medical Colleges* **91**, 1136–1142 (2016).
209. Tamblyn, R., Girard, N., Qian, C. J. & Hanley, J. Assessment of potential bias in research grant peer review in Canada. *CMAJ* **190**, E489–E499 (2018).
210. Bedi, G., Van Dam, N. T. & Munafo, M. Gender inequality in awarded research grants. *Lancet (London, England)* **380**, 474 (2012).
211. Shen, Y. A., Webster, J. M., Shoda, Y. & Fine, I. Persistent Underrepresentation of Women's Science in High Profile Journals. *bioRxiv*, 275362 (2018).
212. Hengel, E. *Publishing while Female. Are women held to higher standards? Evidence from peer review* 1753 (Faculty of Economics, University of Cambridge, 2017).
213. Dupas, P., Modestino, A. S., Niederle, M., Wolfers, J. & Collective, T. S. D. *Gender and the Dynamics of Economics Seminars* w28494 (National Bureau of Economic Research, 2021).
214. Rossiter, M. W. The Matthew Matilda Effect in Science. *Social Studies of Science* **23**, 325–341 (1993).
215. Ward, K. *Faculty Service Roles and the Scholarship of Engagement. ASHE-ERIC Higher Education Report. Jossey-Bass Higher and Adult Education Series* (Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741 (\$24 plus shipping; \$108 per year), 2003).
216. Guarino, C. M. & Borden, V. M. H. Faculty Service Loads and Gender: Are Women Taking Care of the Academic Family? *Research in Higher Education* **58**, 672–694 (2017).

217. Hunter, L. A. & Leahey, E. Parenting and research productivity: New evidence and methods. *Soc Stud Sci* **40**, 433–451 (2010).
218. *These 6 graphs show that Black scientists are underrepresented at every level* Science News.
219. *Diversity Gaps in Computer Science:Exploring the Underrepresentation of Girls, Blacks and Hispanics* (Gallup, 2016).
220. *Google Diversity Annual Report 2020* (Google, 2020), 33.
221. Valentine, H. A. *Data Science: Meet Diversity — SWD at NIH*
222. Ginther, D. K., Kahn, S. & Schaffer, W. T. Gender, Race/Ethnicity, and National Institutes of Health R01 Research Awards: Is There Evidence of a Double Bind for Women of Color? *Academic medicine : journal of the Association of American Medical Colleges* **91**, 1098–1107 (2016).
223. Smith, B. P. & Hawkins, B. Examining Student Evaluations of Black College Faculty: Does Race Matter? *The Journal of Negro Education* **80**, 149–162 (2011).
224. Laland, K. N. Racism in academia, and why the ‘little things’ matter. *Nature* **584**, 653–654 (2020).
225. Basalla, G. The Spread of Western Science. *Science* **156**, 611–622 (1967).
226. Zhou, P. & Leydesdorff, L. The emergence of China as a leading nation in science. *Research Policy* **35**, 83–104 (2006).
227. Wagner, C., Brahmakulam, I. T., Jackson, B. A., Wong, A. & Yoda, T. *Science and Technology Collaboration: Building Capacity in Developing Countries?* MR-1357.0-WB (RAND Publications, Santa Monica, California, USA, 2001).
228. Boudry, C. *et al.* Worldwide inequality in access to full text scientific articles: the example of ophthalmology. *PeerJ* **7** (2019).
229. Bohannon, J. Who’s downloading pirated papers? Everyone. *Science* (2016).

230. Gordin, M. D. *Scientific Babel: How Science Was Done Before and After Global English* 1 edition (University of Chicago Press, Chicago ; London, 2015).
231. Man, J. P., Weinkauf, J. G., Tsang, M. & Sin, J. H. D. D. Why do Some Countries Publish More Than Others? An International Comparison of Research Funding, English Proficiency and Publication Output in Highly Ranked General Medical Journals. *European Journal of Epidemiology* **19**, 811–817 (2004).
232. Leeuwen, V. *et al.* First evidence of serious language-bias in the use of citation analysis for the evaluation of national science systems. *Research Evaluation* **9**, 155–156 (2000).
233. Flowerdew, J. Scholarly writers who use English as an Additional Language: What can Goffman’s “Stigma” tell us? *Journal of English for Academic Purposes. English for Research Publication Purposes* **7**, 77–86 (2008).
234. Harris, M. *et al.* Explicit Bias Toward High-Income-Country Research: A Randomized, Blinded, Crossover Experiment Of English Clinicians. *Health Affairs (Project Hope)* **36**, 1997–2004 (2017).
235. Bozeman, B., Dietz, J. S. & Gaughan, M. Scientific and technical human capital: An alternative model for research evaluation. *International Journal of Technology Management* **22**, 716–740 (2001).
236. Corley, E. A., Bozeman, B., Zhang, X. & Tsai, C.-C. The expanded scientific and technical human capital model: the addition of a cultural dimension. *The Journal of Technology Transfer* **44**, 681–699 (2019).
237. Polanyi, M. & Sen, A. *The Tacit Dimension* Revised ed. edition (University of Chicago Press, Chicago ; London, 2009).
238. Epstein, D. *Range: Why Generalists Triumph in a Specialized World* Illustrated edition (Riverhead Books, New York, 2019).

239. Granovetter, M. S. The Strength of Weak Ties. *American Journal of Sociology* **78**, 1360–1380 (1973).
240. Li, E. Y., Liao, C. H. & Yen, H. R. Co-authorship networks and research impact: A social capital perspective. *Research Policy* **42**, 1515–1530 (2013).
241. Teplitskiy, M., Acuna, D., Elamrani-Raoult, A., Körding, K. & Evans, J. The sociology of scientific validity: How professional networks shape judgement in peer review. *Research Policy* **47**, 1825–1841 (2018).
242. Wennerås, C. & Wold, A. Nepotism and sexism in peer-review. *Nature* **387**, 341–343 (1997).
243. Sandström, U. & Hällsten, M. Persistent nepotism in peer-review. *Scientometrics* **74**, 175–189 (2008).
244. Burris, V. The Academic Caste System: Prestige Hierarchies in PhD Exchange Networks. *American Sociological Review* **69**, 239–264 (2004).
245. Ross, J. S. *et al.* Effect of blinded peer review on abstract acceptance. *JAMA* **295**, 1675–1680 (2006).
246. Okike, K., Hug, K. T., Kocher, M. S. & Leopold, S. S. Single-blind vs Double-blind Peer Review in the Setting of Author Prestige. *JAMA* **316**, 1315–1316 (2016).
247. Tomkins, A., Zhang, M. & Heavlin, W. D. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 201707323 (2017).
248. Bourdieu, P. *Homo Academicus* (Stanford University Press, 1988).
249. Larson, R. C., Ghaffarzadegan, N. & Xue, Y. Too Many PhD Graduates or Too Few Academic Job Openings: The Basic Reproductive Number R_0 in Academia. *Systems Research and Behavioral Science* **31**, 745–750 (2014).
250. Gross, K. & Bergstrom, C. T. Contest models highlight inherent inefficiencies of scientific funding competitions. *PLOS Biology* **17**, e3000065 (2019).

251. Wang, J., Lee, Y.-N. & Walsh, J. P. Funding model and creativity in science: Competitive versus block funding and status contingency effects. *Research Policy* **47**, 1070–1083 (2018).
252. Shen, H.-W. & Barabási, A.-L. Collective credit allocation in science. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 12325–12330 (2014).
253. McCain, K. W. Assessing obliteration by incorporation in a full-text database: JSTOR, Economics, and the concept of “bounded rationality”. *Scientometrics* **101**, 1445–1459 (2014).
254. Stigler, S. M. *Statistics on the Table: The History of Statistical Concepts and Methods* 1st edition (Harvard University Press, Cambridge, Mass., 2002).
255. Balietti, S. & Riedl, C. Incentives, competition, and inequality in markets for creative production. *Research Policy* **50**, 104212 (2021).
256. Jeong, H., Néda, Z. & Barabási, A. L. Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)* **61**, 567 (2003).
257. Redner, S. Citation Statistics from 110 Years of Physical Review. *Physics Today* **58**, 49–54 (2005).
258. Larivière, V. & Gingras, Y. The impact factor’s Matthew Effect: A natural experiment in bibliometrics. *Journal of the American Society for Information Science and Technology* **61**, 424–427 (2010).
259. Bol, T., Vaan, M. d. & Rijt, A. v. d. The Matthew effect in science funding. *Proceedings of the National Academy of Sciences* **115**, 4887–4890 (2018).
260. Medoff, M. H. Evidence of a Harvard and Chicago Matthew Effect. *Journal of Economic Methodology* **13**, 485–506 (2006).
261. Bonitz, M., Bruckner, E. & Scharnhorst, A. Characteristics and impact of the matthew effect for countries. *Scientometrics* **40**, 407–422 (1997).
262. Bonitz, M. Ten years Matthew effect for countries. *Scientometrics* **64**, 375–379 (2005).

263. Cox, D. How mRNA went from a scientific backwater to a pandemic crusher. *Wired UK* (2020).
264. Merton, R. K. in *The Sociology of Science: Theoretical and Empirical Investigations* (ed Storer, N. W.) (University of Chicago Press, Chicago, 1942).
265. Ghiasi, G., Larivière, V. & Sugimoto, C. R. On the Compliance of Women Engineers with a Gendered Scientific System. *PLOS ONE* **10**, e0145931 (2015).
266. Ghiasi, G., Mongeon, P., Sugimoto, C. & Larivière, V. in *STI 2018 Conference Proceedings* 1519–1525 (Centre for Science and Technology Studies (CWTS), 2018).
267. Azoulay, P. & Lynn, F. B. Self-Citation, Cumulative Advantage, and Gender Inequality in Science. *Sociological Science* **7**, 152–186 (2020).
268. West, J. D., Jacquet, J., King, M. M., Correll, S. J. & Bergstrom, C. T. The Role of Gender in Scholarly Authorship. *PLOS ONE* **8**, e66212 (2013).
269. Primack, R. B. & O’Leary, V. Cumulative disadvantages in the careers of women ecologists. *BioScience* **43**, 158–165 (1993).
270. Day, T. E. The big consequences of small biases: A simulation of peer review. *Research Policy* **44**, 1266–1270 (2015).
271. Ginther, D. K. *et al.* RACE, ETHNICITY, AND NIH RESEARCH AWARDS. *Science (New York, N.Y.)* **333**, 1015–1019 (2011).
272. Ginther, D. K. *et al.* Publications as predictors of racial and ethnic differences in NIH research awards. *PLOS ONE* **13**, e0205929 (2018).
273. Peng, H., Teplitskiy, M. & Jurgens, D. Author Mentions in Science News Reveal Wide-Spread Ethnic Bias. *arXiv:2009.01896 [cs]* (2020).

274. Sugimoto, C. R., Ahn, Y.-Y., Smith, E., Macaluso, B. & Larivière, V. Factors affecting sex-related reporting in medical research: a cross-disciplinary bibliometric analysis. *The Lancet* **393**, 550–559 (2019).
275. Ensmenger, N. Is chess the drosophila of artificial intelligence? A social history of an algorithm. *Social Studies of Science* **42**, 5–30 (2012).
276. Mewa, T. in *Identifying Gender and Sexuality of Data Subjects* (PubPub, 2020).
277. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a Feather: Homophily in Social Networks (2001).
278. Holman, L. & Morandin, C. Researchers collaborate with same-gendered colleagues more often than expected across the life sciences. *PLOS ONE* **14**, e0216128 (2019).
279. AlShebli, B. K., Rahwan, T. & Woon, W. L. The preeminence of ethnic diversity in scientific collaboration. *Nature Communications* **9**, 1–10 (2018).
280. Kwiek, M. & Roszka, W. Gender-Based Homophily in Research: A Large-Scale Study of Man-Woman Collaboration. *arXiv:2006.03935 [cs]* (2020).
281. Yegros-Yegros, A., Rafols, I. & D'Este, P. Does Interdisciplinary Research Lead to Higher Citation Impact? The Different Effect of Proximal and Distal Interdisciplinarity. *PLOS ONE* **10**, e0135095 (2015).
282. Okamura, K. Interdisciplinarity revisited: evidence for research impact and dynamism. *Palgrave Communications* **5**, 1–9 (2019).
283. Siegel, D. A. Social Networks and Collective Action. *American Journal of Political Science* **53**, 122–138 (2009).
284. Banerjee, A., Chandrasekhar, A. G., Duflo, E. & Jackson, M. O. The Diffusion of Microfinance. *Science* **341** (2013).

285. Yong, K., Sauer, S. J. & Mannix, E. A. Conflict and Creativity in Interdisciplinary Teams. *Small Group Research* **45**, 266–289 (2014).
286. Catalini, C., Fons-Rosen, C. & Gaulé, P. *Did cheaper flights change the direction of science?* 1520 (Department of Economics and Business, Universitat Pompeu Fabra, 2016).
287. Zucker, L. G., Darby, M. R. & Armstrong, J. GEOGRAPHICALLY LOCALIZED KNOWLEDGE SPILLOVERS OR MARKETS? *Economic Inquiry* **36**, 65–86 (1998).
288. Aghion, P. & Jaravel, X. Knowledge Spillovers, Innovation and Growth. *The Economic Journal* **125**, 533–573 (2015).
289. Boschma, R. Proximity and Innovation: A Critical Assessment. *Regional Studies* **39**, 61–74 (2005).
290. Moretti, E. *The New Geography of Jobs* Reprint edition (Mariner Books, Boston, Mass, 2013).
291. Gruber, J. & Johnson, S. *Jump-Starting America: How Breakthrough Science Can Revive Economic Growth and the American Dream* (PublicAffairs, New York, NY, 2019).
292. Chinchilla-Rodríguez, Z. *et al.* A Global Comparison of Scientific Mobility and Collaboration According to National Scientific Capacities. *Frontiers in Research Metrics and Analytics* **3** (2018).
293. Wang, J., Hooi, R., Li, A. X. & Chou, M.-h. Collaboration patterns of mobile academics: The impact of international mobility. *Science and Public Policy* **46**, 450–462 (2019).
294. Petersen, A. M. Multiscale impact of researcher mobility. *Journal of The Royal Society Interface* **15**, 20180580 (2018).
295. Sugimoto, C. R. *et al.* Scientists have most impact when they're free to move. *Nature News* **550**, 29 (2017).

296. Kogut, B. & Macpherson, J. M. The mobility of economists and the diffusion of policy ideas: The influence of economics on national policies. *Research Policy* **40**, 1307–1320 (2011).
297. Kaiser, U., Kongsted, H. C., Laursen, K. & Ejsing, A.-K. Experience matters: The role of academic scientist mobility for industrial innovation. *Strategic Management Journal* **39**, 1935–1958 (2018).
298. Chellaraj, G., Maskus, K. E. & Mattoo, A. The Contribution of International Graduate Students to US Innovation. *Review of International Economics* **16**, 444–462 (2008).
299. Aman, V. Transfer of knowledge through international scientific mobility: Introduction of a network-based bibliometric approach to study different knowledge types. *Quantitative Science Studies* **1**, 565–581 (2020).
300. Kurka, B., Trippl, M. & Maier, G. *Understanding Scientific Mobility: Characteristics, Location Decisions, and Knowledge Circulation. A Case Study of Internationally Mobile Austrian Scientists and Researchers* DYNREG30 (Economic and Social Research Institute (ESRI), 2008).
301. Liu, X., Wright, M., Filatotchev, I., Dai, O. & Lu, J. Human mobility and international knowledge spillovers: evidence from high-tech small and medium enterprises in an emerging market. *Strategic Entrepreneurship Journal* **4**, 340–355 (2010).
302. Abbott, A. *Department and Discipline: Chicago Sociology at One Hundred* 1 edition (University of Chicago Press, Chicago, IL, 1999).
303. Miguelez, E. & Noumedem Temgoua, C. Inventor migration and knowledge flows: A two-way communication channel? *Research Policy. STEM migration, research, and innovation* **49**, 103914 (2020).
304. Agrawal, A., Cockburn, I. & McHale, J. Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships. *Journal of Economic Geography* **6**, 571–591 (2006).

305. Head, K., Li, Y. A. & Minondo, A. Geography, Ties, and Knowledge Flows: Evidence from Citations in Mathematics. *The Review of Economics and Statistics* **101**, 713–727 (2018).
306. Box, S. & Barsi, E. *The Global Competition for Talent: Mobility of the Highly Skilled* (OECD, 2008).
307. Moed, H. F. & Halevi, G. A Bibliometric Approach to Tracking International Scientific Migration. *Scientometrics* **101**, 1987–2001 (2014).
308. Stephan, P., Scellato, G. & Franzoni, C. International Competition for PhDs and Postdoctoral Scholars: What Does (and Does Not) Matter. *Innovation Policy and the Economy* **15**, 73–113 (2015).
309. Børning, P., Flanagan, K., Gagliardi, D., Kaloudis, A. & Karakasidou, A. International mobility: Findings from a survey of researchers in the EU. *Science and Public Policy* **42**, 811–826 (2015).
310. Allison, P. D. & Long, J. S. Interuniversity Mobility of Academic Scientists. *American Sociological Review* **52**, 643–652 (1987).
311. Azoulay, P., Ganguli, I. & Graff Zivin, J. The mobility of elite life scientists: Professional and personal determinants. *Research Policy* **46**, 573–590 (2017).
312. Kato, M. & Ando, A. National ties of international scientific collaboration and researcher mobility found in Nature and Science. *Scientometrics* **110**, 673–694 (2017).
313. Hunter, R. S., Oswald, A. J. & Charlton, B. G. The Elite Brain Drain*. *The Economic Journal* **119**, F231–F251 (2009).
314. Baruffaldi, S. H. & Landoni, P. Return mobility and scientific productivity of researchers working abroad: The role of home country linkages. *Research Policy* **41**, 1655–1665 (2012).
315. Bauder, H. The International Mobility of Academics: A Labour Market Perspective. *International Migration* **53**, 83–96 (2015).

316. Ackers, L. Internationalisation, Mobility and Metrics: A New Form of Indirect Discrimination? *Minerva* **46**, 411–435 (2008).
317. Markova, Y. V., Shmatko, N. A. & Katchanov, Y. L. Synchronous international scientific mobility in the space of affiliations: evidence from Russia. *SpringerPlus* **5** (2016).
318. Robinson-Garcia, N. *et al.* The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics* **13**, 50–63 (2019).
319. Moed, H. F., Aisati, M. & Plume, A. Studying scientific migration in Scopus. *Scientometrics* **94**, 929–942 (2013).
320. Brandi, M. C., Avveduto, S. & Cerbara, L. *The reasons of scientists mobility: results from the comparison of outgoing and ingoing fluxes of researchers in Italy* 44 (AlmaLaurea Inter-University Consortium, 2011).
321. Kerr, W. America, don't throw global talent away. *Nature* **563**, 445–445 (2018).
322. Spier, R. The history of the peer-review process. *Trends in Biotechnology* **20**, 357–358 (2002).
323. Nature will publish peer review reports as a trial. *Nature* **578**, 8–8 (2020).
324. Ó Faoleán, G. *Frontiers Collaborative Peer Review: criteria to accept and reject manuscripts* Science & research news — Frontiers.
325. King, S. R. Peer Review: Consultative review is worth the wait. *eLife* **6**, e32012 (2017).
326. Lamont, M. *How Professors Think: Inside the Curious World of Academic Judgment* (Harvard University Press, Cambridge, 2009).
327. Siler, K. & Strang, D. Peer Review and Scholarly Originality: Let 1,000 Flowers Bloom, but Don't Step on Any. *Science, Technology, & Human Values* **42**, 29–61 (2017).
328. Cowley, S. J. How peer-review constrains cognition: on the frontline in the knowledge sector. *Frontiers in Psychology* **6** (2015).

329. Haffar, S., Bazerbachi, F. & Murad, M. H. Peer Review Bias: A Critical Review. *Mayo Clinic Proceedings* **94**, 670–676 (2019).
330. Gieryn, T. F. Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists. *American Sociological Review* **48**, 781–795 (1983).
331. Garfield, E. Citation analysis as a tool in journal evaluation. *Science (New York, N.Y.)* **178**, 471–479 (1972).
332. Leydesdorff, L. & Shin, J. C. How to Evaluate Universities in Terms of Their Relative Citation Impacts: Fractional Counting of Citations and the Normalization of Differences Among Disciplines. *J. Am. Soc. Inf. Sci. Technol.* **62**, 1146–1155 (2011).
333. Bogt, H. J. t. & Scapens, R. W. Performance Management in Universities: Effects of the Transition to More Quantitative Measurement Systems. *European Accounting Review* **21**, 451–497 (2012).
334. *Statement on Student Evaluations of Teaching* (American Sociological Association, 2019).
335. Zabaleta, F. The use and misuse of student evaluations of teaching. *Teaching in Higher Education* **12**, 55–76 (2007).
336. Fire, M. & Guestrin, C. Over-optimization of academic publishing metrics: observing Goodhart's Law in action. *GigaScience* **8** (giz053 2019).
337. Wilhite, A. W. & Fong, E. A. Coercive Citation in Academic Publishing. *Science* **335**, 542–543 (2012).
338. Fister, I. J., Fister, I. & Perc, M. Toward the Discovery of Citation Cartels in Citation Networks. *Frontiers in Physics* **4** (2016).
339. Luther, F. Publication ethics and scientific misconduct: the role of authors. *Journal of Orthodontics* **35**, 1–4 (2008).

340. Wang, J., Veugelers, R. & Stephan, P. *Bias against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators* Working Paper 22180 (National Bureau of Economic Research, 2016).
341. scraper, I. a. *How Northeastern University Gamed the College Rankings* Boston Magazine.
342. Gingras, Y. *How to boost your university up the rankings - University World News*
343. Vertesi, J. *Seeing Like a Rover* (The University of Chicago Press, Chicago, 2014).
344. O'Reilly, J. *The Technocratic Antarctic: An Ethnography of Scientific Expertise and Environmental Governance* 1st edition (Cornell University Press, Ithaca, 2017).
345. Chebib, J., Jackson, B. C., López-Cortegano, E., Tautz, D. & Keightley, P. D. Inbred lab mice are not isogenic: genetic variation within inbred strains used to infer the mutation rate per nucleotide site. *Heredity* **126**, 107–116 (2021).
346. Fahey, J. R., Katoh, H., Malcolm, R. & Perez, A. V. The case for genetic monitoring of mice and rats used in biomedical research. *Mammalian Genome* **24**, 89–94 (2013).
347. Tannen, D. Agonism in academic discourse. *Journal of Pragmatics. Negation and Disagreement* **34**, 1651–1669 (2002).
348. Kerr, S. P. & Kerr, W. R. *Economic Impacts of Immigration: A Survey* Working Paper 16736 (National Bureau of Economic Research, 2011).
349. Moser, P., Voena, A. & Waldinger, F. German Jewish Émigrés and US Invention. *American Economic Review* **104**, 3222–3255 (2014).
350. Chinchilla-Rodríguez, Z., Bu, Y., Robinson-García, N., Costas, R. & Sugimoto, C. R. Travel bans and scientific mobility: utility of asymmetry and affinity indexes to inform science policy. *Scientometrics* **116**, 569–590 (2018).
351. Salter, A. J. & Martin, B. R. The economic benefits of publicly funded basic research: a critical review. *Research Policy* **30**, 509–532 (2001).

352. Hatemi-J, A., Ajmi, A. N., El Montasser, G., Inglesi-Lotz, R. & Gupta, R. Research output and economic growth in G7 countries: new evidence from asymmetric panel causality testing. *Applied Economics* **48**, 2301–2308 (2016).
353. Inglesi-Lotz, R., Chang, T. & Gupta, R. Causality between research output and economic growth in BRICS. *Quality & Quantity* **49**, 167–176 (2015).
354. Inglesi-Lotz, R. & Pouris, A. The influence of scientific research output of academics on economic growth in South Africa: an autoregressive distributed lag (ARDL) application. *Scientometrics* **95**, 129–139 (2013).
355. Inglesi-Lotz, R., Balcilar, M. & Gupta, R. Time-varying causality between research output and economic growth in US. *Scientometrics* **100**, 203–216 (2014).
356. Iaria, A., Schwarz, C. & Waldinger, F. Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science*. *The Quarterly Journal of Economics* **133**, 927–991 (2018).
357. Arora, A., Belenzon, S. & Suh, J. *Science and the Market for Technology* w28534 (National Bureau of Economic Research, 2021).
358. Horbach, S. P. J. M. Pandemic Publishing: Medical journals drastically speed up their publication process for Covid-19. *bioRxiv*, 2020.04.18.045963 (2020).
359. Kwon, D. How swamped preprint servers are blocking bad coronavirus research. *Nature* **581**, 130–131 (2020).
360. Oreskes, N. & Conway, E. M. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Climate Change* Reprint edition (Bloomsbury Publishing, New York, NY, 2011).
361. Oreskes, N. The Scientific Consensus on Climate Change. *Science* **306**, 1686–1686 (2004).

362. Levy, K. E. & Johns, D. M. When open data is a Trojan Horse: The weaponization of transparency in science and governance. *Big Data & Society* **3**, 2053951715621568 (2016).
363. Wei, T. *et al.* Do scientists trace hot topics? *Scientific Reports* **3**, 1–5 (2013).
364. Tabakovic, H. & Wollmann, T. G. The impact of money on science: Evidence from unexpected NCAA football outcomes. *Journal of Public Economics* **178**, 104066 (2019).
365. Kennefick, D. Einstein Versus the Physical Review. *Physics Today* **58**, 43–48 (2005).
366. Huisman, J. & Smits, J. Duration and quality of the peer review process: the author's perspective. *Scientometrics* **113**, 633–650 (2017).
367. Smith, R. Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine* **99**, 178–182 (2006).
368. Lee, C. J., Sugimoto, C. R., Zhang, G. & Cronin, B. Bias in peer review. *Journal of the American Society for Information Science and Technology* **64**, 2–17 (2013).
369. Walker, R., Barros, B., Conejo, R., Neumann, K. & Telefont, M. Personal attributes of authors and reviewers, social bias and the outcomes of peer review: a case study. *F1000Research* (2015).
370. Pinholster, G. Journals and funders confront implicit bias in peer review. *Science* **352**, 1067–1068 (2016).
371. Kaatz, A., Gutierrez, B. & Carnes, M. Threats to objectivity in peer review: the case of gender. *Trends in pharmacological sciences* **35**, 371–373 (2014).
372. *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2017 Special Report* NSF 17-310 (National Science Foundation, National Center for Science and Engineering Statistics, Arlington, VA, 2017).
373. *Gender in the Global Research Landscape* (Elsevier, 2017).

374. Larivière, V. & Sugimoto, C. R. *The end of gender disparities in science? If only it were true...* CWTS.
375. Bendels, M. H. K., Müller, R., Brueggmann, D. & Groneberg, D. A. Gender disparities in high-quality research revealed by Nature Index journals. *PLOS ONE* **13**, e0189136 (2018).
376. Bernard, C. Editorial: Gender Bias in Publishing: Double-Blind Reviewing as a Solution? *eNeuro* **5**, ENEURO.0225–18.2018 (2018).
377. Nature's under-representation of women. *Nature* **558**, 344 (2018).
378. King, D. A. The scientific impact of nations. *Nature* **430**, 311 (2004).
379. Li, D. *Gender Bias in NIH Peer Review: Does it Exist and Does it Matter?* 2011.
380. Grant, J., Burden, S. & Breen, G. No evidence of sexism in peer review. *Nature* **390**, 438 (1997).
381. Gilbert, J. R., Williams, E. S. & Lundberg, G. D. Is there gender bias in JAMA's peer review process? *JAMA* **272**, 139–142 (1994).
382. Mutz, R., Bornmann, L. & Daniel, H.-D. Does Gender Matter in Grant Peer Review?: An Empirical Investigation Using the Example of the Austrian Science Fund. *Zeitschrift Fur Psychologie* **220**, 121–129 (2012).
383. Beck, R. & Halloin, V. Gender and research funding success: Case of the Belgian F.R.S.-FNRS. *Research Evaluation* **26**, 115–123 (2017).
384. Edwards, H. A., Schroeder, J. & Dugdale, H. L. Gender differences in authorships are not associated with publication bias in an evolutionary journal. *PLOS ONE* **13**, e0201725 (2018).
385. Coates, R., Sturgeon, B., Bohannan, J. & Pasini, E. Language and publication in Cardiovascular Research articles. *Cardiovascular Research* **53**, 279–285 (2002).

386. Primack, R. B., Ellwood, E., Miller-Rushing, A. J., Marrs, R. & Mulligan, A. Do gender, nationality, or academic age affect review decisions? An analysis of submissions to the journal Biological Conservation. *Biological Conservation* **142**, 2415–2418 (2009).
387. Berger, J., Fisek, H. F., Normal, R. Z. & Zelditch, N. *Status characteristics and social interaction* (Elsevier, New York, 1977).
388. Correll, S. J. & Ridgeway, C. L. in *Handbook of Social Psychology* 29–51 (Springer, Boston, MA, 2006).
389. Podolny, J. M. *Status Signals: A Sociological Study of Market Competition* (Princeton University Press, Princeton, N.J Woodstock, 2008).
390. Long, J. S. & Fox, M. F. Scientific Careers: Universalism and Particularism. *Annual Review of Sociology* **21**, 45–71 (1995).
391. Pfeffer, J., Leong, A. & Strehl, K. Paradigm Development and Particularism: Journal Publication in Three Scientific Disciplines. *Social Forces* **55**, 938–951 (1977).
392. Cole, S. The Hierarchy of the Sciences? *American Journal of Sociology* **89**, 111–139 (1983).
393. Jacobs, J. A. Gender and the Stratification of Colleges. *The Journal of Higher Education* **70**, 161–187 (1999).
394. Weeden, K. A., Thebaud, S. & Gelbgiser, D. Degrees of Difference: Gender Segregation of U.S. Doctorates by Field and Program Prestige. *Sociological Science* **4**, 123–150 (2017).
395. Travis, G. D. L. & Collins, H. M. New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System. *Science, Technology, & Human Values* **16**, 322–341 (1991).
396. Demarest, B., Freeman, G. & Sugimoto, C. R. The reviewer in the mirror: examining gendered and ethnicized notions of reciprocity in peer review. *Scientometrics* **101**, 717–735 (2014).

397. Bagues, M., Sylos-Labini, M. & Zinovyeva, N. Does the Gender Composition of Scientific Committees Matter? *American Economic Review* **107**, 1207–1238 (2017).
398. Ceci, S. J. & Williams, W. M. Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences* **108**, 3157–3162 (2011).
399. Ceci, S. J., Ginther, D. K., Kahn, S. & Williams, W. M. Women in Academic Science: A Changing Landscape. *Psychological Science in the Public Interest: A Journal of the American Psychological Society* **15**, 75–141 (2014).
400. Niederle, M. & Vesterlund, L. Gender and Competition. *Annual Review of Economics* **3**, 601–630 (2011).
401. May, R. M. The Scientific Wealth of Nations. *Science* **275**, 793–796 (1997).
402. Duszak, A. & Lewkowicz, J. Publishing academic texts in English: A Polish perspective. *Journal of English for Academic Purposes. English for Research Publication Purposes* **7**, 108–120 (2008).
403. Salager-Meyer, F. Scientific publishing in developing countries: Challenges for the future. *Journal of English for Academic Purposes. English for Research Publication Purposes* **7**, 121–132 (2008).
404. Yang, W. Policy: Boost basic research in China. *Nature News* **534**, 467 (2016).
405. Langfeldt, L. The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome. *Social Studies of Science* **31**, 820–841 (2001).
406. Schekman, R., Watt, F. & Weigel, D. Scientific Publishing: The eLife approach to peer review. *eLife* **2**, e00799 (2013).
407. Costas, R. & Bordons, M. Do age and professional rank influence the order of authorship in scientific publications? Some evidence from a micro-level perspective. *Scientometrics* **88**, 145–161 (2011).

408. Baerlocher, M. O., Newton, M., Gautam, T., Tomlinson, G. & Detsky, A. S. The meaning of author order in medical research. *Journal of Investigative Medicine: The Official Publication of the American Federation for Clinical Research* **55**, 174–180 (2007).
409. Tscharntke, T., Hochberg, M. E., Rand, T. A., Resh, V. H. & Krauss, J. Author Sequence and Credit for Contributions in Multiauthored Publications. *PLOS Biology* **5**, e18 (2007).
410. Winkler, W. E. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage* (1990).
411. Harvard WorldMap
412. Kliewer, M. A. *et al.* Peer review at the American Journal of Roentgenology: how reviewer and manuscript characteristics affected editorial decisions on 196 major papers. *AJR. American journal of roentgenology* **183**, 1545–1550 (2004).
413. Amrein, K., Langmann, A., Fahrleitner-Pammer, A., Pieber, T. R. & Zollner-Schwetz, I. Women Underrepresented on Editorial Boards of 60 Major Medical Journals. *Gender Medicine* **8**, 378–387 (2011).
414. Cho, A. H. *et al.* Women are underrepresented on the editorial boards of journals in environmental biology and natural resource management. *PeerJ* **2**, e542 (2014).
415. Metz, I. & Harzing, A.-W. Gender Diversity in Editorial Boards of Management Journals. *Academy of Management Learning & Education* **8**, 540–557 (2009).
416. Metz, I. & Harzing, A.-W. An update of gender diversity in editorial boards: a longitudinal study of management journals. *Personnel Review* **41**, 283–300 (2012).
417. Morton, M. J. & Sonnad, S. S. Women on professional society and journal editorial boards. *Journal of the National Medical Association* **99**, 764–771 (2007).

418. Stegmaier, M., Palmer, B. & Assendelft, L. v. Getting on the Board: The Presence of Women in Political Science Journal Editorial Positions. *PS: Political Science & Politics* **44**, 799–804 (2011).
419. Topaz, C. M. & Sen, S. Gender Representation on Journal Editorial Boards in the Mathematical Sciences. *PLOS ONE* **11**, e0161357 (2016).
420. Addis, E. & Villa, P. The Editorial Boards of Italian Economics Journals: Women, Gender, and Social Networking. *Feminist Economics* **9**, 75–91 (2003).
421. Avin, C. *et al.* *Homophily and the Glass Ceiling Effect in Social Networks* in *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science* (ACM, New York, NY, USA, 2015), 41–50.
422. Szell, M. & Thurner, S. How women organize social networks different from men. *Scientific Reports* **3**, 1214 (2013).
423. Lerback, J. & Hanson, B. Journals invite too few women to referee. *Nature News* **541**, 455 (2017).
424. Tite, L. & Schroter, S. Why do peer reviewers decline to review? A survey. *Journal of Epidemiology & Community Health* **61**, 9–12 (2007).
425. O'Brien, K. R. & Hapgood, K. P. The academic jungle: ecosystem modelling reveals why women are driven out of research. *Oikos* **121**, 999–1004 (2012).
426. Valkonen, L. & Brooks, J. Gender balance in Cortex acceptance rates. *Cortex* **47**, 763–770 (2011).
427. Fox, C. W., Burns, C. S., Meyer, J. A. & Thompson, K. Editor and reviewer gender influence the peer review process but not peer review outcomes at an ecology journal. *Functional Ecology* **30**, 140–153 (2015).

428. Borsuk, R. M. *et al.* To Name or Not to Name: The Effect of Changing Author Gender on Peer Review. *BioScience* **59**, 985–989 (2009).
429. Kassis, T. How do research faculty in the biosciences evaluate paper authorship criteria? *PLOS ONE* **12**, e0183632 (2017).
430. Tregenza, T. Gender bias in the refereeing process? *Trends in Ecology & Evolution* **17**, 349–350 (2002).
431. Women in neuroscience: a numbers game. *Nature Neuroscience* **9**, 853 (2006).
432. Gannon, F., Quirk, S. & Guest, S. Searching for discrimination: Are women treated fairly in the EMBO postdoctoral fellowship scheme? *EMBO reports* **2**, 655–657 (2001).
433. Lee, R. v. d. & Ellemers, N. Gender contributes to personal research funding success in The Netherlands. *Proceedings of the National Academy of Sciences* **112**, 12349–12353 (2015).
434. Shen, H. Inequality quantified: Mind the gender gap. *Nature* **495**, 22–24 (2013).
435. Pohlhaus, J. R., Jiang, H., Wagner, R. M., Schaffer, W. T. & Pinn, V. W. Sex Differences in Application, Success, and Funding Rates for Nih Extramural Programs. *Academic Medicine* **86**, 759–767 (2011).
436. Waisbren, S. E. *et al.* Gender differences in research grant applications and funding outcomes for medical school faculty. *Journal of Women's Health (2002)* **17**, 207–214 (2008).
437. Marsh, H. W., Jayasinghe, U. W. & Bond, N. W. Gender differences in peer reviews of grant applications: A substantive-methodological synergy in support of the null hypothesis model. *Journal of Informetrics* **5**, 167–180 (2011).
438. Larivière, V., Gingras, Y., Sugimoto, C. R. & Tsou, A. Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology* **66**, 1323–1332 (2015).

439. Bornmann, L. & Daniel, H.-D. Gatekeepers of science—Effects of external reviewers' attributes on the assessments of fellowship applications. *Journal of Informetrics* **1**, 83–91 (2007).
440. Zhang, X. Effect of reviewer's origin on peer review: China vs. non-China. *Learned Publishing* **25**, 265–270 (2012).
441. Lloyd, M. E. Gender factors in reviewer recommendations for manuscript publication. *Journal of Applied Behavior Analysis* **23**, 539–543 (1990).
442. Wing, D. A., Benner, R. S., Petersen, R., Newcomb, R. & Scott, J. R. Differences in editorial board reviewer behavior based on gender. *Journal of Women's Health (2002)* **19**, 1919–1923 (2010).
443. Petty, R. E., Fleming, M. A. & Fabrigar, L. R. The Review Process at PSPB: Correlates of Interreviewer Agreement and Manuscript Acceptance. *Personality and Social Psychology Bulletin* **25**, 188–203 (1999).
444. Gelman, A. & Stern, H. The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant. *The American Statistician* **60**, 328–331 (2006).
445. Adamo, S. A. Attrition of Women in the Biological Sciences: Workload, Motherhood, and Other Explanations Revisited. *BioScience* **63**, 43–48 (2013).
446. Ceci, S. J., Williams, W. M. & Barnett, S. M. Women's underrepresentation in science: sociocultural and biological considerations. *Psychological Bulletin* **135**, 218–261 (2009).
447. Ceci, S. J. & Williams, W. M. Sex Differences in Math-Intensive Fields. *Current Directions in Psychological Science* **19**, 275–279 (2010).
448. Xie, Y. & Shauman, K. A. Sex Differences in Research Productivity: New Evidence about an Old Puzzle. *American Sociological Review* **63**, 847–870 (1998).

449. Page, S. E. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. (Princeton University Press, Princeton, 2008).
450. Campbell, L. G., Mehtani, S., Dozier, M. E. & Rinehart, J. Gender-Heterogeneous Working Groups Produce Higher Quality Science. *PLOS ONE* **8**, e79147 (2013).
451. Giordan, M., Csikasz-Nagy, A., Collings, A. M. & Vaggi, F. The effects of an editor serving as one of the reviewers during the peer-review process. *F1000Research* **5**, 683 (2016).
452. PEERE policy on data sharing on peer review (PEERE, 2017).
453. Squazzoni, F., Grimaldo, F. & Marušić, A. Publishing: Journals could share peer-review data. *Nature* **546**, 352 (2017).
454. Nature journals offer double-blind review. *Nature News* **518**, 274 (2015).
455. McGillivray, B. & De Ranieri, E. Uptake and outcome of manuscripts in Nature journals by review model and author characteristics. *arXiv:1802.02188 [cs]* (2018).
456. Ware, M. Peer Review in Scholarly Journals: Perspective of the Scholarly Community - Results from an International Study. *Inf. Serv. Use* **28**, 109–112 (2008).
457. Kmietowicz, Z. Double blind peer reviews are fairer and more objective, say academics. *BMJ : British Medical Journal* **336**, 241 (2008).
458. Blank, R. M. The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review. *The American Economic Review* **81**, 1041–1067 (1991).
459. Bravo, G., Grimaldo, F., López-Iñesta, E., Mehmani, B. & Squazzoni, F. The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nature Communications* **10**, 322 (2019).
460. Pulverer, B. A transparent black box. *The EMBO Journal* **29**, 3891–3892 (2010).

461. Merchant, S. & Eckardt, N. A. The Plant Cell Begins Opt-in Publishing of Peer Review Reports. *The Plant Cell* **28**, 2343 (2016).
462. Pourquié, O. & Brown, K. Future developments: your thoughts and our plans. *Development (Cambridge, England)* **143**, 1–2 (2016).
463. Rodgers, P. Peer Review: Decisions, decisions. *eLife* **6**, e32011 (2017).
464. Abdill, R. J. & Blekhman, R. Tracking the popularity and outcomes of all bioRxiv preprints. *bioRxiv*, 515643 (2019).
465. Campbell, F. M. National bias: a comparison of citation practices by health professionals. *Bulletin of the Medical Library Association* **78**, 376–382 (1990).
466. Cramer, K. M. & Alexitch, L. R. Student Evaluations of College Professors: Identifying Sources of Bias. *Canadian Journal of Higher Education* **30**, 143–164 (2000).
467. Joye, S. W. & Wilson, J. H. Professor Age and Gender Affect Student Perceptions and Grades. *Journal of the Scholarship of Teaching and Learning* **15**, 126–138 (2015).
468. MacNell, L., Driscoll, A. & Hunt, A. N. What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innov High Educ* **40**, 291–303 (2015).
469. Reid, L. D. The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.Com. *Journal of Diversity in Higher Education* **3**, 137–152 (2010).
470. Boyer, E. L. *Scholarship Reconsidered: Priorities of the Professoriate* 1 edition (Jossey-Bass, Princeton, NJ, 1997).
471. Boring, A., Ottoboni, K. & Stark, P. Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness. *ScienceOpen Research* (2016).
472. Miles, P. & House, D. The Tail Wagging the Dog; An Overdue Examination of Student Teaching Evaluations. *International Journal of Higher Education* **4**, 116–126 (2015).

473. Mengel, F., Sauermann, J. & Zölitz, U. Gender bias in teaching evaluations. *Journal of the European Economic Association* (2017).
474. Wilson, J. H., Beyer, D. & Monteiro, H. Professor Age Affects Student Ratings: Halo Effect for Younger Teachers. *College Teaching* **62**, 20–24 (2014).
475. Sohr-Preston, S. L., Boswell, S. S., McCaleb, K. & Robertson, D. Professor Gender, Age, and "Hotness" in Influencing College Students' Generation and Interpretation of Professor Ratings. *Higher Learning Research Communications* **6** (2016).
476. Stonebraker, R. J. & Stone, G. S. Too Old to Teach? The Effect of Age on College and University Professors. *Res High Educ* **56**, 793–812 (2015).
477. Uttl, B. & Smibert, D. Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. *PeerJ* **5**, e3299 (2017).
478. Boysen, G. A., Kelly, T. J., Raesly, H. N. & Casner, R. W. The (mis)interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education* **39**, 641–656 (2014).
479. Adams, M. J. D. & Umbach, P. D. Nonresponse and Online Student Evaluations of Teaching: Understanding the Influence of Salience, Fatigue, and Academic Environments. *Research in Higher Education* **53**, 576–591 (2012).
480. Gruber, T. *et al.* Investigating the Influence of Professor Characteristics on Student Satisfaction and Dissatisfaction: A Comparative Study. *Journal of Marketing Education* **34**, 165–178 (2012).
481. Sheehan, D. S. On the Invalidity of Student Ratings for Administrative Personnel Decisions. *The Journal of Higher Education* **46**, 687–700 (1975).
482. Bunge, N. Students Evaluating Teachers Doesn't Just Hurt Teachers. It Hurts Students. *The Chronicle of Higher Education* (2018).

483. Falkoff, M. Why We Must Stop Relying on Student Ratings of Teaching. *The Chronicle of Higher Education* (2018).
484. Flaherty, C. Most institutions say they value teaching but how they assess it tells a different story. *Inside Higher Ed* (2018).
485. Habermas, J. & Blazek, J. R. The Idea of the University: Learning Processes. *New German Critique*, 3–22 (1987).
486. Neumann, R. Perceptions of the Teaching-Research Nexus: A Framework for Analysis. *Higher Education* **23**, 159–171 (1992).
487. Neumann, R. The Teaching-Research Nexus: Applying a Framework to University Students' Learning Experiences. *European Journal of Education* **29**, 323–338 (1994).
488. Turner, N., Wuetherick, B. & Healey, M. International perspectives on student awareness, experiences and perceptions of research: implications for academic developers in implementing research-based teaching and learning. *International Journal for Academic Development* **13**, 199–211 (2008).
489. Brennan, L., Cusack, T., Delahunt, E., Kuznesof, S. & Donnelly, S. Academics' conceptualisations of the research-teaching nexus in a research-intensive Irish university: A dynamic framework for growth & development. *Learning and Instruction* **60**, 301–309 (2019).
490. Galbraith, C. S. & Merrill, G. B. Faculty Research Productivity and Standardized Student Learning Outcomes in a University Teaching Environment: A Bayesian Analysis of Relationships. *Studies in Higher Education* **37**, 469–480 (2012).
491. Taylor, J. The teaching:research nexus : a model for institutional management. *Higher Education* **54**, 867–884 (2007).

492. Carter, R. E. Faculty Scholarship Has a Profound Positive Association With Student Evaluations of Teaching—Except When It Doesn't. *Journal of Marketing Education* **38**, 18–36 (2016).
493. Courant, P. N. & Turner, S. *Faculty Deployment in Research Universities* Working Paper 23025 (National Bureau of Economic Research, 2017).
494. Coate, K., Barnett, R. & Williams, G. Relationships Between Teaching and Research in Higher Education in England. *Higher Education Quarterly* **55**, 158–174 (2001).
495. Gomez-Mejia, L. R. & Balkin, D. B. Determinants of Faculty Pay: An Agency Theory Perspective. *The Academy of Management Journal* **35**, 921–955 (1992).
496. Hattie, J. & Marsh, H. W. The Relationship Between Research and Teaching: A Meta-analysis. *Review of Educational Research* **66**, 507–542 (1996).
497. Miller, J. D. *How To Fight RateMyProfessors.com — Inside Higher Ed* Inside Higher Ed.
498. Davison, E. & Price, J. How do we rate? An evaluation of online student evaluations. *Assessment & Evaluation in Higher Education* **34**, 51–65 (2009).
499. Gregory, K. M. How Undergraduates Perceive Their Professors: A Corpus Analysis of Rate My Professor. *Journal of Educational Technology Systems* **40**, 169–193 (2011).
500. Kindred, J. & Mohammed, S. N. "He Will Crush You Like an Academic Ninja!" Exploring Teacher Ratings on Ratemyprofessors.com. *Journal of Computer-Mediated Communication* **10**, 00–00 (2005).
501. Coladarci, T. & Kornfield, I. RateMyProfessors. com versus formal in-class student evaluations of teaching. *Practical Assessment & Research Evaluation* **12** (2007).
502. Silva, K. M. *et al.* Rate My Professor: Online Evaluations of Psychology Instructors. *Teaching of Psychology* **35**, 71–80 (2008).
503. *About RateMyProfessors.com* Rate My Professor.

504. Winkler, W. E. *The State of Record Linkage and Current Research Problems* (Statistical Research Division, U.S. Bureau of the Census, Washington, DC, 1999).
505. Winkler, W. E. *Overview of Record Linkage and Current Research Directions #2006-2* (Statistical Research Division, U.S. Bureau of the Census, Washington, DC, 2006).
506. Jaro, M. A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* **84**, 414–420 (1989).
507. Rosen, A. S. Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors.com data. *Assessment & Evaluation in Higher Education* **43**, 31–44 (2018).
508. DeMeis, D. K. & Turner, R. R. Effects of students' race, physical attractiveness, and dialect on teachers' evaluations. *Contemporary Educational Psychology* **3**, 77–86 (1978).
509. Tsou, A., Bowman, T. D., Sugimoto, T., Lariviere, V. & Sugimoto, C. R. Self-presentation in scholarly profiles: Characteristics of images and perceptions of professionalism and attractiveness on academic social networking sites. *First Monday* **21** (2016).
510. Feeley, T. Evidence of Halo Effects in Student Evaluations of Communication Instruction. *Communication Education* **51**, 225–236 (2002).
511. Lewis, M. B. Who is the fairest of them all? Race, attractiveness and skin color sexual dimorphism. *Personality and Individual Differences* **50**, 159–162 (2011).
512. Schmit, B. *Gendered Language in Teaching Evaluations*
513. Chávez, K. & Mitchell, K. M. W. Exploring Bias in Student Evaluations: Gender, Race, and Ethnicity. *PS: Political Science & Politics*, 1–5 (2019).
514. Kavas, A. & Kavas, A. An Exploratory Study of Undergraduate College Students' Perceptions and Attitudes toward Foreign Accented Faculty. *College Student Journal* **42**, 879–890 (2008).

515. Gill, M. M. Accent and stereotypes: Their effect on perceptions of teachers and lecture comprehension. *Journal of Applied Communication Research* **22**, 348–361 (1994).
516. Subtirelu, N. C. “She does have an accent but...”: Race and language ideology in students’ evaluations of mathematics instructors on RateMyProfessors.com. *Language in Society* **44**, 35–62 (2015).
517. Figlio, D. N., Schapiro, M. O. & Soter, K. B. *Are Tenure Track Professors Better Teachers?* Working Paper 19406 (National Bureau of Economic Research, 2013).
518. Centra, J. A. Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Course Work? *Research in Higher Education* **44**, 495–518 (2003).
519. Otto, J., Jr, D. A. S. & Ross, D. N. Does ratemyprofessor.com really rate my professor? *Assessment & Evaluation in Higher Education* **33**, 355–368 (2008).
520. Clayson, D. E. & Haley, D. A. Student Evaluations in Marketing: What is Actually being Measured? *Journal of Marketing Education* **12**, 9–17 (1990).
521. Marsh, H. W. The Influence of Student, Course, and Instructor Characteristics in Evaluations of University Teaching. *American Educational Research Journal* **17**, 219–237 (1980).
522. Darby, J. A. The effects of the elective or required status of courses on student evaluations. *Journal of Vocational Education & Training* **58**, 19–29 (2006).
523. Feldman, K. A. Class size and college students’ evaluations of teachers and courses: A closer look. *Research in Higher Education* **21**, 45–116 (1984).
524. Mateo, M. A. & Fernandez, J. Incidence of Class Size on the Evaluation of University Teaching Quality. *Educational and Psychological Measurement* **56**, 771–778 (1996).
525. Rojstaczer, S. & Healy, C. Where A Is Ordinary: The Evolution of American College and University Grading, 1940–2009. *Teachers College Record* **114**, 1–23 (2012).

