NORTHWESTERN UNIVERSITY

The Effect of Gender Diversity in Creative Teams

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Chemical and Biological Engineering

By

João Amado Gomez Moreira

EVANSTON, ILLINOIS

September 2017

ABSTRACT

The Effect of Gender Diversity in Creative Teams

João Amado Gomez Moreira

Individuals commonly engage in collaborative behavior to more easily produce works of high societal impact. The effect of many individual characteristics such as age or gender on the effectiveness of a team is still unclear. Gender is especially pertinent because many professional settings are still far from gender parity, despite ongoing controversy about innate differences between males and females.

In this dissertation, I use a rigorous mathematical approach validated with large datasets to study the effect of gender diversity in scientific collaborations and movie productions, and the impact of scientific works.

First, I analyze the publication records of thousand of researchers in science, technology, engineering, and math disciplines and show that previous contradicting findings of gender differences in collaboration patterns are a by-product of females' historic disadvantages in academia. I also present evidence of gender segregation in some sub-disciplines of molecular biology.

While there have been claims that males may be better suited for research than females, the same cannot be said for the movie industry. Therefore, to ensure the generality of my

findings, I also study gender diversity in U.S. movie casts. I demonstrate that a period of concentration of power at the hands of a small group of male leaders had a severe negative influence on female representation in the U.S. movie industry. Moreover, I find gender diversity among movie producers, directors, and actors to be strongly interdependent which can exacerbate female under-representation in movie casts.

The