526. Ewing, A. M. Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review* **31**, 141–154 (2012).
527. Hattie, J. & Marsh, H. *One Journey to Unravel the Relationship between Research and Teaching - Semantic Scholar* in. Research and Teaching: Closing the Divide? An international Colloquium (Hampshire, UK, 2004).
528. Figlio, D. & Schapiro, M. O. *Are great teachers poor scholars?* (Brookings Institute, 2017).
529. Euwals, R. & Ward, M. E. What matters most: teaching or research? Empirical evidence on the remuneration of British academics. *Applied Economics* **37**, 1655–1672 (2005).
530. Gottlieb, E. E. & Keith, B. The academic research-teaching nexus in eight advanced-industrialized countries. *Higher Education* **34**, 397–419 (1997).
531. Arnold, I. J. M. Course Level and the Relationship between Research Productivity and Teaching Effectiveness. *The Journal of Economic Education* **39**, 307–321 (2008).
532. Stack, S. Research Productivity and Student Evaluation of Teaching in Social Science Classes: A Research Note. *Research in Higher Education* **44**, 539–556 (2003).
533. Brew, A. & Boud, D. Teaching and research: Establishing the vital link with learning. *High Educ* **29**, 261–273 (1995).
534. Hornstein, H. A. Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education* **4** (ed Law, H. F. E.) 1304016 (2017).
535. Smeby, J.-C. Knowledge Production and Knowledge Transmission. The interaction between research and teaching at universities. *Teaching in Higher Education* **3**, 5–20 (1998).
536. Bak, H.-J. & Kim, D. H. Too much Emphasis on Research? An Empirical Examination of the Relationship Between Research and Teaching in Multitasking Environments. *Res High Educ* **56**, 843–860 (2015).

537. *The Annual Report on the Economic Status of the Profession, 2018–19* (American Association of University Professors, Washington D.C., U.S.A., 2019).
538. *National Center for Science and Engineering Statistics, Scientists and Engineers Statistical Data System (SESTAT)* (National Science Foundation, 2016).
539. Friedman, J. 10 Universities Where TAs Teach the Most Classes. *US News & World Report* (2017).
540. Murray, D. S. The precarious new faculty majority: communication and instruction research and contingent labor in higher education. *Communication Education* **68**, 235–245 (2019).
541. Fong, C., Dillard, J. & Hatcher, M. Teaching Self-Efficacy of Graduate Student Instructors: Exploring Faculty Motivation, Perceptions of Autonomy Support, and Undergraduate Student Engagement. *International Journal of Educational Research* **98** (2019).
542. Patridge, E. V., Barthelemy, R. S. & Rankin, S. R. FACTORS IMPACTING THE ACADEMIC CLIMATE FOR LGBQ STEM FACULTY. *Journal of Women and Minorities in Science and Engineering* **20** (2014).
543. Nielsen, E.-J. & Alderson, K. G. Lesbian and Queer Women Professors Disclosing in the Classroom: An Act of Authenticity. *The Counseling Psychologist* (2014).
544. Dilley, P. LGBTQ Research in Higher Education: A Review of Journal Articles, 2000–2003. *Journal of Gay & Lesbian Issues in Education* **2**, 105–115 (2004).
545. Kitcher, P. *Advancement of Science: Science Without Legend, Objectivity Without Illusions* (Oxford University Press, New York, 1995).
546. Popper, K. *Conjectures and Refutations: The Growth of Scientific Knowledge* 2nd edition (Routledge, London ; New York, 1963).
547. Sarewitz, D. The voice of science: let's agree to disagree. *Nature* **478**, 7–7 (2011).
548. The power of disagreement. *Nature Methods* **13**, 185–185 (2016).

549. Comte, A. *The Positive Philosophy of Auguste Comte* (Calvin Blanchard, 1856).
550. Collins, H. *Gravity's Kiss: The Detection of Gravitational Waves* 1st edition (The MIT Press, Cambridge, Massachusetts, 2017).
551. Castelvecchi, D. Mystery over Universe's expansion deepens with fresh data. *Nature* **583**, 500–501 (2020).
552. Bruggeman, J., Traag, V. A. & Uitermark, J. Detecting Communities through Network Data. *American Sociological Review* **77**, 1050–1063 (2012).
553. Shwed, U. & Bearman, P. S. Symmetry Is Beautiful. *American Sociological Review* **77**, 1064–1069 (2012).
554. Catalini, C., Lacetera, N. & Oettl, A. The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences* **112**, 13823–13826 (2015).
555. Chen, C., Song, M. & Heo, G. E. A Scalable and Adaptive Method for Finding Semantically Equivalent Cue Words of Uncertainty. *Journal of Informetrics* **12**, 158–180 (2018).
556. Nicholson, J. M. *et al.* scite: a smart citation index that displays the context of citations and classifies their intent using deep learning. *bioRxiv*, 2021.03.15.435418 (2021).
557. Bertin, M., Atanassova, I., Sugimoto, C. R. & Lariviere, V. The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. *Scientometrics* **109**, 1417–1434 (2016).
558. Balietti, S., Mäs, M. & Helbing, D. On Disciplinary Fragmentation and Scientific Progress. *PLOS ONE* **10**, e0118747 (2015).
559. Hargens, L. L. Scholarly Consensus and Journal Rejection Rates. *American Sociological Review* **53**, 139–51 (1988).
560. Cole, S., Simon, G. & Cole, J. R. Do Journal Rejection Rates Index Consensus? *American Sociological Review* **53**, 152–156 (1988).

561. Fanelli, D. “Positive” Results Increase Down the Hierarchy of the Sciences. *PLOS ONE* **5**, e10068 (2010).
562. Nicolaisen, J. & Frandsen, T. F. Consensus formation in science modeled by aggregated bibliographic coupling. *Journal of Informetrics* **6**, 276–284 (2012).
563. Evans, J. H. Consensus and knowledge production in an academic field. *Poetics* **35**, 1–21 (2007).
564. Szarvas, G., Vincze, V., Farkas, R., Móra, G. & Gurevych, I. Cross-Genre and Cross-Domain Detection of Semantic Uncertainty. *Computational Linguistics* **38**, 335–367 (2012).
565. Yang, L. Y., Yue, T., Ding, J. L. & Han, T. A comparison of disciplinary structure in science between the G7 and the BRIC countries by bibliometric methods. *Scientometrics* **93**, 497–516 (2012).
566. Hyland, K. *Hedging in Scientific Research Articles* (John Benjamins Publishing Company, Amsterdam, 1998).
567. Small, H. Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *Journal of Informetrics* **12**, 461–480 (2018).
568. Small, H., Boyack, K. W. & Klavans, R. Citations and certainty: a new interpretation of citation counts. *Scientometrics* **118**, 1079–1092 (2019).
569. Bornmann, L., Wray, K. B. & Haunschild, R. Citation concept analysis (CCA): a new form of citation analysis revealing the usefulness of concepts for other researchers illustrated by exemplary case studies including classic books by Thomas S. Kuhn and Karl R. Popper. *Scientometrics* **122**, 1051–1074 (2020).
570. Moravcsik, M. J. & Murugesan, P. Some Results on the Function and Quality of Citations. *Social Studies of Science* **5**, 86–92 (1975).

571. Teufel, S., Siddharthan, A. & Tidhar, D. *An Annotation Scheme for Citation Function* in *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2006), 80–87.
572. Teufel, S., Siddharthan, A. & Tidhar, D. *Automatic Classification of Citation Function* in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2006), 103–110.
573. Bertin, M., Atanassova, I., Gingras, Y. & Larivière, V. The invariant distribution of references in scientific articles. *Journal of the Association for Information Science and Technology* **67**, 164–177 (2016).
574. Valenzuela, M., Ha, V. & Etzioni, O. *Identifying Meaningful Citations* in *AAAI Workshop: Scholarly Big Data* (2015).
575. Waltman, L. An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics* **6**, 700–711 (2012).
576. Van Eck, N. J. & Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **84**, 523–538 (2010).
577. Dieckmann, N. F. & Johnson, B. B. Why do scientists disagree? Explaining and improving measures of the perceived causes of scientific disputes. *PLOS ONE* **14**, e0211269 (2019).
578. Radicchi, F. In science “there is no bad publicity”: Papers criticized in comments have high scientific impact. *Scientific Reports* **2** (2012).
579. Latour, B. *Science in Action: How to Follow Scientists and Engineers Through Society* Reprint edition (Harvard University Press, Cambridge, Mass, 1988).
580. Fanelli, D. & Glänzel, W. Bibliometric Evidence for a Hierarchy of the Sciences. *PLOS ONE* **8**, e66938 (2013).

581. Biglan, A. The characteristics of subject matter in different academic areas. *Journal of Applied Psychology* **57**, 195–203 (1973).
582. Smolin, L. *The Trouble With Physics: The Rise of String Theory, The Fall of a Science, and What Comes Next by Lee Smolin* (Mariner Books, 2007).
583. Baron, M. G., Norman, D. B. & Barrett, P. M. A new hypothesis of dinosaur relationships and early dinosaur evolution. *Nature* **543**, 501–506 (2017).
584. Langer, M. C. *et al.* Untangling the dinosaur family tree. *Nature* **551**, E1–E3 (2017).
585. Rife, S. C., Rosati, D. & Nicholson, J. M. scite: The next generation of citations. *Learned Publishing* (2021).
586. Small, H., Tseng, H. & Patek, M. Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics* **11**, 46–62 (2017).
587. Armano, G. & Javarone, M. A. The Beneficial Role of Mobility for the Emergence of Innovation. *Scientific Reports* **7** (2017).
588. NW, 1. L. S., 800Washington, S. & Inquiries, D. 2.-4.-4.
bibinitperiod M.-8.-8.
bibinitperiod F.-4.-4.
bibinitperiod M. *Global Migration Map: Origins and Destinations, 1990-2017* Pew Research Center's Global Attitudes Project.
589. Zipf, G. K. The P1 P2/D Hypothesis: On the Intercity Movement of Persons. *American Sociological Review* **11**, 677–686 (1946).
590. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100 (2012).
591. Brown, L. A., Odland, J. & Golledge, R. G. Migration, Functional Distance, and the Urban Hierarchy. *Economic Geography* **46**, 472–485 (1970).

592. Kim, J., Park, J. & Lee, W. Why do people move? Enhancing human mobility prediction using local functions based on public records and SNS data. *PLOS ONE* **13**, e0192698 (2018).
593. Deville, P. *et al.* Career on the Move: Geography, Stratification, and Scientific Impact. *Scientific Reports* **4**, 4770 (2014).
594. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
595. Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* **115**, E3635–E3644 (2018).
596. Kozlowski, A. C., Taddy, M. & Evans, J. A. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review* **84**, 905–949 (2019).
597. Nakandala, S., Ciampaglia, G. L., Su, N. M. & Ahn, Y.-Y. *Gendered Conversation in a Social Game-Streaming Platform* in. Eleventh International AAAI Conference on Web and Social Media. (Montreal, Canada, 2017), 10.
598. Hamilton, W. L., Leskovec, J. & Jurafsky, D. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *arXiv:1605.09096 [cs]* (2018).
599. Le, Q. & Mikolov, T. *Distributed Representations of Sentences and Documents* in *International Conference on Machine Learning* International Conference on Machine Learning (PMLR, 2014), 1188–1196.
600. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. *arXiv:1607.00653 [cs, stat]* (2016).
601. Linzhuo, L., Lingfei, W. & James, E. Social centralization and semantic collapse: Hyperbolic embeddings of networks and text. *Poetics*, 101428 (2020).

602. Liu, X., Liu, Y. & Li, X. *Exploring the context of locations for personalized location recommendations* in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (AAAI Press, New York, New York, USA, 2016), 1188–1194.
603. Feng, S., Cong, G., An, B. & Chee, Y. M. POI2Vec: Geographical Latent Representation for Predicting Future Visitors. *AAAI 2017*, 7 (2017).
604. Yao, Z., Fu, Y., Liu, B., Hu, W. & Xiong, H. Representing Urban Functions through Zone Embedding with Human Mobility Patterns, 3919–3925 (2018).
605. Cao, H., Xu, F., Sankaranarayanan, J., Li, Y. & Samet, H. Habit2vec: Trajectory Semantic Embedding for Living Pattern Recognition in Population. *IEEE Transactions on Mobile Computing* **19**, 1096–1108 (2020).
606. Crivellari, A. & Beinat, E. From Motion Activity to Geo-Embeddings: Generating and Exploring Vector Representations of Locations, Traces and Visitors through Large-Scale Mobility Data. *ISPRS International Journal of Geo-Information* **8**, 134 (2019).
607. Solomon, A., Bar, A., Yanai, C., Shapira, B. & Rokach, L. *Predict Demographic Information Using Word2vec on Spatial Trajectories* in *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* (Association for Computing Machinery, New York, NY, USA, 2018), 331–339.
608. Levy, O. & Goldberg, Y. *Neural word embedding as implicit matrix factorization* in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (MIT Press, Montreal, Canada, 2014), 2177–2185.
609. An, J., Kwak, H. & Ahn, Y.-Y. *SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment* in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* ACL 2018 (Association for Computational Linguistics, Melbourne, Australia, 2018), 2450–2461.

610. Curiel, R. P., Pappalardo, L., Gabrielli, L. & Bishop, S. R. Gravity and scaling laws of city to city migration. *PLOS ONE* **13**, e0199892 (2018).
611. Jung, W.-S., Wang, F. & Stanley, H. E. Gravity model in the Korean highway. *EPL (Euro-physics Letters)* **81**, 48005 (2008).
612. Hong, I. & Jung, W.-S. Application of gravity model on the Korean urban bus network. *Physica A: Statistical Mechanics and its Applications* **462**, 48–55 (2016).
613. Truscott, J. & Ferguson, N. M. Evaluating the Adequacy of Gravity Models as a Description of Human Mobility for Epidemic Modelling. *PLoS Computational Biology* **8** (2012).
614. Barthélémy, M. Spatial networks. *Physics Reports* **499**, 1–101 (2011).
615. Chen, Y. The distance-decay function of geographical gravity model: Power law or exponential law? *Chaos, Solitons & Fractals* **77**, 174–189 (C 2015).
616. Belkin, M. & Niyogi, P. *Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering in Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* (MIT Press, Cambridge, MA, USA, 2001), 585–591.
617. Morin, F. & Bengio, Y. *Hierarchical Probabilistic Neural Network Language Model* in *International Workshop on Artificial Intelligence and Statistics* International Workshop on Artificial Intelligence and Statistics (PMLR, 2005), 246–252.
618. Gutmann, M. & Hyvärinen, A. *Noise-contrastive estimation: A new estimation principle for unnormalized statistical models* in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (JMLR Workshop and Conference Proceedings, 2010), 297–304.
619. Dyer, C. Notes on Noise Contrastive Estimation and Negative Sampling. *arXiv:1410.8251 [cs]* (2014).

620. Levy, O., Goldberg, Y. & Dagan, I. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics* **3**, 211–225 (2015).
621. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]* (2018).
622. Gates, A. J., Ke, Q., Varol, O. & Barabási, A.-L. Nature’s reach: narrow work has broad impact. *Nature* **575**, 32–34 (2019).
623. Schakel, A. M. J. & Wilson, B. J. Measuring Word Significance using Distributed Representations of Words. *arXiv:1508.02297 [cs]* (2015).
624. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (2018).
625. Kraemer, M. U. G. *et al.* The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020).
626. Cao, C., Baas, J., Wagner, C. S. & Jonkers, K. Returning scientists and the emergence of China’s science system. *Science and Public Policy* **47**, 172–183 (2020).
627. Lerman, G. & Shakhnovich, B. E. Defining functional distance using manifold embeddings of gene ontology annotations. *Proceedings of the National Academy of Sciences* **104**, 11334–11339 (2007).
628. Levy, O. & Goldberg, Y. *Linguistic Regularities in Sparse and Explicit Word Representations* in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning* (Association for Computational Linguistics, Ann Arbor, Michigan, 2014), 171–180.
629. Wilson, B., Hoffman, J. & Morgenstern, J. Predictive Inequity in Object Detection. *arXiv:1902.11097 [cs, stat]* (2019).

630. Evans, J. H. Stratification in knowledge production: Author prestige and the influence of an American academic debate. *Poetics* **33**, 111–133 (2005).
631. Tebbett, N., Jons, H. & Hoyler, M. Openness towards diversity? Cultural homophily in student perceptions of teaching and learning provided by international and home academics (2020).
632. Beaver, D. G. *Teacher fashion, classroom homophily, and the impact on student evaluations* Thesis (Texas Tech University, 1999).
633. Melguizo, I. Homophily and the Persistence of Disagreement. *The Economic Journal* (2018).
634. Lee, E. *et al.* Homophily and minority size explain perception biases in social networks. *Nature human behaviour* **3**, 1078–1087 (2019).
635. Dandekar, P., Goel, A. & Lee, D. T. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences* **110**, 5791–5796 (2013).
636. Dong, M., Jiao, J. & Xia, J. Consequences of homophily: does social status similarity enhance project performance? *Asian Business & Management* (2020).
637. Abramo, G., D'Angelo, C. A. & Di Costa, F. Does the geographic proximity effect on knowledge spillovers vary across research fields? *Scientometrics* **123**, 1021–1036 (2020).
638. "Raw Data" Is an Oxymoron (ed Gitelman, L.) in collab. with Bowker, G. C. *et al.* (The MIT Press, Cambridge, Massachusetts ; London, England, 2013).
639. Silberzahn, R. *et al.* Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science* **1**, 337–356 (2018).
640. Cañibano, C., Otamendi, J. & Andújar, I. Measuring and assessing researcher mobility from CV analysis: the case of the Ramón y Cajal programme in Spain. *Research Evaluation* **17**, 17–31 (2008).

641. Cañibano, C., Otamendi, F. J. & Solís, F. International temporary mobility of researchers: a cross-discipline study. *Scientometrics* **89**, 653 (2011).
642. Jacob, B. A. & Lefgren, L. The impact of research grant funding on scientific productivity. *Journal of Public Economics. Special Issue: The Role of Firms in Tax Systems* **95**, 1168–1177 (2011).
643. García-Carpintero, E., Granadino, B. & Plaza, L. The representation of nationalities on the editorial boards of international journals and the promotion of the scientific output of the same countries. *Scientometrics* **84**, 799–811 (2010).
644. Collins, S. L. & Verdier, J. M. Editorial Boards Must Be Internationally Representative. *BioScience* **68**, 235–235 (2018).
645. Babcock, P. & Marks, M. *Leisure College, USA: The Decline in Student Study Time. Education Outlook. No. 7* (American Enterprise Institute for Public Policy Research, 2010).
646. Costas, R. & Bordons, M. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics. The Hirsch Index* **1**, 193–203 (2007).
647. Seeber, M., Cattaneo, M., Meoli, M. & Malighetti, P. Self-citations as strategic response to the use of metrics for career decisions. *Research Policy. Academic Misconduct, Misrepresentation, and Gaming* **48**, 478–491 (2019).
648. Raan, A. F. J. v. Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics* **67**, 491–502 (2006).
649. Petersen, A. M., Majeti, D., Kwon, K., Ahmed, M. E. & Pavlidis, I. Cross-disciplinary evolution of the genomics revolution. *Science Advances* **4**, eaat4211 (2018).

650. Aksnes, D. W., Rørstad, K., Piro, F. N. & Sivertsen, G. Are mobile researchers more productive and cited than non-mobile researchers? A large-scale study of Norwegian scientists. *Research Evaluation* **22**, 215–223 (2013).
651. Bernstein, S., Diamond, R., McQuade, T. & Pousada, B. The Contribution of High-Skilled Immigrants to Innovation in the United States. *Stanford Graduate School of Business Working Paper* (2019).
652. *High-Skilled Immigrants — IGM Forum*
653. Jaffe, K. *et al.* Productivity in Physical and Chemical Science Predicts the Future Economic Growth of Developing Countries Better than Other Popular Indices. *PLoS ONE* **8** (2013).
654. Kumar, R. R., Stauvermann, P. J. & Patel, A. Exploring the link between research and economic growth: an empirical study of China and USA. *Quality & Quantity* **50**, 1073–1091 (2016).
655. Laverde-Rojas, H. & Correa, J. C. Can scientific productivity impact the economic complexity of countries? *Scientometrics* **120**, 267–282 (2019).
656. Kindler, A., Golosovsky, M. & Solomon, S. Early prediction of the outcome of Kickstarter campaigns: is the success due to virality? *Palgrave Communications* **5**, 1–6 (2019).
657. Rijt, A. v. d., Kang, S. M., Restivo, M. & Patil, A. Field experiments of success-breeds-success dynamics. *Proceedings of the National Academy of Sciences* **111**, 6934–6939 (2014).
658. Salganik, M. J., Dodds, P. S. & Watts, D. J. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science* **311**, 854–856 (2006).
659. Fontana, M., Iori, M., Montobbio, F. & Sinatra, R. New and atypical combinations: An assessment of novelty and interdisciplinarity. *Research Policy* **49**, 104063 (2020).

660. Petersen, A. M., Riccaboni, M., Stanley, H. E. & Pammolli, F. Persistence and uncertainty in the academic career. *Proceedings of the National Academy of Sciences* **109**, 5213–5218 (2012).
661. Janosov, M., Battiston, F. & Sinatra, R. Success and luck in creative careers. *EPJ Data Science* **9**, 1–12 (2020).
662. Strevens, M. *The Knowledge Machine: How Irrationality Created Modern Science* Illustrated edition (Liveright, New York, 2020).
663. Merton, R. K. *The Sociology of Science: Theoretical and Empirical Investigations* (ed Storer, N. W.) (University of Chicago Press, Chicago, 1979).
664. Chang, H. *Inventing Temperature: Measurement and Scientific Progress* 1 edition (Oxford University Press, Oxford ; New York, 2007).
665. Gelman, A., Gregg, M. & Simpson, D. Gaydar and the Fallacy of Decontextualized Measurement. *Sociological Science* **5**, 270–280 (2018).
666. Larivière, V. *et al.* A simple proposal for the publication of journal citation distributions. *bioRxiv*, 062109 (2016).
667. Lozano, G. A., Larivière, V. & Gingras, Y. The weakening relationship between the impact factor and papers' citations in the digital age. *Journal of the American Society for Information Science and Technology* **63**, 2140–2145 (2012).
668. Leydesdorff, L. & Bornmann, L. The Operationalization of "Fields" as WoS Subject Categories (WCs) in Evaluative Bibliometrics: The cases of "Library and Information Science" and "Science & Technology Studies". *arXiv:1407.7849 [cs]* (2014).
669. Ioannidis, J. P. A., Boyack, K. & Wouters, P. F. Citation Metrics: A Primer on How (Not) to Normalize. *PLoS Biology* **14** (2016).

670. Jha, R., Jbara, A.-A., Qazvinian, V. & Radev, D. R. NLP-driven citation analysis for scientometrics. *Natural Language Engineering* **23**, 93–130 (2017).
671. Zhu, X., Turney, P., Lemire, D. & Vellino, A. Measuring academic influence: Not all citations are equal. *J Assn Inf Sci Tec* **66**, 408–427 (2015).
672. Sandström, U. Combining curriculum vitae and bibliometric analysis: mobility, gender and research performance. *Research Evaluation* **18**, 135–142 (2009).
673. Bertrand, M. & Mullainathan, S. *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination* Working Paper 9873 (National Bureau of Economic Research, 2003).
674. Clayson, D. E. What does ratemyprofessors.com actually rate? *Assessment & Evaluation in Higher Education* **39**, 678–698 (2014).
675. West, G. *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies* First Edition (Penguin Press, New York, 2017).
676. West, G. B., Woodruff, W. H. & Brown, J. H. Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals. *Proceedings of the National Academy of Sciences* **99**, 2473–2478 (suppl 1 2002).
677. Milojević, S. Modes of collaboration in modern science: Beyond power laws and preferential attachment. *Journal of the American Society for Information Science and Technology* **61**, 1410–1423 (2010).
678. Bettencourt, L. M. A. The Origins of Scaling in Cities. *Science* **340**, 1438–1441 (2013).
679. Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C. & West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences* **104**, 7301–7306 (2007).

680. Börner, K., Sanyal, S. & Vespignani, A. Network science. *Annual Review of Information Science and Technology* **41**, 537–607 (2007).
681. Folke, C. *et al.* Regime Shifts, Resilience, and Biodiversity in Ecosystem Management. *Annual Review of Ecology, Evolution, and Systematics* **35**, 557–581 (2004).
682. Folke, C. *et al.* Resilience and Sustainable Development: Building Adaptive Capacity in a World of Transformations. *AMBIO: A Journal of the Human Environment* **31**, 437–440 (2002).
683. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
684. Alberts, B. *et al.* Self-correction in science at work. *Science* **348**, 1420–1422 (2015).
685. Bruner, J. P. POLICING EPISTEMIC COMMUNITIES. *Episteme* **10**, 403–416 (2013).
686. Ioannidis, J. P. A. Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* **7**, 645–654 (2012).
687. Taleb, N. N. *Antifragile: Things That Gain from Disorder* Reprint edition (Random House Publishing Group, New York, 2014).
688. Lamberson, P. J. & Page, S. E. Tipping points. *Quarterly Journal of Political Science* **7**, 175–208 (2012).
689. Scheffer, M. Foreseeing tipping points. *Nature* **467**, 411–412 (2010).
690. Cunningham, S. W. & Kwakkel, J. H. Tipping points in science: A catastrophe model of scientific change. *Journal of Engineering and Technology Management. Special Issue on Emergence of Technologies: Methods and Tools for Management* **32**, 185–205 (2014).
691. Zhang, W., Lim, C. & Szymanski, B. K. Analytic treatment of tipping points for social consensus in large random networks. *Physical Review E* **86**, 061134 (2012).

692. Doyle, C., Sreenivasan, S., Szymanski, B. K. & Korniss, G. Social consensus and tipping points with opinion inertia. *Physica A: Statistical Mechanics and its Applications* **443**, 316–323 (2016).
693. Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E. & Havlin, S. Catastrophic cascade of failures in interdependent networks. *Nature* **464**, 1025–1028 (2010).
694. Berezin, Y., Bashan, A., Danziger, M. M., Li, D. & Havlin, S. Localized attacks on spatially embedded networks with dependencies. *Scientific Reports* **5**, 8934 (2015).
695. Majdandzic, A. *et al.* Multiple tipping points and optimal repairing in interacting networks. *Nature Communications* **7**, 10850 (2016).
696. Majdandzic, A. *et al.* Spontaneous recovery in dynamical networks. *Nature Physics* **10**, 34–38 (2014).
697. Siegenfeld, A. F. & Bar-Yam, Y. An Introduction to Complex Systems Science and Its Applications. *Complexity* **2020**, e6105872 (2020).
698. Heng, H. H. Q. The Conflict Between Complex Systems and Reductionism. *JAMA* **300**, 1580–1581 (2008).
699. Slingo, J. & Palmer, T. Uncertainty in weather and climate prediction. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* **369**, 4751–4767 (2011).
700. Batty, M. & Torrens, P. M. Modelling complexity : The limits to prediction. *Cybergeo : European Journal of Geography* (2001).
701. Tetlock, P. E. & Gardner, D. *Superforecasting: The Art and Science of Prediction* (2015).
702. Edwards, M. A. & Roy, S. Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. *Environmental Engineering Science* **34**, 51–61 (2017).

703. Kvivilang, N., Bjurström, E. & Almqvist, R. Making sense of complexity in governance: the case of local public management in the City of Stockholm. *Policy Studies* **41**, 623–640 (2020).
704. Khan, S. *et al.* Embracing uncertainty, managing complexity: applying complexity thinking principles to transformation efforts in healthcare systems. *BMC Health Services Research* **18**, 192 (2018).
705. Sargut, G. & McGrath, R. G. Learning to Live with Complexity. *Harvard Business Review* (2011).
706. Larsen, P. O. & von Ins, M. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* **84**, 575–603 (2010).
707. Woelfle, M., Olliari, P. & Todd, M. H. Open science is a research accelerator. *Nature Chemistry* **3**, 745–748 (2011).
708. Wolfram, D., Wang, P., Hembree, A. & Park, H. Open peer review: promoting transparency in open science. *Scientometrics* **125**, 1033–1051 (2020).
709. Flaherty, C. *Rutgers Graduate School faculty takes a stand against Academic Analytics Inside Higher Ed.*
710. Van Eck, N. J., Waltman, L., Larivière, V. & Sugimoto, C. R. *Crossref as a new source of citation data: A comparison with Web of Science and Scopus CWTS.*
711. Sonntag, M. E., Bassett, J. F. & Snyder, T. An Empirical Test of the Validity of Student Evaluations of Teaching Made on RateMyProfessors.com. *Assessment & Evaluation in Higher Education* **34**, 499–504 (2009).
712. Boyle, P. J., Smith, L. K., Cooper, N. J., Williams, K. S. & O'Connor, H. Gender balance: Women are funded more fairly in social science. *Nature News* **525**, 181 (2015).
713. Leslie, S.-J., Cimpian, A., Meyer, M. & Freeland, E. Expectations of brilliance underlie gender distributions across academic disciplines. *Science (New York, N.Y.)* **347**, 262–265 (2015).

714. in. *Wikipedia* (2020).
715. Miranda, R. & Garcia-Carpintero, E. Overcitation and overrepresentation of review papers in the most cited papers. *Journal of Informetrics* **12**, 1015–1030 (2018).
716. Smith, J. W. N., Surridge, B. W. J., Haxton, T. H. & Lerner, D. N. Pollutant attenuation at the groundwater–surface water interface: A classification scheme and statistical analysis using national-scale nitrate data. *Journal of Hydrology. Transfer of pollutants in soils, sediments and water systems: From small to large scale (AquaTerra)* **369**, 392–402 (2009).
717. Brannon-Peppas, L. & Blanchette, J. Nanoparticle and targeted systems for cancer therapy. *Advanced Drug Delivery Reviews* **56**, 1649–1659 (2004).
718. Gottlieb, M. & Bailitz, J. Comparison of Early Goal-Directed Therapy With Usual Care for Severe Sepsis and Septic Shock. *Annals of Emergency Medicine* **66**, 632–634 (2015).
719. Zhang, T. & Johansson, J. S. A calorimetric study on the binding of six general anesthetics to the hydrophobic core of a model protein. *Biophysical Chemistry* **113**, 169–174 (2005).
720. Rui, J. R. & Wang, H. Social network sites and international students' cross-cultural adaptation. *Computers in Human Behavior* **49**, 400–411 (2015).
721. Papp, P. *et al.* Analytical continuation in coupling constant method; application to the calculation of resonance energies and widths for organic molecules: Glycine, alanine and valine and dimer of formic acid. *Chemical Physics* **418**, 8–13 (2013).
722. Husson, J. M., Schoene, B., Bluher, S. & Maloof, A. C. Chemostratigraphic and U–Pb geochronologic constraints on carbon cycling across the Silurian–Devonian boundary. *Earth and Planetary Science Letters* **436**, 108–120 (2016).
723. Treanor, J. J. Prospects for Broadly Protective Influenza Vaccines. *American Journal of Preventive Medicine* **49**, S355–S363 (2015).

724. Stephenson, P. J. *et al.* Unblocking the flow of biodiversity data for decision-making in Africa. *Biological Conservation. SI:Measures of biodiversity* **213**, 335–340 (2017).
725. Alén, E., Domínguez, T. & de Carlos, P. University students perceptions of the use of academic debates as a teaching methodology. *Journal of Hospitality, Leisure, Sport & Tourism Education* **16**, 15–21 (2015).
726. French, B. M. & Koeberl, C. The convincing identification of terrestrial meteorite impact structures: What works, what doesn't, and why. *Earth-Science Reviews* **98**, 123–170 (2010).
727. Doody, O. & Condon, M. Increasing student involvement and learning through using debate as an assessment. *Nurse Education in Practice* **12**, 232–237 (2012).
728. Ersoy, A. F. Social studies teacher candidates' views on the controversial issues incorporated into their courses in Turkey. *Teaching and Teacher Education* **26**, 323–334 (2010).
729. Nam, C. W. The effects of trust and constructive controversy on student achievement and attitude in online cooperative learning environments. *Computers in Human Behavior* **37**, 237–248 (2014).
730. Kalter, H. Teratology in the 20th century: Environmental causes of congenital malformations in humans and how they were established. *Neurotoxicology and Teratology. Special Issue: Teratology in the Twentieth Century. Congenital malformations in humans and how their environmental causes were established.* **25**, 131–282 (2003).
731. Millan, M. J. Multi-target strategies for the improved treatment of depressive states: Conceptual foundations and neuronal substrates, drug discovery and therapeutic application. *Pharmacology & Therapeutics. Multi-Target Strategies for Treating Depression* **110**, 135–370 (2006).

732. Bruschke, J. & Divine, L. Debunking Nixon's radio victory in the 1960 election: Re-analyzing the historical record and considering currently unexamined polling data. *The Social Science Journal* **54**, 67–75 (2017).
733. Stepanova, O. & Bruckmeier, K. The relevance of environmental conflict research for coastal management. A review of concepts, approaches and methods with a focus on Europe. *Ocean & Coastal Management* **75**, 20–32 (2013).
734. Colston, N. M. & Vadjunec, J. M. A critical political ecology of consensus: On “Teaching Both Sides” of climate change controversies. *Geoforum* **65**, 255–265 (2015).
735. Munro, S. Lipid Rafts: Elusive or Illusive? *Cell* **115**, 377–388 (2003).
736. Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B. & Bandettini, P. A. The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? *NeuroImage* **44**, 893–905 (2009).
737. Zhao, G., Sun, M., Wilde, S. A. & Sanzhong, L. Late Archean to Paleoproterozoic evolution of the North China Craton: key issues revisited. *Precambrian Research* **136**, 177–202 (2005).
738. Li, Z. X. *et al.* Assembly, configuration, and break-up history of Rodinia: A synthesis. *Precambrian Research. Testing the Rodinia Hypothesis: Records in its Building Blocks* **160**, 179–210 (2008).
739. Kusky, T. M. & Li, J. Paleoproterozoic tectonic evolution of the North China Craton. *Journal of Asian Earth Sciences* **22**, 383–397 (2003).
740. Zhao, G., Wilde, S. A., Cawood, P. A. & Sun, M. Archean blocks and their boundaries in the North China Craton: lithological, geochemical, structural and P–T path constraints and tectonic evolution. *Precambrian Research* **107**, 45–73 (2001).

741. Zhai, M.-G. & Santosh, M. The early Precambrian odyssey of the North China Craton: A synoptic overview. *Gondwana Research. Precambrian geology and tectonic evolution of the North China Craton* **20**, 6–25 (2011).
742. Debat, P. *et al.* A new metamorphic constraint for the Eburnean orogeny from Paleoproterozoic formations of the Man shield (Aribinda and Tampelga countries, Burkina Faso). *Precambrian Research* **123**, 47–65 (2003).
743. Wilde, S. A., Zhao, G. & Sun, M. Development of the North China Craton During the Late Archaean and its Final Amalgamation at 1.8 Ga: Some Speculations on its Position Within a Global Palaeoproterozoic Supercontinent. *Gondwana Research* **5**, 85–94 (2002).
744. Kusky, T. M. Geophysical and geological tests of tectonic models of the North China Craton. *Gondwana Research. Precambrian geology and tectonic evolution of the North China Craton* **20**, 26–35 (2011).
745. Franzoni, C., Scellato, G. & Stephan, P. Foreign-born scientists: mobility patterns for 16 countries. *Nature Biotechnology* **30**, 1250–1253 (2012).
746. Meyer, J.-B. Network Approach versus Brain Drain: Lessons from the Diaspora. *International Migration* **39**, 91–110 (2001).
747. Ioannidis, J. P. A. Global estimates of high-level brain drain and deficit. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* **18**, 936–939 (2004).
748. GAILLARD, A. M. & GAILLARD, J. The International Circulation of Scientists and Technologists: A Win-Lose or Win-Win Situation? *Science Communication* **20**, 106–115 (1998).
749. *The Growing Role of Knowledge in the Global Economy* (UNESCO, 2010).
750. Scellato, G., Franzoni, C. & Stephan, P. Migrant scientists and international networks. *Research Policy* **44**, 108–120 (2015).

751. Robinson-Garcia, N. *et al.* Scientific mobility indicators in practice: International mobility profiles at the country level. *El Profesional de la Información* **27**, 511 (2018).
752. Vaccario, G., Verginer, L. & Schweitzer, F. The Mobility Network of Scientists: Analyzing Temporal Correlations in Scientific Careers. *arXiv:1905.06142 [physics]* (2019).
753. Albarrán, P., Carrasco, R. & Ruiz-Castillo, J. Geographic mobility and research productivity in a selection of top world economics departments. *Scientometrics* **111**, 241–265 (2017).
754. Woolley, R. & Turpin, T. CV analysis as a complementary methodological approach : investigating the mobility of Australian scientists. *Research Evaluation* **18**, 143–151 (2009).
755. Jeh, G. & Widom, J. *Scaling personalized web search* in *Proceedings of the 12th international conference on World Wide Web* (Association for Computing Machinery, New York, NY, USA, 2003), 271–279.
756. Qiu, J. *et al.* *Network Embedding as Matrix Factorization: Unifying DeepWalk, LINE, PTE, and node2vec* in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Association for Computing Machinery, New York, NY, USA, 2018), 459–467.
757. Xu, L. & Yuille, A. Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Transactions on Neural Networks* **6**, 131–143 (1995).
758. Huber. *Robust Statistics, 2nd Edition* (2009).
759. Xu, J., Tao, Y. & Lin, H. *Semantic word cloud generation based on word embeddings* in *2016 IEEE Pacific Visualization Symposium (PacificVis) 2016 IEEE Pacific Visualization Symposium (PacificVis)* (2016), 239–243.
760. Chandrasekaran, V., Sanghavi, S., Parrilo, P. A. & Willsky, A. S. Rank-Sparsity Incoherence for Matrix Decomposition. *SIAM Journal on Optimization* **21**, 572–596 (2011).

761. Candès, E. J., Li, X., Ma, Y. & Wright, J. Robust principal component analysis? *Journal of the ACM* **58**, 11:1–11:37 (2011).
762. Ma, S., Bassily, R. & Belkin, M. *The Power of Interpolation: Understanding the Effectiveness of SGD in Modern Over-parametrized Learning* in. ICML (2018).
763. Smith, S., Elsen, E. & De, S. *On the Generalization Benefit of Noise in Stochastic Gradient Descent* in *International Conference on Machine Learning* International Conference on Machine Learning (PMLR, 2020), 9058–9067.
764. Zhang, G. *et al.* Which Algorithmic Choices Matter at Which Batch Sizes? Insights From a Noisy Quadratic Model. *arXiv:1907.04164 [cs, stat]* (2019).
765. D'Angelo, C. A. & van Eck, N. J. Collecting large-scale publication data at the level of individual researchers: a practical proposal for author name disambiguation. *Scientometrics* (2020).

Appendix A

Study1: Peer review at *eLife*

A.1 Text

Modelling homogeneity using main effects with interaction term.

We used logistic regression to model the degree to which gender equity in peer review outcomes differed based on the composition of the reviewer team in order to find support for the inequity observed in Fig 6. Fig 7.A demonstrates that last author gender inequity persisted even when controlling for the gender composition of the reviewer team, but did not address the degree to which this equity manifests in submissions reviewed by all-male vs. mixed-gender reviewer teams. Given that there is no established method of addressing this question, we considered several approaches. The first approach modelled the interaction between last author gender and the gender-composition of the reviewer team (see S9 Table, column 2), however this approach proved difficult to interpret: adding the interaction term appeared to suppress the main effects of last author gender and reviewer team composition observed in Fig 7.A, though the corresponding ANOVA table demonstrated these effects to still account for a significant amount of deviance (see S11 Table). There were no significant interaction term, conflicting with Fig 6; main effects are often made less interpretable by the addition of interaction terms. A low sample size across interaction groups further complicates interpretation. Moreover, this approach modelled individual-level interactions between the author and reviewer composition on a per-submission basis, not differences in group-level estimates of inequity.

Modelling homogeneity using separately trained models.

S9 Table, columns 3 and 4 shows the results of two logistic regression models of percentage of full submissions accepted, constructed as in Fig 7.A, but each calculated using only full submissions reviewed by either all-male or mixed-gender reviewer teams. In the all-male model, a male last author was associated with a 1.23 times increased odds of acceptance (95% CI = [1.05, 1.41], $p = 0.027$) compared to a female last author; in contrast, a smaller non-significant effect was observed between male and female last authors in the model containing only mixed-gender reviewer teams. This approach shows a larger positive effect favoring male last authors under the condition of all-male teams than for mixed-reviewer teams, affirming results of models in Fig 7.B, but this approach has several limitations that favor the approach from Fig 7. The confidence intervals for the effect of the regression for submissions reviewed by mixed reviewer teams are wide, making precise comparisons difficult. Interpretation of S9 Table is further complicated by possible population differences between groups as well as the different amount of data used to fit each model, $n=3,090$ for the all-male reviewer model and $n = 3,280$ for the mixed-gender reviewer model.

A.2 Figures

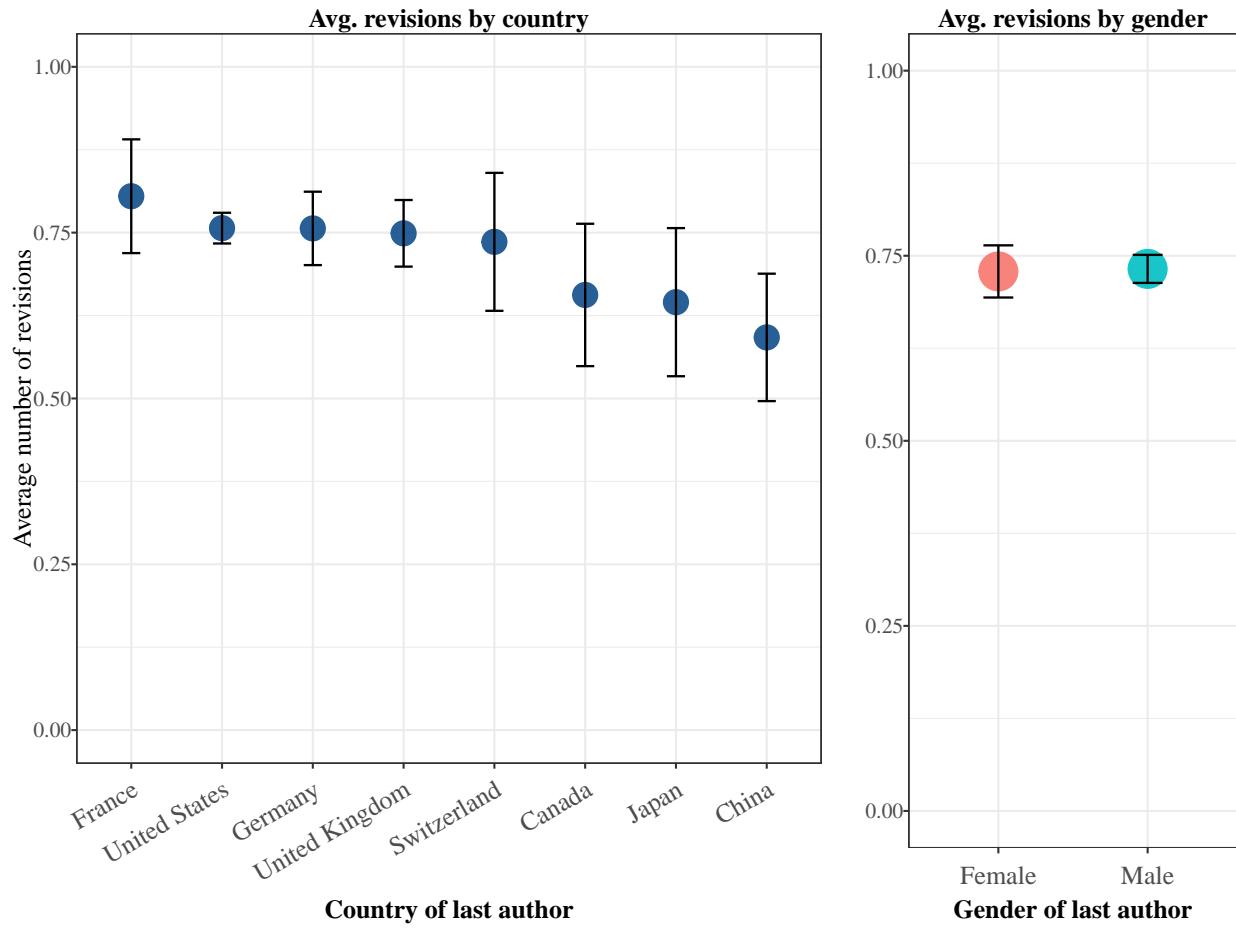
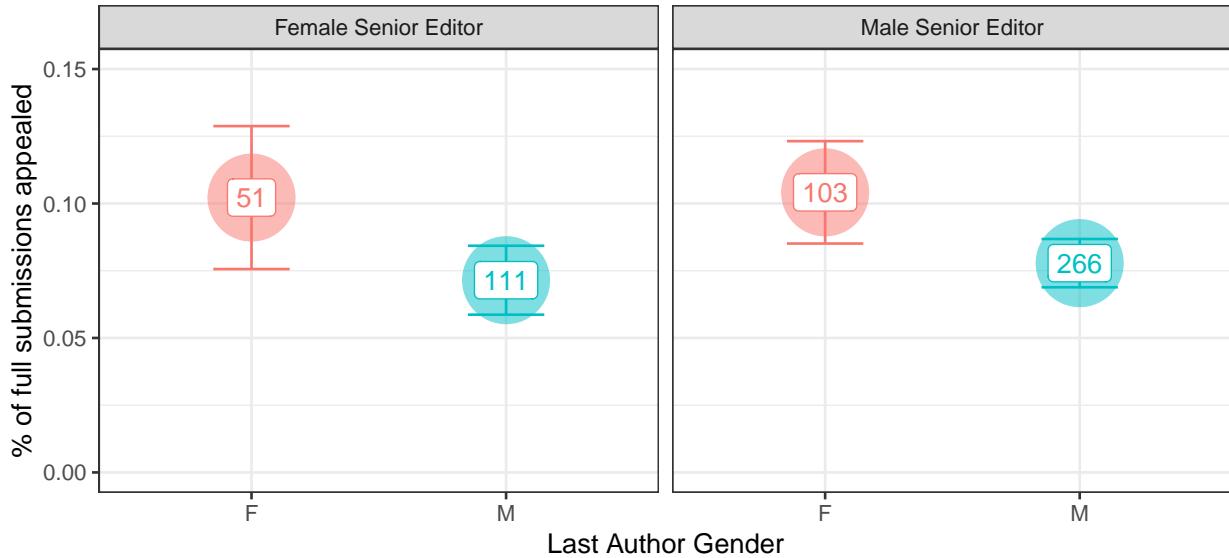


Figure A.1: **Number of revisions by author gender and country of affiliation.** Average number of revisions a full submissions undergoes before a final decision of accept or reject is made. In this case, zero revisions occurs when a full submission is accepted or rejected without a request for any revisions. The dataset records at maximum two revisions, though only a small number of manuscripts remain in revision after two submissions (see Fig 3.1). For this figure, we only include manuscripts for which a final decision is made after zero, one, or two revisions. The left panel shows differences in the average number of revisions by the country of the last author. The right shows the average revisions by the gender of the last author. Code to reproduce this figure can be found on the linked Github repository at the path `figures/revision_information/average_revisions.rmd`.

Full Submissions appealed by gender of last author and senior editor



Full Submissions appealed by gender of last author and reviewing editor

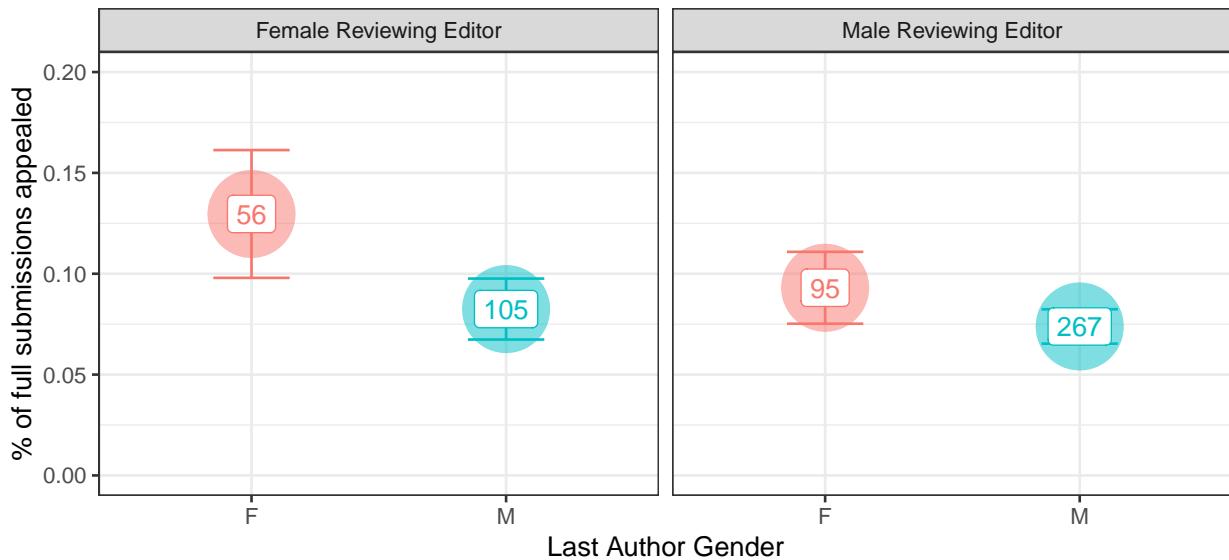


Figure A.2: **Number of appeals by gender of author and reviewing editor.** Count of submissions appealed, at any review stage, by the gender of the last author gender and Senior Editor (top) and reviewing editor (bottom). Code to reproduce this figure can be found on the linked Github repository at the path figures/appeals/gender_and_appeals.rmd.

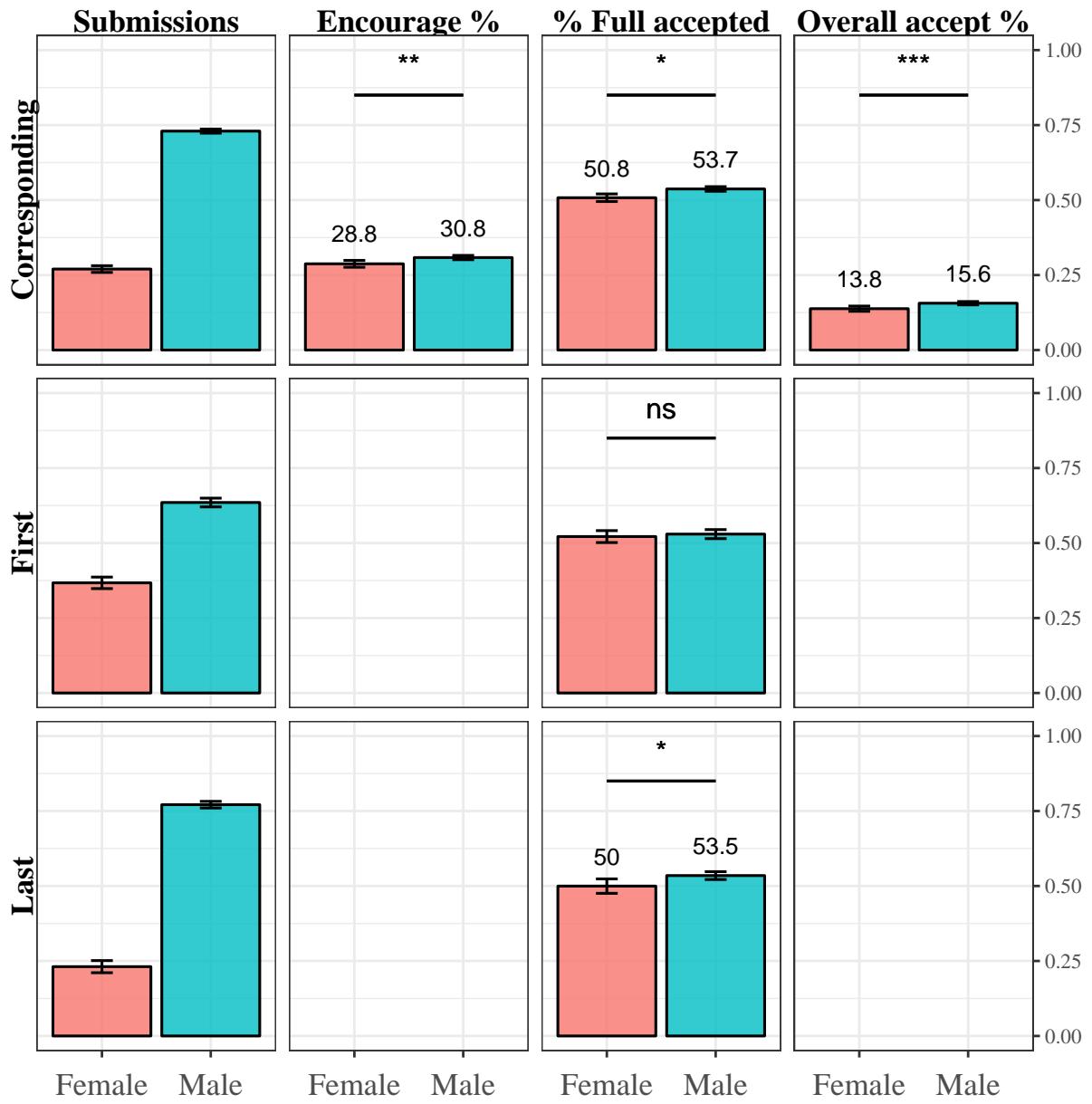


Figure A.3: Submission and success rates by gender of corresponding, first, and last author. Proportion of initial submissions, encourage rate, overall acceptance rate, and acceptance rate of full submissions by the gender of the corresponding author, first author, and last author. Gender data is unavailable for first and last authors of initial submissions that were never submitted as full submissions, therefore these cells remain blank. Authors whose gender is unknown are excluded from analysis. Vertical error bars indicate 95% confidence intervals of the proportion of submitted, encouraged, and accepted initial and full submissions. Asterisks indicate significance level of χ^2 tests of independence of frequency of encourage and acceptance by gender; “***” = $p < 0.001$; “**” = $p < 0.01$; “*” = $p < 0.05$; “-” = $p < 0.1$; “ns” = $p \geq 0.1$. Code to reproduce this figure can be found on the linked Github repository at the path figures/author_outcomes/supp_submission_outcomes.rmd.

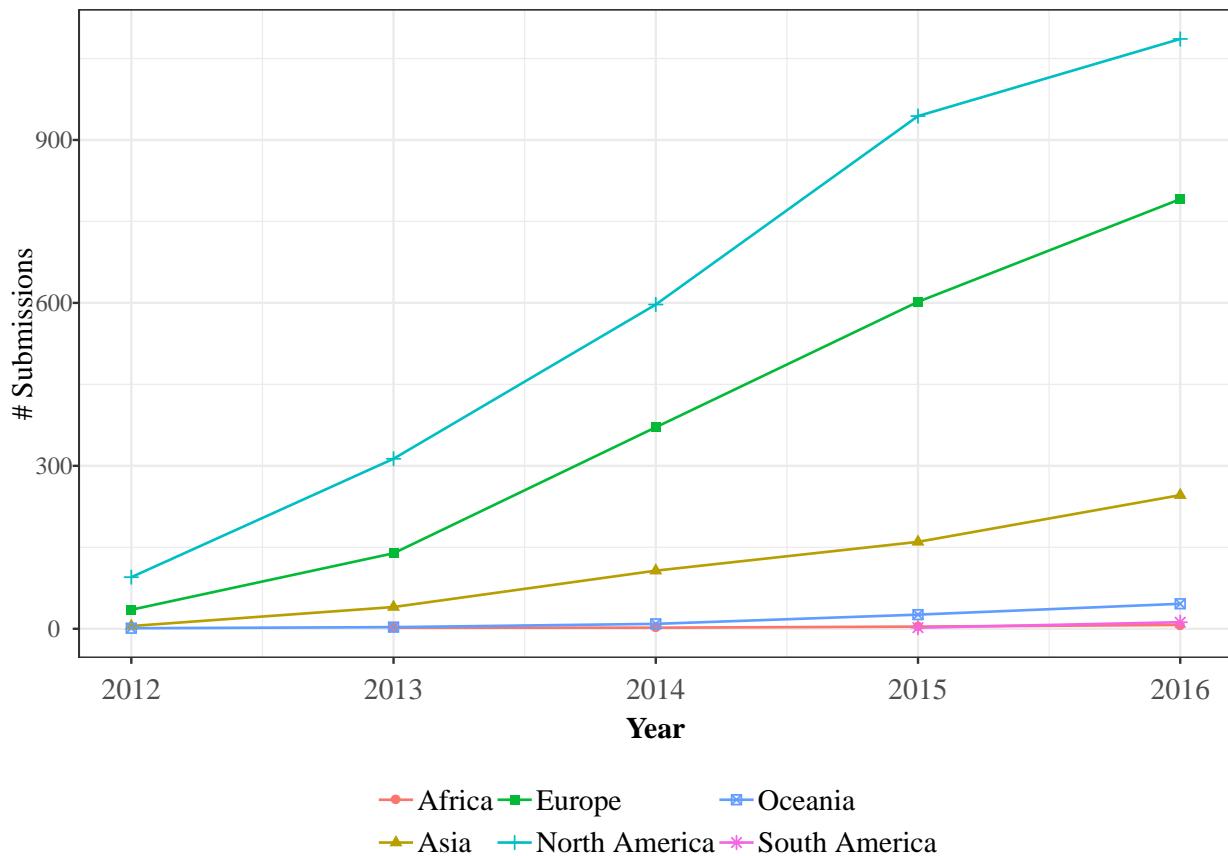


Figure A.4: **Geographic composition over time.** Count of initial submissions by country of corresponding authors over time. Code to reproduce this figure can be found on the linked Github repository at the path `figures/selectivity_over_time/country_composition_shift.rmd`.

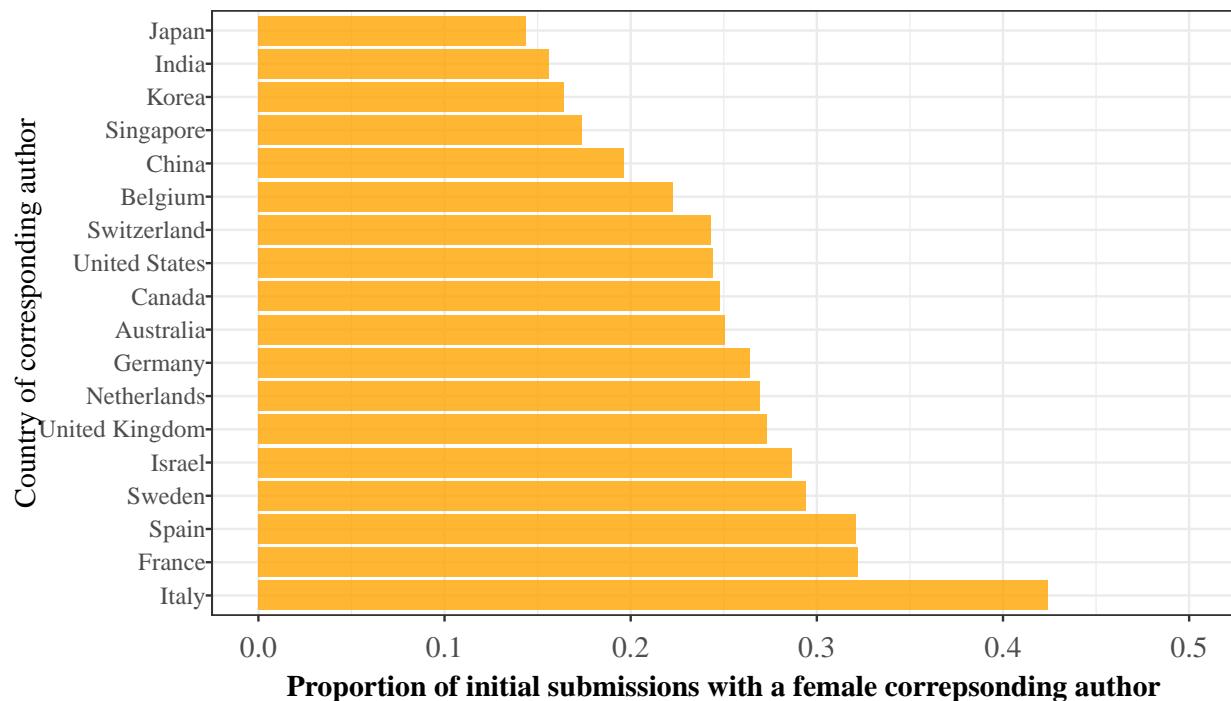


Figure A.5: **Proportion of women corresponding authors by country.** Proportion of female corresponding authors on initial submissions for each country having more than 200 initial submissions during the period of study. Code to reproduce this figure can be found on the linked Github repository at the path `figures/general_infomation/supp_gender_prop_by_country.rmd`.

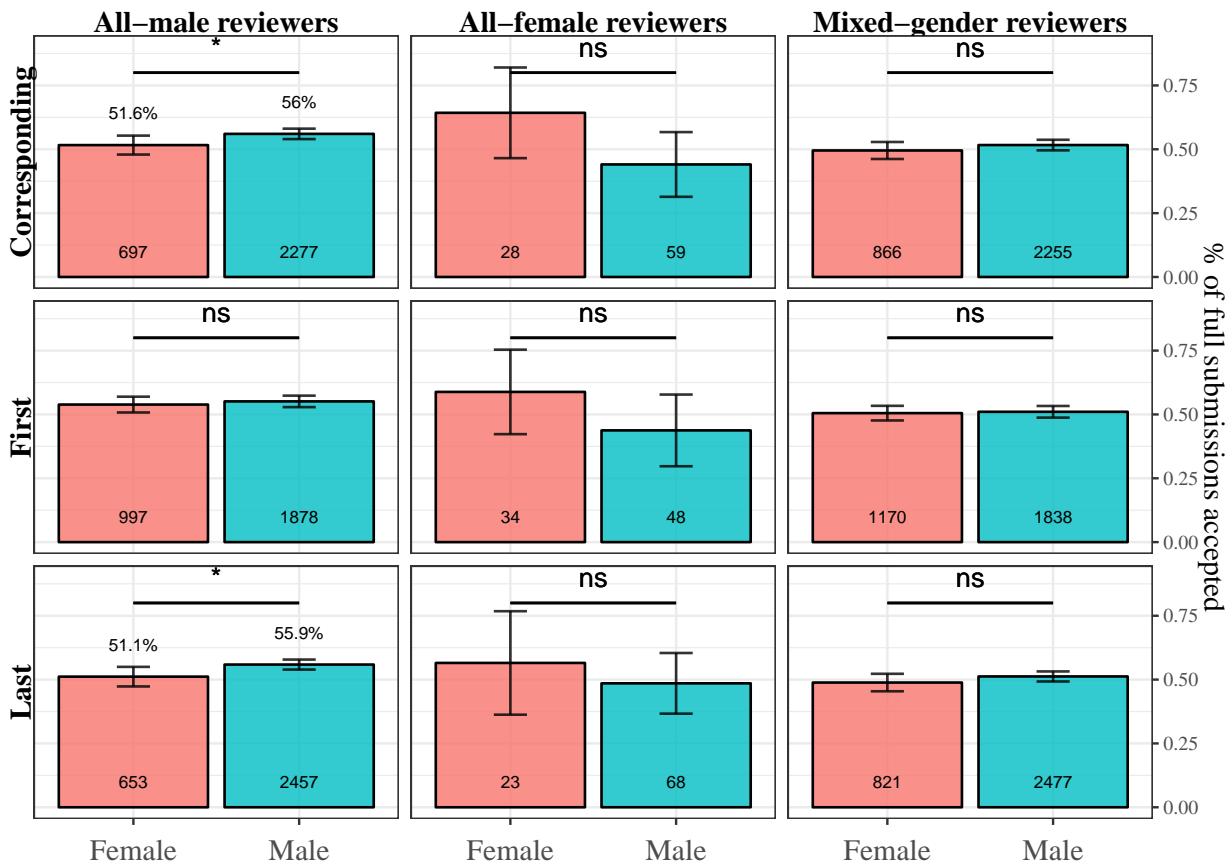


Figure A.6: Submission and success rates by authorship role and gatekeeper gender composition. Percentage of full submissions that were accepted, shown by the gender of the corresponding, first, and last author, and by the gender composition of the peer reviewers. Text at the base of each bar indicate the number full submissions within each category of reviewer team and authorship gender. Vertical error bars indicate 95% percentile confidence intervals of the proportion of accepted full submissions. Asterisks indicate significance level of χ^2 tests of independence on frequency of acceptance by gender of author given each team composition.”***” = $p < 0.001$; ”**” = $p < 0.01$; ”*” = $p < 0.05$; ”-” = $p < 0.1$; “ns” = $p \geq 0.1$. Code to reproduce this figure can be found on the linked Github repository at the path figures/gatekeeper_author_outcomes/supp_homophily_outcomes.rmd.

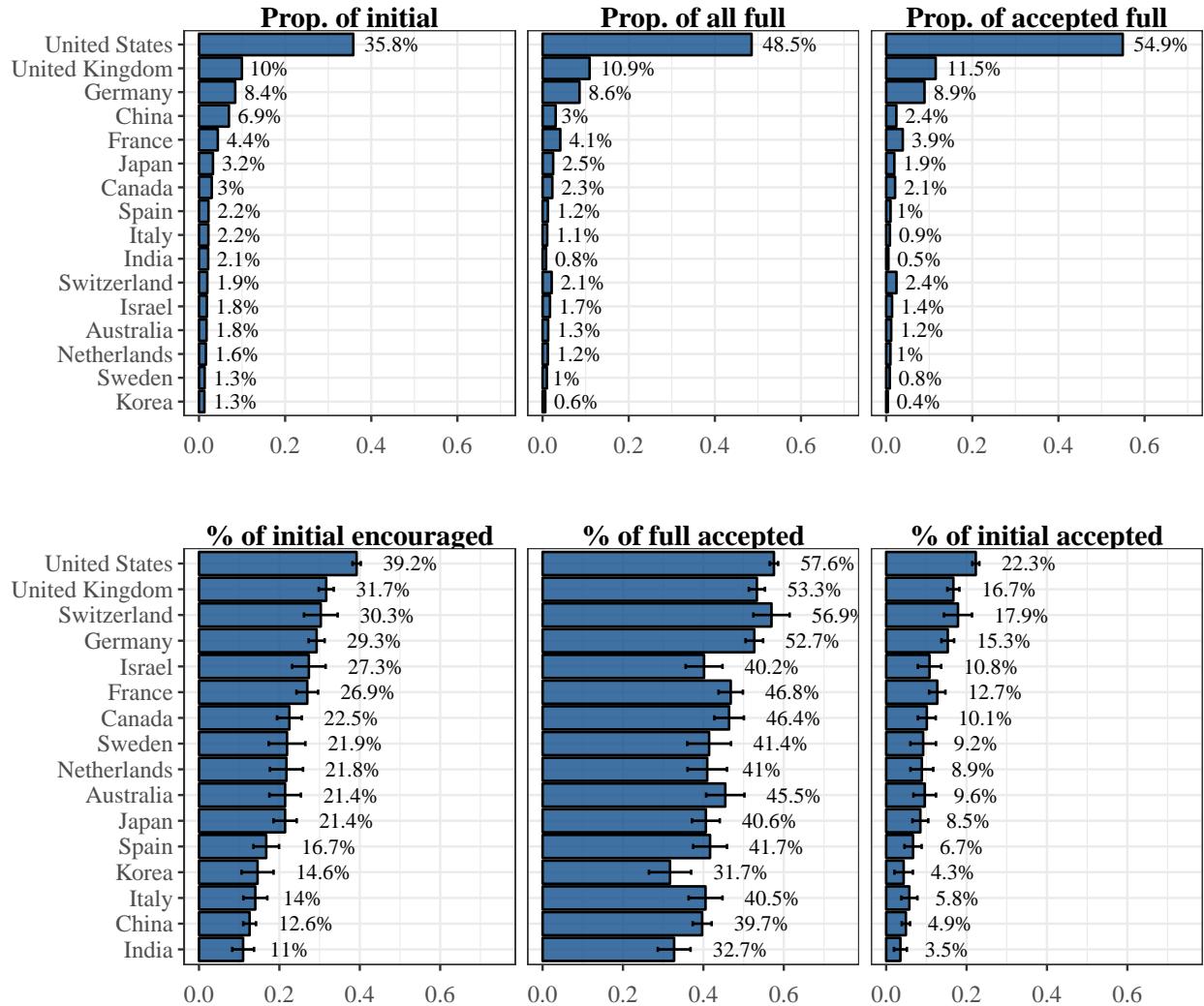


Figure A.7: Submission and success rates by country for top 16 most prolific countries. Top: proportion of all initial submissions, encouraged initial submissions, and accepted full submissions comprised by the country of affiliation of the corresponding author for the top sixteen most prolific countries in terms of initial submissions. Bottom: acceptance rate of full submissions, encourage rate of full submissions, and overall acceptance rate of full submissions by country of affiliation of the corresponding author for the top eight more prolific countries in terms of initial submissions. Error bars on bottom panel indicate standard error of proportion of encouraged initial submissions and accepted initial and full submissions for each country. Code to reproduce this figure can be found on the linked Github repository at the path figures/author_outcomes/supp_outcomes_16_countries.rmd.

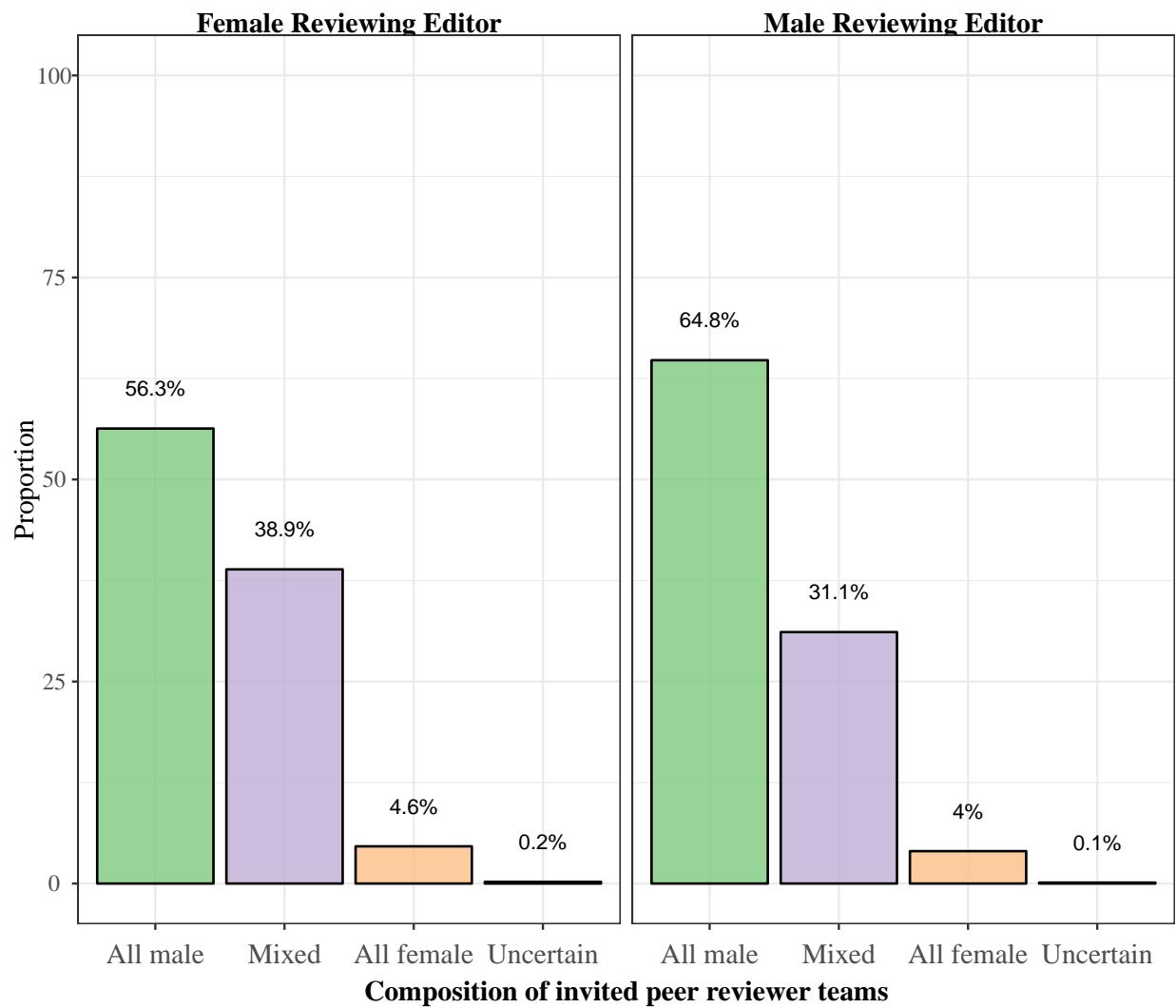


Figure A.8: **Proportion of peer reviewer team's gender compositions by gender of the Reviewing Editor.** Compositions are determined while excluding the Reviewing Editor from team membership, if they are listed as a peer reviewer. Code to reproduce this figure can be found on the linked Github repository at the path figures/gatekeeper_representation/supp_reviewing_editor_composition.rmd.

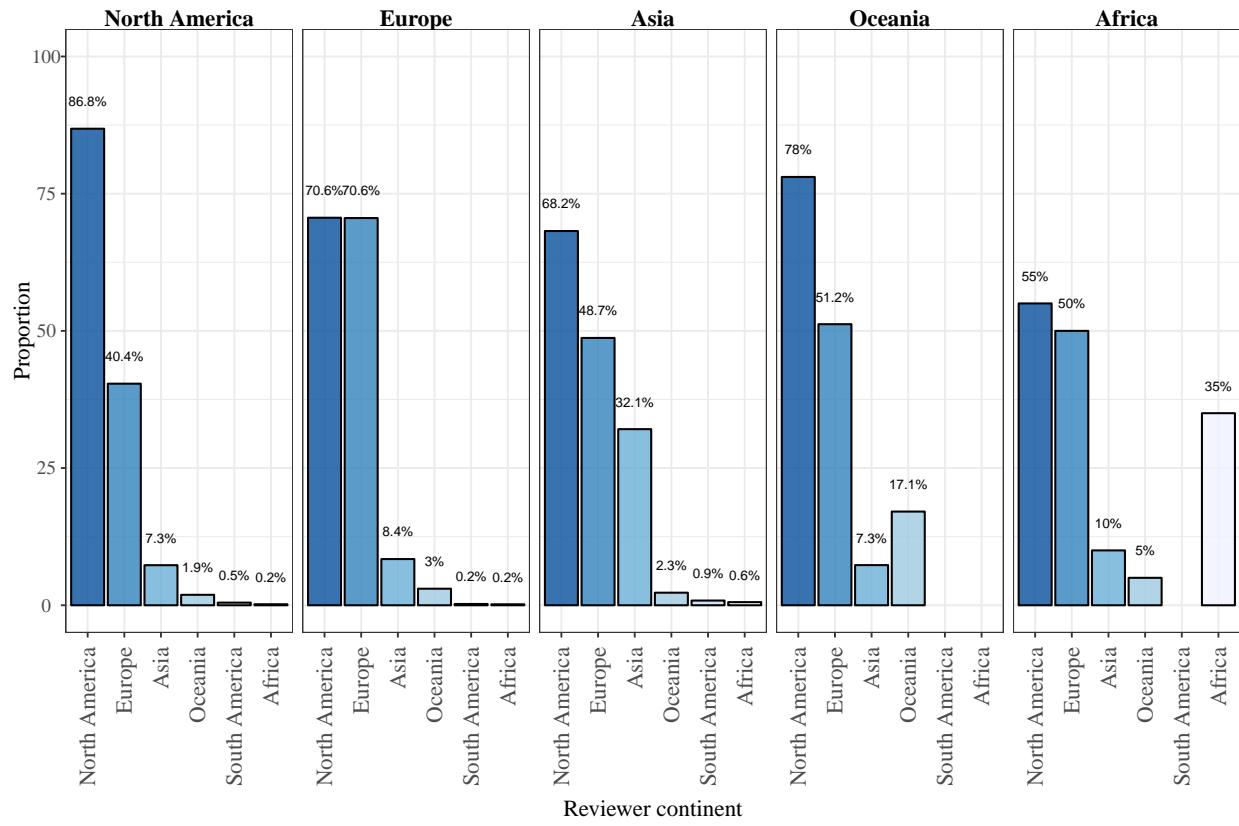


Figure A.9: **Proportion of peer review teams containing at least one peer reviewer of each continent, by continent of Reviewing Editor.** Compositions are determined while excluding the Reviewing Editor from team membership, if they are listed as a peer reviewer. Code to reproduce this figure can be found on the linked Github repository at the path figures/gatekeeper_representation/reviewing_editor_continental_comp.rmd.

A.3 Tables

Table A.1: **Gender demographics of *eLife*.** Counts of distinct male and female corresponding authors, first authors, last authors, and gatekeepers. Includes counts on all initial and full submissions submitted between 2012 and 2017. First and last authors and gatekeepers appeared only on full submissions, whereas corresponding authors appeared on rejected or in-progress initial submissions as well. This table contains the same values as visualized in Fig 3.3.A.

Role	Gender	#	%
Corr. Author (Initial)	F	4846	0.266
Corr. Author (Initial)	M	12243	0.673
Corr. Author (Initial)	UNK	1106	0.061
Corr. Author (Full)		1	0
Corr. Author (Full)	F	1437	0.253
Corr. Author (Full)	M	3944	0.695
Corr. Author (Full)	UNK	296	0.052
First Author	F	2263	0.339
First Author	M	3859	0.578
First Author	UNK	552	0.083
Gatekeeper	F	1440	0.216
Gatekeeper	M	5207	0.781
Gatekeeper	UNK	22	0.003
Last Author	F	1341	0.24
Last Author	M	4250	0.76
Last Author	UNK	4	0.001

Table A.2: **Summary demographic characteristics of distinct *eLife* reviewers and editors.** The count of Senior Editors includes former editors, as well as the Deputy Editors and Editor-in-Chief, who also serve as Senior Editors. The count of BREs includes former editors and guest editors. Reviewers are only relevant for publications that were submitted for full review, thus leading to lower total counts. Includes all individuals involved in processing manuscripts at *eLife* between 2012 and 2017.

Reviewership	Female		Male		Unassigned		All
	N	%	N	%	N	%	
Senior Editors	15	26.3	42	73.7	0	0.0	57
Reviewing Editors	209	24.0	661	76.0	0.0	0.0	870
Peer Reviewers	1,526	21.5	5,572	78.4	7	0.1	7,222

Table A.3: **Summary country of affiliation demographics of unique *eLife* reviewers and editors.** The count of Senior Editors includes former editors, as well as the Deputy Editors and Editor-in-Chief, who also serve as Senior Editors. The count of reviewing editors includes former editors and guest editors. Reviewers are only relevant for publications that were submitted for full review, thus leading to lower total counts than the number of initial submissions. Includes all individuals involved in processing manuscripts at *eLife* between 2012 and 2017.

Country	# Peer Rev.	% Peer Rev.	# Rev. Editor	% Rev. Editor	# Sen. Editor	% Sen. Editor
United States	11,313	0.600	536	0.620	32	0.561
United Kingdom	1,896	0.101	88	0.102	7	0.123
Germany	1,416	0.075	69	0.080	6	0.105
Canada	627	0.033	22	0.025	3	0.053
Switzerland	444	0.024	19	0.022	2	0.035
China	140	0.007	10	0.012	2	0.035
Israel	214	0.011	19	0.022	1	0.018
Netherlands	270	0.014	11	0.013	1	0.018
Spain	201	0.011	10	0.012	1	0.018
Japan	296	0.016	9	0.010	1	0.018
India	89	0.005	6	0.007	1	0.018
France	571	0.030	21	0.024		
Australia	198	0.011	7	0.008		
South Africa	28	0.001	5	0.006		
Austria	118	0.006	4	0.005		
Belgium	114	0.006	3	0.003		
Finland	82	0.004	3	0.003		
Italy	133	0.007	3	0.003		
Singapore	82	0.004	3	0.003		
Thailand	16	0.001	3	0.003		
Denmark	78	0.004	2	0.002		
Korea	59	0.003	2	0.002		
Estonia	2	0.0001	1	0.001		
Hong Kong	7	0.0004	1	0.001		
Hungary	20	0.001	1	0.001		
Ireland	38	0.002	1	0.001		
Kenya	7	0.0004	1	0.001		
Mexico	23	0.001	1	0.001		
New Zealand	19	0.001	1	0.001		
Poland	26	0.001	1	0.001		
Sweden	128	0.007	1	0.001		
Albania	2	0.0001				
Andorra	2	0.0001				
Argentina	21	0.001				
Brazil	9	0.0005				
Chile	10	0.001				
Croatia	3	0.0002				
Czech Rep.	8	0.0004				
Greece	15	0.001				
Guyana	2	0.0001				
Iceland	2	0.0001				
Madagascar	2	0.0001				
Malaysia	2	0.0001				
Monaco	1	0.0001				
Norway	20	0.001				
Portugal	55	0.003				
Puerto Rico	2	0.0001				
Russia	1	0.0001				
Saudi Arabia	2	0.0001				
Slovenia	1	0.0001				
Taiwan	18	0.001				
Turkey	4	0.0002				
United Arab Emirates	3	0.0002				
Uruguay	2	0.0001				
Vietnam	1	0.0001				

Table A.4: **Geographic demographics of *eLife*.** Counts of distinct corresponding authors, first authors, last authors, and gatekeepers, by continent of affiliation. Includes counts on all initial and full submissions submitted between 2012 and 2017. First and last authors and gatekeepers appeared only on full submissions, whereas corresponding authors appeared on rejected or in-progress initial submissions as well. This table contains the same values as visualized in Fig 3.3.B.

Role	Continent	#	%
Corr. Author (Initial)	Africa	61	0.003
Corr. Author (Initial)	Asia	3238	0.178
Corr. Author (Initial)	Europe	7264	0.399
Corr. Author (Initial)	North America	7045	0.387
Corr. Author (Initial)	Oceania	399	0.022
Corr. Author (Initial)	South America	188	0.01
Corr. Author (Full)	Africa	10	0.002
Corr. Author (Full)	Asia	624	0.11
Corr. Author (Full)	Europe	2078	0.366
Corr. Author (Full)	North America	2854	0.503
Corr. Author (Full)	Oceania	95	0.017
Corr. Author (Full)	South America	17	0.003
First Author	Africa	14	0.002
First Author	Asia	751	0.113
First Author	Europe	2373	0.356
First Author	North America	3412	0.511
First Author	Oceania	102	0.015
First Author	South America	22	0.003
Gatekeeper	Africa	17	0.003
Gatekeeper	Asia	378	0.057
Gatekeeper	Europe	2162	0.324
Gatekeeper	North America	3992	0.599
Gatekeeper	Oceania	98	0.015
Gatekeeper	South America	22	0.003
Last Author	Africa	13	0.002
Last Author	Asia	619	0.111
Last Author	Europe	2063	0.369
Last Author	North America	2789	0.498
Last Author	Oceania	94	0.017
Last Author	South America	17	0.003

Table A.5: **Submissions and proportion of author/gatekeeper homogeneity by country.** Includes number of full submissions submitted with corresponding authors from each of 20 countries, and proportion of these full submissions with the condition of author/reviewer homogeneity such that at least one involved gatekeeper from the same country. Countries listed are in order of the proportion of author/reviewer homogeneity, and contain the top 20 countries with the highest homogeneity.

Country	# Submissions	# Homogeneity	% Country Homogeneity
United States	3605	3185	0.883
United Kingdom	803	236	0.294
Germany	641	168	0.262
Mexico	5	1	0.2
Korea	45	8	0.178
Canada	176	27	0.153
Japan	184	19	0.103
Australia	101	10	0.099
China	233	23	0.099
Switzerland	163	16	0.098
Ireland	11	1	0.091
South Africa	11	1	0.091
France	310	28	0.09
Poland	12	1	0.083
Belgium	41	3	0.073
Finland	14	1	0.071
Norway	14	1	0.071
India	59	4	0.068
Denmark	32	2	0.062

Table A.6: **Model coefficients of initial submissions—author characteristics:** Odds ratio, associated confidence intervals, and model diagnostics for logistic regression model using the encouragement of initial submission as a response variable. Predictor variables include control variables of the submission year and type, and variables capturing author characteristics. For continent of affiliation, "North America" was used as the reference level. For submission type, "RA" (research article) was used as the reference level; the submission type "SR" means "Short Reports", and "TR" means "Tools and Resources". This table contains the same values as visualized in Fig 3.5.A.

	ENCOURAGED <i>logistic</i>
Submission Year	.918*** (.894,.942)
Submission Type = SR	.742*** (.638,.847)
Submission Type = TR	.740*** (.567,.913)
Corr. Author is Male	1.118** (1.051,1.185)
Corr. Author Gender UNK	.932 (.795,1.070)
Corr. Author Inst. Top	1.726*** (1.663,1.789)
Corr. Author from Africa	.535* (-.018,1.088)
Corr. Author from Asia	.395*** (.301,.488)
Corr. Author from Europe	.676*** (.611,.740)
Corr. Author from Oceania	.559*** (.336,.783)
Corr. Author from South America	.205*** (-.269,.679)
Constant	.638*** (.526,.749)
Observations	23,615
Log Likelihood	-13,778.170
Akaike Inf. Crit.	27,580.330

Notes:

*P < .05

**P < .01

***P < .001

Table A.7: **Model coefficients of full submissions—author characteristics:** Odds ratio, associated confidence intervals, and model diagnostics for logistic regression model using the acceptance of full submission as a response variable. Predictor variables include control variables of the submission year and type, and variables capturing author characteristics. For continent of affiliation, "North America" was used as the reference level. For submission type, "RA" (research article) was used as the reference level; the submission type "SR" means "Short Reports", and "TR" means "Tools and Resources". This table contains the same values as visualized in Fig 3.5.B.

	ACCEPTED <i>logistic</i>
Submission Year	.888*** (.847,.929)
Submission Type = SR	.897 (.711,1.082)
Submission Type = TR	1.117 (.800,1.434)
First Author is Male	1.022 (.914,1.129)
First Author is Unknown Gender	1.033 (.840,1.226)
Last Author is Male	1.145* (1.027,1.263)
Last Author Inst. Top	1.379*** (1.272,1.486)
Last Author from Africa	1.484 (.464,2.503)
Last Author from Asia	.585*** (.408,.763)
Last Author from Europe	.860** (.749,.972)
Last Author from Oceania	.906 (.490,1.323)
Last Author from South America	.839 (-.098,1.776)
Constant	1.430*** (1.230,1.629)
Observations	6,461
Log Likelihood	-4,390.813
Akaike Inf. Crit.	8,807.626
<i>Notes:</i>	*P < .05 **P < .01 ***P < .001

Table A.8: **Model coefficients of initial submissions—author characteristics and homogeneity:** Odds ratio, associated confidence intervals, and model diagnostics for logistic regression model using the encouragement of initial submission as a response variable. Predictor variables include control variables of the submission year and type, and variables capturing author characteristics and author-reviewer homogeneity. For continent of affiliation, "North America" was used as the reference level. For submission type, "RA" (research article) was used as the reference level; the submission type "SR" means "Short Reports", and "TR" means "Tools and Resources".

	ENCOURAGED <i>logistic</i>
Submission Year	.918*** (.894,.942)
Submission Type = SR	.742*** (.638,.847)
Submission Type = TR	.741*** (.568,.914)
Corr. Author is Male	1.115** (1.048,1.182)
Corr. Author is Unknown Gender	.930 (.792,1.068)
Corr. Author Inst. Top	1.709*** (1.645,1.772)
Corr. Author from Africa	.579 (.021,1.137)
Corr. Author from Asia	.443*** (.337,.549)
Corr. Author from Europe	.800*** (.724,.877)
Corr. Author from Oceania	.570*** (.328,.813)
Corr. Author from South America	.225*** (-.254,.703)
Corr. Author-Editor Geo. Distance	1.022*** (1.010,1.034)
Corr. Author-Editor Geo. Distance = 0	1.560*** (1.448,1.673)
Constant	.465*** (.320,.610)
Observations	23,615
Log Likelihood	-13,742.830
Akaike Inf. Crit.	27,513.650

Notes:

*P < .05
**P < .01
***P < .001

Table A.9: Model coefficients of regressions on full submissions: Odds ratio, associated confidence intervals, and model diagnostics for logistic regression model using the acceptance of full submission as the response variable. Control variables include the submission year, submission type, last author institutional prestige, and the gender of the first author. Other predictor variables include the gender of the last author, continent of affiliation of the last author, gender-composition of the reviewers, the last author-reviewers geographic distance, and variables attempting to capture the gender equity by reviewer-team composition group. Five models are presented: the first (Main Effects) shows only the main effects for the model including all full submissions without any additional manipulation or variables (1); the second model (2, Standard Interaction) models the main effects as well as an interaction term between last author gender and the gender composition of the reviewer team (an ANOVA table for this model has been provided in **ANOVA table for author-reviewer interaction model:** Results of ANOVA test run on the fitted model containing main effects for author and reviewer characteristics for full submissions as well as the interaction between last author gender and reviewer team composition; the next two models were separately trained on only submissions reviewed by all-male reviewer teams (3) and only submission trained on mixed-gender reviewer teams (4), respectively; the last model (5) models gender equity between reviewer-composition groups using a new variable with all combinations of author and reviewer gender (see Fig 3.7). Columns (1) and (5) contain the same values as Fig 3.7A and Fig 3.7.B, respectively. For continent of affiliation, "North America" was used as the reference level. For submission type, "RA" (research article) was used as the reference level; the submission type "SR" means "Short Reports", and "TR" means "Tools and Resources". For the combination variable of last author gender and reviewer team composition, we held "last author female, all rev. male" as the reference level. Missing cells indicates that the corresponding variable was not part of that model.

	ACCEPTED <i>logistic</i>	
	All Male 1	Mixed-Gender 2
Submission Year	.907** (.848,.966)	.881*** (.823,.940)
Submission Type = SR	.993 (.727,1.259)	.770 (.503,1.038)
Submission Type = TR	1.035 (.574,1.496)	1.139 (.692,1.586)
First Author is Male	1.034 (.875,1.193)	1.022 (.873,1.172)
First Author is Unknown Gender	1.163 (.869,1.456)	.967 (.704,1.230)
Last Author Inst. Top	1.519*** (1.362,1.676)	1.330*** (1.180,1.480)
Last Author Male	1.228* (1.051,1.405)	1.088 (.926,1.249)
Last author from Africa	2.212 (.477,3.948)	2.276 (.972,3.581)
Last author from Asia	.758 (.447,1.068)	.851 (.551,1.152)
Last author from Europe	1.020 (.835,1.205)	.951 (.776,1.125)
Last author from Oceania	.974 (.312,1.636)	2.516** (1.826,3.205)
Last author from South America	.975 (-.543,2.492)	1.656 (.390,2.923)
Sum of geo. distance (1000s km)	.992 299 (.983,1.001)	.982*** (.973,.991)
Sum of geo. distance is zero	1.240 (.996,1.483)	.797 (.558,1.037)
All Reviewers Male	1.271	1.872***

Table A.10: **Continent-level table of initial and full submission counts:** The number of initial and full submissions for each of seven continents including Antarctica (which was excluded from other analysis).

	Continent	# Initial submissions	# Full submissions
1	North America	9591	3785
2	Europe	9106	2527
3	Asia	4382	735
4	Oceania	472	107
5	South America	215	20
6	Africa	78	17
7	Antarctica	1	NA

Table A.11: **ANOVA table for author-reviewer interaction model:** Results of ANOVA test run on the fitted model containing main effects for author and reviewer characteristics for full submissions as well as the interaction between last author gender and reviewer team composition.

	term	df	Deviance	Resid..Df	Resid..Dev	p.value
1	Submission Year	1	47.997	6459	8889.773	<0.0001
2	Submission Type	2	2.397	6457	8887.377	0.30172
3	First Author Gender	2	0.306	6455	8887.071	0.85814
4	Last Author Inst. Prestige	1	62.855	6454	8824.216	<0.0001
5	Last Author Gender	1	5.194	6453	8819.022	0.02266
6	Last author Continent	5	37.397	6448	8781.626	<0.0001
7	Last Author-Reviewers Geographic Distance	1	22.679	6447	8758.946	<0.0001
8	Sum of geo. distance is zero	1	0.018	6446	8758.928	0.89338
9	Reviewer Gender Composition	2	7.797	6444	8751.131	0.02027
10	Last Author Gender*Reviewer Gender Composition	2	1.767	6442	8749.365	0.4134

Table A.12: **Model coefficients of full submissions—author characteristics and reviewing-editor only homogeneity:** Odds ratio, associated confidence intervals, and model diagnostics for logistic regression model using the encouragement of full submission as a response variable. Predictor variables include control variables of the submission year and type, and variables capturing author characteristics and homogeneity between the author and reviewing editor only. For continent of affiliation, "North America" was used as the reference level. For submission type, "RA" (research article) was used as the reference level; the submission type "SR" means "Short Reports", and "TR" means "Tools and Resources". This regression models gender equity between reviewer composition groups using a new variable containing all combinations of last author gender and reviewer team composition; for this new categorical variable, we used "last author female - female rev. editor" as the reference level.

	ACCEPTED
Submission Year	.897*** (.856,.939)
Submission Type = SR	.890 (.703,1.078)
Submission Type = TR	1.090 (.767,1.413)
First Author is Male	1.010 (.901,1.119)
First Author is Unknown Gender	1.057 (.862,1.253)
Last Author Inst. Top 50	1.383*** (1.275,1.492)
Last author from Africa	2.239 (1.201,3.278)
Last author from Asia	.805 (.586,1.024)
Last author from Europe	1.002 (.863,1.141)
Last author from Oceania	1.520 (1.039,2.000)
Last author from South America	1.194 (.246,2.142)
Dist. between author and rev. editor (1000km)	1.018 (.990,1.045)
Sum of author-reviewer distance (1000km)	.984*** (.975,.993)
Total dist. between author and reviewers is zero	.977 (.792,1.162)
Dist. between author and rev. editor is zero	1.095 (.882,1.308)
Last author female - male rev. editor	1.204 (.975,1.433)
Last author male - female rev. editor	1.178 (.955,1.400)
Last author male - male rev. editor	1.352** (1.148,1.556)
All Female Reviewers	.962 (.706,1.218)
Mixed Reviewers	.951 (.843,1.060)
Constant	1.306 (.991,1.621)
Observations	6,320
Log Likelihood	-4,280.736
Akaike Inf. Crit.	8,603.471
<i>Notes:</i>	
	*P < .05
	**P < .01
	***P < .001

Appendix B

Study 2: Teaching evaluations

B.1 Text

Values no longer used by RateMyProfessor.com.

Before summer, 2017, when a user left a review for a professor on *RateMyProfessor.com*, they were prompted to select their interest level prior to attending class. The available interest levels were grouped into five ordinal categories, ranging from “Low” at the bottom to “It’s My Life” at the top (further description available in Table B.3 and Table B.2). The selected interest level was then displayed on each user’s review. Since the summer of 2017, this feature is no longer present on the site. We however have data on student interest for every review beforehand, and so we do not exclude this variable from our analysis.

Clarity and helpfulness were two other quantitative metrics in use by *RateMyProfessor.com*, but were recently removed from the site. Though we have these data, we exclude them from our present analysis because past research, and our own analysis, demonstrates that they are strongly correlated with overall quality [502].

The “chili pepper” or “hotness” rating, while never explicitly defined as such, was implicitly associated with the physical attractiveness of the professor. This rating was removed from *RateMyProfessor.com* as of 2018, however, it was present throughout the period of data collection and so we include it in our analysis.

Validity of Academic Analytics and RateMyProfessor.com.

As *Academic Analytics* is used by more institutions, its validity has been called into question [709]. However, these allegations have largely been anecdotal given that the data is proprietary and thus

not available for public scrutiny. A large-scale validation exercise remains necessary to thoroughly assess the accuracy of research indicators in AA2017. One such analysis has already been conducted on CrossRef—the source of AA2017’s publication and citation network [710]; this analysis found considerable overlap of both citations and publications with more widely accepted research evaluation databases such as Scopus and Web of Science. Our own small-scale analysis, compared the counts of items listed on the CVs of professors to their counts listed in AA2017 and demonstrated reasonably accurate coverage of publications. In light of this evidence, we believe that AA2017, while not thoroughly vetted, is sufficient for large-scale analysis. An advantage of using AA2017 is that they collect data for faculty in a variety of disciplines, potentially leading to greater coverage of Humanities and Social Sciences than traditional bibliometric sources such as Scopus and Web of Science [125].

Concerns have been raised over the validity of *RateMyProfessor.com*. The website lacks external validity as a result of its open and anonymous nature, allowing students to rate a course on their first day of attendance, or even years after [498]. Additionally, the content on the site is entirely user-generated with little to no gatekeeping to ensure that real students are in fact reviewing real professors for courses they actually took. One result of this is entirely fabricated records; for example, *RateMyProfessor.com* included a profile for “Albus Dumbledore, professor of transfiguration at Hogwarts School of Witchcraft and Wizardry”, a popular fictitious character from the *Harry Potter* franchise, who had 143 student ratings at the time of writing. By merging records with *Academic Analytics*, which contains a known list of active tenure and tenure-track faculty, we mitigate the impact of fabricated or otherwise misleading profiles. Despite criticism, evidence supports that *RateMyProfessor.com* ratings correlate with traditional student-evaluations of teachers [499–502, 711], suggesting that findings from our analysis are likely to generalize to other faculty evaluations.

Representativeness of matched vs. unmatched records in Academic Analytics.

Table B.3 details descriptive information for individuals from the *Academic Analytics* dataset (AA2017) who were matched versus not matched with records in *RateMyProfessor.com*. The unmatched statistics includes only tenure and tenure track faculty. This comparison allows us to assess the extent of bias in our matching process.

The largest differences we observed between the matched and unmatched dataset relate to the discipline of faculty and the control of the university (public vs. private), and the rank of the professor. Under the assumption that all students in a course are equally likely to leave their instructor on *RateMyProfessor.com*, professors who teach more classes and are exposed more students would be more likely to appear in RMP2018. Given this, one possible contributing factor to differences between matched and unmatched data is the relative exposure of these faculty; this exposure will likely differ across disciplinary and university contexts. For example, medical scientists were underrepresented in the matched data; a cursory investigation of CVs from these faculty revealed that while they often held affiliations with university medical schools, they focus on research and teach comparatively little. Similarly, public universities tend to be much larger, on average, than private universities, and so faculty are likely to teach larger classes and be exposed to more students, leading to an over-representation of public schools. Even within a university, associate faculty may have greater teaching loads than full faculty, and so would be more likely, on average, to have a review on *RateMyProfessor.com*, resulting in an over-representation of associate faculty in the matched data. Moreover, since Table B.3 includes only tenure and tenure-track faculty, departments that use lecturers to teach large courses may be underrepresented; our analysis only concerns tenure and tenure-track faculty and so while lecturers may have many students, they are excluded from analysis.

We observed only small differences in research indicators between matched and unmatched faculty; research performance may differ by faculty's institutional affiliation, seniority, and discipline;

we partially control for this last factor through a simple field-normalizing, but this form of normalization is flawed [669]. Minor differences that we observed in indicators of research performance between and unmatched faculty may have resulted from the differences in the distribution of these faculty across university, disciplinary, and professional contexts.

Representativeness of matched vs. unmatched records in RateMyProfessor.com.

Past analyses of faculty rating data from *RateMyProfessor.com* have typically followed one of two approaches: in the first approach researchers examine large samples of profiles sampled from the website [476, 507, 674]; however, results from this approach may be confounded by fake profiles and by mixing profiles of full-time research faculty with those of graduate instructors, lecturers, or part-time faculty. Other studies have instead examined smaller known population, typically limited to faculty from a small number of departments and universities whose profiles can be manually extracted from *RateMyProfessor.com* [469, 492, 498, 501, 502]; however, these studies may lack external validity. The present study attempted a balance between these approaches by examining a large and diverse set of known tenure and tenure-track faculty.

Fig. B.9 shows how ratings from RMP2018 differ between the population of matched and unmatched faculty. Matched faculty tended to have slightly lower ratings of overall quality (median = 3.7, mean = 3.5) than unmatched faculty (median = 4.0, mean = 3.8). Matched faculty tended to be rated as more difficult (median = 3.2, mean = 3.2) than unmatched faculty (median = 2.8, mean = 2.9). Ratings of student interest were roughly the same between matched and unmatched faculty. The distribution of the number of comments was highly skewed, though matched faculty tended to have more comments (median = 8, mean = 10.2) than unmatched faculty (median = 7, mean = 9.5). However, these values include only individuals who had 25 or fewer reviews, as was used in the main analysis. The distribution of reviews tended to be highly positively skewed, with a maximum value of 268 for matched faculty, and 2,365 for unmatched faculty.

The largest difference between matched and unmatched faculty in RMP2018 was that a larger proportion of unmatched faculty had a chili pepper (30.5 percent with) than matched faculty (19.9 percent with). Assignment of the chili pepper, indicating the attractiveness, is associated with scientific age (see Fig. B.4); one possible reason for this difference may be that matched faculty are older than unmatched faculty, however since RMP2018 does not record age this cannot be assessed. Matched faculty more often were mentioned as having an accent (8.7 percent) than unmatched faculty (5.4 percent) though this difference was relatively small. Matched faculty were also more likely to have had a teaching assistant mentioned (5.3 percent) compared to unmatched faculty (1.1 percent).

Differences between matched and unmatched faculty likely resulted from differences in university, disciplinary, and professional contexts that shape who is most likely to be reviewed. Since we matched RMP profiles with records from AA, matched faculty were all tenure and tenure-track faculty at research-oriented universities. The unmatched RMP profiles however contained many faculty from small liberal arts colleges, community colleges, and other teaching-oriented institutions; these institutions may have had different faculty demographics than larger research-oriented institutions. Related to institutional context is discipline—for example, larger research-oriented universities may have been able to host more faculty and students in disciplines requiring lab space or special facilities (e.g.: medicine) or disciplines requiring accreditation (e.g.: civil engineering). Some of these disciplines will be more likely to use teaching assistants. There are differences in faculty demographics between disciplines, based on gender and nationality [28, 712, 713] which might relate to mentions of accent. Unmatched faculty consisted of many non-tenure faculty and so would include part-time faculty, non-research instructors, and graduate instructors. These various teacher roles may be associated with distinct demographics and teaching contexts, which may have contributed to the observed differences between matched and unmatched faculty.

B.2 Figures

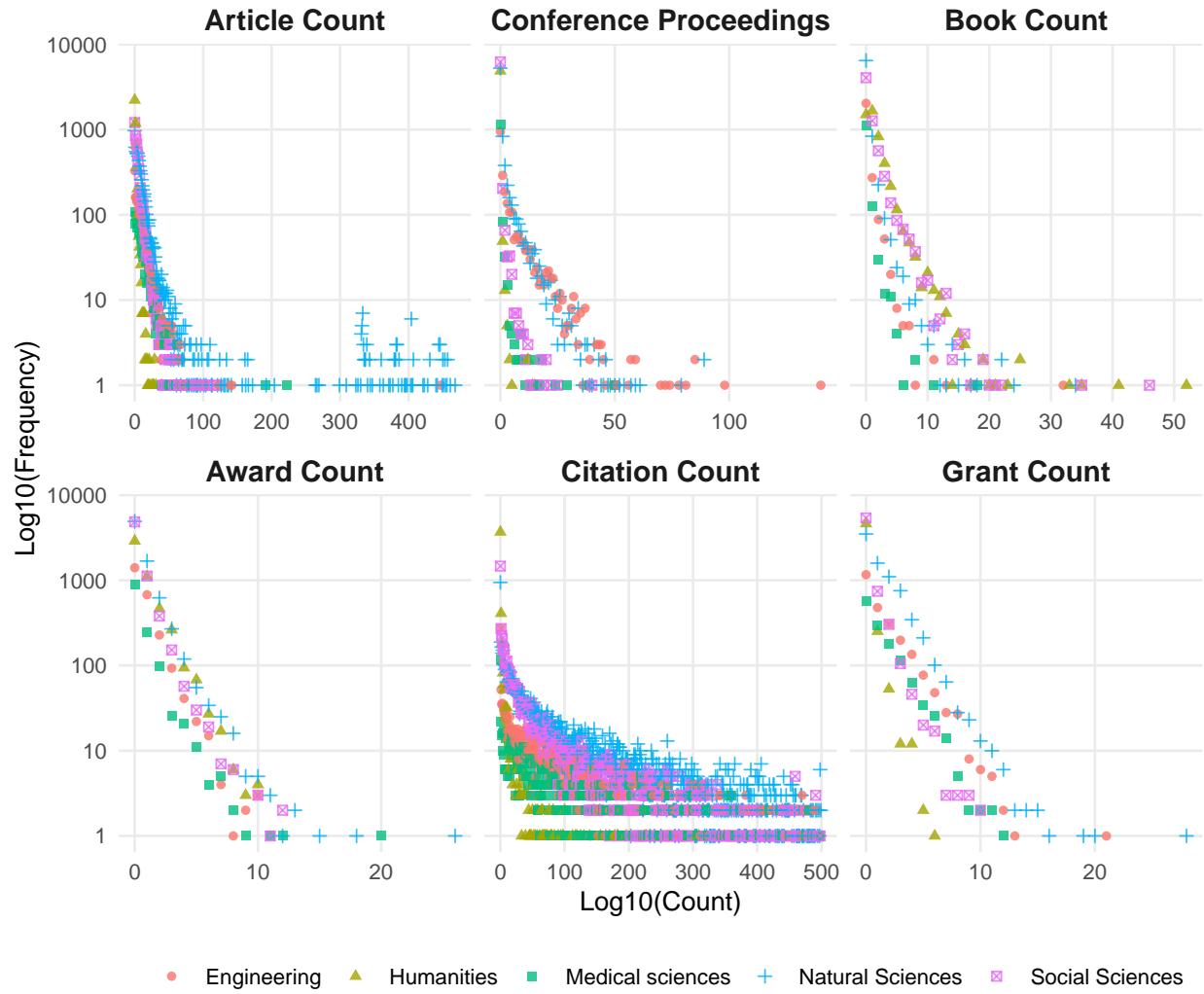


Figure B.1: **Distribution of research indicators.** A point-based histogram of frequencies of research indicator values in the dataset placed on a LogLog scale. Each point plots the frequency of professors with a given “count” of research items. Non-normalized raw counts are used. Points are grouped by discipline, specified by color and shape. Aggregate values by discipline can be found in Table B.3

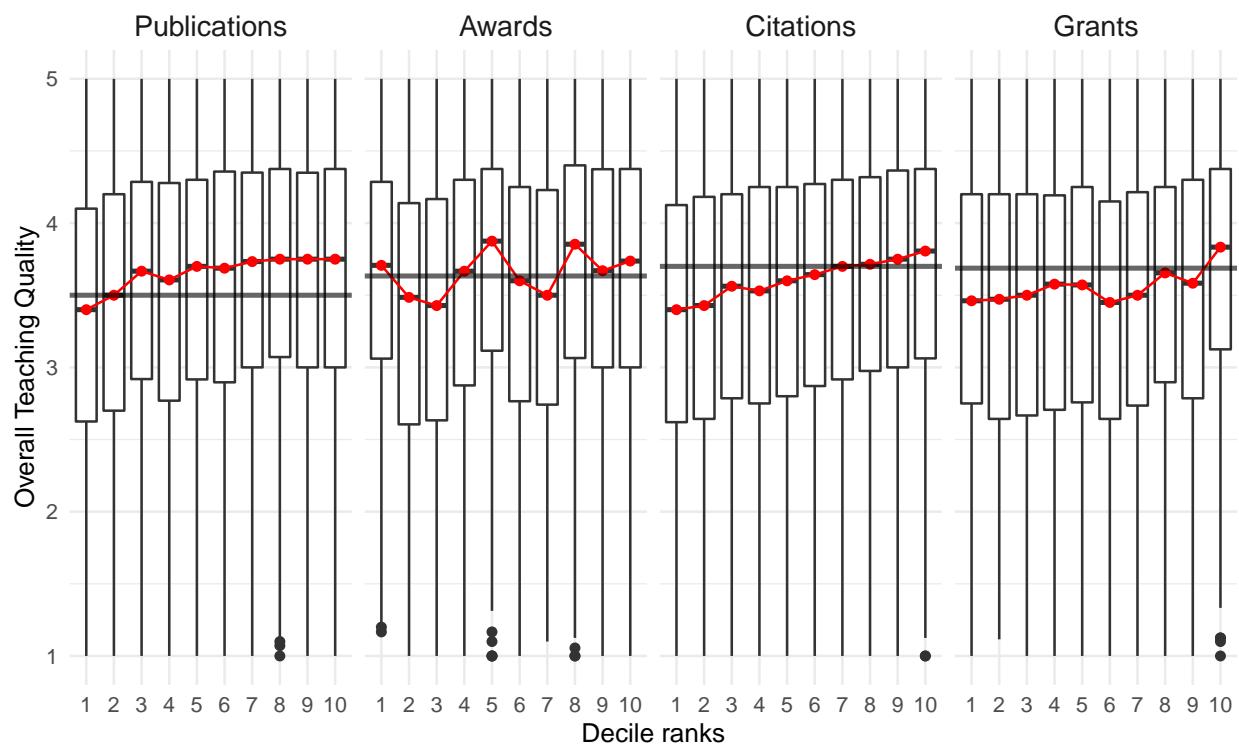


Figure B.2: **Ratings of teaching quality by research performance.** Boxplots of ratings of overall teaching quality for faculty having a positive non-zero value for field-normalized research indicators. Indicator performance is binned into deciles (x-axis). The horizontal grey line is the median for faculty with a value of zero in each indicator. The red line corresponds to the median rating of overall teaching quality for faculty in each decile bin.

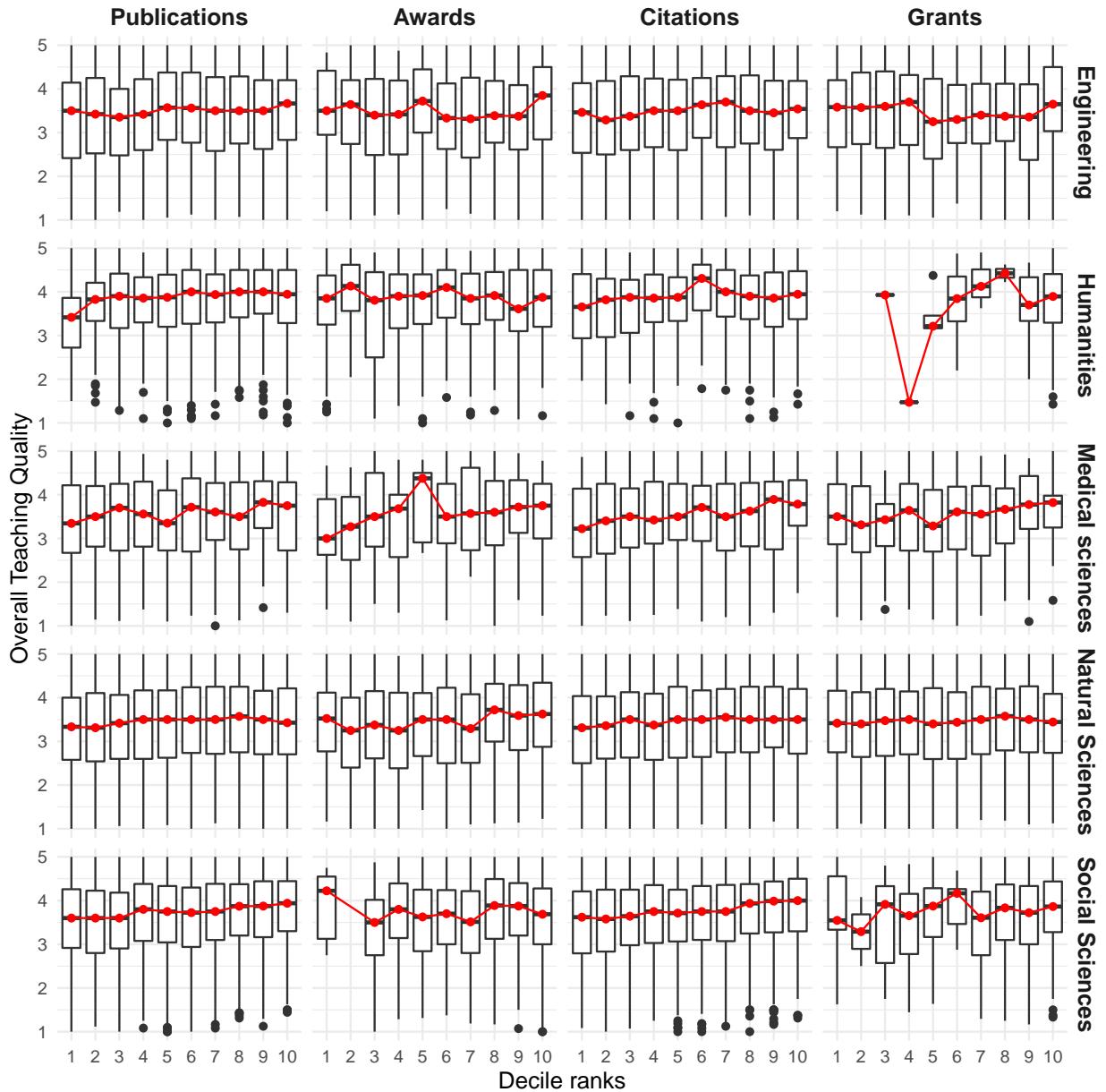


Figure B.3: Ratings of teaching quality by research performance and discipline. Boxplots of ratings of overall teaching quality for faculty having a positive non-zero value for field-normalized research indicators. Indicator performance is binned into deciles (x-axis), repeated for faculty in each of the five discipline categories. The red line corresponds to the median rating of overall teaching quality for faculty in each decile bin.

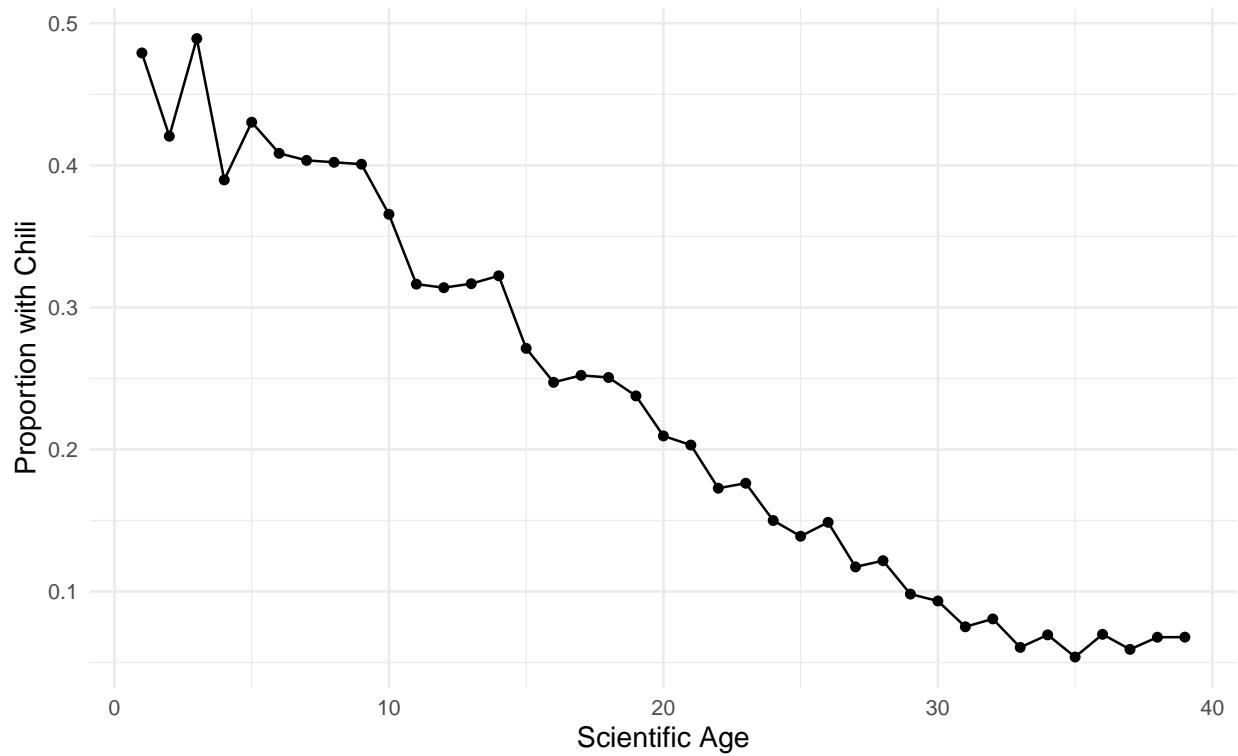


Figure B.4: **Younger faculty more often assigned chili pepper.** The proportion of faculty in the matched dataset that were assigned a chili pepper (y-axis), implicitly suggesting attractiveness, by scientific age (x-axis).

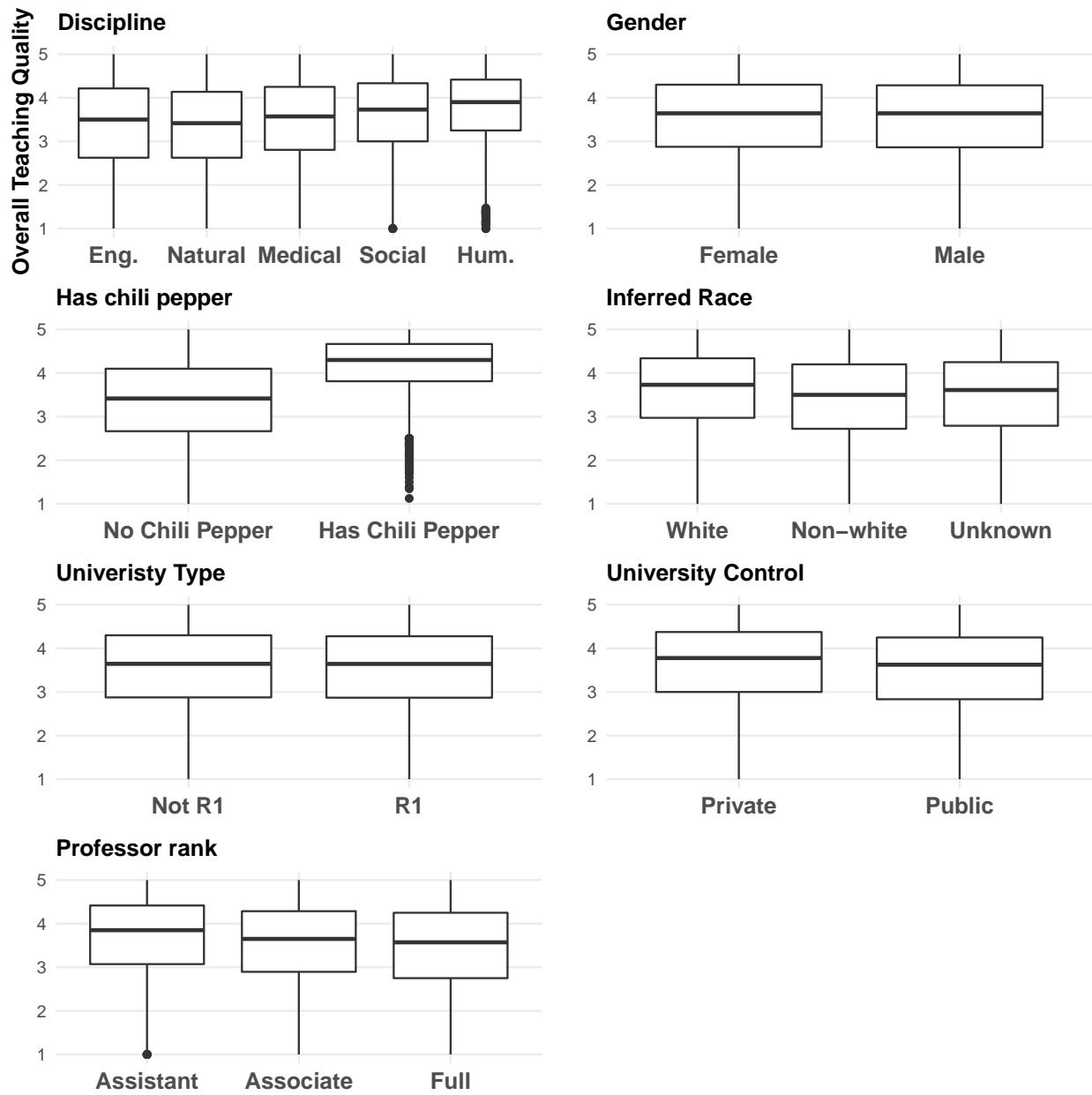


Figure B.5: **Distribution of teaching quality across categorical variables.** The distribution of ratings of overall teaching quality (y-axis) for values of each categorical variable (x-axis) from the matched dataset. Includes discipline, gender, whether the faculty has a chili pepper, inferred race, the university type, the university control, and the professor's rank.

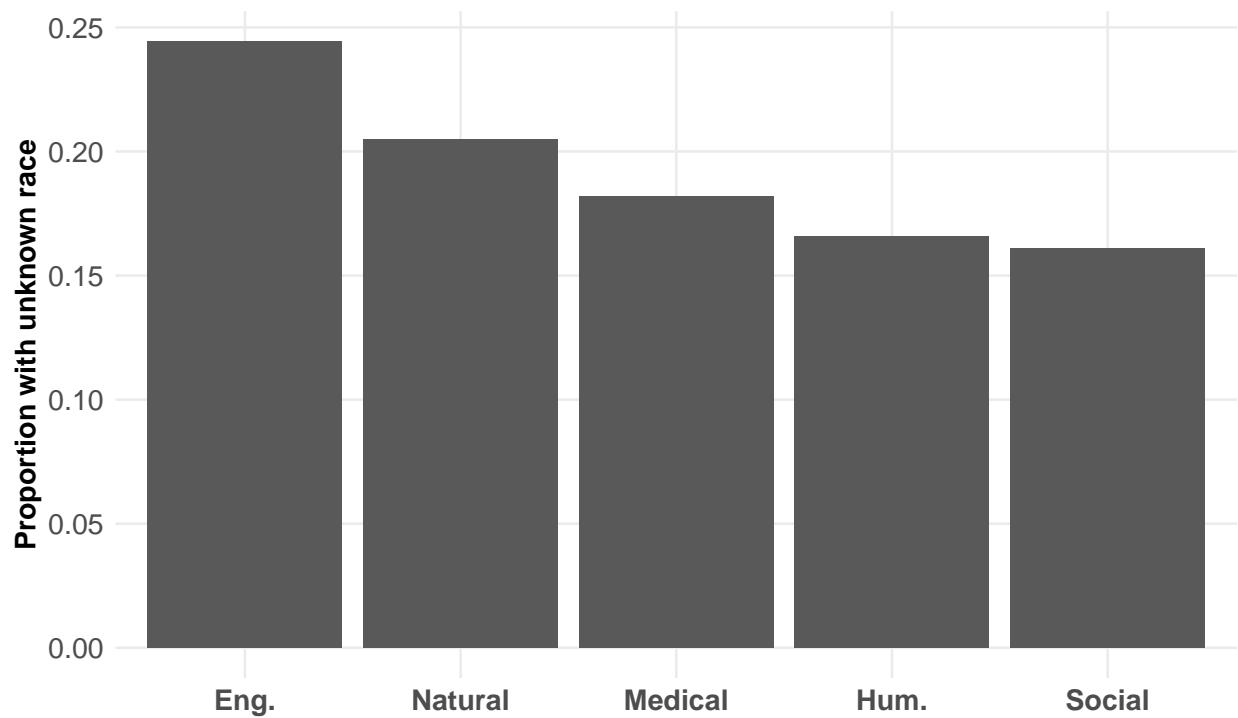


Figure B.6: Proportion of faculty with unknown race by discipline.

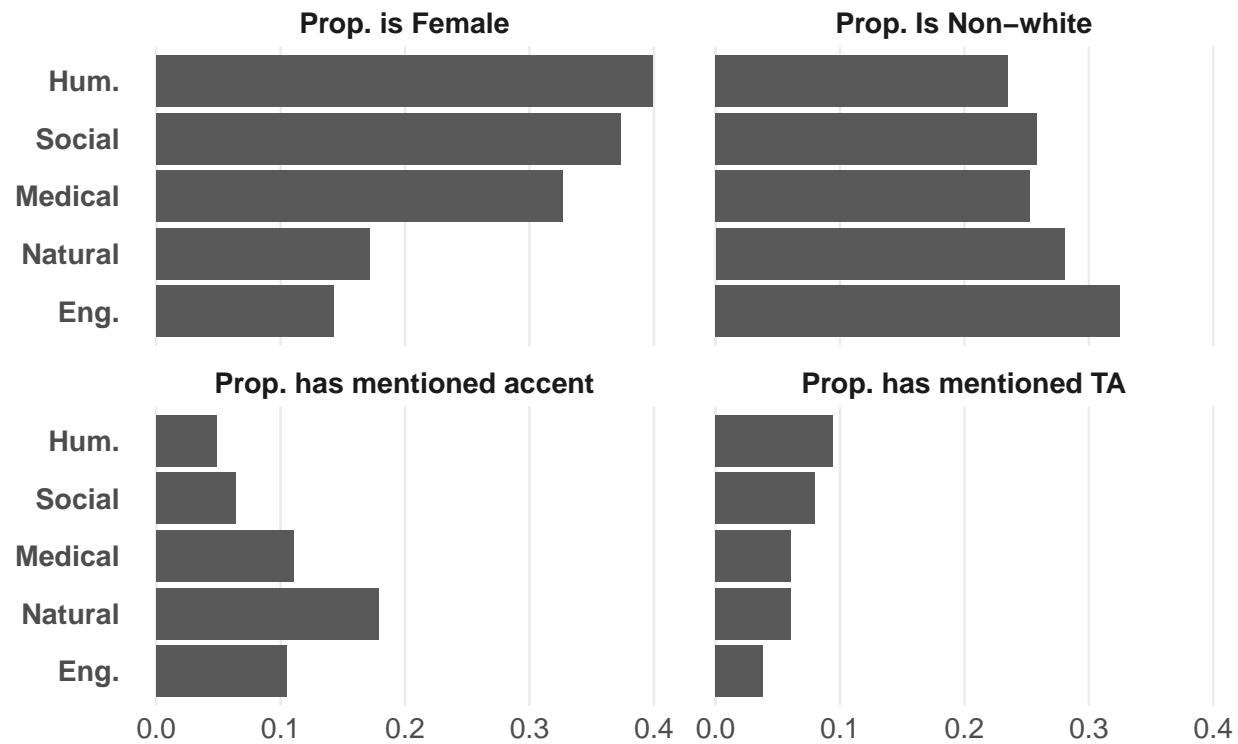


Figure B.7: Faculty demographics by discipline.

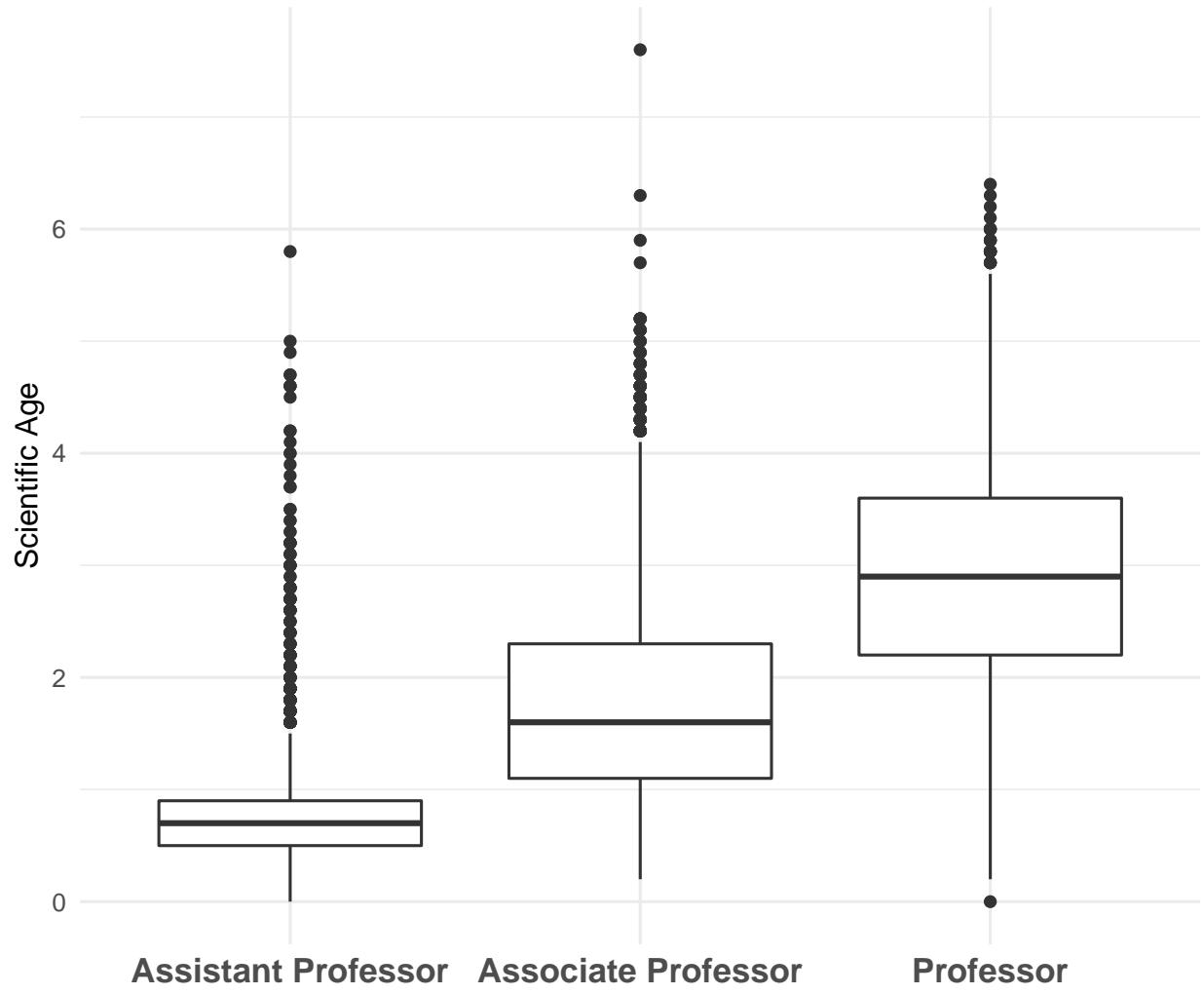


Figure B.8: **Distribution of faculty scientific age by rank.** Boxplots for the distribution of scientific age (years since earning PhD or other terminal degree) and the rank of faculty, as indexed in AA2017.

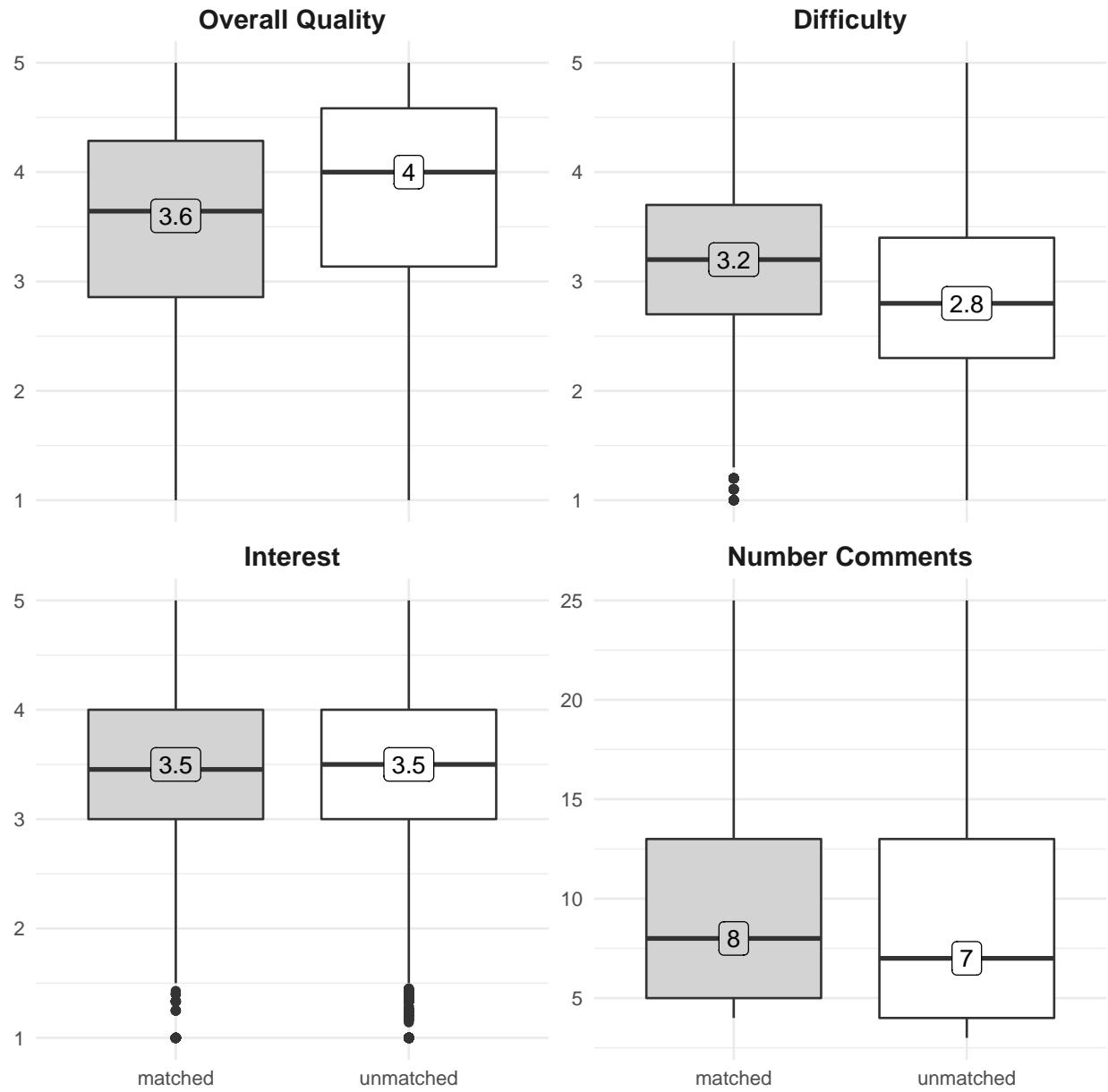


Figure B.9: **Matched rated more poorly, more difficult, with more comments.** Boxplots detailing the distribution of overall quality, difficulty, interest, and the number of comments for individuals from RMP2018 were unmatched (white) vs. matched to records in AA2017 (dark grey). Labels in each boxplot state the median.

B.3 Tables

Table B.1: **Description of relevant variables from Academic Analytics 2016 dataset.** Source indicates where the variable was collected by Academic Analytics. “Submitted” indicates that the variable was given to Academic Analytics by the institution.

Variable	Source	Description
Person name	Submitted; Cross-Ref, Institutional Websites	Full name of individual, as collected or submitted to Academic Analytics.
Institution Name	Submitted, Institutional Websites	Name of the institution in which the individual holds an affiliation
Degree Year	Submitted; Institutional Websites; CVs	Year in which the individual obtained their terminal degree (usually a doctorate)
Program Name	Submitted; Institutional websites	The name of the program to which the individual is affiliated, usually the name of their home department
Level Name	1 Assigned by AA	Disciplinary classification of the individual, one of 172 low-level classifications created by AA and assigned based on the individuals’ program affiliation
Article Count	CrossRef	Count of peer-reviewed journal articles published in 2013, 2014, 2015 and 2016. For coauthored articles, all authors are credited
Conf. Proceeding Count	CrossRef	Count of indexed conference proceedings published in 2013, 2014, 2015, and 2016. For coauthored articles, all authors are credited
Citation Count	CrossRef	Citations to articles and proceedings that were published in 2012, 2013, 2014, 2015, and 2016; data is derived from the CrossRef citation-linking network. Self-citations are included
Book Count	British Library, Baker & Taylor, and AA’s internal collection efforts	Count of time appearing as author, co-author, editor, co-editor, and translator of books published in 2006-2016 (inclusive). Introductions, forewords, afterwards, and citations are not included in the 2014 Academic Analytics dataset. Due to their limited use of DOIs, chapters are limited.
Grant Count	Publicly available databases and FOIA requests	Count of grants data from 13 federal agencies and two non-federal sources matched to principal investigators. For NIH, NSF and NOAA grants, matching includes co-principal/multi-principal investigators
Award Count	AA’s Internal collection efforts	The count of awards among honorific awards from 821 governing societies that are open to all people in a discipline, sub discipline, or a large subset of people at a national or international level, and that can be matched to individuals appearing in the Academic Analytics dataset

Table B.2: Description of relevant variables extracted from RateMyProfessor.com reviews.

Variable	Description
Comment	An optional textual review of the teacher—in the newest iteration can contain up to 350 characters
Course	A code indicating the course that the student took with the teacher. Can be entered manually or selected from a list of options from a history of course codes used in previous reviews of the professor
Overall quality	A rating of the overall quality of the professor on an ordinal scale of 1 to 5. When posting a review on the RMP website the user is prompted with the question "How would you rate this professor as an instructor?". A rating of 1 indicates that the professor is poor quality while a rating of 5 indicates high quality.
Level of difficulty	A rating of the level of difficulty of the professor on an ordinal scale of 1 to 5. When posting a review a user is prompted with the question "How hard did you have to work for this class?". A rating of 1 indicates that the course is not difficult whereas a rating of 5 indicates that the course is very difficult.
Tags	The user is prompted to select at most three of a list of 23 predefined tags that describe the characteristics of the professor and their course
Interest	A legacy variable, this feature appears to have been removed at some point in the middle of 2017 and no longer features on reviews posted to the website. In older reviews this variable was intended to measure the interest that the student had in the subject of the course on a qualitative ordinal scale containing values of "Low", "Meh", "Sorta Interested", "Really into It" and "It's My Life"

B5 Table. Results of Kendall Rank Tau. Results of non-parametric test of dependence between the RMP2018 overall quality rating of the population, tested against each of the AA2017 research indicators.

Table B.3: Description of relevant variables extracted from RateMyProfessor.com teacher profiles.

Variable	Description
Name	Name of the teacher, as listed on the website, added by the users
Department	Department to which the teacher is affiliated as listed on the website and added by users
School	The School that the teacher is affiliated with, added by users
Overall Quality	An average of the individual ratings of a teacher's overall quality, intended as a general quality indicator. Individual ratings are on an ordinal scale between one and five, where one is considered poor quality, and five is considered high quality
Level of Difficulty	An average of the individual ratings of a teacher's level of difficulty, intended as an indicator of the difficulty of the teacher's courses. Individual ratings are on an ordinal scale between one and five, where one is considered easy and five is considered difficult
Chili	Presence of a "chili pepper" on the website, which indicates "hotness", or "attractiveness". When submitting a review, the user is asked to select between a positive and negative "hotness" rating. When the number of positive hotness ratings is greater than the number of negative hotness ratings, then that professor's profile is marked with a chili pepper
Tags	Individual reviewers can select from at most three of twenty pre-defined TAGS relating to characteristics of the professor or the course. These tags are then aggregated at the level of the professor. While the RMP website includes counts for the number of times these tags have been applied, we only capture a Boolean value indicating the presence of a tag

S7 Table Results of multiple linear regression model using continuous research performance indicators. For binary variables, "false" is always used as the reference level. For Gender, "female" is used as the reference. For "Rank", "Assistant" is used as the reference. For Discipline, "Engineering" is set as the reference. For Uni. Control, "Private" is used as the reference. For Uni. Type, "Not R1" is used as the reference. Continuous field-normalized values are used for research performance indicators. Estimates are listed first, followed by 95th percentile

Table B.4: **Results of multiple linear regression model.** For binary variables, “false” is always used as the reference level. For Gender, “female” is used as the reference. For race, ”Non-White” is used as the reference. For rank, “Assistant” is used as the reference. For Discipline, “Engineering” is set as the reference. For Uni. Control, “Private” is used as the reference. For Uni. Type, “Not R1” is used as the reference. For all research indicators, “Low” is used as the reference. Estimates are listed first, followed by 95th percentile confidence intervals. Results here are identical to those in Fig. 4.1.A

	<i>Dependent variable:</i>
	overall
Is Male	0.106*** (0.084, 0.127)
Scientific Age	-0.133*** (-0.147, -0.119)
Mentions Accent = True	-0.172*** (-0.203, -0.142)
Has Chili Pepper	0.417*** (0.393, 0.442)
Rank = Associate	0.047*** (0.017, 0.076)
Rank = Full	0.137*** (0.100, 0.174)
Race unknown	0.046*** (0.019, 0.074)
Race Lilely White	0.118*** (0.096, 0.140)
Difficulty	-0.391*** (-0.401, -0.381)
Student Interest	0.329*** (0.319, 0.339)
Mentions TA = True	-0.184*** (-0.219, -0.149)
Citedness = Moderate	-0.024 (-0.054, 0.006)
Citedness = High	-0.033 (-0.085, 0.018)
Output = Moderate	0.034* (-0.001, 0.069)
Output = High	0.022 (-0.033, 0.077)
Grants Held = Moderate	-0.002 (-0.026, 0.023)
Grants Held = High	0.031 (-0.024, 0.086)
Awards Won = Moderate	0.010 (-0.011, 0.031)
Awards Won = High	0.010 (-0.071, 0.091)
Humanities	0.184*** (0.142, 0.225)
Medical Sci.	0.105*** (0.056, 0.153)
Natural Sci.	0.070*** (0.037, 0.103)
Social Sci.	0.044** (0.008, 0.079)
Uni. Type = R1	-0.030*** (-0.051, -0.010)
Uni. Control = Public	-0.084*** (-0.110, -0.059)
#Reviews	0.006*** (0.004, 0.007)
Constant	3.173*** (3.114, 3.231)
Observations	18,973
R ²	0.514
Adjusted R ²	0.514
Residual Std. Error	0.643 (df = 18946)
F Statistic	771.537*** (df = 26; 18946) (p = 0.000)

Note:

*p<0.1; **p<0.05; ***p<0.01

Tested Variable	Z-Value	Estimated Tau	P-Value
Article Count	11.349	0.0453	<2.2e-16
Award Count	2.1285	0.0093	0.0333
Book Count	10.122	0.0444	<2.2e-16
Citation Count	3.2322	0.0130	0.00123
Proceeding Count	-8.990	-0.0411	<2.2e-16
Grant Count	-5.9901	-0.0264	2.097e-09

Table B.5: **Results of multiple linear regression model with interactions.** For binary variables, “false” is always used as the reference level. For Gender, “female” is used as the reference. For race, ”Non-White” is used as the reference. For rank, “Assistant” is used as the reference. For Discipline, “Engineering” is set as the reference. For Uni. Control, “Private” is used as the reference. For Uni. Type, “Not R1” is used as the reference. For all research indicators, “Low” is used as the reference. Estimates are listed first, followed by 95th percentile confidence intervals. Interactions are marked as variables separated by a ”**”.

	<i>Dependent variable:</i>
	overall
Is Male	0.133** (0.027, 0.239)
Scientific Age	-0.124*** (-0.153, -0.095)
Mentions Accent = True	-0.161*** (-0.225, -0.097)
Has Chili Pepper	0.424*** (0.381, 0.467)
Rank = Associate	0.021 (-0.029, 0.071)
Rank = Full	0.131*** (0.065, 0.197)
Race unknown	0.017 (-0.035, 0.069)
Race Lilely White	0.082*** (0.041, 0.123)
Difficulty	-0.392*** (-0.402, -0.382)
Student Interest	0.329*** (0.319, 0.339)
Mentions TA = True	-0.183*** (-0.218, -0.148)
Citedness = Moderate	-0.023 (-0.053, 0.007)
Citedness = High	-0.034 (-0.086, 0.018)
Output = Moderate	0.035* (-0.0003, 0.070)
Output = High	0.023 (-0.031, 0.078)
Grants Held = Moderate	-0.0003 (-0.025, 0.024)
Grants Held = High	0.032 (-0.023, 0.087)
Awards Won = Moderate	0.010 (-0.011, 0.030)
Awards Won = High	0.006 (-0.076, 0.087)
Humanities	0.292*** (0.206, 0.378)
Medical Sci.	0.182*** (0.079, 0.286)
Natural Sci.	0.140*** (0.054, 0.226)
Social Sci.	0.093** (0.011, 0.176)
Uni. Type = R1	-0.031*** (-0.051, -0.010)
Uni. Control = Public	-0.086*** (-0.111, -0.060)
#Reviews	0.006*** (0.004, 0.007)
Is Male * Has Chili	-0.008 (-0.059, 0.044)
Is Male * Humanities	-0.145*** (-0.237, -0.052)
Is Male * Medical	-0.092 (-0.210, 0.026)
Is Male * Natural	-0.082* (-0.175, 0.011)
Is Male * Social	-0.050 (-0.140, 0.041)
Is Male * Scientific Age	-0.011 (-0.043, 0.021)
Is Male * Race unknown	0.041 (-0.020, 0.103)
Is Male * Race Likely White	0.051** (0.002, 0.100)
Is Male * Mentions Accent	-0.014 (-0.086, 0.059)
Is Male * Rank Associate	0.039 (-0.023, 0.100)
Is Male * Rank Full	0.011 (-0.066, 0.088)
Constant	3.140*** (3.039, 3.242)
Observations	18,973
R ²	0.515
Adjusted R ²	0.514
Residual Std. Error	0.643 (df = 18935)
F Statistic	543.120*** (df = 37; 18935) (p = 0.000)

Note:

*p<0.1; **p<0.05; ***p<0.01

confidence intervals.

	<i>Dependent variable:</i>
	overall
Is Male	0.106*** (0.083, 0.128)
Scientific Age	-0.128*** (-0.143, -0.114)
Mentions Accent = True	-0.201 *** (-0.232, -0.170)
Has Chili Pepper	0.405*** (0.379, 0.431)
Rank = Associate	0.016 (-0.013, 0.045)
Rank = Full	0.108*** (0.071, 0.145)
Difficulty	-0.405*** (-0.416, -0.394)
Student Interest	0.365*** (0.354, 0.376)
Mentions TA = True	-0.181*** (-0.217, -0.145)
Citations	-0.002 (-0.007, 0.002)
Publications	0.006 (-0.004, 0.017)
Awards	0.008** (0.001, 0.014)
Grants	0.001 (-0.003, 0.005)
Humanities	0.213*** (0.176, 0.250)
Medical Sci.	0.116*** (0.064, 0.167)
Natural Sci.	0.079*** (0.044, 0.113)
Social Sci.	0.065*** (0.029, 0.101)
Uni. Type = R1	-0.031*** (-0.052, -0.010)
Uni. Control = Public	-0.079*** (-0.106, -0.052)
#Reviews	0.006*** (0.004, 0.008)
Constant	3.279*** (3.227, 3.331)
Observations	17,600
R ²	0.513
Adjusted R ²	0.512
Residual Std. Error	0.643 (df = 17579)
F Statistic	925.473*** (df = 20; 17579) (p = 0.000)

Note:

*p<0.1; **p<0.05; ***p<0.01

S8 Table. Little evidence of multicollinearity in discrete regression model. Generalized and adjusted variance inflation factor scores for the regression model of overall teaching quality with discrete research indicators.

	GVIF	Df	GVIF^(1/(2*Df))
Gender	1.115	1	1.056
Scientific Age	2.312	1	1.521
Mentions Accent	1.101	1	1.049
Has Chili Pepper	1.178	1	1.085
Rank	2.146	2	1.210
Race	1.067	2	1.016
Difficulty	1.121	1	1.059
Interest	1.121	1	1.059
Mentions TA	1.097	1	1.047
Citedness	2.766	2	1.290
Output	1.988	2	1.187
Grants	1.582	2	1.121
Awards Won	1.190	2	1.044
Discipline	2.251	4	1.107
Uni. Type	1.054	1	1.027
Uni. Control	1.019	1	1.010
Review Count	1.125	1	1.061

S9 Table. Little evidence of multicollinearity in continuous regression model. Generalized and adjusted variance inflation factor scores for the regression model of overall teaching quality with continuous research indicators.

	GVIF	Df	GVIF^(1/(2*Df))
Gender	1.115	1	1.056
Scientific Age	2.073	1	1.440
Mentions Accent	1.061	1	1.030
Has Chili Papper	1.175	1	1.084
Rank	2.029	2	1.193
Difficulty	1.119	1	1.058
Interest	1.120	1	1.058
Mentions TA	1.096	1	1.047
Norm. Citations	1.423	1	1.193
Norm. Publications	1.525	1	1.235
Norm. Awards	1.123	1	1.060
Norm. Grants	1.063	1	1.031
Discipline	1.200	4	1.023
Uni. Type	1.034	1	1.017
Uni. Control	1.019	1	1.010
Review Count	1.121	1	1.059

S10 Table Little difference between population of matched and unmatched Academic Analytics faculty. Shown are characteristics of the faculty in the matched dataset and of the tenure and tenure-track faculty of the unmatched Academic Analytics dataset.

Variable	Measure	Matched	Unmatched
Gender	% Male	71.5	64.2
	% Female	28.5	35.8
Inferred Race	% White	54.4	56.9
	% Non-White	26.9	25.5
	% Unknown	18.8	17.5
Professional Rank	% Assistant	17.8	23.4
	% Associate	35.9	27.3
	% Full	46.4	49.3
Scientific Age	1st Quartile	11	12
	Median	20	21
	3rd Quartile	30	21
University Type	% R1	69.9	39.7
	% Not R1	30.1	30.3
University Control	% Private	15.7	28.4
	% Public	84.3	71.6
Discipline	% Engineering	10.7	13.2
	% Social Science	21.9	25
	% Medical Science	5.5	26.7
	% Natural Science	33.6	22.9
	% Humanities	21.3	12.3
Publications	% No Publications	19.6	14.2
	% Moderate Publications	73.1	76.8
	% High Publications	7.2	9.5
Citations	% No Citations	28	19
	% Moderate Citations	66.2	72.7
	% High Citations	5.8	8.4
Grants	% No Grants	65.8	60.6
	% Moderate Grants	31.1	35.4
	% High Grants	3.1	4
Awards	% No Awards	64.6	67.5
	% Moderate Awards	33.9	28.9
	% High Awards	1.5	3.6

S11 Table Average counts of research item, by discipline

Discipline	Articles	Proceedings	Books	Awards	Citations	Grants
Engineering	12.034	5.921	0.348	0.805	189.504	1.566
Humanities	1.451	0.044	1.721	0.846	3.428	0.114
Medical Sciences	11.917	0.398	0.304	0.616	207.930	1.345
Natural Sciences	15.319	1.931	0.340	0.755	474.638	1.452
Social Sciences	5.854	0.182	0.990	0.489	62.174	0.357

Appendix C

Study 3: Measuring disagreement

C.1 Text

Disciplinary artefacts in queries

The incidence of citation sentences returned by each query are not uniform across all signal and filter term combinations. There are far more publications from the Biomedical and Health Sciences (Bio & Health) than in other fields, accounting for a total of 47.5% of all publications indexed in the Web of Science Database; in contrast, publications in Math and Computer Sciences (Math & Comp) comprise a far smaller proportion of the database, accounting for only 3.1 percent.

Even accounting for the different number of publications per field, we still observe that some signal terms appear more in certain fields than expected, often as a result of differences in disciplinary jargon, topics, and norms (Figure C.2b). For example, there are more conflict* citances than expected in Social Sciences and Humanities (Soc & Hum), where it often appears in relation to conflict as a topic of study, such as the study of international conflict, conflict theory, or other interpersonal conflicts (Table C.3, I). Similarly, disprov* citances appear more often in Math & Comp, where disprove is often used in relation to proving or disproving theorems and other mathematical proofs (Table C.3, II). Other notable differences are *controvers** citances appearing more often in Bio & Health, *debat** appearing most often in the Life and Earth Sciences (Life & Earth), and *disagree** appearing most in Physical Sciences and Engineering (Phys & Engr).

Filter terms are also non-randomly distributed across fields (Figure C.2c). For example, the *+ideas* filter term appears more often than expected in Soc & Hum, possibly as a result of disciplinary norms around use and discussion of abstract theories and concepts (Table C.3, III). In contrast, *+methods* is over-represented in Phys & Engr and Math & Comp, likely a result of

these field's focus on methods and technique (Table C.3, IV). Notably, *+studies* and *+results* are under-represented among Math & Comp publications, whereas *+ideas* and *+methods* are under-represented among papers in the Biomedical and Health Sciences.

The complexity of disciplinary differences between queries is made apparent when examining combinations of signal and filter phrases (Figure C.2d). While there are no obvious or consistent patterns between fields, there are particular differences by field. For example, compared to all other fields *contravers** citances are over-represented in Bio & Health (Table C.3, V), except for *contravers** *+ideas*, which is instead slightly over-represented in Life & Earth. In contrast, *disagree** citances are under-represented in Bio & Health, but over-represented in Life & Earth and Phys & Engr (Table C.3, VI). In some cases, the specific signal +filter term combination has a massive effect, such as *no consensus* *+ideas*, which is heavily over-represented in Soc & Hum (Table C.3, VII), whereas all other signal and filter term combinations are under-represented. Similarly, *contradict** *+ideas* and *contradict** *+methods* are over-represented in Math & Comp (Table C.3, VIII), whereas *+results* and *+studies* are underrepresented. Similar intricacies abound across the 325 combinations of signal term, filter term, and field, and demonstrate the importance that field plays in the utility and significance of our signal and filter terms.

Especially at the fine-grained field level, methodological artefacts can drive differences we observe between meso-fields. For example, in Soc & Hum, one of the meso-fields with the most disagreement was composed of papers from journals such as “Political Studies” and “International Relations”—journals and fields for which “debates” and “conflicts” are objects of study, which could confound the *debat** and *conflict** signal terms. This is demonstrated by the following invalid citation sentences,

1. “Since the late-1990s, there has been even less room for **debate** within the party (...).”
2. “Indeed, this whole idea harkens back to the badges of slavery of the 13th Amendment and the **debate** in (...).”

3. “In political behaviour literature, we refer to such **conflictive** opinions as “ambivalence” (...).”
4. “In politics as usual, people often do not like to see the **conflicts** and disagreements common to partisan debate (...).”

Even though terms such as “public debate”, and “parliamentary debate” were excluded (Table 5.1), the *debat** signal terms were over-represented in Soc & Hum (Figure C.2); conflict* was also over-represented to a lesser extent. Interpretation of the results for main and meso-fields needs to be moderated by these, and other confounding artefacts (see the Supporting Information for further discussion).

Robustness

To test the robustness of our results, we compare findings using the 23 queries with greater than 80 percent validity to those using the 36 queries with greater than 70 percent validity. The new queries include *contradict* _standalone_, contrary +studies, contrary + +methods, conflict* +results, disagree* +methods, disagree* +ideas, disprov* +methods, disprov* +ideas, refut* +studies, refut* +results, refut* +ideas, debat* +ideas, and questionable +ideas*. Queries above the 80 percent validity cutoff account for about 450,000 citances; the addition of 13 queries above the 70 percent cutoff bring this total to about 650,000.

We find that our findings are robust whether using an 80 percent or 70 percent validity cutoff. Relaxing the validity cutoff results in including more citances, inflating the share of disagreement across all results. However, the qualitative interpretation of these results does not change (Table C.4). The 80 percent and 70 percent cutoffs both produce the same ordering of fields from most to least disagreement. Similarly, the ordering of fields from high-to-low disagreement holds between the 80 percent and 70 percent cutoff for all quantities presented here, including the average change per year, the ratio of disagreement between non-self-citation and self-citation, and the average change in disagreement per age bin. Some fields gain more from these new queries more

than others, manifesting in more or less intense field differences in findings. For example, Soc & Hum gains a full 17 percentage points in overall disagreement with the 70 percent threshold, with the increase across all fields at only 8 points. Similarly, the ratio of non-self-citation to self-citation is 2.2x for Math & Comp with the 80 percent cutoff, but only 1.3x for the 70 percent cutoff. This likely stems from the relative skew of the added queries to certain fields, leading to larger gains in disagreement.

Disagreement by contextual factors

Paper age

Other contextual factors may relate to disagreement. For example, authors may disagree with more recent papers at different rates than older ones. We quantify disagreement based on the age of a cited paper, relative to the citing paper. Following a brief bump, or increase in disagreement (at 05-09 years), older papers tend to be disagreed with less (Figure C.4), a pattern driven by field differences. Low consensus, high complexity fields such as Soc & Hum and Bio & Health both exhibit a clear decreasing pattern, with falling disagreement as the paper ages. Life & Earth, in the middle of the hierarchy, repeats this pattern, but only after a period of stability in disagreement in the first ten years. Disagreement instead steadily increases over time in high consensus and low complexity fields such as Phys & Engr and Math & Comp.

Position in the paper

Disagreement is not equally likely to occur throughout a paper. Investigating the distribution of disagreement citations across papers, we find that they are far more likely to occur in the beginning of a paper, likely in the introduction, and then towards the end, likely the discussion section (Figure C.5). However, the precise patterns differ by field. For example, in Soc & Hum, disagreement citations are more evenly distributed through the first 40 percent of the paper, likely reflecting

longer introductions and literate reviews. In contrast, disagreement in Bio & Health and Life & Earth are more likely to appear near the end of a paper, potentially reflecting the distinct article structure in these fields.

Gender of citing-paper author

Men and women authors may be more likely disagree at different rates. To test this, we infer a gender for the first and last authors of papers with a disagreement citance published after 2008, determined based on the author's first name as in past work [28]. Overall, there is little difference in the rate of disagreement between men and women first and last authors (Figure C.6). The one exception is Math and Computer science, where women last authors disagree 1.2 times more often than men, though the rate of disagreement is small, and driven by a small number of instances.

Self citation

We investigate the extent to which disagreement differs based on self-citation, that is when there is in the presence of overlap between the authors in a citing and the cited papers. We would expect that authors will be less likely to cite their own work in the context of disagreement, which is affirmed by our indicator. Overall, the rate of disagreement among non-self-citing papers is 2.4 times greater than for self-citation citances (Figure C.3). The field with the largest difference is Bio & Health (2.5 times greater), followed by Phys & Engr (2.2 times greater), Math & Comp (2.2 times greater), Life & Earth (1.9 times greater), and finally, Soc & Hum (1.6 times greater).

Papers with most disagreement citances

While the extensive manual validation of our queries and results ensures the robustness of our analysis at an aggregate scale, the list of publications issuing most disagreement citations does reveal that it remains difficult to separate research object from commentary on cited material. Table C.6 shows the papers that issue the most disagreement citations; going through these papers

reveals several artefacts. “Debating” and “debates” are, for example, the object of study in Alén, Domínguez, & De Carlos (2015) and Doody & Condon (2012), and the citances in the paper reflect this, e.g. “students also seem to both enjoy debates and recognise their value” and “debate is effective in helping students learn a discipline and demonstrate the ability to read and write critically.” Controversy, likewise, is the subject in Nam (2014) and Colston & Vadjunec (2015), as are environmental conflicts in Stepanova & Bruckmeier (2013). Bruschke & Divine (2017) makes frequent mention of the first televised US presidential debate. This leaves French & Koeberl (2010), Kalter (2003) and Millan (2006), three publications that do not immediately appear focused on subjects that would trigger our queries.

French & Koeberl (2010) discusses methods for identifying meteorite impact structures on earth, “as well as an overview of doubtful criteria or ambiguous lines of evidence that have erroneously been applied in the past”, and this paper indeed cites many sources in the context of controversy, e.g. “the identification of such glasses as impact or non-impact products is difficult and commonly controversial,” “the impact origin for many glasses still remains controversial and unconfirmed,” “there are also debates about the formation of maskelynite itself” and “the nature, characteristics, and causes of these changes have been widely studied and are still being debated.”

Kalter (2003), while classified as a full-length article, is a book that was also included in a special issue of the journal Neurotoxicology and Teratology. Considered a pillar in the study of congenital abnormalities, its exceptional length may account in part for its high number of disagreement citances. One of these citances describes “a discussion—debate better characterizes it—that took place in 1953,” but many others indeed refer to scientific disagreements within the field of study, e.g. “a Mayo Clinic study seemed to agree, despite conflicting evidence,” “a contrary finding came from Scotland, another high-risk region,” “an early analysis, as well as a later one, disagreed” and “earlier retrospective and prospective studies had been contradictory.”

Millan (2006) is a review article making a case for multi-target agents to treat depressive

states. It likewise introduces a number of citations that indeed signify disagreement in the scientific literature, e.g. “the gravity of cognitive impairment in young patients is still debated,” “this notion remains somewhat controversial,” “its precise degree of efficacy in this regard is still debated” and “for recent critical discussions of these controversial issues—from a variety of viewpoints—see [...].” However, a few false positives also occur, when citing the work of an author by the name of DeBattisa, whose name was caught by our *debate** query.

Papers that received the most disagreement citations

We also examine disagreement from the cited paper perspective, that is by looking at those papers that received the most disagreement citations. Table C.5 lists these papers, and again we reveal issues with methodological artefacts, but also highlight interesting instances of controversy in the literature. The majority of these publications relate to plate tectonics, and in particular, the North China Craton. These papers include Zhao, Sun, Wilde, & Sanzhong (2005), Kusky & Li (2003), Zhao, Wilde, Cawood, & Sun (2001), Zhai & Santosh (2011), Wilde, Zhao, & Sun (2002) and Kusky (2011). Li et al. (2008) also appears to be closely related. Several of these publications have authors in common, notably Zhao (3), Wilde (3), Sun (3), Kusky (2) and Li (2). Zhai & Santosh (2011) summarizes the situation as follows:

”A long controversy and debate surround the evolution of the NCC, particularly the timing and tectonic processes involved in the amalgamation of the Eastern (Yanliao) and Western (Ordos and Yinshan) Blocks along the Central Orogenic Belt. One school of thought proposes an east-directed subduction of an old ocean, with final collision between the two blocks at 1.85 Ga [several citations to Zhao’s work]. In contrast, some others suggest a westward subduction, with final collision between the two blocks to form the NCC at 2.5 Ga [several citations to Kusky’s work].”

The majority of disagreement citations to these papers also mention this long-standing scientific controversy, including phrases such as,

- “the number of continental blocks and the mechanism by which they were welded together to form the coherent basement remain controversial,”

- “tectonic history of this central region is in debate,”
- “it is currently debated how the collisional processes proceeded,”
- “controversy still remains as to the timing and tectonic processes involved,”
- “it still remains controversial as to how the craton should be subdivided and where the collisional boundaries are located,”
- “models that evaluate the Paleoproterozoic crustal evolution of the NCC remain controversial,”
- “the timing of the collision between these blocks remains controversial,”
- “this controversy is also reflected in various tectonic models for the NCC” and
- “the time of the amalgamation between the Eastern Block and the Western block is still debated,”

These phrases are often accompanied by a large number of citations to papers by Zhao and Kusky, with both authors either explicitly posed as opposing one another like in the example from Zhai & Santosh (2011) above, or citations to their papers grouped together as if to present a large body of conflicting literature. It is clear that this scientific controversy dominates scholarship on the North China Cranton, to the point that even Wikipedia mentions it extensively [714]. It should be noted that the works by Zhao and Kusky also receive a fair share of citations that mention their empirical and theoretical contributions, without the context of disagreement, making material research contributions to both their preferred model and the research on the cranton at large. The high share of disagreement citations to these papers appears to stem from the highly divided nature of their research field.

The three remaining papers cover different topics. Munro (2003) reviews literature on lipids in cell membranes of eukaryotic cells and discusses the (non)existence of lipid rafts. It points out that, regardless of enthusiasm for the lipid raft model in the research community, observations of these raft structures are problematic and several factors exist that cast doubt on the model. The existence of lipid rafts should therefore be treated as hypothetical rather than established

fact. Citances that mention this paper in the context of disagreement also appear to use it as an exemplar of controversy in the field, with phrasing such as,

- “the entity of lipid microdomains is controversial,”
- “the existence of noncaveolar lipid rafts in vivo is still debated,”
- “this early operative definition of lipid rafts was subject of much debate,”
- “the lipid raft concept has been controversial since it was introduced several decades ago”

and

- “a number of points have been recently debated in the literature.”

Even many citances that do not qualify as disagreement per our operationalization appear to still signal it, e.g.

- “rafts still remain a hypothesis,”
- “evidence that lipid rafts exist in living cells remains elusive,”
- “the classical perception of rafts as stable entities within the fluid bilayer has provoked some opposition,”
- “the biological substrate for this notion is not clearly defined” and
- “isolation of DRM or lipid rafts is however a delicate matter.”

This paper, specifically, serves as a focal point of controversy within the community researching lipid rafts precisely because its primary purpose appears to be to create this controversy; raising concerns about the lipid raft model and calling for a reevaluation of its canonicity in the face of shaky foundational evidence of the existence of these rafts. This is different from the North China Cranton papers, where controversy is long established in the field.

Murphy, Birn, Handwerker, Jones, & Bandettini (2009) focuses on a pre-processing method used in low-frequency fMRI research, called global signal regression. This paper alleges that the method is inadequate and “may cause spurious findings of negatively correlated regions in the brain.” This paper appears to have heralded a change in how data is handled in this research field, with many citances mentioning it to explain why data was processed differently, e.g.

- “the global signal was not regressed out due to its controversial biological interpretations,”
- “given the controversy of removing the global signal in the preprocessed rs-fMRI data, we did not regress the global signal out in the present study,”
- “global signal regression is a somewhat controversial part of the preprocessing pipeline for resting state MRI data, and was not performed in this study” and
- “given the controversy of removing the global signal in the preprocessing of R-fMRI data [...], we did not regress the global signal out.”

Other citances in the context of disagreement simply point out this controversy, e.g.

- “there is ongoing debate as to the nature of anti-correlations introduced by global signal regression”
- “in recent years there has been an ongoing debate on global signal removal in the preprocessing.”

As with the lipid raft paper, many citances that do not qualify as disagreement also embrace the controversy, as evident in phrasing such as “global-signal was not included in the model for its effects on brain–behavior correlations” and “we decided against this approach as several recent studies showed that global signal regression may significantly bias connectivity analyses.”

Finally, Debat et al. (2003) is another example of a false positive result of our approach, in which the lead author’s name activated the debate query. While this is unfortunate, our extensive manual validation of our query results shows that despite this prominent false positive, there were no large systemic flaws in our approach that might otherwise color our analysis at the aggregate level. While eliminating cited author names from citances at scale is not trivial, this example serves to stress the importance of text pre-processing.

Disagreement and citation impact

To address whether publications that received a disagreement citation were cited differently than their counterparts, we compared the number of citations received in year $t + 1$ for papers that

featured in a disagreement citation for the first time in year t , with the average number of citations received in year $t + 1$ by papers that received the exact same number of citations in year t . This over- or undercitation of individual papers that encountered disagreement can then be aggregated to arrive at the average under- or overcitation of ‘disagreed-with’ papers in general.

We define t as the time in years since publication and c as the number of citations a paper received at time t . We calculate for each combination of t and c the number of papers $p(c, t)$ that received their first disagreement citation at time t when they held c citations. Using these, we calculate the number of citations received by these papers in the year following publication, averaged across all combinations of t and c ,

$$\bar{c}_{\text{next,disagreement}} = \frac{\sum_c \sum_t p_{c,t} \bar{c}_{\text{next,disagreement},c,t}}{\sum_c \sum_t p_{c,t}}$$

In the same way, we also calculate the expected number of citations, defined using the average number of citations received by papers that received c citations in year t , regardless of whether they received a disagreement citation.

$$\bar{c}_{\text{next,expected}} = \frac{\sum_c \sum_t p_{c,t} \bar{c}_{\text{next,expected},c,t}}{\sum_c \sum_t p_{c,t}}$$

We calculated d as the ratio of these two values. When greater than one, it indicates that papers received more citations than expected in the year after having received a disagreement citation. A value less than one indicates that papers with a disagreement citation received fewer citations in the year following.

$$d = \frac{\bar{c}_{\text{next,disagreement}}}{\bar{c}_{\text{next,expected}}}$$

The results of this analysis (Table C.5) show that being cited in a context of disagreement has little to no effect on the citations received by papers in the year following them receiving (or not

receiving) a disagreement citation.

We also investigate whether disagreement relates to the number of citations a paper receives. First, we examine the citing paper perspective, identifying the 3.5 percent ($n = 126,250$) of publications that contain at least one disagreement citation in their text. Across all publications, those with at least one disagreement citation tended to receive more citations than those without disagreement in the first four years, beginning with one additional citation in the first year following publication, and expanding to a difference of about 4.7 citations by the fourth year (Figure ??), an effect that varies, yet qualitatively consistent across all fields. This effect may be confounded by article type—for example, review articles are over-represented in terms of disagreement—24.6 percent of all review articles contain a disagreement citation—and review articles are also known to be more highly cited [715]. While excluding review articles does shrink this gap, the citation count for full research articles (85 percent of all publications) remains 2.5 citations higher for those with a disagreement than for those without.

C.2 Figures

C.3 Tables

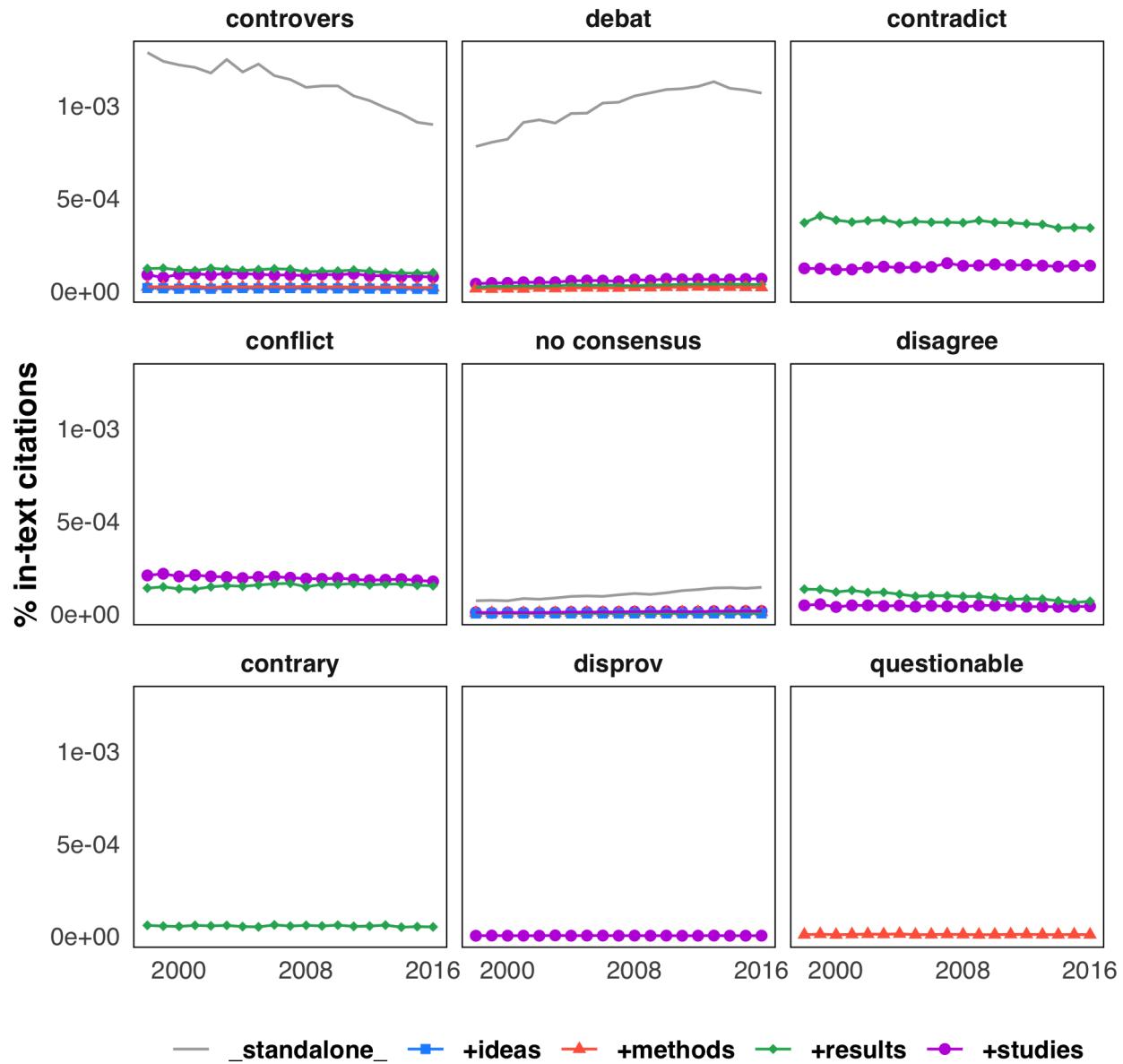


Figure C.1: Percent of all citances returned by each of the 23 queries with validity over 80 percent. Each panel corresponds to the signal phrase, and lines within each panel to filter phrases.

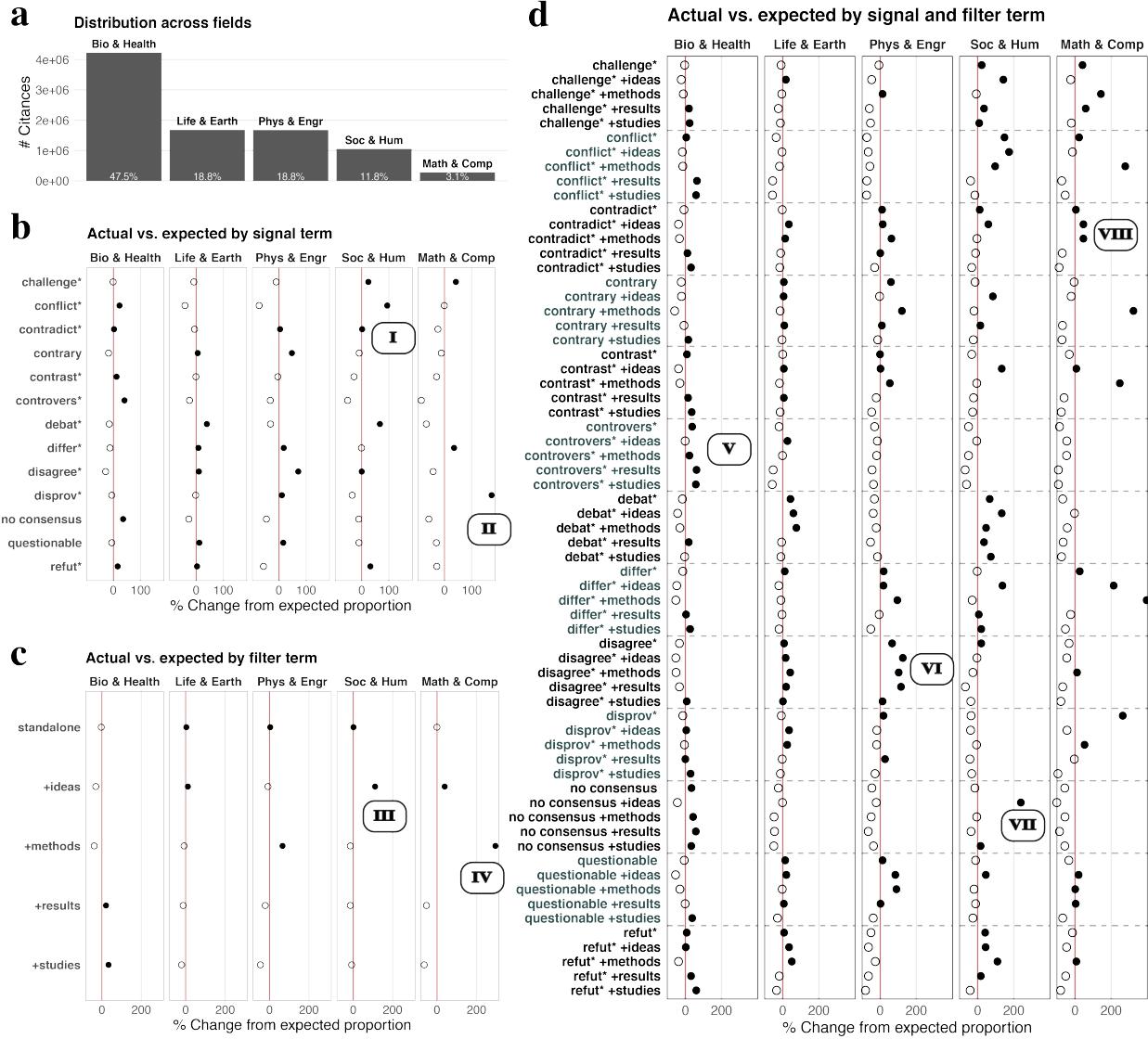


Figure C.2: Distribution of citances returned by signal/filter term queries. Callouts (I, II, ..., VIII) map to examples in Table C.3. **a.** Distribution of all disagreement citances appearing in papers across five fields: Biomedical and Health Sciences, Life and Earth sciences, Physical Sciences and Engineering, Social Sciences and Humanities, and Math and Computer Science. **b-d.** Percentage change between the actual number of citances per field and signal/filter term combination compared to the expected given the disciplinary distribution (from a). The red line corresponds to 0 percent increase between the actual and expected. White dots indicate that the citances for that signal/filter term are under-represented (lower than expected, ratio less than zero), whereas black dots indicate that citances are over-represented (more than expected). Shown aggregated across signal terms (**b**), filter terms (**c**), and for all signal/filter term combinations (**d**).

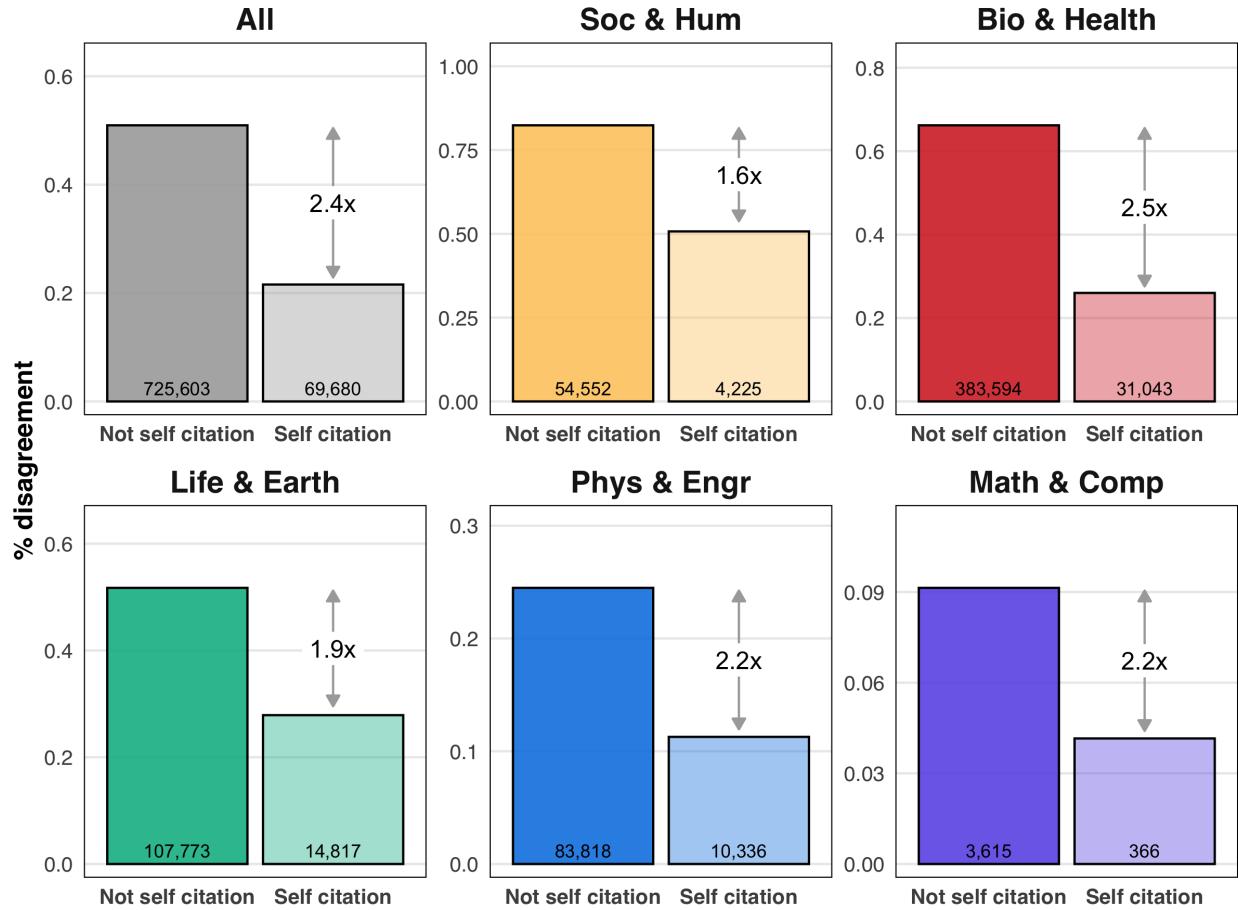


Figure C.3: Authors disagree less when citing their own work. Percentage of disagreement citances among instances of non-self and self-citation, 2000-2015. A citance is defined as a self-citation when the citing and cited paper have at least one name in common. Results are shown by field. Numbers below each bar are the number of disagreement citances. Overall, disagreement is 2.4 times more common for non-self citation than for self-citation, with variance between major fields.

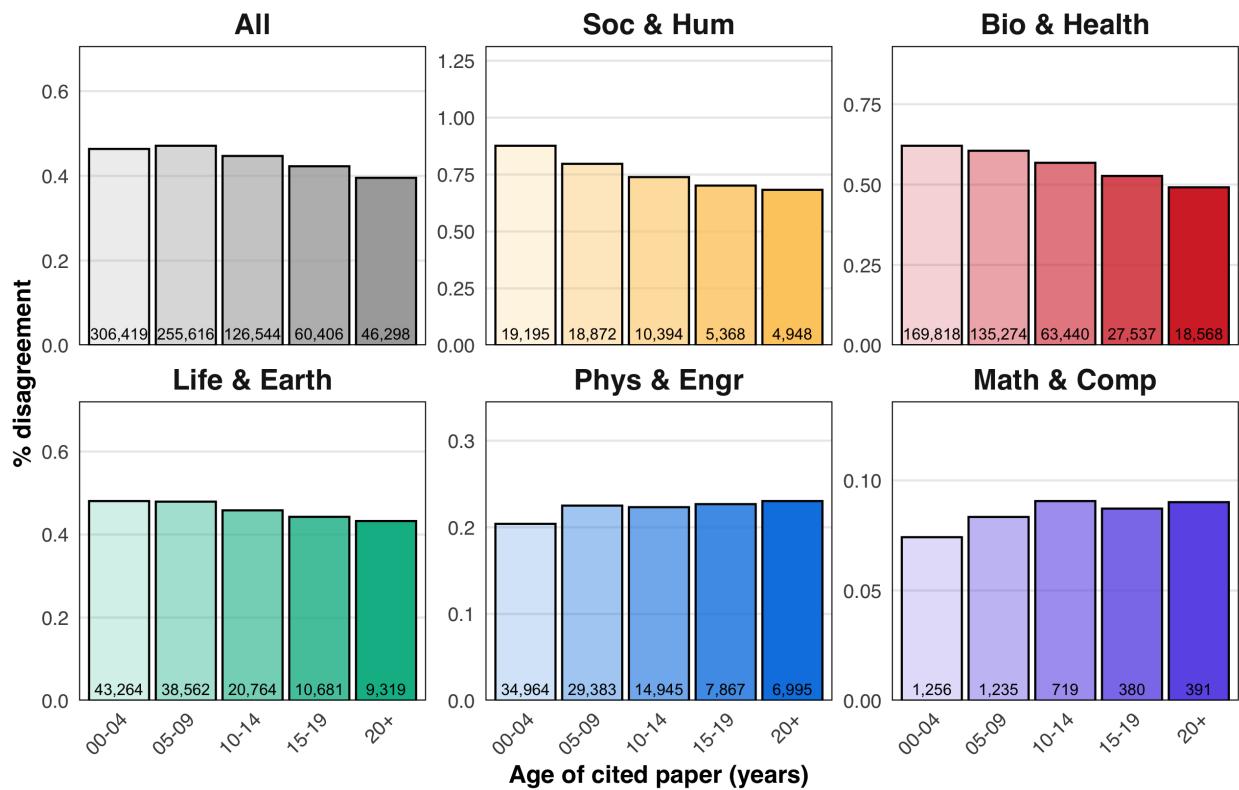


Figure C.4: Older papers are disagreed with more often. Percentage of disagreement citations by the relative age of the citing to the cited paper, in years, and high-level field, for papers published between 2000 and 2015. Intensity of color corresponds to the age category of the cited paper.

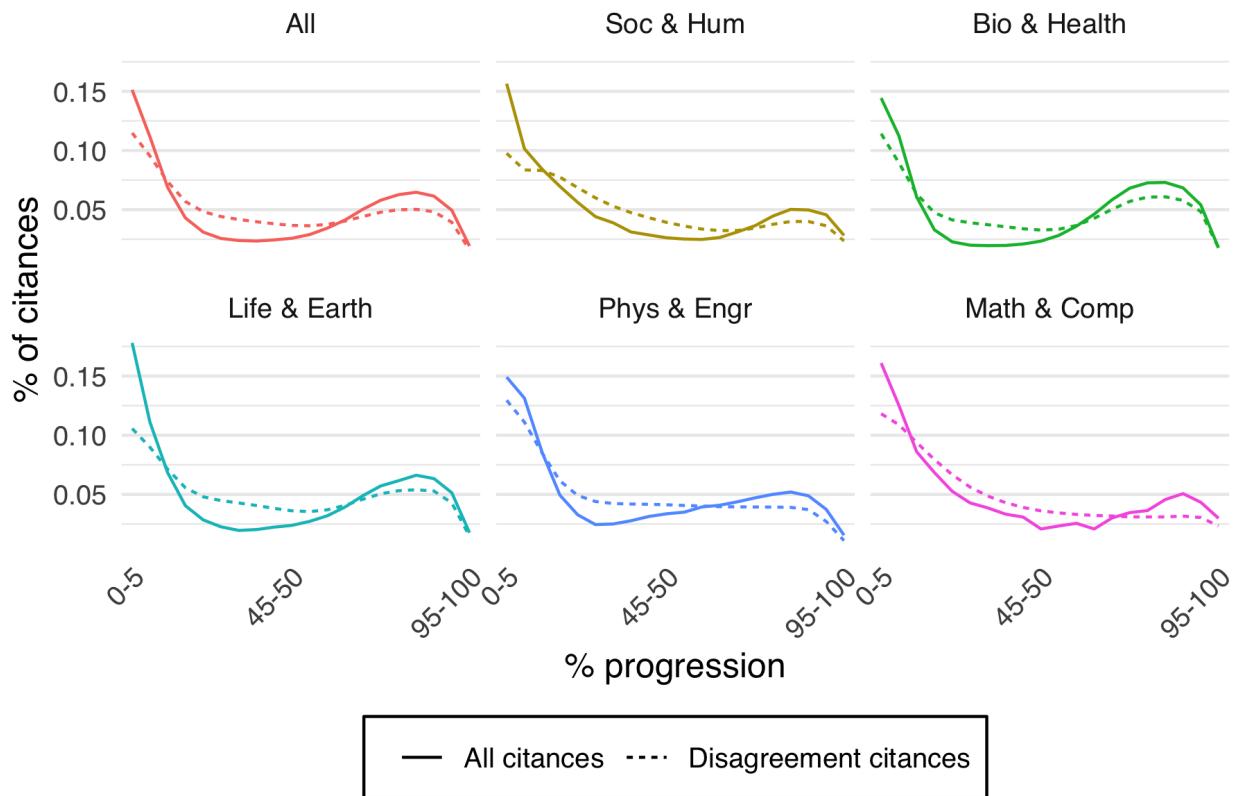


Figure C.5: Distribution of citances by their position in the text of the manuscript, and by field. Shown for all citances (solid line) and disagreement citances (dotted line). For example, about 15 percent of disagreement citances in Physical Sciences and Engineering appear in the first 0-5 percent of the sentences in documents.

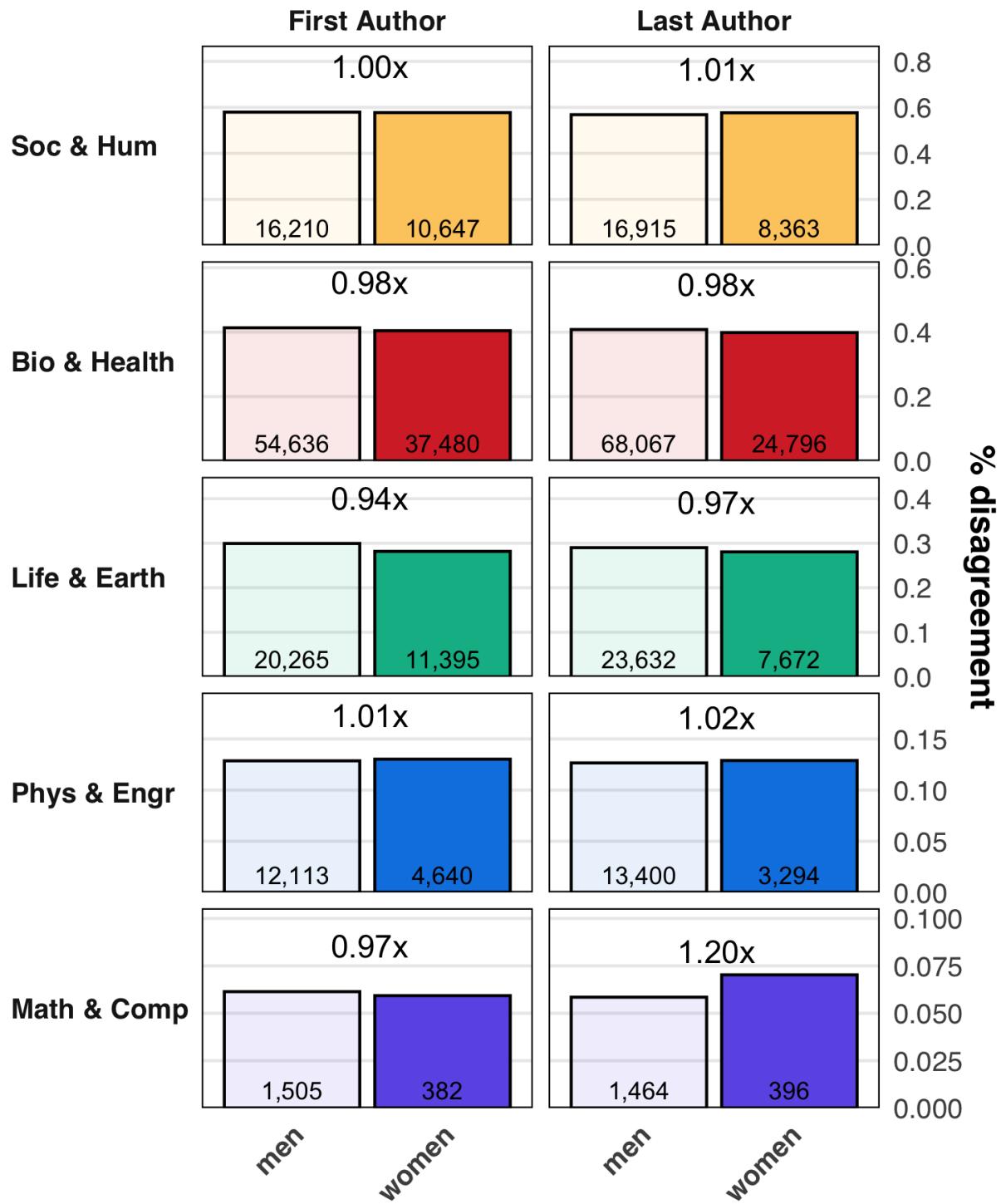


Figure C.6: Little difference in disagreement between men and women. Percentage of disagreement citations by gender of the citing-paper author, their authorship position (first or last), and the high-level field. Numbers above each bar corresponds to the ratio difference between the percentage of disagreement between women and men. The number below each bar corresponds to the number of disagreement citations. we infer a gender for the first and last authors of papers with a disagreement citation published after 2008, determined based on the author's first name as in past work [28].

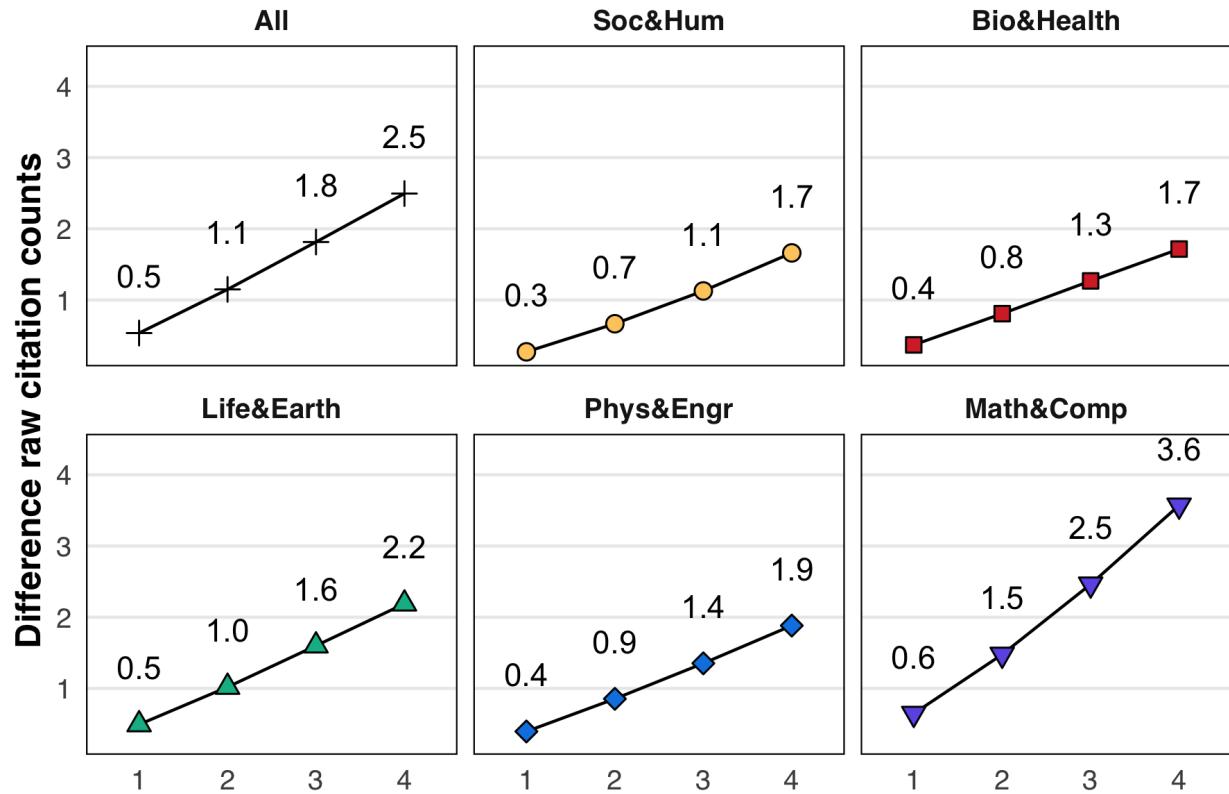


Figure C.7: Full research articles with a disagreement citance are cited more. The y-axis shows the difference in average citation counts for papers containing at least one disagreement citance, and for papers without. Positive values indicate that publications with disagreement received more citations than those without. Values are shown for the population of publications in each year following publication (x-axis). Shown here for only articles labeled in the Web of Science database as full research articles.

Table C.1: Number of citations in the Elsevier ScienceDirect database containing signal term (rows) and filter term (columns) combination. For some signal terms, variants are excluded; for example, “not contradict” is not matched. For filter terms, “*standalone*” indicates that only the signal term was used to query. The remaining columns, +studies, +ideas, +methods, and +results correspond to sets of filter terms outlined in Table C.2.

	<i>-standalone</i>	+studies	+ideas	+methods	+results
challenge*	405,613	16,120	7,114	13,806	15,352
conflict*	212,246	22,190	3,603	5,560	49,961
contradict*	115,375	19,509	5,482	2,793	52,648
contrary	171,711	17,207	3,651	4,699	27,273
contrast*	1,257,866	116,450	7,774	37,372	119,181
controvers*	154,608	12,187	1,840	3,028	15,473
debat*	150,617	8,509	1,774	2,678	4,663
differ*	2,003,677	100,764	9,531	85,309	110,599
disagree*	52,615	5,724	1,142	1,682	12,459
disprov*	2,938	278	528	100	358
no consensus	16,632	1,424	37	830	421
questionable	24,244	1,045	852	1,175	2,050
refut*	10,322	1,399	1,564	338	2,262

Table C.2: Examples of valid and invalid citations returned for "challenge*" and filter term combinations.

Filter term	Valid	Sentence	Reference
<code>_standalone_-</code>	Yes	Although phosphorus has traditionally been seen as the limiting nutrient in freshwater ecosystems (e.g. Blake et al., 1997; Karl, 2000), more recent evidence has begun to challenge this view and has demonstrated that both nitrogen and phosphorus can limit, or at least co-limit, primary production in freshwaters (e.g. Elser et al., 2007; Francoeur, 2001).	Smith et al., 2009 [716]
<code>+studies</code>	No	Analogs of these molecules have shown up to 1000-fold higher activity but are a great challenge to delivery because of their extreme hydrophobicity [33].	Brannon-Peppas & Blanchette, 2004 [717]
<code>+studies</code>	Yes	However, recent studies have challenged this survival benefit in comparison with current usual care.4-6	Gottlieb et al., 2015 [718]
<code>+ideas</code>	No	The low affinity with which volatile general anesthetics bind to macromolecules has made conclusive identification of the in vivo targets by direct binding studies a challenge [1].	Zhang & Johansson, 2005 [719]
<code>+ideas</code>	Yes	This result challenges AUM theory (Gudykunst, 2005) and some prior research (e.g. Gao & Gudykunst, 1990).	Rui & Wang, 2015 [720]
<code>+ideas</code>	No	The description of the resonant electron capture by molecules connected with the formation of negative ions represents still the challenge for the theory [1].	Papp et al., 2013 [721]
<code>+methods</code>	Yes	This model has since been challenged by claims that Helderberg formation boundaries are isochronous across the basin (Anderson et al., 1984; Demicco and Smith, 2009).	Husson et al., 2015 [722]
<code>+methods</code>	No	Subsequent studies in the human challenge model have also supported the role of NA-specific antibody in protection. [34]	Treanor, 2015 [723]
<code>+results</code>	Yes	It has been reported that the prevalence of autoimmune disorders in celiac disease is related to the duration of exposure to gluten [34], although this result has been challenged [35].	Skovbjerg et al., 2004
<code>+results</code>	No	Some of the larger challenges identified in Africa include data collection, access and management, infrastructure and capacity (Han et al., 2014).	Stephenson et al., 2017 [724]

Table C.3: Examples of citances from notable signal/filter phrase combinations labeled in Figure C.2. “[...]” has been used in places of reference names or numbers. Relevant signal have been bolded whereas filter terms have been italicized.

Label	Signal	Filter	Valid	Example
I	conflict*	None	No	“For instance, [...] study conflicts based on ethnicity where ethnic identity works as a device to enforce coalition membership.”
II	disprove*	None	No	“These techniques are typically used to confirm or disprove an a priori hypothesized model, i.e. to test the statistical adequacy of a proposed causal model [...]”
III	challenge*	+ideas	Yes	“Reversal theory challenges the <i>idea</i> of personality traits in suggesting that people fluctuate between metamotivational states that are opposite and mutually exclusive [...]”
IV	contradict*	+methods	Yes	“The existence of topological singularities is in contradiction with [...]’s <i>method</i> of continuous transformations of a rectangular Cartesian frame of coordinates into a curvilinear grid without singularities [...]”
V	controversy*	+ideas	Yes	“There is still a considerable controversy about the <i>idea</i> of homology between specific areas of rat and primate PFC [...]”
VI	disagree*	+results	Yes	“These <i>results</i> were in disagreement with much of the literature, where there is consensus that, H2 is almost exclusively catalyzed by SRB at low COD/SO42- ratios [...]”
VII	no consensus	+ideas	Yes	“Because of the controversial data collected, no consensus about this <i>theory</i> has been reached to date [...]”
VIII	contradict*	+methods	Yes	“Thus, the <i>approach</i> is somewhat contradictory to certain approaches in robotics that strive to develop highly autonomous robots capable of performing independent decisions based on sensory data [...]”

Table C.4: Results are robust to both the 80 percent and 70 percent validity cutoffs. Quantities of interest using the 23 queries above the 80 percent validity cutoff, and the 36 queries above the 70 percent validity cutoff. Shown are the overall rates of disagreement and the change in the share of disagreement per year.

Quantity	Validity cutoff	All Fields	Soc & Hum	Bio & Health	Life & Earth	Phys & Engr	Math & Comp
Overall	80%	0.32%	0.61%	0.41%	0.29%	0.15%	0.06%
	70%	0.40%	0.78%	0.50%	0.36%	0.20%	0.15%
Change by year	80%	-0.0005	-0.0033	+0.0017	+0.0018	-0.0045	-0.0019
	70%	-0.0005	-0.0042	+0.0022	+0.0019	-0.0061	-0.0028

Table C.5: Receiving a disagreement citation has little impact on citations in the year following. For each field, shown are the number of cited papers, the average citations received by papers in the year following receiving a disagreement citation, average citations received by all papers of the same age and citation count, and the ratio, d, between these values. When d is greater than one, papers receiving a disagreement citation receive more citations in the following year than expected; when d is less than one, they receive fewer citations than expected.

Scientific field	# Records	Avg. citations in year following disagreement	Expected citations in year following disagreement	d
All	109545	3.03	3.08	0.983
Bio & Health	60707	2.73	2.81	0.969
Life & Earth	20581	3.43	3.35	1.023
Math & Comp	770	3.36	3.34	1.005
Phys & Engr	18011	3.55	3.52	1.006
Soc & Hum	9476	3.04	3.11	0.979

Table C.6: Papers that introduced the most disagreement citations

Total citations	Disagreement citations	Publication	Title	Document type
50	27	Alén, Domínguez, & De Carlos (2015) [725]	University students' perceptions of the use of academic debates as a teaching methodology	Full-length Article
400	27	French & Koeberl (2010) [726]	The convincing identification of terrestrial meteorite impact structures: What works, what doesn't, and why	Review Article
66	26	Doody & Condon (2012) [727]	Increasing student involvement and learning through using debate as an assessment	Full-length Article
64	25	Ersoy (2010) [728]	Social studies teacher candidates' views on the controversial issues incorporated into their courses in Turkey	Full-length Article
91	24	Nam (2014) [729]	The effects of trust and constructive controversy on student achievement and attitude in online cooperative learning environments	Full-length Article
1292	23	Kalter (2003) [730]	Teratology in the 20th century Environmental causes of congenital malformations in humans and how they were established	Full-length Article
1708	23	Millan (2006) [731]	Multi-target strategies for the improved treatment of depressive states: Conceptual foundations and neuronal substrates, drug discovery and therapeutic application	Review Article
50	21	Bruschke & Divine (2017) [732]	Debunking Nixon's radio victory in the 1960 election: Re-analyzing the historical record and considering currently unexamined polling data	Full-length Article
107	21	Stepanova & Bruckmeier (2013) [733]	The relevance of environmental conflict research for coastal management. A review of concepts, approaches and methods with a focus on Europe	Review Article
74	20	Colston & Vadjunec (2015) [734]	A critical political ecology of consensus: On "Teaching Both Sides" of climate change controversies	Full-length Article

Table C.7: Publications that received the most disagreement citations.

Total citations	Disagreement citations	Publication Title	Document type
389	99	Munro (2003) [735] Lipid Rafts Elusive or Illusive?	Review Article
477	99	Murphy, Birn, Handwerker, Jones, & Bandettini (2009) [736] The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced?	Full-Length Article
1753	83	Zhao, Sun, Wilde, & Sanzhong (2005) [737] Late Archean to Paleoproterozoic evolution of the North China Craton: key issues revisited	Full-Length Article
1344	69	Li et al. (2008) [738] Assembly, configuration, and break-up history of Rodinia: A synthesis	Full-Length Article
663	65	Kusky & Li (2003) [739] Paleoproterozoic tectonic evolution of the North China Craton	Full-Length Article
1377	64	Zhao, Wilde, Cawood, & Sun (2001) [740] Archean blocks and their boundaries in the North China Craton: lithological, geochemical, structural and P-T path constraints and tectonic evolution	Full-Length Article
972	51	Zhai & Santosh (2011) [741] The early Precambrian odyssey of the North China Craton: A synoptic overview	Full-Length Article
50	45	Debat et al. (2003) [742] A new metamorphic constraint for the Eburnean orogeny from Paleoproterozoic formations of the Man shield (Aribinda and Tampelga countries, Burkina Faso)	Full-Length Article
448	45	Wilde, Zhao, & Sun (2002) [743] Development of the North China Craton During the Late Archaean and its Final Amalgamation at 1.8 Ga: Some Speculations on its Position Within a Global Palaeoproterozoic Supercontinent	Full-Length Article
247	43	Kusky (2011) [744] Geophysical and geological tests of tectonic models of the North China Craton	Full-Length Article

Appendix D

Study 4: Embedding mobility

D.1 Text

Mobility and science

As scholars move, they bring their knowledge, skills, and social connections with them—collectively the movements of researchers shape the structure and direction of the global scientific enterprise. For example, prestige-driven mobility between doctoral-granting and employing institution is highly unequal [184, 593], which affects the diffusion of ideas across academia [116]. By placing researchers in new social settings, mobility can lead to the formation of new collaborative relationships [84], which in turn spurs the further diffusion of knowledge and innovations [194, 195, 297, 587]. Perhaps resulting from the selection effects of who gets to move, or the reconfiguring of social and epistemic networks, movement is associated with increased scientific impact [83, 294, 295, 745]. At the national level, the understanding of mobility has progressed beyond simplistic narratives of brain drain and brain gain, and instead adopts a new perspective of *flows* of talent [746–748]. Under this flow model, a mobile researcher is viewed as contributing to both their origin and destination countries, a perspective that fosters that is evidenced by the strong science of open countries [22]. Perhaps because of these individual and national benefits, policy-makers have come to recognize the importance of global mobility [306, 749]. Movement is a key mechanism that has clear impacts on the composition and direction of the global scientific workforce and our collective scientific understanding. Understanding the structure and dynamics of mobility is thus essential for understanding global science.

Modeling scientific mobility There are many ways of modeling scientific mobility from bibliographic data, the first consideration being the unit of analysis. Most studies of mobility

have focused on *country-level* mobility—the flows of researchers across nations [295, 745, 750, 751]. Practically, country-level analyses benefit from higher reliability, such that idiosyncrasies and errors inherent to bibliographic databases are mitigated by this higher level of aggregation. Epistemically, country-level analysis is useful for national science governance who aims to understand the status of their country in the global landscape and make informed policy decisions. Analyses at lower levels of analysis are far less common. *Regional*-level scientific mobility—the flow of researchers between regions or cities within or across countries has been only minimally studied [752], possibly due to lack of reliable long-term data and lack of policy relevance to national-level lawmakers. *Organization*-level mobility has the potential to inform institutional policy and to understand the composition of mobility within a single country or region, especially as it relates to organization performance, prestige, and inequality [116, 184, 593, 753]. However, affiliation disambiguation and noise in bibliometric data makes large-scale organization-level analysis challenging. Here, we learn neural-network embeddings of scientific mobility at the level of organizations using a curated bibliographic database. These embeddings are robust to noise, and so are capable of representing clear structure even amid issues with organizational disambiguation. In doing so, embeddings also capture a more detailed understanding of mobility than has been previously studied.

Another consideration when analyzing scientific mobility is what kinds of mobility to study. Typical understandings of mobility are directional: movement is always *from* one place and *to* another. However, scientific mobility is more complicated. For example, scientists often hold multiple affiliations at a time [317], listing them as co-affiliations on a single paper, or even choosing a subset of affiliations to use for multiple simultaneous projects [318]. Even clearly-directional migration to another institution is complex—researchers may continue to publish with an old affiliation for projects that began before their move, and they may maintain social and organizational links to their old institution (e.g., collaborators, projects, graduate students) such that there is no clear breakage after migrating. There is also a whole range of short-term scientific mobility, such as visit-

ing scholarships and short-term visits that are only visible through intensive efforts such as manual extraction from CVs [641, 672, 754]. Here, we focus on more long-term mobility that can be derived from bibliographic data. Due to the complexity of scientific mobility, we make the simplifying assumption that all scientific mobility is *symmetric* or without direction such that any move from an organization A to organization B is equivalent to a move from B to A . By assuming non-directional mobility, all mobility events are commensurate, meaning that they can be treated identically in our analysis—this allows us to represent the complexity of mobility without making decisions about the directional of their mobility or which is their main affiliation. Moreover, this assumption has the practical advantage of matching the data format expected by the *word2vec* model, as well as the theoretical advantage of adhering to the symmetricity assumption of the gravity model of mobility.

Building affiliation trajectories

For each mobile researcher who has at least two distinct affiliations, we construct an affiliation trajectory based on the affiliations listed on their published papers indexed in the Web of Science database between 2008 and 2019. An author is considered mobile if they published with at least two distinct affiliations during the time period of study. Affiliation names were manually disambiguated, and each was mapped to a unique organization identifier. An affiliation trajectory for an individual researcher is a sequence of organizations in ascending order of year of publication. If a researcher published papers with affiliation A in year t , B in $t + 1$, C in $t + 2$ and A again in $t + 3$, then the affiliation trajectory is expressed as (A, B, C, A) .

In the case that an individual lists multiple affiliations in a single year, affiliations listed on publications published in that year are shuffled between each iteration of the *word2vec* training process (each epoch). For example, an author who published with affiliation A in t_0 , and affiliations B and C in t_1 could appear as one of (A, B, C) or (A, C, B) in each training iteration. This effectively removes the effect of order within a year, as the order cannot be meaningfully established based on

co-affiliations in a single paper, or on different affiliations listed on separate papers, for which its date of publication may not be representative of the actual completion of the project.

Other than restricting to only mobile researchers, we do not perform any filtering or reductions to affiliation trajectories. In the case than an author publishes with organization A four times in t_0 , and affiliation B two times in t_1 , then their trajectory will be (A, A, A, A, B, B) . Although mobile authors who publish more papers will have longer trajectories, *word2vec* will skip duplicate consecutive organization IDs, mitigating the impact of long repetitive trajectories.

Network-based personalized page rank distances

We examine the gravity model on the Personalized Page Rank (PPR)[755] as a benchmark on the network. We construct the co-occurrence network of N organizations, in which each edge between organizations i and j represents a co-occurrence of i and j in the same affiliation trajectory, with weight W_{ij} given by the sum of the co-occurrences over all researchers. and edges are co-occurrence between two organizations. The Personalized Page Rank is a ranking algorithm for nodes based on a random walk process on networks. The walker visiting at a node moves to a neighboring node chosen randomly with a probability proportionally to the weight of the edge in one step. Furthermore, with probability α , the walker is teleported back to the starting node. The rank of a node is determined by the probability that the walker visits the node in the stationary state. The stationary distribution of the random walker starting from node i , denoted by $p_i = (p_{ik})$, is given by

$$p_i = (1 - \alpha)v_i + \alpha p_i W, \quad (\text{D.1})$$

where v_i is a column vector of length N with entries that are all zero except the i th entry that equals to one, $W = (W_{ij})$ is the weighted adjacency matrix. We used $\alpha = 0.9$ here.

We can think p_i as a representation vector of the organization i , and calculate the distance

between organizations i and j , d_{ij} with measuring distance between p_i and p_j to examine the gravity law. We consider two distance measures in this analysis. The first one is cosine distance which is used for our embedding method, $d_{ij} = 1 - \frac{\mathbf{p}_i \cdot \mathbf{p}_j}{\|\mathbf{p}_i\| \|\mathbf{p}_j\|}$. Also, if we think p_i as a discrete probability distribution, then we can consider Jensen–Shannon divergence (JSD), can be written as,

$$d_{ij} = JSD(p_i || p_j) = \frac{1}{2} D_{KL}(p_i || m) + \frac{1}{2} D_{KL}(p_j || m), \quad (\text{D.2})$$

$$D_{KL}(p_i || m) = \sum_x^x p_{ix} \log \frac{p_{ix}}{m_x}, \quad (\text{D.3})$$

where $m = \frac{1}{2}(p_i + p_j)$. We report the result with cosine distance ($R^2 = 0.14$, Fig. D.11) and Jensen–Shannon divergence ($R^2 = 0.19$, Fig. D.12). In both cases, the performance is under the performance of the model with geographical distance. Even though the length of the PPR vectors is extremely larger than the length of our embedding vectors, result with the embedding distance outperforms both of them.

Singular value decomposition distance

We use the truncated singular value decomposition (SVD) on the underlying mobility co-occurrence matrix as a baseline embedding. In short, truncated SVD performs linear low-rank approximation of the matrix with given dimensions, d . First, we construct the co-occurrence matrix of N organizations, A_{ij} given by the co-occurrence of organizations i and j in the same affiliation trajectory. Then, we apply truncated singular value decomposition with $d = 300$ on the flow matrix A directly.

We calculate distance between organizations in the SVD embedding space using cosine distance, finding that it explains slightly more of the flux between organizations than does geographic distance ($R^2 = 0.247$, Table D.3). When used as an input to the gravity model, this distance produces better

predictions than geographic distance using both the exponential (RMSE = 0.859, Table D.4) and power-law models (RMSE = 0.839, Table D.4), performing slightly better with the power-law formulation.

Laplacian Eigenmap distance

We also consider Laplacian Eigenmap embeddings [616] as a baseline, which is one of the most fundamental approaches for graph embedding. First, we construct the co-occurrence matrix of N organizations, A_{ij} given by the co-occurrence of organizations i and j in the same affiliation trajectory and degree matrix D which is the diagonal matrix for which $D_{ii} = \sum_j A_{ij}$. Then we construct graph Lapalcian matrix $L = D - A$ and apply truncated singular value decomposition in the matrix L with $d = 300$.

We only report results based on the cosine distance between Laplacian embedding vectors, finding that it explains less of the total flux than geographic distance ($R^2 = 0.212$, Table D.3). When used as an input to the gravity model, the Laplacian cosine distance produces marginally-better predictions than geographic distance using both the exponential (RMSE = 0.878, Table D.4) and power-law models (RMSE = 0.87, Table D.5), performing slightly better with the power-law formulation.

Levy's symmetric SVD *word2vec* distance

We also compare the *word2vec* embedding distnace against a baseline of direct matrix factorization approach, using the symmetric SVD *word2vec* method [608]. Based on idea of *word2vec* is just implicit matrix factorization, Levy proposed symmetric SVD *word2vec* embedding, which should directly compute the embedding that *word2vec* only attempts to efficiently approximate. First, we construct the matrix of N organization

$$M_{ij} = \log \left(\frac{N(i,j)|D|}{N(i)N(j)} \right) - \log k, \quad (\text{D.4})$$

where $N(i,j)$ is the number of times the location pair (i,j) appears given the window size w in the total corpus D , $N(i) = \sum_{j=N}^i N(i,j)$ as the number of items i occurred given the window size w in D , and k is the number of negative samples. We used $w = 1$ and $k = 5$ which is same setting in the our main result. Then, we factorized matrix M with truncated singular value decomposition in the matrix with $d = 300$ into $U_d \Sigma_d V_d$, and used the embedding vector as $U_d \sqrt{\Sigma_d}$.

For this baseline, we report results using the cosine distance, Euclidean distance, and dot product between the embedding vectors. We find that the dot product performs by far the worst, worse than any other baseline considered ($R^2 = 0.004$, Table D.3). The cosine distance performs better, but worse than geographic distance ($R^2 = 0.212$, Table D.3). The Euclidean distance performs best, explaining more of the flux than geographic distance, and only being below the embedding distance ($R^2 = 0.341$, Table D.3). Focusing on the Euclidean distance, we find that using it as input to the gravity model results in better predictions than geographic distance using both the exponential model (RMSE = 0.803, Table D.4) and power law models, (RMSE = 0.78, Table D.5), though it performs slightly better with the power law formulation.

Why do neural embeddings outperform direct matrix factorization?

Why do neural-embedding approaches, which rely on stochastic gradient descent, outperform Levy's direct matrix factorization [608], especially given that *word2vec* is implicitly approximating factorization? We speculate that is stems, in part, from the sensitivity of matrix factorization to small flows between locations. Levy's matrix factorization embeds affiliations i and j such that their dot similarity is as close as possible to $\log(T_{ij}/P(i)P(j))$. If the flow T_{ij} is considerably small or zero, the dot similarity goes to $-\infty$, pushing i and j very far from other affiliations in the embedding space [608, 756]. This is particularly problematic when the window size is small because most

affiliation pairs would have no flow, which is indeed the case in our experiments. To circumvent this problem, previous studies [608, 756] added a constant flow between the affiliation pairs with no flow. However, in addition to altering the underlying data, these small flows can still have a strong impact on the embedding.

It is also well known that singular value decomposition (SVD) is vulnerable to outliers [757–761]. The stochastic gradient descent algorithm, which is employed in SGNS *word2vec*, is more robust than SVD and can enhance generalization and effectiveness of *word2vec* model [762–764].

Organization disambiguation and metadata

Affiliations mapped to one of 8,661 organizations, disambiguated following that originally designed for the Leiden Rankings of World Universities [151]. Organizational records were associated with a full name, a type indicating the sector (e.g., University, Government, Industry), and an identifier for the country and city of the organization. Sixteen different sector types were included in the analysis, which we aggregated to four high-level codes: *University*, *Hospital*, *Government*, and *Other*. Each record was also associated with a latitude and longitude. However, for many organizations, these geographic coordinates were missing or incorrect. We manually updated the coordinates of 2,267 organizations by searching the institution name and city on Google Maps; in cases where a precise location of the organization could not be identified, we used the coordinates returned when searching the name of the city. The data was further enriched with country-level information, including region, most widely-spoken language, and its language family (e.g., the language family of *Spanish* is *Italic*). State/province-level information was added using the reverse geocoding service LocationIQ using each organization’s latitude and longitude as input. Regional census classifications were added for states in the United States. For each organization, we calculated size as the average number of unique authors (mobile and non-mobile) who published with that organization across each year of our dataset; in the case that authors publish with multiple affiliations in a single year, they are

counted towards each.

As a result of our disambiguation procedure, some affiliations are mapped to two organizations, one specific, and one more general. For example, any author affiliated with “Indiana University Bloomington” will also be listed as being affiliated with the “Indiana University System”, a more general designation for all public universities in Indiana. However, a more general organization may not always occur alongside the more specific one. For example, a researcher affiliated with the smaller regional school “Indiana University South Bend” will be listed as affiliated with only the “Indiana University System”. We identify all specific organizations that always co-occur along with a more general one. For every career trajectory that includes one of these specific organizations, we remove all occurrences of the more general organization; trajectories containing only a general designation are not altered.

Author name disambiguation

Author-name disambiguation, the problem of associating names on papers with individuals authors, remains difficult for the use of bibliometric data [765]. Authors in our dataset have been disambiguated using a rule-based algorithm that makes use of author and paper metadata, such as physical addresses, co-authors, and journal, to score papers on the likelihood of belonging to an author cluster—a cluster of publications believed to have been authored by the same individual [97]. We limit our period of analysis to the period of 2008 to 2019, as in 2008 the Web of Science began indexing additional author-level metadata such as full names and email addresses. The disambiguation algorithm is conservative, favoring splitting clusters over merging. Past studies have validated this data and shown that the disambiguated authors are comparable to ground-truth records such as those from ORCID and useful for a wide range of bibliometric studies [292, 295, 318, 350].

Derivation of Eq.D.11 - Noise Contrastive Estimation

The noise contrastive estimation (NCE) [618, 619]. NCE is an unbiased estimator for a probability model P_m of the form:

$$P_m(x) = \frac{f(x)}{\sum_{x' \in X} f(x')}, \quad (\text{D.5})$$

where f is a non-negative likelihood function of data x , and X is the set of all data. This general form includes the word2vec model (Eq. (6.8)), where $f(x) = \exp(x)$ and $x = \mathbf{u}_j \cdot \mathbf{v}_i$. NCE fits the probability model using a binary classification task in the same way as in negative sampling but using a Bayesian formalism for logistic regression. Specifically, before the training, we know that 1 in $1 + k$ words is sampled from the given data, which can be modeled as prior probabilities

$$P(Y_j = 1) = \frac{1}{k+1}, \quad P(Y_j = 0) = \frac{k}{k+1}. \quad (\text{D.6})$$

Using the Bayes rule, the posterior probability for Y_j given word j is given by

$$P(Y_j|j) = \frac{P(j|Y_j) P(Y_j)}{P(j|Y_j = 0) P(Y_j = 0) + P(j|Y_j = 1) P(Y_j = 1)}. \quad (\text{D.7})$$

Bearing in mind that word j is sampled from the given data if $Y_j = 1$ and from the noise distribution p_0 if $Y_j = 0$. Assuming that the given data is generated from the probability model to fit, the class-conditional probability, $P(j|Y_j)$, is given by

$$P(j|Y_j = 1) = P_m(\mathbf{u}_j \cdot \mathbf{v}_i), \quad P(j|Y_j = 0) = p_0(j). \quad (\text{D.8})$$

Putting Eqs. (D.6), (D.7) and (D.8) together, the posterior probability for Y_j is given by

$$P(Y_j = 1|j) = \frac{P_m(\mathbf{u}_j \cdot \mathbf{v}_i)/(k+1)}{P_m(\mathbf{u}_j \cdot \mathbf{v}_i)/(k+1) + kp_0(j)/(k+1)} \quad (\text{D.9})$$

$$= \frac{P_m(\mathbf{u}_j \cdot \mathbf{v}_i)}{P_m(\mathbf{u}_j \cdot \mathbf{v}_i) + kp_0(j)}, \quad (\text{D.10})$$

which can be rewritten in form of sigmoid function:

$$P^{\text{NCE}}(Y_j = 1|j) = \frac{1}{1 + kp_0(j)/P_m(\mathbf{u}_j \cdot \mathbf{v}_i)} \quad (\text{D.11})$$

$$= \frac{1}{1 + \exp[\ln kp_0(j) - \ln P_m(\mathbf{u}_j \cdot \mathbf{v}_i)]} \quad (\text{D.12})$$

$$= \frac{1}{1 + \exp[-\ln f(\mathbf{u}_j \cdot \mathbf{v}_i) + \ln p_0(j) + c]}, \quad (\text{D.13})$$

where $c = \ln k + \ln \sum_{x' \in \mathcal{X}} f(x')$ is a constant. NCE maximizes the log-likelihood

$$\mathcal{J}^{\text{NCE}} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{D}} [Y_j \log P^{\text{NCE}}(Y_j = 1|j) + (1 - Y_j) \log P^{\text{NCE}}(Y_j = 0|j)]. \quad (\text{D.14})$$

by calculating the gradients for embedding vectors \mathbf{u}_j , \mathbf{v}_i and iteratively updating them. An important consequence of this framework is that NCE is an unbiased estimator that has convergence to the optimal embedding in terms of the original word2vec's objective function, \mathcal{J} if we increase the number of words to sample and the training iterations [618, 619].

Reconstructing Times ranking with network measure

The performance of the embedding ranking in reconstructing the Times ranking is comparable to that of network-derived measures such as degree strength (Spearman's $\rho = 0.73$, Fig. D.19a) and eigencentrality centrality (Spearman's $\rho = 0.76$, Fig. D.19b). However, while both embedding- and network-based measures relate to university prestige, they are qualitatively and quantitatively different. The embedding-ranking of U.S. universities is less correlated with degree strength

(Spearman's $\rho = 0.45$, Fig. D.20a) and eigenvector centrality (Spearman's $\rho = 0.55$) than with the Times ranking itself (Spearman's $\rho = 0.73$, Fig. D.20b). The embedding ranking over-ranks large research-intensive universities such as North Carolina State University, University of Florida, and Texas A&M University, whereas the network-derived ranking over-ranks smaller, more specialized universities such as Brandeis University, Yeshiva University, and University of San Francisco. This suggests that the embedding encodes information on prestige hierarchy at least as well as a network representation, with some noticeable qualitative differences.

Speculation on variations of the convex-curve pattern

The convex-curve pattern observed in Fig. 6.6 repeats across many countries, with variations. For example, the representative vector of Chinese organizations has a larger norm than that of the U.S. ($\bar{l} = 2.97$ vs $\bar{l} = 2.39$, Table D.2), causing its curve to be shifted upwards with a larger peak vector norm; this may reflect a tendency for organizations in the U.S. to appear more frequently in different contexts than Chinese organizations. Other nations such as Poland, Iran, and Turkey show a linear relationship between an organization's number of researchers and the vector norm, indicating that their largest organizations belong to very specific contexts (Fig. D.24). The organization-level distribution of vector norms reveals deeper heterogeneity. The distribution of the vector norms for the U.S. is relatively skewed, suggesting their large norm is driven by a small and tight community of organizations ($skew = -0.82$, Fig. D.25). Germany and the U.K. have comparable representative vector norms to the U.S. ($\bar{l} = 2.6$ and $\bar{l} = 2.61$, respectively), with lower skewness ($skew = -0.63$ and $skew = -0.55$), suggesting more tight community of organizations. The vector norms of organizations in some countries are even more skewed, such as in Iran ($\bar{l} = 3.57$, $skew = -2.13$) and China ($\bar{l} = 2.97$, $skew = -1.08$), indicating the strong difference between their most- and least-connected organizations. For some countries, their organizations are positively-skewed, though seemingly for different reasons. For example, Austria has a balanced distribution of organization vector norms,

suggesting a diverse range of organizations with most being well connected ($\bar{l} = 2.64$, $s = 0.18$); Russia, in contrast, has a number of organization vectors of moderate norms, but also several isolated organizations with large vector norms ($\bar{l} = 3.08$, $s = 0.67$).

D.2 Figures

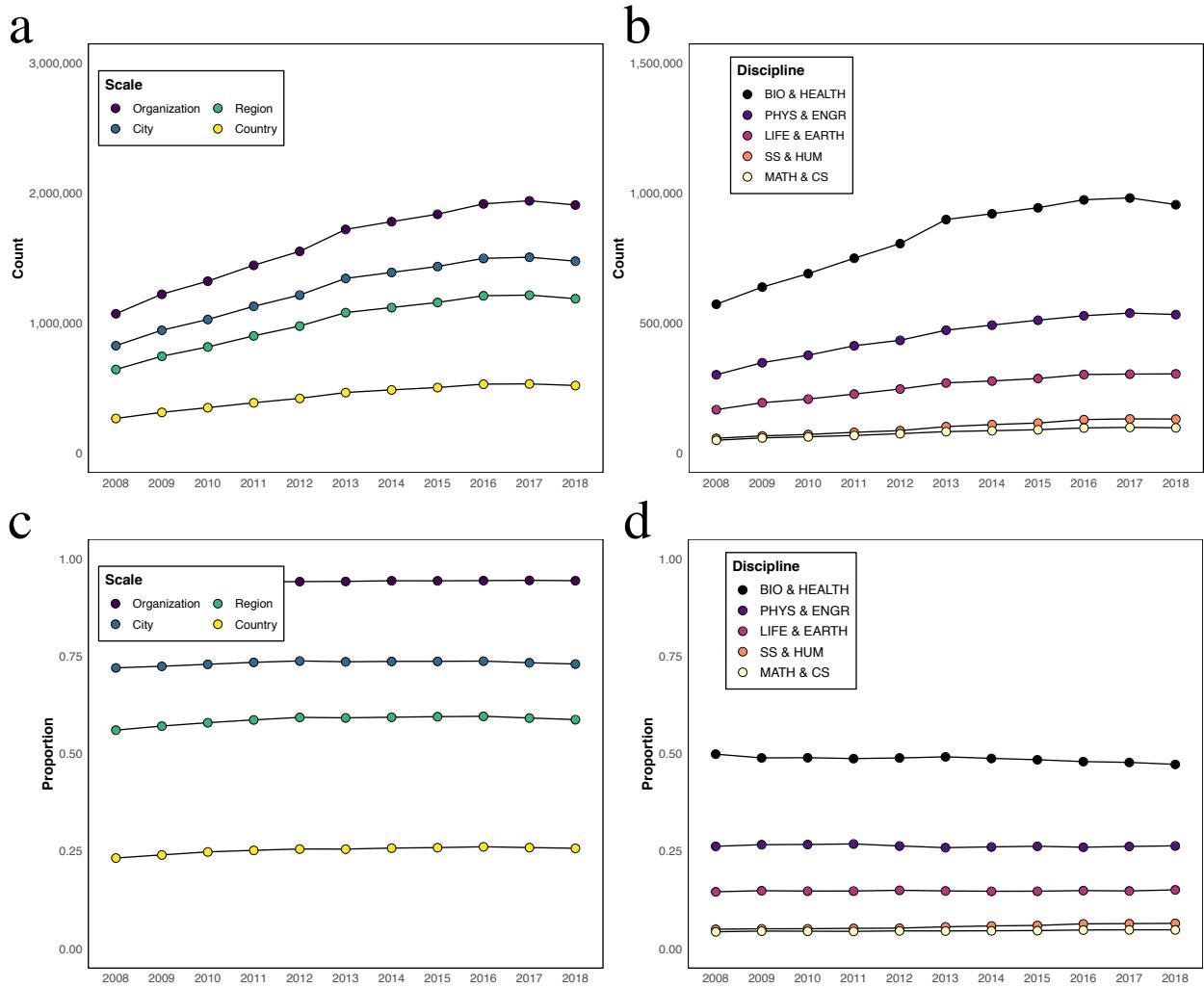


Figure D.1: Publications over time. **a.** The number of papers published by mobile authors has been steadily increasing from 2008 to 2017, with a small decrease in 2018, which may be due to an artifact of the Web of Science indexing process. Lines correspond to publications by mobile authors, by authors with affiliations in at least two cities, at least two regions, and at least two countries. We did not find major changes in the publication patterns of mobile authors during this time period. **b.** Lines correspond to the proportion of publications classified as Biology and Health (black), Physics and Engineering (purple), Life and Earth Science (magenta), Social Science and Humanities (orange), and Math and Computer Science (yellow). The rate of publication in Biology and Health has leveled since about 2013, whereas the rate of publication in other fields has steadily increased. **c.** While the absolute count of publications has increased, the percentage of mobile scholars, and those with affiliations in at least two cities, regions, or countries, as a proportion of all publications, has remained stable over time. **d.** The proportion of authors' publications across fields has largely remained steady. Biology and Health Science has comprised the majority of publications across nearly all years but has steadily declined in proportion. However, the proportion of Social Science and Humanities publications has been steadily increasing.

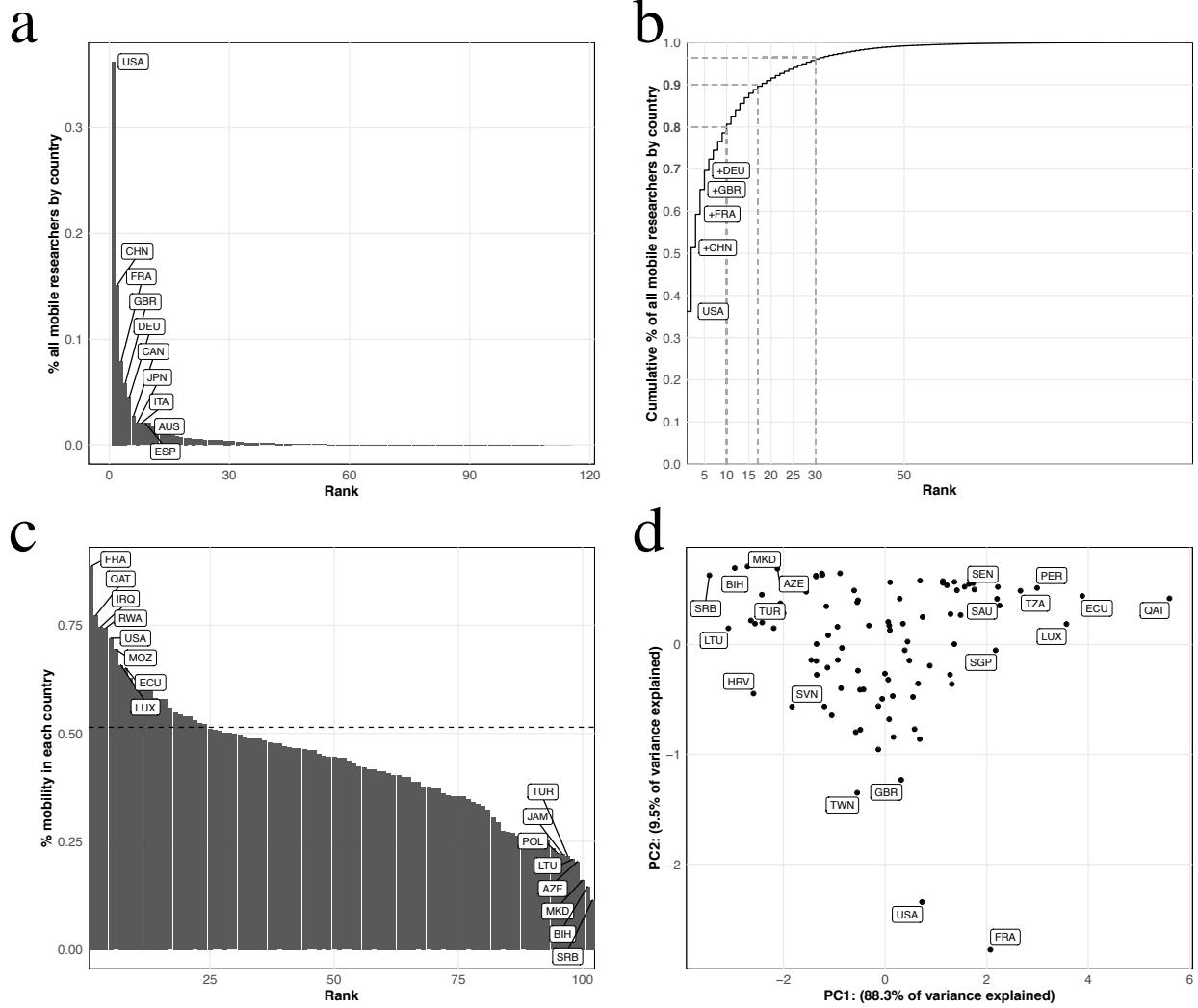


Figure D.2: Extent and nature of mobility by country. **a.** The proportion of all mobile researchers contributed by each country. Over 30% of all mobile researchers have been affiliated with organizations in the U.S. during the period of study. **b.** Cumulative distribution of data shown in (a). The U.S., China, and France, the U.K., and Germany comprise about 70% of all mobile researchers. **c.** The proportion of each country’s researchers who are mobile. The dashed line indicates the proportion of all researchers in the data who are mobile. France, followed by Qatar and the U.S. have the highest proportion of mobile researchers. **d.** First two principal components of four variables: proportion of researchers in each country mobile across organizations, proportion mobile across cities, proportion mobile across regions, and proportion mobile across countries. The countries are roughly sorted in order of the number of mobile researchers and the fraction of international mobile researchers in the first and second principal components, which are indicated by PC1 and PC2, respectively. PC1 explains 88.3% of the total variance, whereas PC2 explains 9.5% of the total variance.

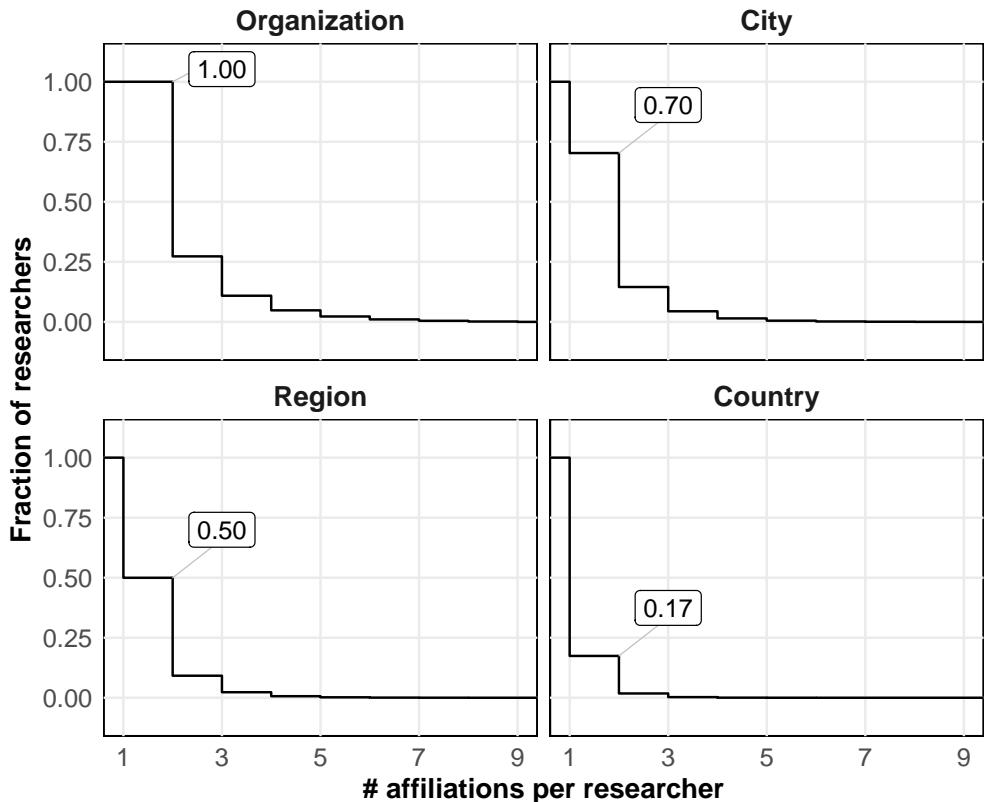


Figure D.3: **Reverse cumulative-distribution function of mobile researchers by geographic scale.** **a.** Survival probability of mobile researchers with respect to the number of organizations in their affiliation trajectory. All mobile authors were affiliated with at least two organizations (i.e., survival probability of one) and about 25.0% were affiliated with three or more. **b.** About 70% of mobile authors listed at least two cities represented in their career trajectories. **c.** 50% of mobile authors have two or more regions represented in their career trajectories. **d.** Only 17% of mobile authors had two or more countries represented in their career trajectories.

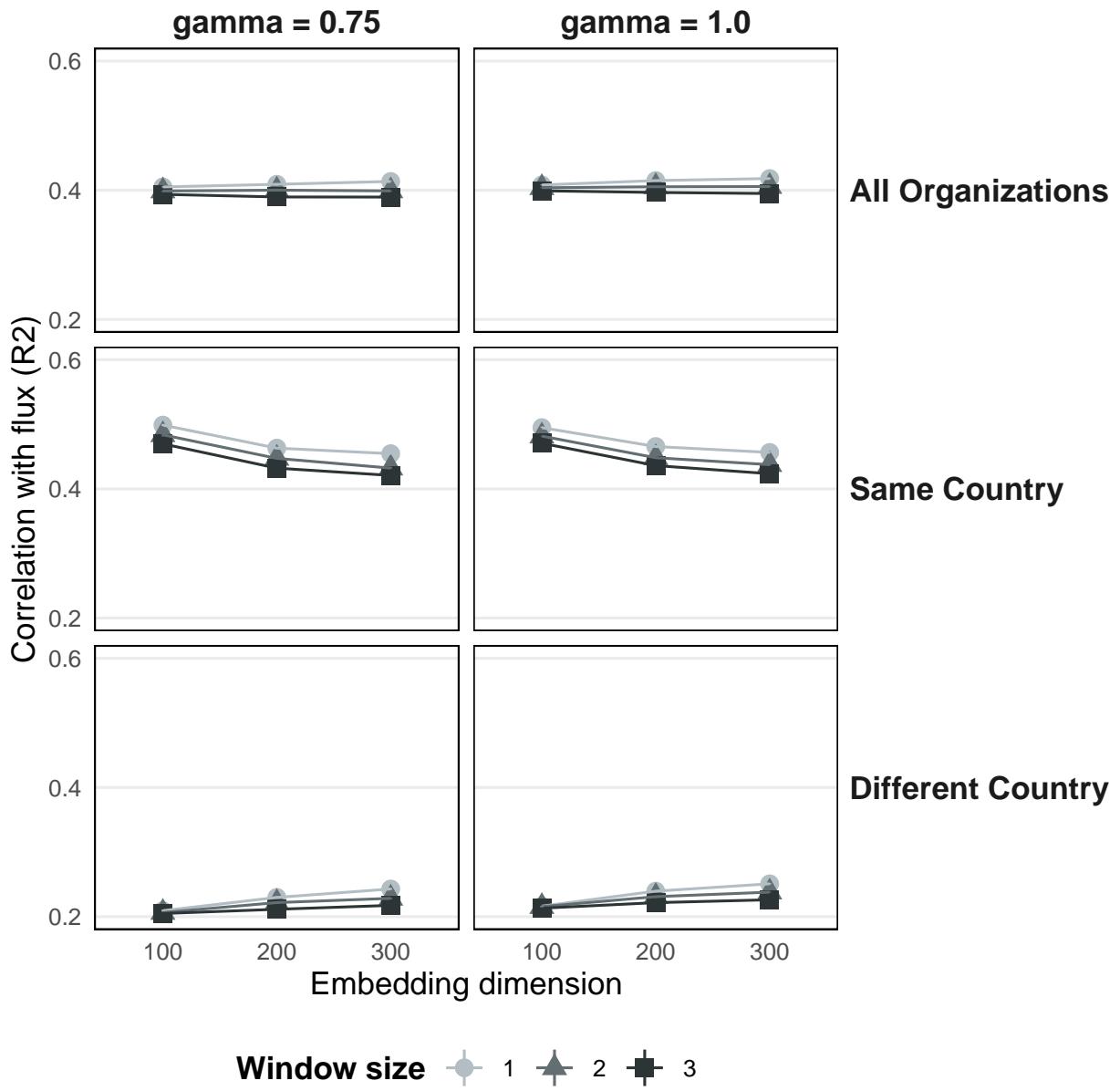


Figure D.4: **Larger dimensions, smaller window size improves embedding performance.** The correlation, or amount of flux explained by the embedding distance with varying skip-gram negative sampling hyperparameters. Window size refers to w , the size of the context window that defines the context in a trajectory. Smaller window sizes result in an embedding that explains more flux. Embedding dimensions refer to the size of the embedding vector. Larger vectors perform better, though with little difference between 200 and 300. Gamma refers to the γ parameter in *word2vec*, which shapes the negative sampling distribution. A value of $\gamma = 0.75$ is the default for *word2vec*. There is virtually no difference in performance based on γ . All variants perform better on same-country organization pairs, and worse on different-country pairs, than on all pairs of organizations than on all pairs. Embeddings with larger dimensions outperform mid-size embeddings for the different-country case.

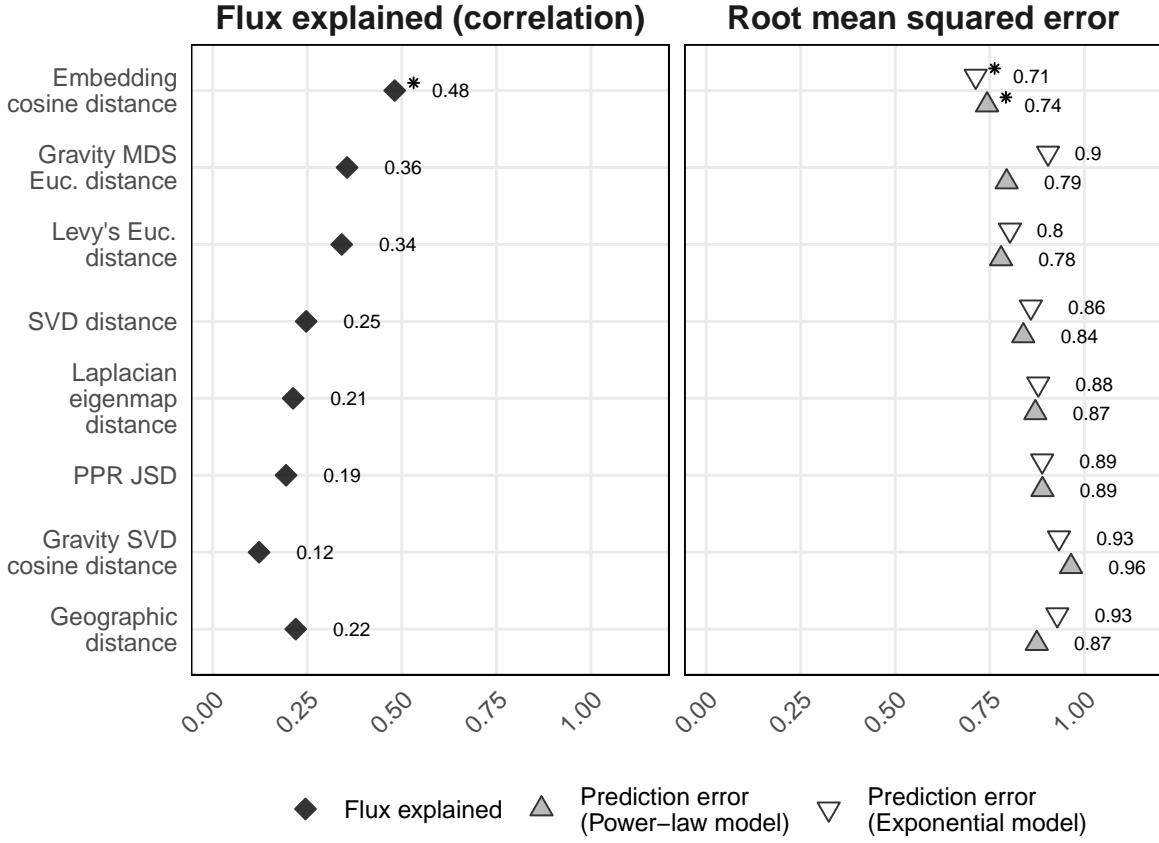


Figure D.5: **Neural embeddings outperform baselines for scientific mobility.** Cosine distance between embedding vectors generated with *word2vec* explains more of the flux, and better predicts flux when used in the gravity model of mobility than geographic distance and other baselines. Shown is the correlation between the flux and embedding distances, measured with R^2 (left), and the prediction error when using the distance as input to the gravity model of mobility. The asterisk denotes the top-performing metric. For prediction error, we show results based on both the exponential and power-law forms of the gravity model. All embedding-based methods use dimensions of 300. Here, embedding distance is obtained from neural embeddings learned with window size of 1 and $\gamma = 1$. In all cases, organization population is defined as the mean annualized number of unique mobile and non-mobile authors, and flux is calculated for all global mobility. Baselines include the top-performing distance metrics calculated between vectors obtained by personalized-page rank (PPR), singular value decomposition (SVD), laplacian eigenmap, direct-factorization following Levy's approach [608], and direct optimization of the gravity model using SVD and multidimensional scaling (MDS), as well as the geographic distance between organizations. Embedding distance better explains and predicts flux than any other baseline, though there is some variation by experimental parameters (Table D.3, Table D.4, and Table D.5).

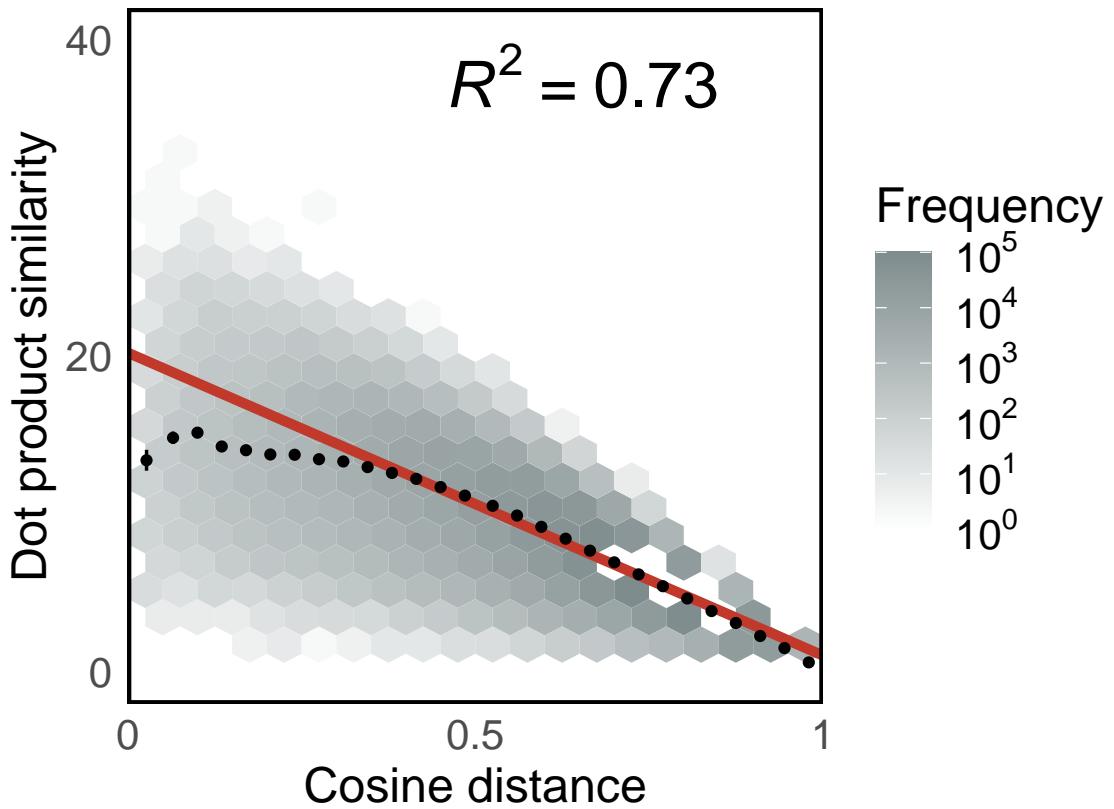


Figure D.6: **Cosine distance is correlated with dot product similarity.** We find a relatively high correlation between the embedding distance—one minus the cosine similarity—and the dot product similarity between organization vectors ($R^2 = 0.73$). Color of each hex bin indicates the frequency of organization pairs. The red line is the line of the best fit. Black dots are mean flux across binned distances. 99% confidence intervals are plotted for the mean flux in each bin based on a normal distribution. Correlation is calculated on the data in the log-log scale ($p < 0.0001$).

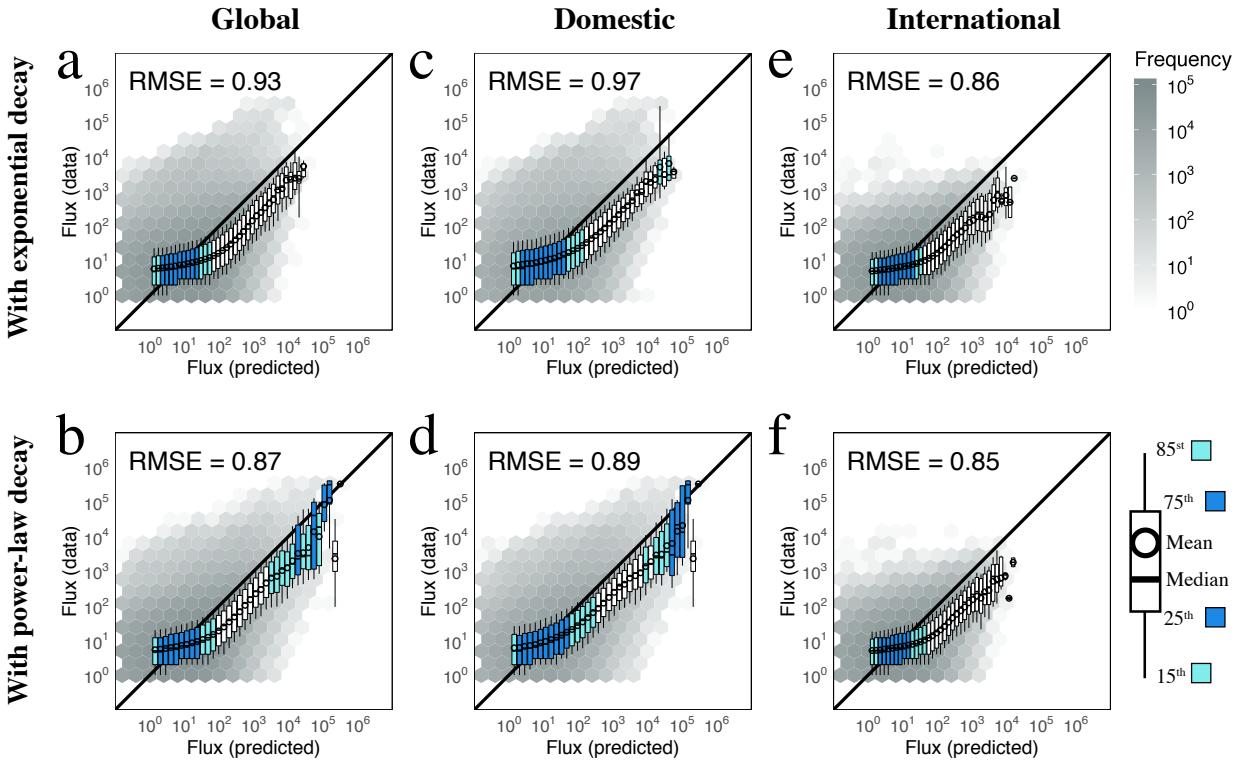


Figure D.7: **For geographic distance, the power-decay gravity model is better.** Flux between organization pairs predicted by the gravity model with different distance decay functions, i.e., exponential decay function (**a**) and power-law decay function (**b**) using geographic distance. Boxplots show distribution of actual flux for binned values of predicted flux. Box color corresponds to the degree to which the distribution overlaps with $x = y$; a perfect prediction yields all points on the black line. “RMSE” is the root-mean-squared error between the actual and predicted values. Shown for all pairs of organization (**a-b**), domestic (**c-d**), and international only (**e-f**) mobility. The gravity model with the power-decay function outperforms that with an exponential decay function.

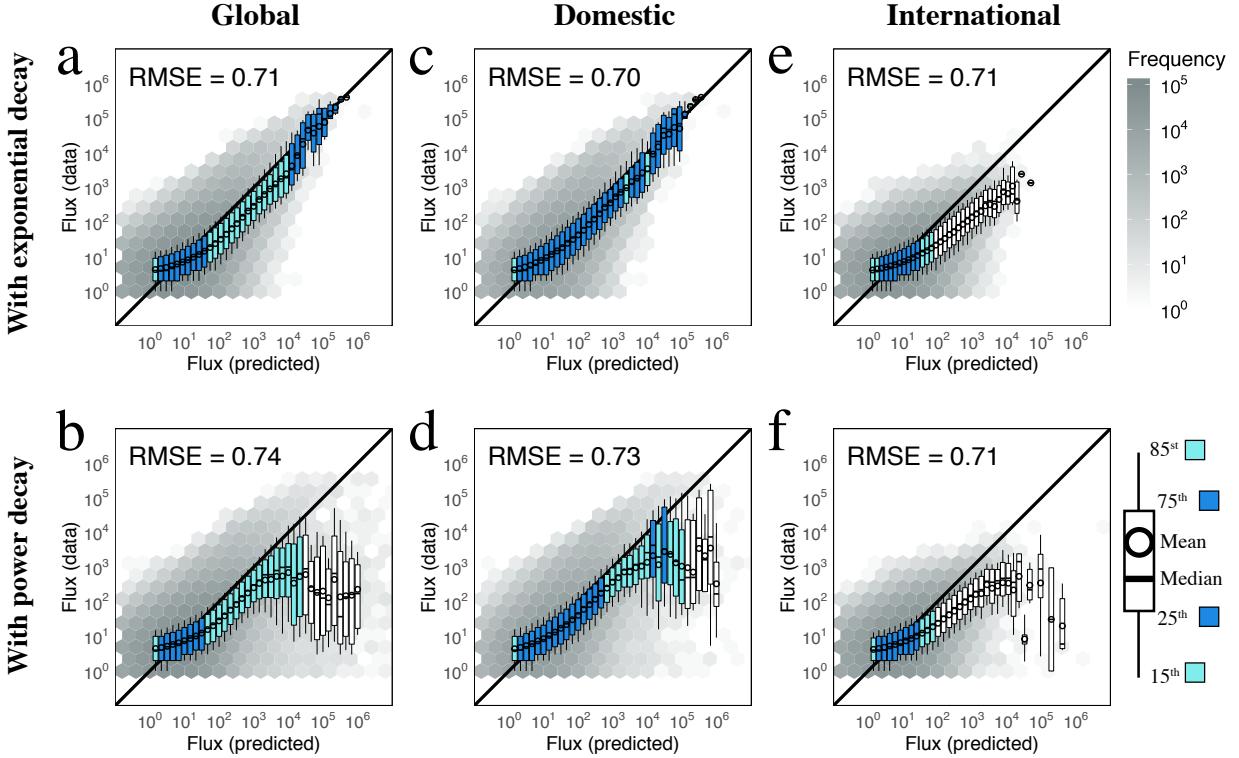


Figure D.8: **For embedding distance, the exponential-decay gravity model is slightly better.** Flux between organization pairs predicted by the gravity model with different distance decay functions, i.e., exponential decay function (a) and power-law decay function (b) using embedding distance. Boxplots show the distribution of actual flux for binned values of predicted flux. Box color corresponds to the degree to which the distribution overlaps with $x = y$; a perfect prediction yields all points on the black line. “RMSE” is the root-mean-squared error between the actual and predicted values. Shown for all pairs of organization (a-b), domestic (c-d), and international only (e-f) mobility. The gravity model with the exponential decay function slightly outperforms that with a power-decay function except in the case of international-only mobility, for which power-decay performs slightly better.

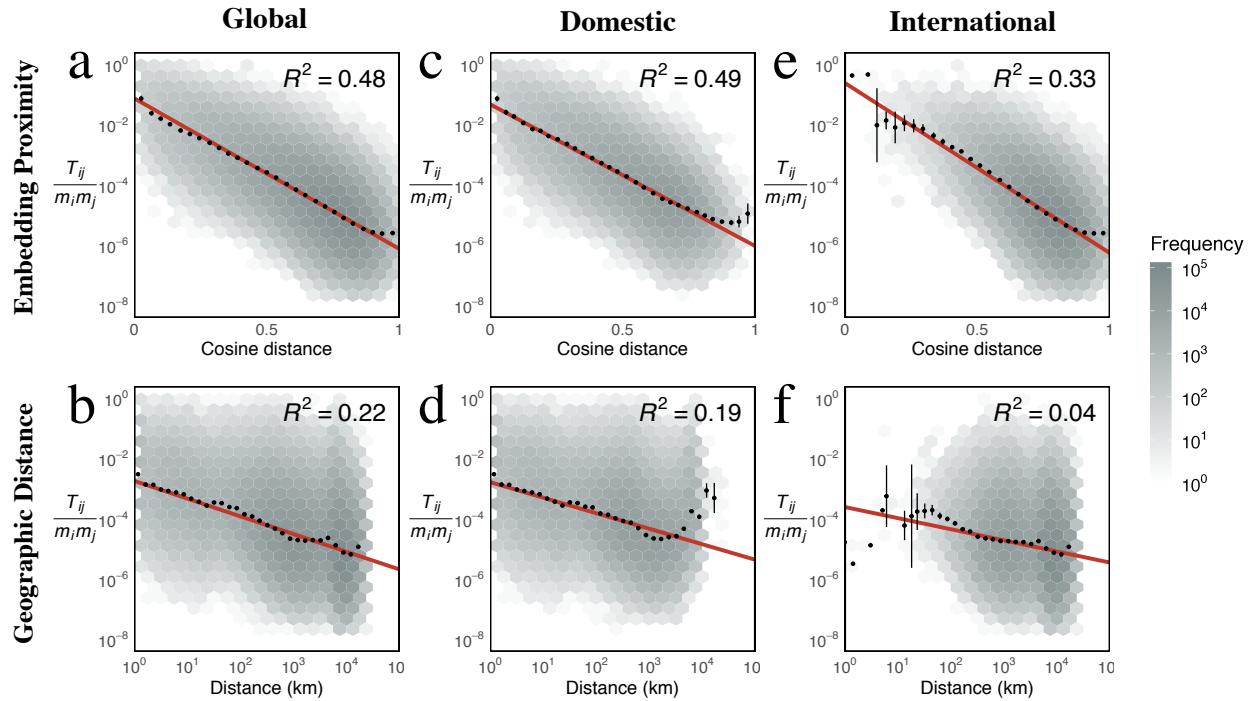


Figure D.9: Embedding distance explains more variance for global, within, and across country flux than geographic distance. **a.** Embedding distance explains more flux than geographic distance (**b**). The red line is the line of the best fit. Black dots are mean flux across binned distances. 99% confidence intervals are plotted for the mean flux in each bin based on a normal distribution. Correlation is calculated on the data in the log-log scale ($p < 0.0001$ across all fits). Color of each hex bin indicates frequency of organization pairs. Results here are identical to those shown in Fig. 6.2. **c-d.** embedding distance explains more variance when considering only within-country organization pairs. **e-f.** embedding distance is more robust than geographic distance when considering only across-country organization pairs.

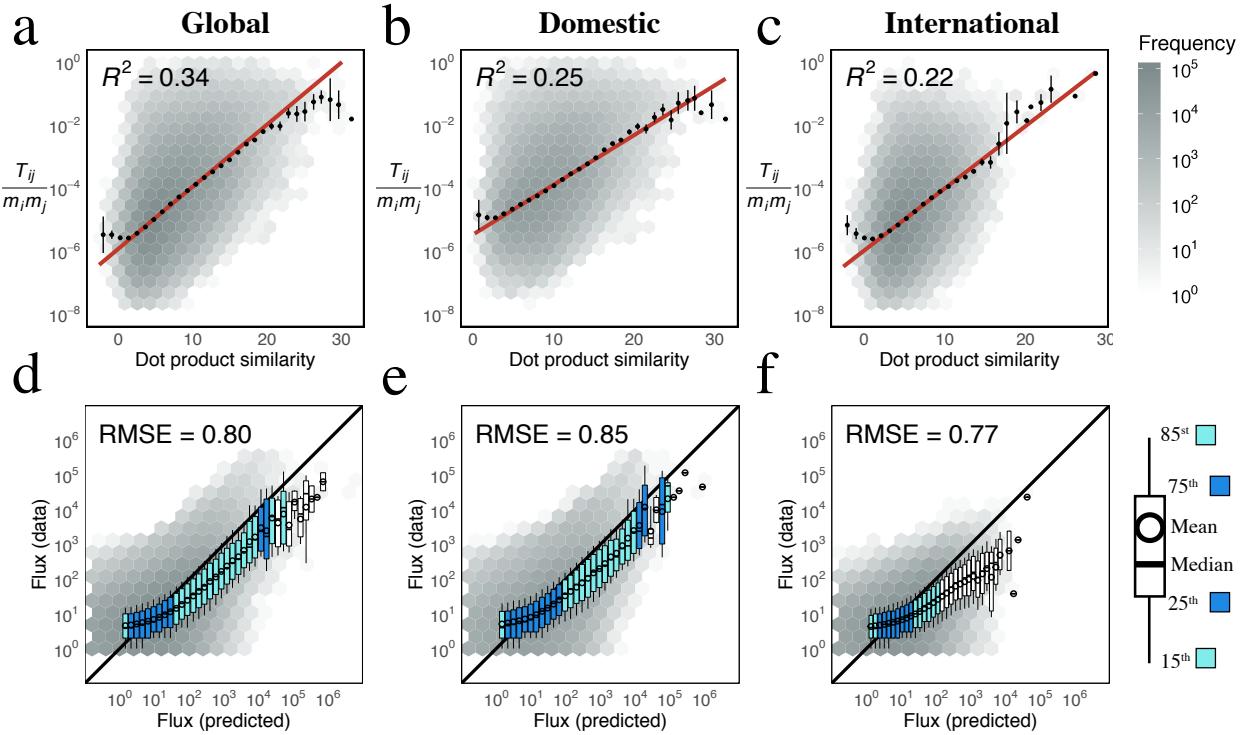


Figure D.10: Examine gravity model with dot product on the embedding space. Performance of dot product similarities in explaining and predicting mobility. Similarity scores are calculated as the pairwise dot product between organizational vectors. Dot product similarity performs better than geographic distance, though worse than cosine similarity in explaining global mobility (**a**), or domestic (**b**) or international (**c**) country mobility. The red line is the line of the best fit. Black dots are mean flux across binned distances. 99% confidence intervals are plotted for the mean flux in each bin based on a normal distribution. Correlation is calculated on the data in the log-log scale ($p < 0.0001$ across all fits). Color indicates frequency of organization pairs within each hex bin. Similarly, PPR distance performs comparably to geographic distance in predicting global (**d**), domestic (**e**) and international (**f**) scientific mobility. Boxplots show distribution of actual flux for binned values of predicted flux. Box color corresponds to the degree to which the distribution overlaps $x = y$; a perfect prediction yields all points on the black line. “RMSE” is the root-mean-squared error between the actual and predicted values.

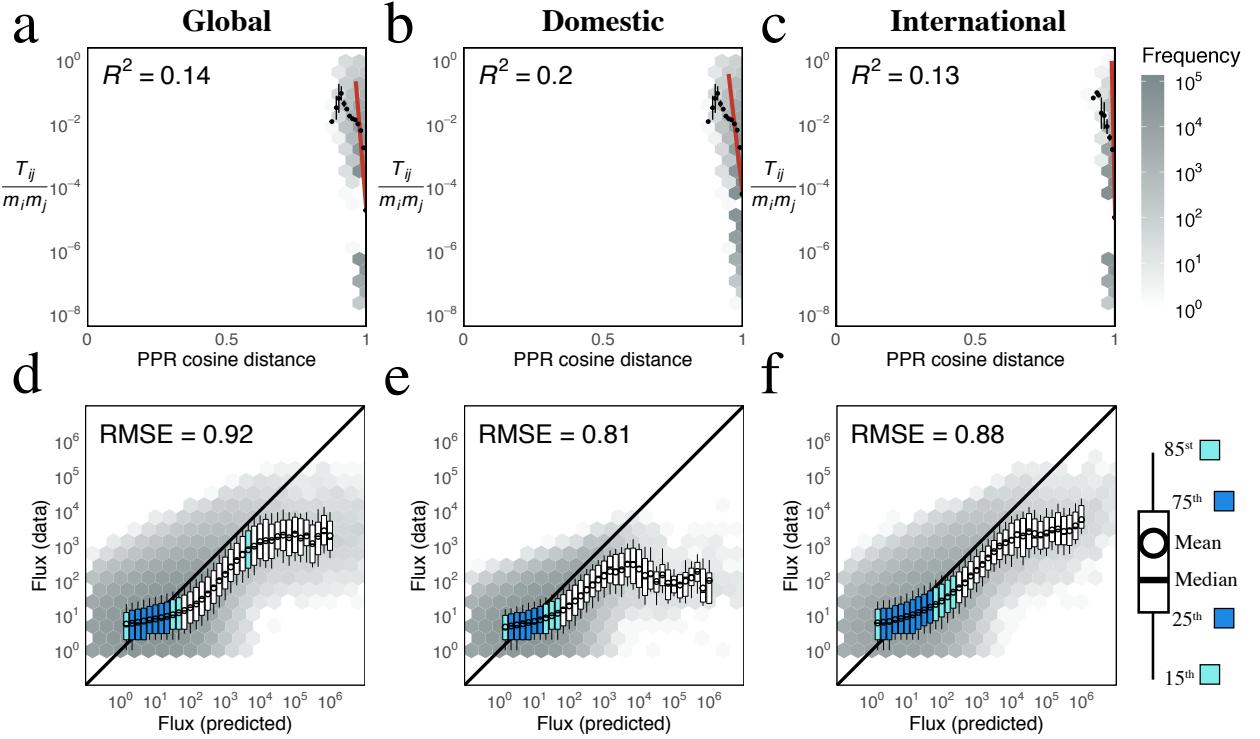


Figure D.11: Personalized page rank with cosine distance. Performance of personalized page rank scores in explaining and predicting mobility. Personalized page rank is calculated for the underlying mobility network, and distance measured as the cosine distance between PPR probability distribution vectors. PPR cosine distance performs roughly similar to geographic distance in explaining global (**a**), domestic (**b**), or international (**c**) country mobility. The red line is the line of the best fit. Black dots are mean flux across binned distances. 99% confidence intervals are plotted for the mean flux in each bin based on a normal distribution. Correlation is calculated on the data in the log-log scale ($p < 0.0001$ across all fits). Color of hex bind indicates frequency of organization pairs. Similarly, PPR distance performs comparably to geographic distance in predicting global (**d**), domestic (**e**) and international (**f**) scientific mobility. Boxplots show distribution of actual flux for binned values of predicted flux. Box color corresponds to the degree to which the distribution overlaps $x = y$; a perfect prediction yields all points on the black line. “RMSE” is the root-mean-squared error between the actual and predicted values.

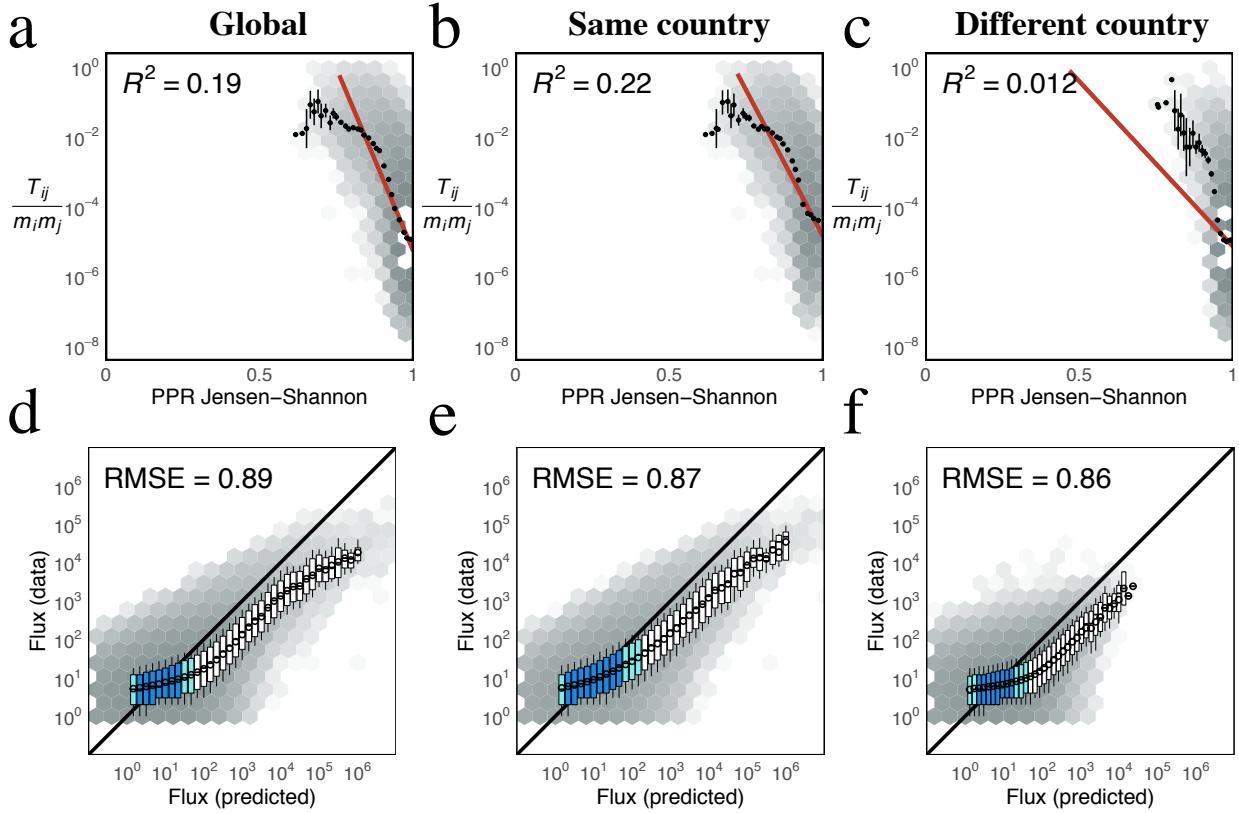


Figure D.12: **Personalized page rank with Jensen-Shannon Divergence.** Performance of personalized page rank scores in explaining and predicting mobility. Personalized page rank is calculated for the underlying mobility network, and distance measured as the Jensen-Shannon Divergence (JSD) between PPR probability distribution vectors. PPR JSD performs roughly similar to geographic distance in explaining global mobility (a), or domestic (b) or international (c) country mobility. Overall, PPR JSD explains more variance in mobility than using cosine distance (Fig. D.11), except for international mobility, for which cosine similarity out-performs JSD. The red line is the line of the best fit. Black dots are mean flux across binned distances. 99% confidence intervals are plotted for the mean flux in each bin based on a normal distribution. Correlation is calculated on the data in the log-log scale ($p < 0.0001$ across all fits). Color of hex bind indicates frequency of organization pairs. Similarly, PPR JSD performs comparably to geographic distance in predicting global (d), domestic (e) and international (f) scientific mobility. Boxplots show distribution of actual flux for binned values of predicted flux. Box color corresponds to the degree to which the distribution overlaps $x = y$; a perfect prediction yields all points on the black line. “RMSE” is the root-mean-squared error between the actual and predicted values.

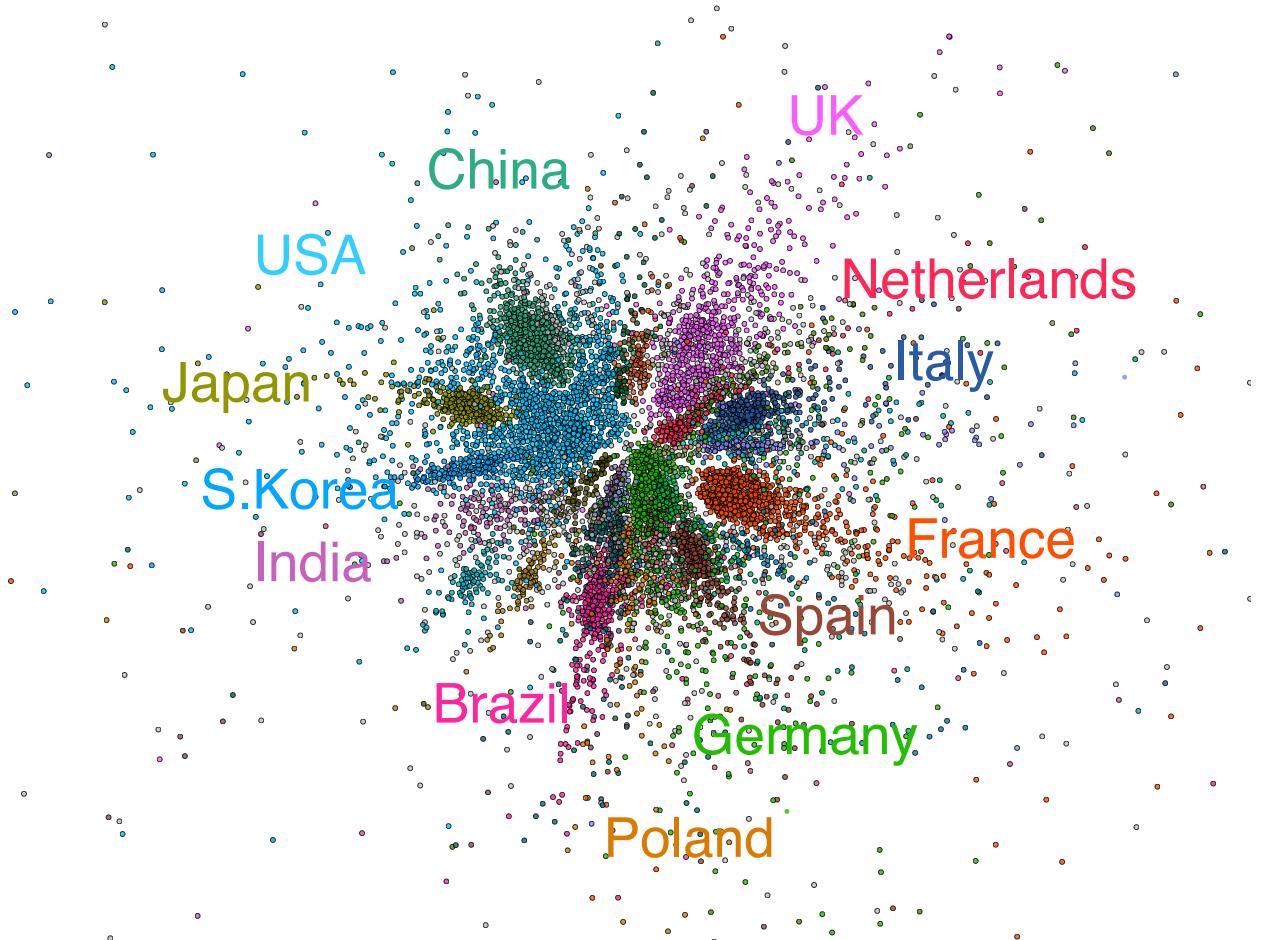


Figure D.13: **Visualization of global mobility network.** The network demonstrates country-level structure, but not at the detail or the extent of the global UMAP projection (Fig. 6.3a). Each node corresponds to an organization, whereas weighted edges (not shown) correspond to the flow of mobile researchers between the two organization. Nodes are colored by the country of the organization. Nodes are positioned using the Force Atlas layout algorithm.

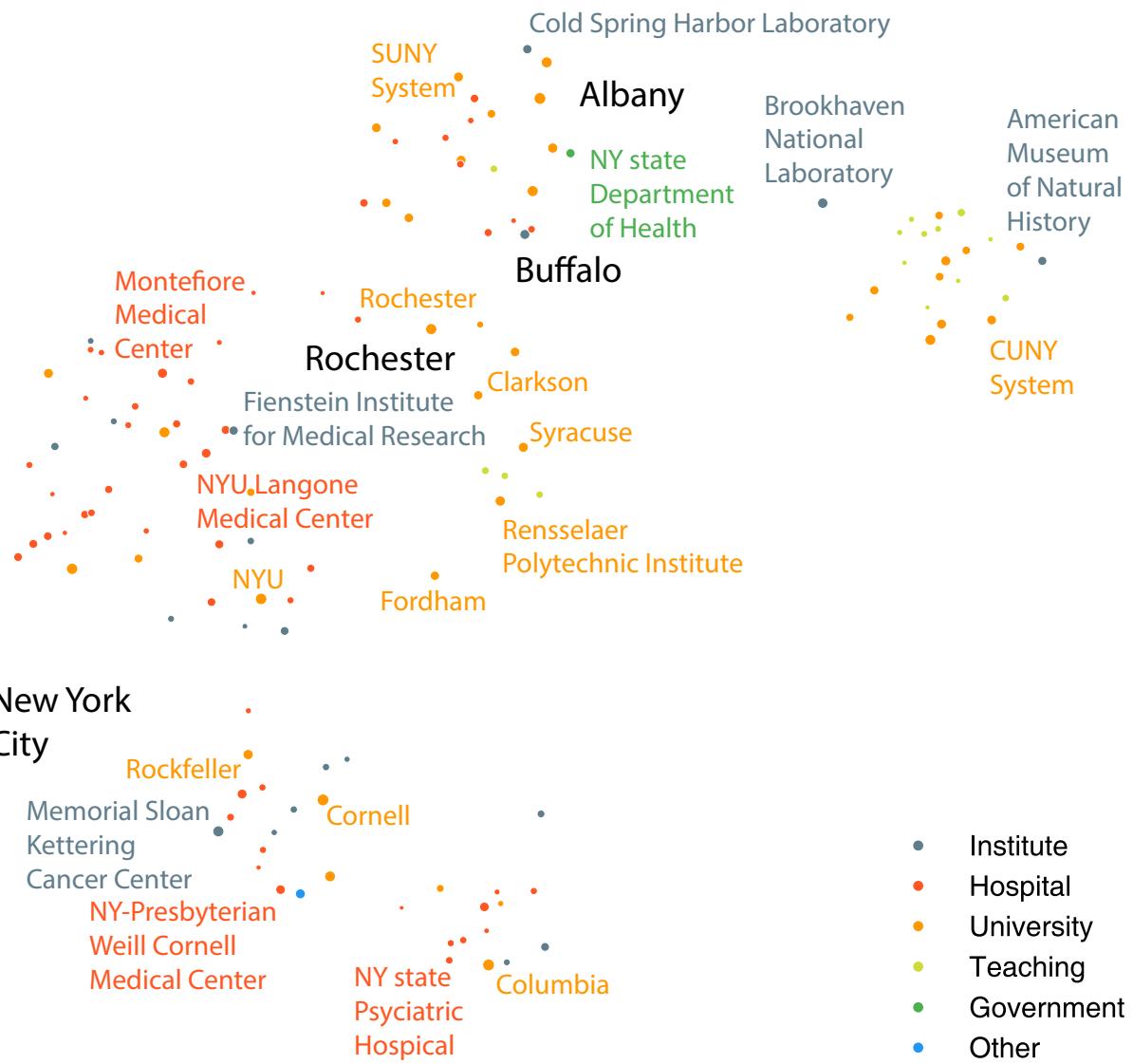


Figure D.14: **UMAP Projection of organizations in New York.** Each point corresponds to an organization and its size indicates the average annual number of mobile and non-mobile authors affiliated with that organization from 2008 to 2019. Color indicates the sector.

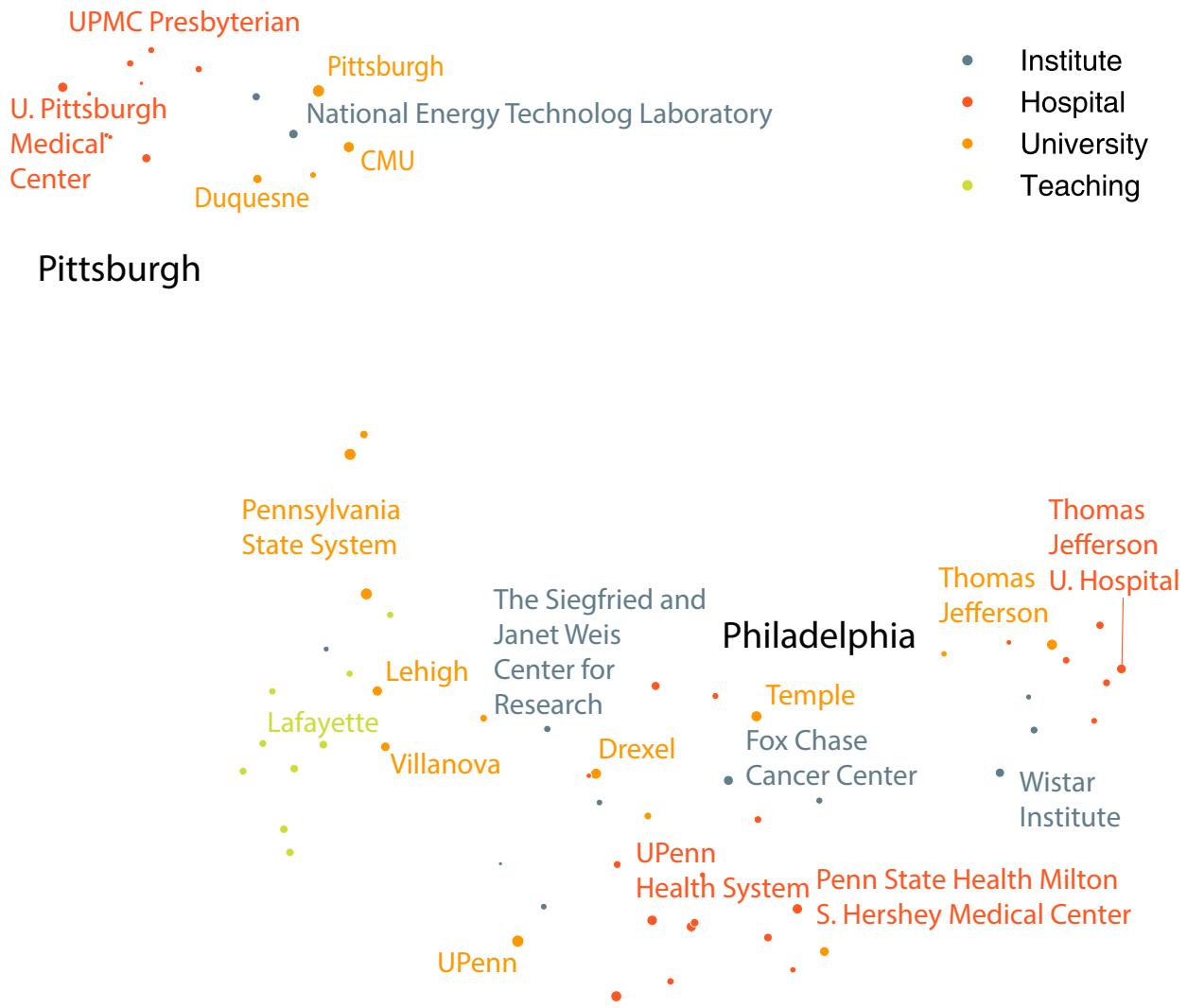


Figure D.15: **UMAP Projection of organizations in Pennsylvania.** UMAP projection of the embedding space of organizations in Pennsylvania reveals clustering based on geography, sector, and academic prestige. Each point corresponds to an organization and its size indicates the average annual number of mobile and non-mobile authors affiliated with that organization from 2008 to 2019. Color indicates the sector.

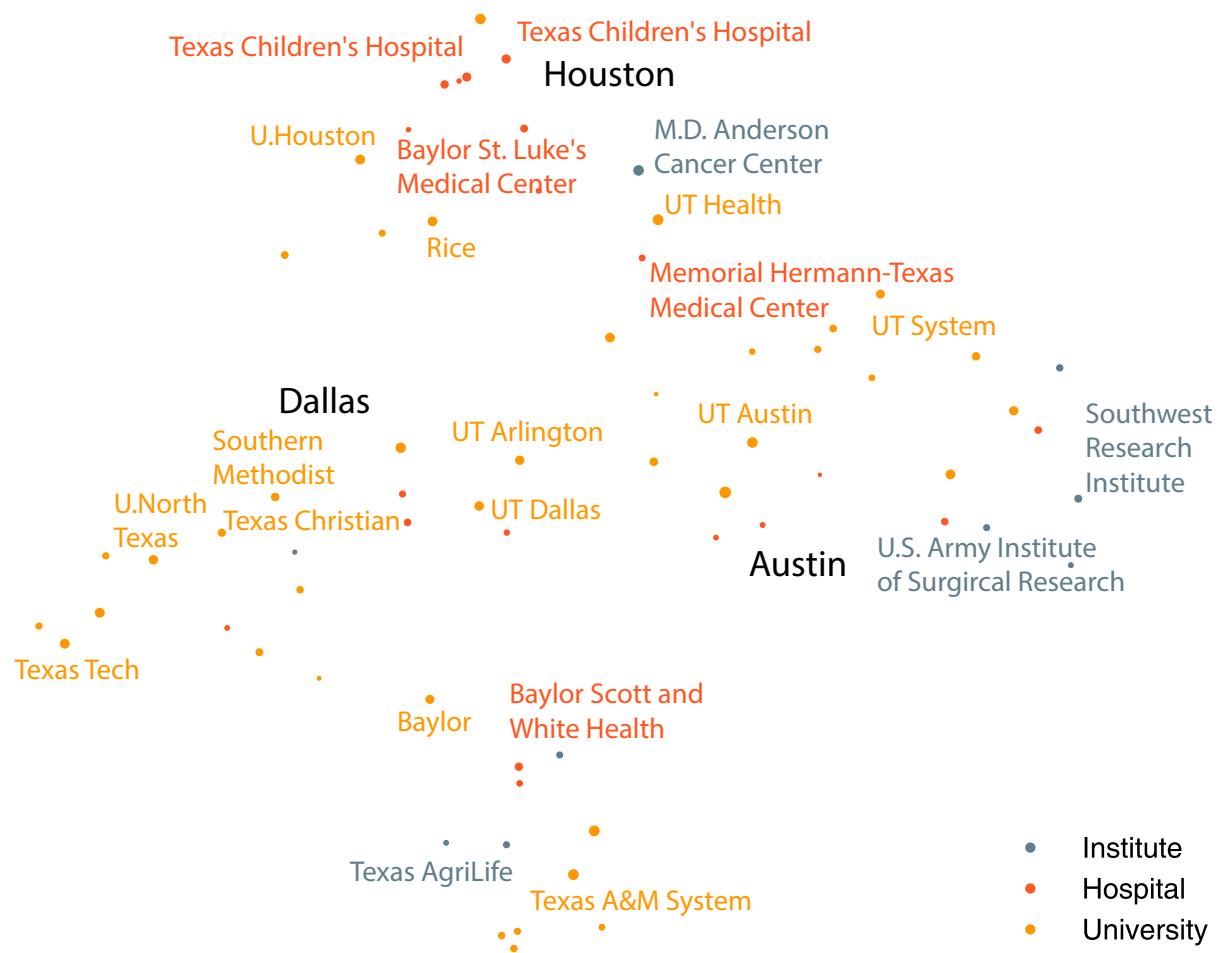


Figure D.16: **UMAP Projection of organizations in Texas.** Each point corresponds to an organization and its size indicates the average annual number of mobile and non-mobile authors affiliated with that organization from 2008 to 2019. Color indicates the sector.

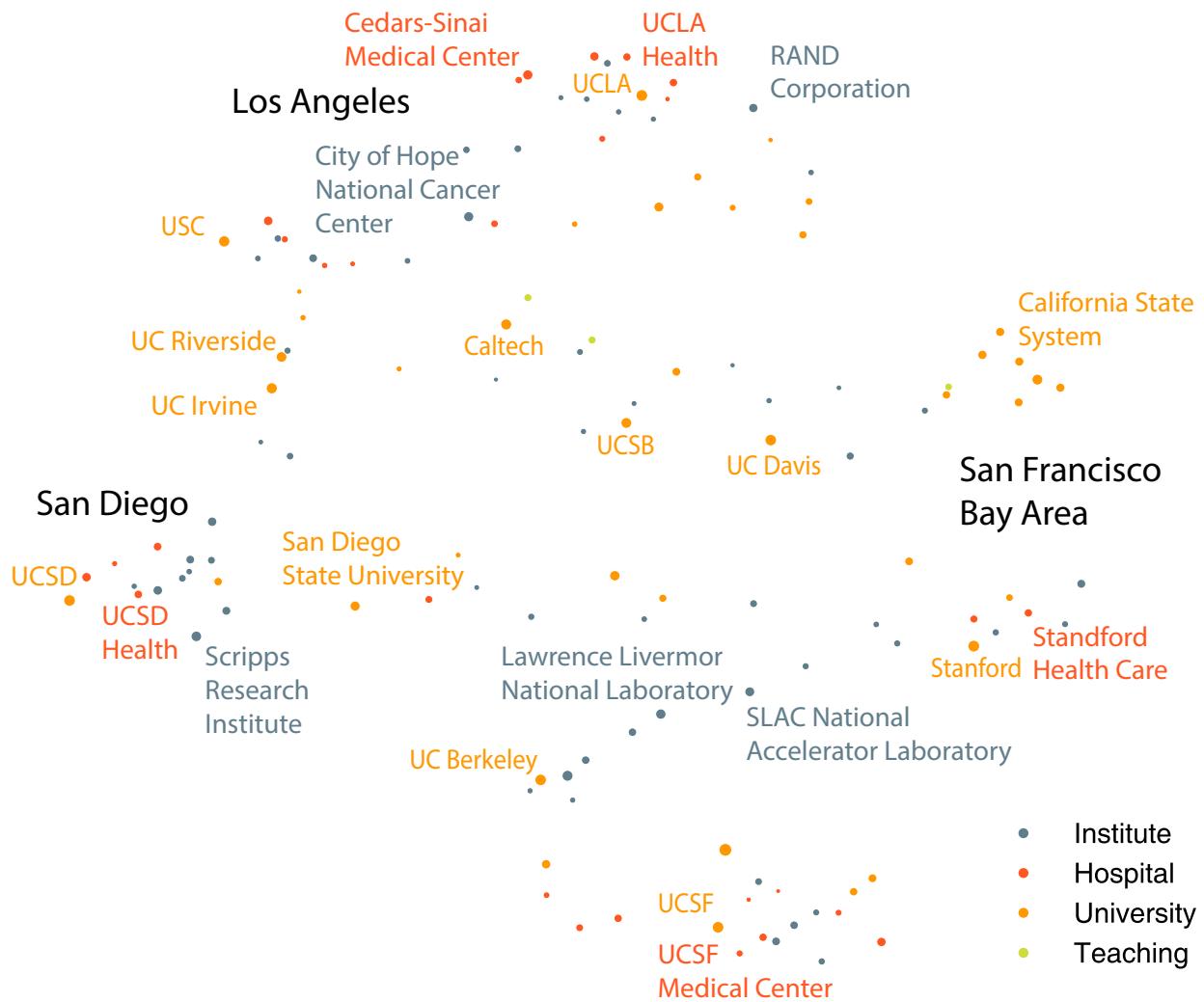


Figure D.17: **UMAP Projection of organizations in California.** Each point corresponds to an organization and its size indicates the average annual number of mobile and non-mobile authors affiliated with that organization from 2008 to 2019. Color indicates the sector.

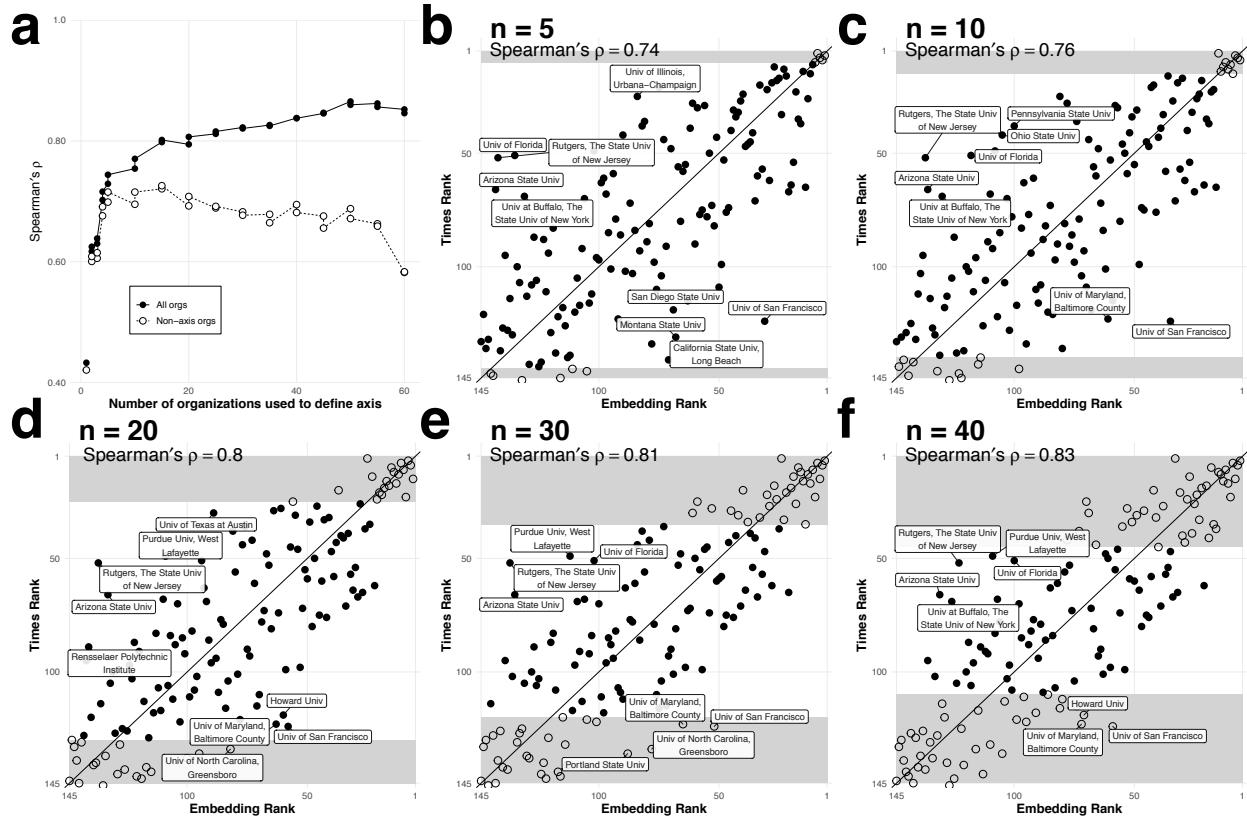


Figure D.18: **SemRank hierarchy is robust.** **a.** Spearman's ρ ($n = 143$) between Times prestige rank and embedding rank derived using SemAxis, with poles defined using the top and bottom (geographically matched) ranked universities. Black points show spearman correlation using all organizations; white points show correlation using only universalizes not aggregated in the poles. Including more universities improves performance, but quickly saturates after around five universities. **b - f.** Comparison between the Times and SemAxis ranks of universities, by the number of universities used to define the poles (n). White points are those top and bottom 20 universities aggregated to define the ends of the axis. The grey box corresponds to the top 20 and bottom 20 ranks. Spearman's ρ details the estimate from Spearman correlation between the two rankings using all universities, including those used to define the ends of each axis. All correlations are significant with $p < 0.0001$.

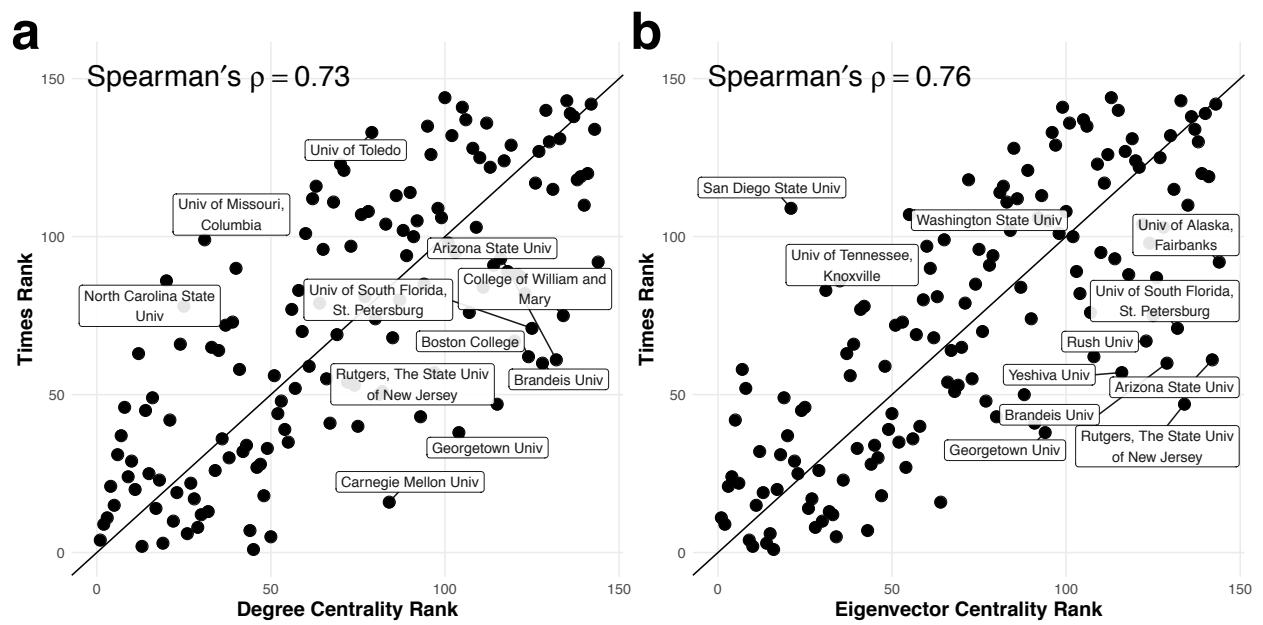


Figure D.19: **Network centrality is strongly correlated with Times ranking.** Comparison between the ranking of organizations by their network-centrality rank and their rank in the 2018 Times Higher Education ranking of U.S. Universities . The Times rank is correlated with degree centrality rank (**a**) with Spearman's $\rho = 0.73$, and is correlated with the eigenvector centrality rank (**b**) with Spearman's $\rho = 0.76$. All correlations are significant with $p < 0.0001$.

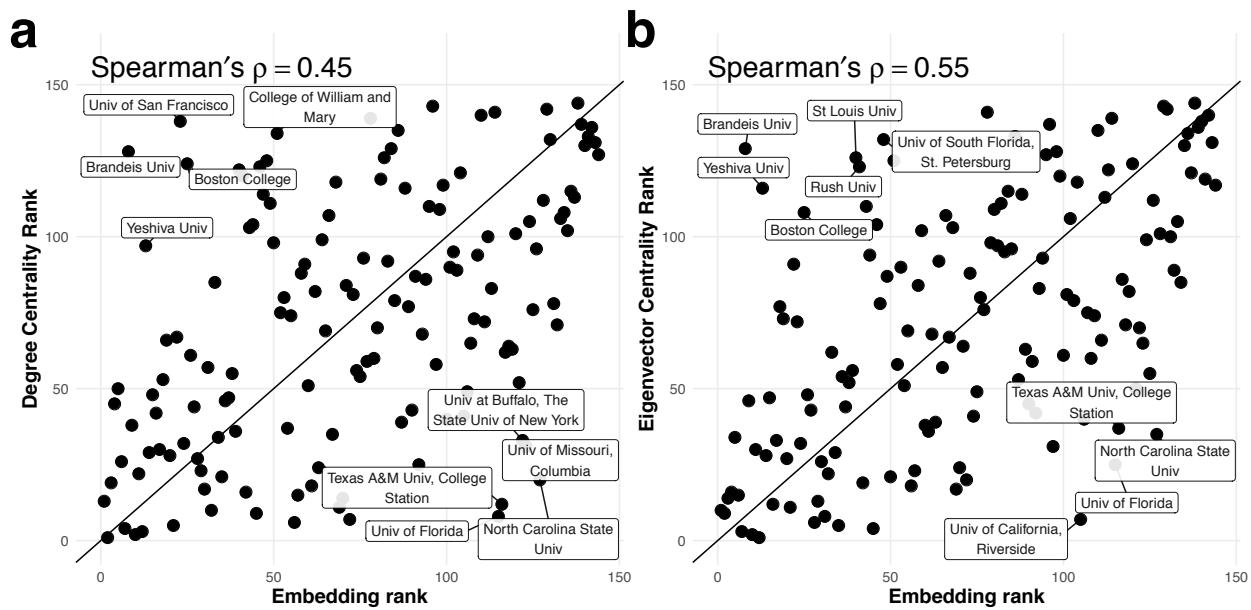


Figure D.20: **Network centrality less correlated with Embedding rank.** Comparison between the ranking of organizations by their network-centrality rank and the embedding rank derived with SemAxis with poles defined using the top five to geographically-matched bottom five universities ranked by the 2018 Times Higher Education ranking of U.S. Universities . Embedding rank is correlated with degree centrality rank (**a**) with Spearman's $\rho = 0.45$, and is correlated with the eigenvector centrality rank (**b**) with Spearman's $\rho = 0.55$. All correlations are significant with $p < 0.0001$.

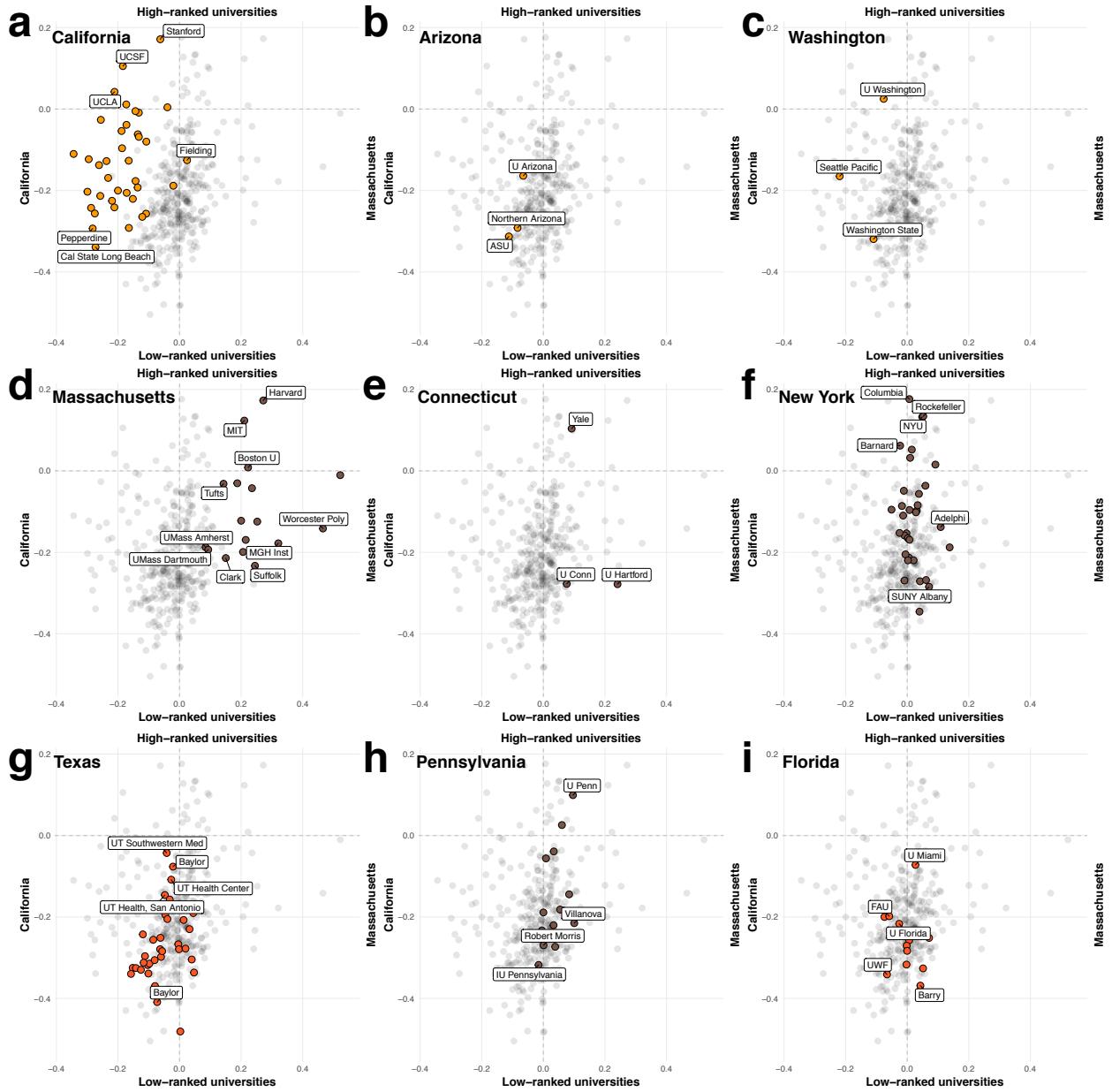


Figure D.21: Geography and prestige SemAxis by U.S. state. SemAxis projection along two axes, comparing California to Massachusetts universities (left to right), and between the top 20 and geographically-matched bottom 20 universities ranked by the 2018 Times Higher Education ranking of U.S. Universities (bottom to top). Points correspond to universities shown for California (a), Arizona (b), Washington (c), Massachusetts (d), Connecticut (e), New York (f), Texas (g), Pennsylvania (h), and Florida (i). Grey points correspond to all other U.S. universities. Full organization names listed in Table D.1.

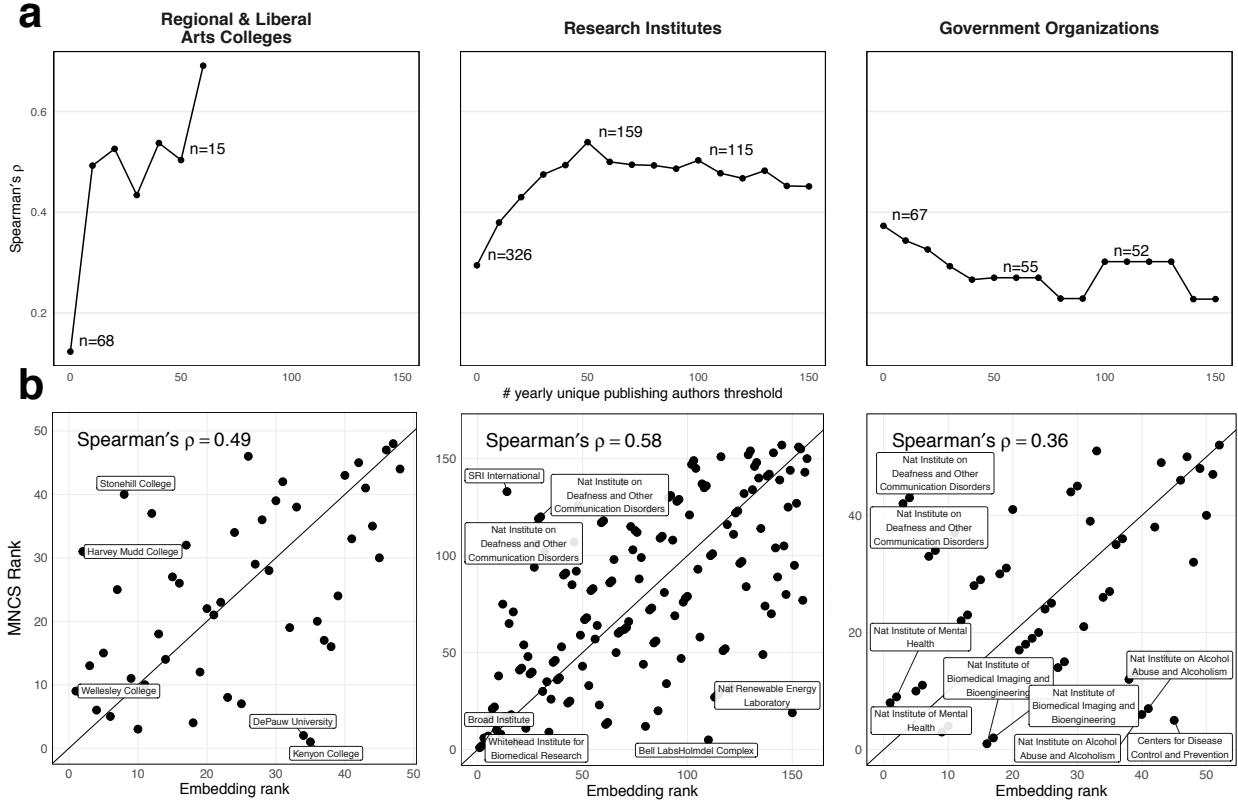


Figure D.22: SemAxis reconstructs publication impact in non-university sectors. Comparison between the ranking of organizations in each non-university sector by their citation impact and the embedding rank. Citation impact is calculated as the mean-normalized citation score using papers published in the Web of Science database between 2008 and 2019. The embedding rank is derived by first projecting non-university organizations onto the SemAxis axis formed with poles defined using the top five to geographically-matched bottom five universities ranked by the 2018 Times Higher Education ranking of U.S. Universities. **a** Shows how the correlation between the citation impact and SemAxis rankings differ while varying the size threshold for including an organization. Size is calculated as the mean annualized number of unique authors publishing with that organization. Annotations show the number of organizations remaining at thresholds of 0, 50, and 100. **b**. Comparison of organizations using a size threshold of 10 for regional and liberal arts colleges, and 50 for research institutes and government organizations; these thresholds were chosen as points thresholds of stability in **a**. The impact rank is correlated with the embedding rank for regional and liberal arts colleges with Spearman's $\rho = 0.49$ ($n = 48$), research institutes with Spearman's $\rho = 0.58$ ($n = 159$), and for government organizations with Spearman's $\rho = 0.36$ ($n = 55$). All correlations are significant with $p < 0.001$.

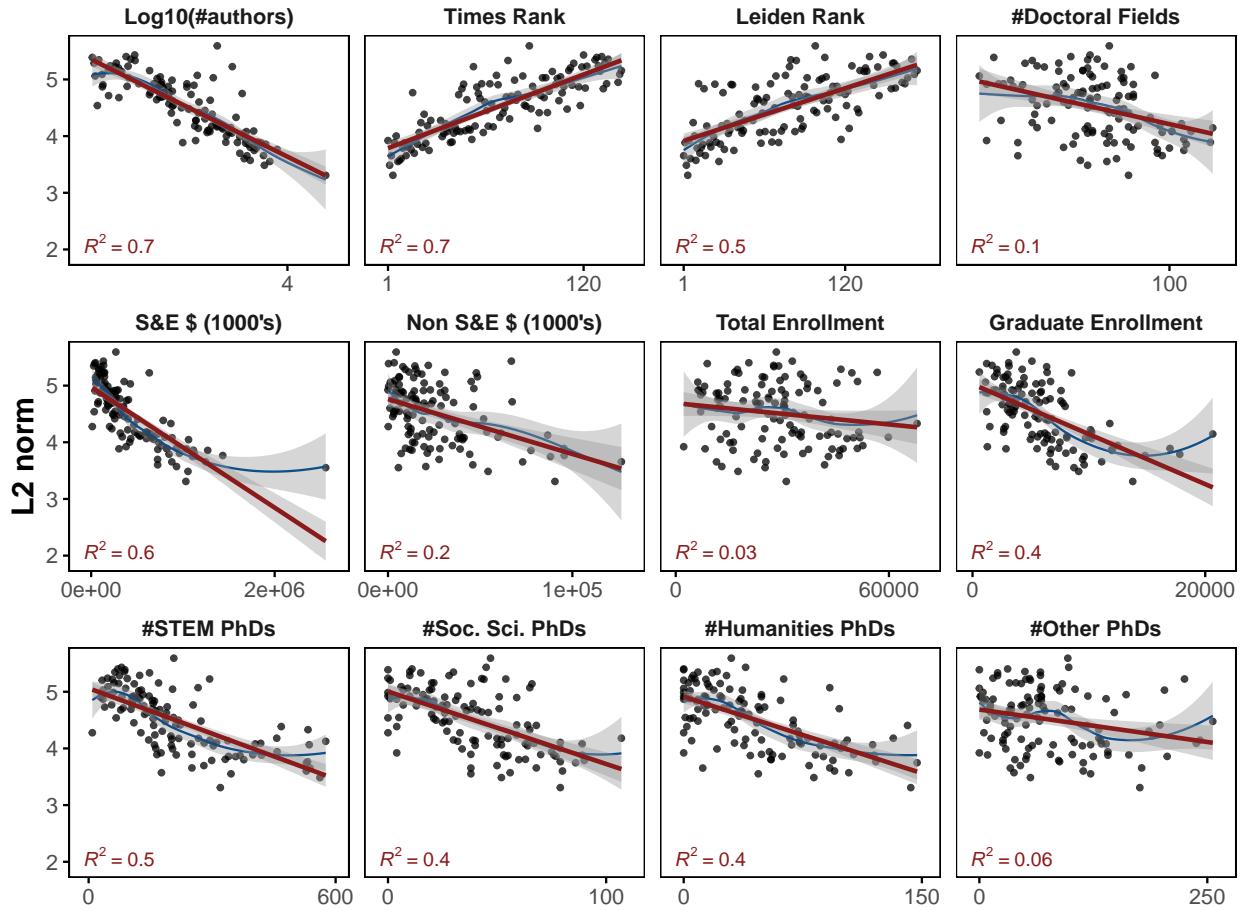


Figure D.23: **Factors relating to the L2 norm of vectors for U.S. universities** Correlation between the L2 norm of organization embedding vectors of U.S. universities and characteristics of U.S. universities. Dots correspond to organizations. The red line is the line of the best fit with corresponding 99% confidence intervals. Red text is the regression estimate. The blue line is the loess regression line with 99% confidence intervals. Number of authors is the average annual count of unique mobile and non-mobile authors. Rankings are derived from the Times Ranking of World Universities, and the Leiden Rankings of Universities. Remaining variables come from the Carnegie Classification of Higher Education Institutions. The factors that best explain s_i are the number of authors, the rank, the amount of Science and Engineering (S&E) funding, and the number of doctorates granted.

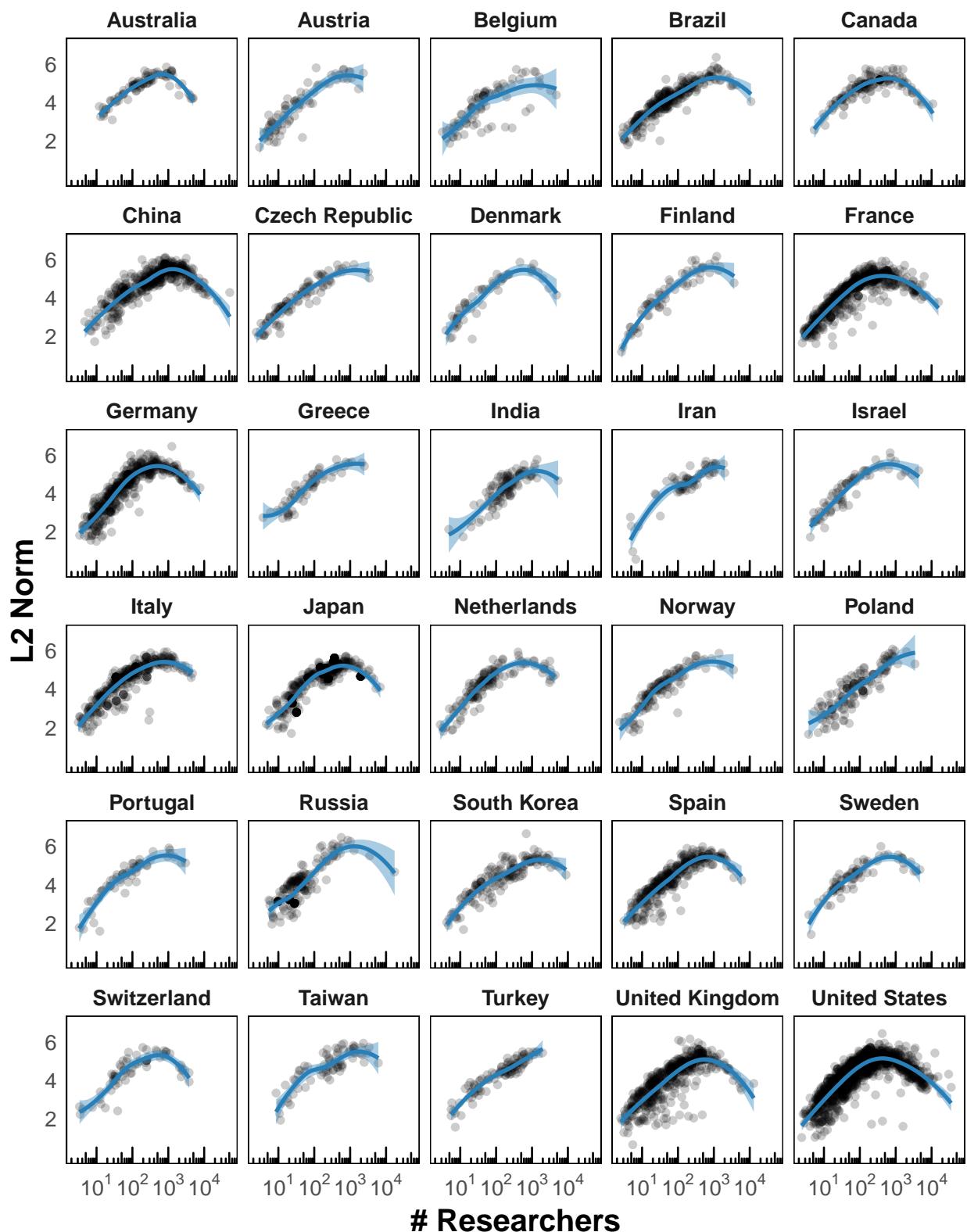


Figure D.24: **Concave-curve repeats across most of 30 countries with most researchers.** Size (L2 norm) of organization embedding vectors compared to their number of researchers for U.S. universities. Loess regression line is shown for each country with 99% confidence intervals. Countries shown are the 30 with the largest number of total unique mobile and non-mobile researchers.

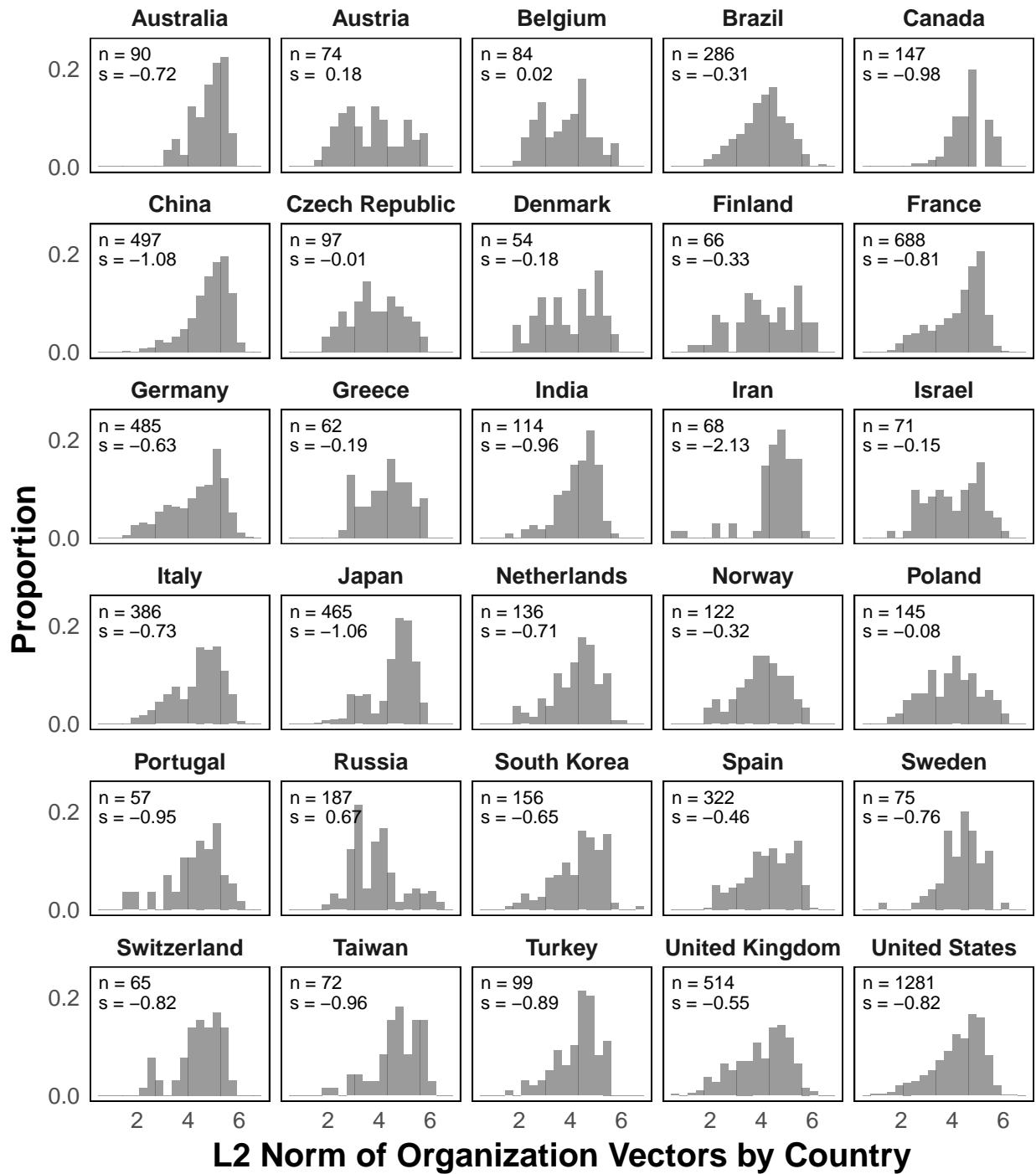


Figure D.25: **Distribution of organization embedding vector norms by country.** Histogram showing the distribution of L2 norm values of organization embedding vectors in each of the 30 countries with the largest number of total unique mobile and non-mobile researchers. Text in each panel shows the number of organizations in the country (n) and the GINI index of inequality of the distribution (g); a small GINI index indicates that the L2 norms of organizations are more balanced, whereas a high GINI value indicates that they are more unequal.

D.3 Tables

Table D.1: Full organization names

Short	Full	Short	Full
Stanford	Stanford Univ	Northwestern	Northwestern Univ
Columbia	Columbia Univ	Ball State	Ball State Univ
Harvard	Harvard Univ	IU Bloomington	Indiana Univ, Bloomington
UCLA	Univ of California, Los Angeles	Stevens Institute	Stevens Institute of Technology
Cal State Long Beach	California State Univ, Long Beach	NJIT	New Jersey Institute of Technology
Wright State	Wright State Univ	NYU	New York Univ
U Toledo	Univ of Toledo	SUNY Albany	Univ at Albany, The State Univ of New York
Boston U	Boston Univ	NY Medical College	New York Medical College
Suffolk	Suffolk Univ	Miami University	Miami Univ
CUNY	City Univ of New York (CUNY)	IU Pennsylvania	Indiana Univ of Pennsylvania
U Arizona	Univ of Arizona	Baylor	Baylor College of Medicine
OSU	Ohio State Univ	UT Health Center	Univ of Texas Health Science Center
MIT	Massachusetts Institute of Technology	Bard College	Bard College
Princeton	Princeton Univ	Stonehill College	Stonehill College
GCU	Grand Canyon Univ	Carleton College	Carleton College
Northcentral	Northcentral Univ	Hanover College	Hanover College
UCSF	Univ of California, San Francisco	Queens College	Queens College
Fielding	Fielding Graduate Univ	DePauw	DePauw College
Pepperdine	Pepperdine Univ	Naval Academy	United States Naval Academy
Argosy	Argosy Univ	Cal State San Marcos	California State Univ San Marcos
Yale	Yale Univ	Broad Inst	Broad Institute
U Hartford	Univ of Hartford	Forsyth Inst	Forsyth Institute
FAU	Florida Atlantic Univ	U Alaska Museum	Univ of Alaska Museum of the North
U Miami	Univ of Miami	Lawrence Berkeley	Lawrence Berkeley Natl Laboratory
UWF	The Univ of West Florida	Allen Institute	Allen Institute for Brain Science
FIT	Florida Institute of Technology	RTI International	RTI InterNatl
Purdue	Purdue Univ, West Lafayette	Fermilab	Fermilab
Notre Dame	Univ of Notre Dame	State of NY	State of New York
Indiana State	Indiana State Univ	Mayo Clinic	Mayo Clinic
Saint Mary's	Saint Mary's College	Fish and Wildlife	Fish and Wildlife Research Institute
Tufts	Tufts Univ	EPA	United States Environmental Protection Agency
Mattel	Mattel Children's Hospital	US Army	United States Army
Clark	Clark Univ	NSF	Natl Science Foundation
UMass Amherst	Univ of Massachusetts Amherst	US Navy	United States Navy
Montclair	Montclair State Univ	US Air Force	United States Air Force
Farleigh Dickinson	Farleigh Dickinson Univ-Metro Campus	Ames Laboratory	Ames Laboratory
Rockefeller	Rockefeller Univ	Olin College	Oin College of Engineering
Adelphi	Adelphi Univ	Scrips Institute	Scrips Institute
Barnard	Barnard College	Idaho Natl Lab	Idaho Natl Laboratory
Saint John Fisher	Saint John Fisher College	Dana Faber	Dana Faber Cancer Institute
U Penn	Univ of Pennsylvania	Dept of Agriculture	United States Department of Agriculture
Villanova	Villanova Univ	DOE	United States Department of Energy
Widener	Widener Univ-Main Campus	NIAMS	Natl Institute of Arthritis, Skin Diseases
Robert Morris	Robert Morris Univ	JMI Labs	JMI Laboratories
U Cincinnati	Univ of Cincinnati	Whitehead Inst	Whitehead Institute of Biomedical Research
Case Western	Case Western Reserve Univ	Wellesley	Wellesley Univ
Ashland	Ashland Univ	UT Health, San Antonio	Univ of Texas Health Science Center, San Antonio
Texas A&M	Texas A&M Univ-Commerce	UNT	Univ of North Texas
Texas Southern	Texas Southern Univ	UT Southwestern Med	Univ of Texas Southwestern Medical Center
Baylor	Univ of Mary Hardin-Baylor	UT El Paso	Univ of Texas, El Paso
U Washington	Univ of Washington - Seattle	USF	Univ of South Florida, Tampa
Washington State	Washington State Univ	Florida A&M	Florida Agricultural and Mechanical Univ
Seattle Pacific	Seattle Pacific Univ	Barry	Barry Univ
Cal State Fresno	California State Univ-Fresno	UMass Dartmouth	Univ of Massachusetts Dartmouth
Northern Arizona	Northern Arizona Univ	Worcester Poly	Worcester Polytechnic Institute
IUPUI	Indiana Univ - Purdue Univ Indianapolis	Umass Boston	Univ of Massachusetts Boston
U Dayton	Univ of Dayton	MGH Inst	MGH Institute of Health Professions
U Conn	Univ of Connecticut	Joseph W. Jones Center	Joseph W. Jones Ecological Research Center
ASU	Arizona State Univ	Vaccine Research Center	Vaccine Research Center, San Diego
U Florida	Univ of Florida	LA Ag Center	Lousiana Agricultural Center
Northern Illinois	Northern Illinois Univ	FL Fish and Wildlife	Florida Fish and Wildlife Conservation Commission
Concordia Chicago	Concordia Univ-Chicago	NHLBI	Natl Heart, Lung, and Blood Institute
U Chicago	Univ of Chicago	NY Dept. of Health	New York Department of Health
SIU Edwardsville	Southern Illinois Univ, Edwardsville	St Michaels	Saint Michaels College
SIU Carbondale	Southern Illinois Univ, Carbondale		

Table D.2: **L2 Norm of country's representative vectors.** Shown for top 30 countries with the most unique mobile and non-mobile researchers

Country	L2 Norm	# Organizations
United States	2.39	1281
Germany	2.6	485
United Kingdom	2.61	514
Austria	2.64	74
France	2.83	688
Belgium	2.84	84
Switzerland	2.85	66
Spain	2.94	322
China	2.97	497
India	2.99	114
Poland	3.02	145
Canada	3.02	147
Italy	3.04	386
Russia	3.08	187
Norway	3.1	122
Netherlands	3.11	136
Sweden	3.16	75
Brazil	3.16	286
Finland	3.17	66
Denmark	3.21	54
Czech Republic	3.23	97
Greece	3.24	62
Australia	3.24	90
Turkey	3.28	99
South Korea	3.28	156
Israel	3.32	71
Portugal	3.33	57
Japan	3.35	465
Iran	3.57	68
Taiwan	3.67	72

Table D.3: Correlation between flux and distance over metrics, experimental parameters. Each cell corresponds to the correlation between the real-world flux between scientific organizations (measured with R^2) and baseline metrics, shown by subsets of mobility data and by definitions of organization population. The asterisk denotes the top-performing distance metric by column. Distance metrics are ordered from highest R^2 to lowest, based on global mobility with organization population defined using all mobile and non-mobile authors. “All” means that population is defined as the average yearly number of unique mobile and non-mobile scholars who published with the organizations’ affiliation; population is defined in the same way for “Mobile only”, except only using unique mobile researchers; “Raw freq” means that organization populations are defined as their frequency across all the trajectories, similar to word frequency in language embedding. Embedding distance, measured as the cosine distance between embedding vectors, explains more of the flux than baselines in nearly every case, except using raw frequency population and domestic and international mobility, where direct optimization of the gravity model works better, as well as Levy’s factorization [608] for domestic and international only mobility.

	All	Mobile only	Global	Domestic	International	Global	Domestic	International
Embedding cosine	*0.481	*0.418	*0.435	*0.492	*0.456	*0.489	0.325	0.251
Gravity MDS euclidean	0.355	0.165	0.161	0.328	0.112	0.115	*0.369	0.174
Levy’s Euclidean	0.341	0.369	0.382	0.213	0.271	0.284	0.305	*0.323
Embedding dot	0.341	0.313	0.316	0.254	0.265	0.267	0.218	0.181
SVD cosine	0.247	0.297	0.309	0.213	0.314	0.325	0.111	0.152
Geographic	0.219	0.174	0.197	0.188	0.157	0.176	0.04	0.019
Laplacian cosine	0.212	0.199	0.218	0.176	0.157	0.18	0.079	0.1
Levy’s cosine	0.208	0.227	0.231	0.169	0.246	0.246	0.057	0.054
PPR JSD	0.194	0.276	0.276	0.218	0.335	0.335	0.012	0.077
PPR cosine	0.138	0.136	0.143	0.196	0.186	0.197	0.13	0.149
Gravity SVD cosine	0.122	0.118	0.122	0.118	0.133	0.138	0.056	0.047
Levy’s dot	0.004	0.002	0.013	0.002	0.002	0.004	0.039	0.034

Raw freq

Table D.4: Prediction error between actual and predicted mobility with exponential gravity model, by metrics and experimental parameters. Each cell corresponds to the prediction error (measured with root mean squared error) when using each distance as input to the exponential form of the gravity model of mobility to predict the flux between organizations, shown by subsets of mobility data, and by definitions of organization population. The asterisk denotes the top-performing distance metric by column (lowest prediction error). Distance metrics are ordered from lowest prediction error to highest, based on global mobility with organization population defined using all mobile and non-mobile authors. “All” means that population is defined as the average yearly number of unique mobile and non-mobile scholars who published with the organizations’ affiliation; population is defined in the same way for “Mobile only”, except only using unique mobile researchers; “Raw freq” means that organization populations are defined as their frequency across all the trajectories, similar to word frequency in language embedding. Embedding distance, measured as the cosine distance between embedding vectors, results in better predictions of mobility than baselines in nearly every case, however Levy’s factorization [608] perform better in the case of international and domestic only mobility when using raw frequency populations.

	All	Mobile only	Raw freq	Global	Domestic	International	Global	Domestic	International
Embedding cosine	*0.713	*0.76	*0.737	*0.702	*0.749	*0.713	*0.715	0.764	0.748
³ Levy’s Euclidean	0.803	0.791	0.771	0.874	0.867	0.844	0.725	*0.726	*0.705
Embedding dot	0.803	0.825	0.811	0.851	0.87	0.854	0.769	0.798	0.784
SVD cosine	0.859	0.835	0.815	0.874	0.841	0.82	0.82	0.812	0.792
Laplacian cosine	0.878	0.891	0.867	0.894	0.933	0.904	0.835	0.837	0.815
Levy’s cosine	0.881	0.875	0.86	0.898	0.882	0.866	0.844	0.858	0.841
PPR JSD	0.888	0.847	0.835	0.871	0.828	0.814	0.865	0.847	0.834
Gravity MDS Euclidean	0.904	0.944	0.93	0.907	0.972	0.955	0.793	0.838	0.821
PPR cosine	0.918	0.925	0.908	0.884	0.916	0.894	0.812	0.814	0.796
Geographic	0.929	0.951	0.928	0.973	1.002	0.983	0.856	0.875	0.853
Gravity SVD cosine	0.933	0.939	0.923	0.92	0.946	0.927	0.853	0.865	0.847
Levy’s dot	0.987	0.995	0.98	0.979	1.014	0.996	0.853	0.867	0.849

Table D.5: Prediction error between actual and predicted mobility with power-law gravity model, by metrics and experimental parameters. Each cell corresponds to the prediction error (measured with root mean squared error) when using each distance as input to the power-law form of the gravity model of mobility to predict the flux between organizations, shown by subsets of mobility data, and by definitions of organization population. The asterisk denotes the top-performing distance metric by column (lowest prediction error). Distance metrics are ordered from lowest prediction error to highest, based on global mobility with organization population defined using all mobile and non-mobile authors. “All” means that population is defined as the average yearly number of unique mobile and non-mobile scholars who published with the organizations’ affiliation; population is defined in the same way for “Mobile only”, except only using unique mobile researchers; “Raw freq” means that organization populations are defined as their frequency across all the trajectories, similar to word frequency in language embedding. Embedding distance, measured as the cosine distance between embedding vectors, results in better predictions for global mobility with “All”; and “Mobile only” definitions of population, though direct gravity-law optimization with MDS performs better when using raw frequencies to measure organization population, and Levy’s factorization [608] perform better here with domestic and international only mobility.

	All	Mobile only	Global	Domestic	International	Global	Domestic	International
Embedding cosine	*0.743	0.784	0.76	*0.73	*0.784	*0.749	0.714	0.762
³⁹ Levy’s Euclidean	0.78	*0.776	*0.755	0.844	0.847	0.822	0.703	*0.711
Gravity MDS Euclidean	0.795	0.91	0.898	0.808	0.957	0.939	*0.691	0.802
Embedding dot	0.822	0.844	0.831	0.864	0.879	0.863	0.783	0.809
SVD cosine	0.839	0.785	0.761	0.89	0.834	0.81	0.792	0.752
Laplacian cosine	0.87	0.837	0.801	0.937	0.919	0.886	0.804	0.772
Geographic	0.874	0.905	0.879	0.888	0.933	0.906	0.852	0.874
PPR JSD	0.889	0.85	0.838	0.871	0.834	0.819	0.865	0.847
PPR cosine	0.92	0.927	0.91	0.885	0.918	0.896	0.812	0.815
Levy’s cosine	0.927	0.926	0.911	0.93	0.924	0.909	0.861	0.872
Gravity SVD cosine	0.965	0.965	0.95	0.955	0.958	0.941	0.873	0.883
Levy’s dot	0.99	0.998	0.983	0.977	1.015	0.997	0.845	0.86

Dakota S. Murray

Phone: +1 (828) 691 5724
email: dakmurra@iu.edu; dakota.s.murray@gmail.com
URL: <http://www.dakotamurray.me>
ORCID ID: [0000-0002-7119-0169](https://orcid.org/0000-0002-7119-0169)
GITHUB: <https://github.com/murrayds>

Current Position

Postdoctoral Associate
Center for Complex Networks Research (Advisor: Albert-Laszlo Barabasi)
Network Science Institute, Northeastern University

Areas of Specialization

Science of Science • Computational Social Science • Data Science

Education

- 2016-2021 PH.D. in Informatics
Track in *Computing, Culture, and Society*; Minor in *Statistics*
School of Informatics, Computing, and Engineering, Indiana University Bloomington
Dissertation: “*Embracing Complexity in the Science of Science*”
Advisors: Drs. Cassidy R. Sugimoto and Yong-Yeol Ahn
- 2012-2016 B.S. in Computer Science
Department of Computer Science, Appalachian State University
Advisor: Dr. Mitch Parry

Appointments and Employment

- 2021-present Postdoctoral Associate
Dr. Albert-Laszlo Barabasi
Northeastern University
- 2019-2021 Research Assistant
Dr. Yong-Yeol Ahn, *Science Genome Project*
School of Informatics, Computing, and Engineering, Indiana University
- 2018-2019 Associate Instructor
School of Informatics, Computing, and Engineering, Indiana University
- 2017 (Summer) Visiting Scholar
Center for Science and Technology Studies, Leiden University
- 2016-2018 Research Assistant
Dr. Cassidy Sugimoto, *Scholarly Communication Lab*

School of Informatics, Computing, and Engineering, Indiana University

2013-2016 Research Assistant and Web Developer
Department of Computer Science, Appalachian State University

2015 NIFS Student Intern
(Summer & Fall) NASA Langley Research Center

Publications

JOURNAL ARTICLES AND INDEXED CONFERENCE PROCEEDINGS

- Boothby, C., **Murray, D.**, Sugimoto, C. R., Waggy, A. P., & Tsou, A. (in press). Credibility of Scientific Information on Social Media: Variation by Platform and Presence of Formal Credibility Cues.
- Miao, L., **Murray, D.**, Jung, W.-S., Larivière, V., Sugimoto, C. R., & Ahn, Y.-Y. (under review). The latent structure of national scientific development.
- Lamers, W. S., Boyack, K., Larivière, V., Sugimoto, C. R., van Eck, N. J., Waltman, L., & **Murray, D.** (2021). Measuring Disagreement in Science.
- **Murray, D.**, Yoon, J., Kojaku, S., Costas, R., Jung, W., Milojević, S., & Ahn, Y. (under review). Unsupervised embedding of trajectories captures the latent structure of mobility.
- 2020 **Murray, D.**, Boothby, C., Zhao, H., Minik, V., Bérubé, N., Larivière, V., & Sugimoto, C. R. (2020). Exploring the personal and professional factors associated with student evaluations of tenure-track faculty. *PLOS ONE*, 15(6), e0233515.
- 2020 Brown, J., **Murray, D.**, Furlong, K., Coco, E., & Dablander, F. (2021). A breeding pool of ideas: Analyzing interdisciplinary collaborations at the Complex Systems Summer School. *PLOS ONE*, 16(2), e0246260.
- 2019 Robinson-Garcia, N., Sugimoto, C. R., **Murray, D.**, Yegros-Yegros, A., Larivière, V., & Costas, R. (2019). The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics*, 13(1), 50–63.
- 2019 Bu, Y., **Murray, D. S.**, Xu, J., Ding, Y., Ai, P., Shen, J., & Yang, F. (2018). Analyzing scientific collaboration with “giants” based on the milestones of career. *Proceedings of the Association for Information Science and Technology*, 55(1), 29–38.
- 2019 **Murray, D.**, Siler, K., Larivière, V., Chan, W. M., Collings, A. M., Raymond, J., & Sugimoto, C. R. (2019). Author-Reviewer Homophily in Peer Review.[Preprint]. BioRxiv.
- 2018 Robinson-Garcia, N., Sugimoto, C. R., **Murray, D.**, Yegros-Yegros, A., Larivière, V., & Costas, R. (2018). Scientific mobility indicators in practice: International mobility profiles at the country level. *El Profesional de La Información*, 27(3), 511.
- 2018 Bu, Y., Ding, Y., Liang, X., & **Murray, D. S.** (2018). Understanding persistent scientific collaboration. *Journal of the Association for Information Science and Technology*, 69(3), 438–448.
- 2018 Bu, Y., **Murray, D. S.**, Ding, Y., Huang, Y., & Zhao, Y. (2018). Measuring the stability of scientific collaboration. *Scientometrics*, 114(2), 463–479.
- 2018 Chinchilla-Rodríguez, Z., Miao, L., **Murray, D.**, Robinson-García, N., Costas, R., & Sugimoto, C. R. (2018). A Global Comparison of Scientific Mobility and Collaboration According to National Scientific Capacities. *Frontiers in Research Metrics and Analytics*, 3.
- 2017 Sugimoto, C. R., Robinson-Garcia, N., **Murray, D. S.**, Yegros-Yegros, A., Costas, R., & Larivière, V. (2017). Scientists have most impact when they’re free to move. *Nature*, 550(7674), 29–31.
- 2017 Das, S., Goard, J., & **Murray, D.** (2017). How Celebrities Feed Tweeple with Personal and Promotional Tweets: Celebrity Twitter Use and Audience Engagement. In *Proceedings of the 8th*

International Conference on Social Media and Society (pp. 30:1–30:5). New York, NY, USA: ACM.

OTHER PUBLICATIONS

- 2019 Sugimoto C, Allen L, Jeroen B et al. Rethinking Impact Factors: New Pathways in Journal Metrics. *F1000Research* 2019, 8:671 (document).
- 2018 Cassidy R, Sugimoto, **Dakota S. Murray**, and Vincent Larivière (April 26, 2018), "Open citations to open science", *ISSI Blog*.
- 2017 Rodrigo Costas, Alfredo Yegros, Vincent Larivière, Cassidy Sugimoto, Nicolas Robinson-Garcia, **Dakota S. Murray** (October 5, 2017), "The global scientific brain: Policy implications of barriers to scientific mobility", *CWTS Blog*.

Conference Works and Talks

REFEREED CONFERENCE WORKS

- 2021 Kozlowski, **D.**, **Murray**, D. S., Bell, A., Hulsey, W., Larivière, V., Monroe-White, T., & Sugimoto, C. R. (2021). Avoiding bias when inferring race using name-based approaches. ISSI 2021. Leuven, Belgium (virtual).
- 2020 **Murray D.**, Larivière V., Sugimoto C.R., & Cabanac G., (2020). Honoring our dead: text mining a century of academic obituaries in *The Lancet*. IC2S2 2020. Boston, USA (virtual).
- 2019 **Murray, D.**, Larivière, V., Sugimoto C. R., (2019). Context matters: how the usage and semantics of hedging terms differs between sections of scientific papers. ISSI2019. Rome, Italy.
- 2019 **Murray, D.**, Lamers, W., Boyack K., Larivière, V., Sugimoto C. R., van Eck, N., Waltman L., (2019). Measuring disagreement in science. ISSI2019. Rome, Italy.
- 2019 Miao, L., **Murray, D.**, Larivière, V., Sugimoto, C., Jung, W-S., & Yeol, Y-Y., (2019). The disciplinary structure of nation's scientific production. NetSci2019. Burlington, Vermont, USA.
- 2017 **Murray, D.**, Zhou, H., Minik, V., Larivière, V., & Sugimoto, C. R., (2017). Are Great Researchers Terrible Teachers? How research and teaching performance relate at U.S. universities. ISSI2017. Wuhan, China
- 2017 Robinson-Garcia, N., Sugimoto, C.R., **Murray, D.**, Yegros-Yegros, A., Larivière, V., & Costas, R. (2017). Unveiling the multiple faces of mobility: Towards a taxonomy of scientific mobility types based on bibliometric data. STI2017. Paris, France.
- 2017 Chinchilla-Rodriguez, Z., Miao, L., **Murray, D.**, Robinson-Garcia, N., Costas, R., & Sugimoto, C.R. (2017). A large-scale comparison of the position of countries in international collaboration and mobility according to their scientific capacities. STI2017. Paris, France.
- 2017 **Murray, D.**, Sugimoto, C.R., & Lariviere, V. (2017). A balanced portfolio? The relationship between gender and funding for US academic professors. STI2017. Paris, France.
- 2017 Miao, L., **Murray, D.**, Chinchilla-Rodriguez, Z., Lariviere, V., & Sugimoto, C.R. (2017). Glass boundaries: Differences in interdisciplinarity between men and women. STI2017. Paris, France.

PANELS

- 2018 Zaida Chinchilla-Rodríguez, presented by **Dakota Murray** , (2018) Moving on up: the relationship between international mobility and global leadership. 4S, Sydney, Australia
- 2018 Cassidy Sugimoto, **Dakota Murray**, & Vincent Larivière, Open panel, *The Global Scientific Brain: International Mobility of the Scientific Workforce*, 4S Sydney, Sydney, Australia

INVITED TALKS

- 2020 **Murray, D.**, Lamers, W., Eck, N., Boyack, K., Larivière, V., Waltman, L., Sugimoto, C.R. (2020). Measuring disagreement in science. CWTS Colloquium. Leiden University, Netherlands.
- 2020 **Murray, D.** Embeddings capture the multiscale structure of scientific mobility (2020). Data Science Student Consortium. University of Michigan, Ann Arbor, USA.

POSTERS

- 2020 **Murray, D.**, Yoon, J., Kojaku, S., Costas, R., Jung, W., Milojević, S., & Ahn, Y. (2020). Unsupervised embedding of trajectories captures the latent structure of mobility. NetSci 2020. Rome, Italy.
- 2020 **Murray, D.**, Yoon, J., Kojaku, S., Costas, R., Jung, W., Milojević, S., & Ahn, Y. (2020). Unsupervised embedding of trajectories captures the latent structure of mobility. IC2S2 2020. Boston, USA.
- 2020 **Murray, D.**, Cabanac, G., Larivière, V., & Sugimoto, C.R. (2020). Honoring our dead: text mining a century of academic obituaries in *The Lancet*. IC2S2 2020. Boston, USA.
- 2020 Miao, L., **Murray, D.**, Larivière, V., Sugimoto, C.R., Jung, W & Ahn, Y. (2020). The scientific development of nations. IC2S2 2020. Boston, USA.

WORKSHOPS AND SCHOOLS

- 2021 Santa Fe Institute's *Graduate Workshop in Computational Social Science*, Santa Fe, New Mexico (virtual)
- 2020 Understanding and Exploring Network Epidemiology in the Time of Coronavirus, (virtual)
- 2019 *Complex Networks Winter Workshop*, Quebec City, Canada
- 2019 Invited Scholar, *MIDAS Data Science Symposium*, Ann Arbor, Michigan
- 2019 Santa Fe Institute's *Summer School in Complex Systems*, Santa Fe, New Mexico
- 2017 Lorentz Center Workshop *Rethinking Impact Factors: New Pathways in Journal Metrics*, Leiden, Netherlands.

Grants, honors, and awards

GRANTS AND OTHER FUNDING

- 2019 GISA Departmental Travel Award, School of Informatics, Computing, and Engineering - \$950
- 2017-2018 Integrated Doctoral Education with Application to Scholarly Communication (IDEASc) fellowship, Institute of Museum and Library Services
- 2017 Cross-track collaboration travel grant, School of Informatics, Computing, and Engineering - \$500

ACADEMIC AWARDS

- 2016 Graduated B.S. *Summa Cum Laude* and with departmental honors
- 2012-2016 Chancellor's List (GPA > 3.85), Appalachian State University
- 2012, 2013 NSF Academy of Science Scholarship at Appalachian State University

Teaching

INDIANA UNIVERSITY BLOOMINGTON

- Fall 2018- Spring 2019 Associate Instructor, Informatics Capstone I & II: Design and Development of an Information System

APPALACHIAN STATE UNIVERSITY

2015-2016 Computer Science Tutor, Appalachian State University
2013 Lab Assistant, Computer Science I and II

GUEST LECTURES

Spring 2021 Introduction to bibliometric indicators in your career. *Article writing for graduate students*. Indiana University Bloomington.
Fall 2019 Introduction to bibliometric indicators in your career. *Article writing for graduate students*. Indiana University Bloomington.
Fall 2019 Excel workshop for beginners. *Understanding Social Data*. Indiana University Bloomington.
Fall 2019 Introduction to Machine Learning. *Sociology: Understanding Social Data*. Indiana University Bloomington.
Fall 2019 A gentle overview of academic conferences. *Informatics: Intro to informatics*. Indiana University Bloomington.
Spring 2018 Invited presenter and discussion leader. *Informatics: intro to informatics*. Indiana University Bloomington.

Service

REVIEWING

(Year of first review): *Science Advances* (2021) *Quantitative Science Studies* (2020); *Frontiers in Research Analytics and Metrics* (2020); *Advances in complex systems* (2020); *International Journal of Information Management* (2019); *Scientometrics* (2019); *PLoS Computational Biology* (2019); *PLoS One* (2017); *Journal of Informetrics* (2017); *Journal of the Association for Information Science & Technology* (2017)

INDIANA UNIVERSITY BLOOMINGTON

2017-2018 Representative, Graduate and Professional Student Government, Indiana University Bloomington
2017 Publicity Chair, Graduate Informatics Student Association, Indiana University Bloomington
2016 Social Chair, Graduate Informatics Student Association, Indiana University Bloomington

PROFESSIONAL ORGANIZATIONS

2018 - present Member, Society for the Social Studies of Science
2017 - present Member, International Society for Scientometrics and Informetrics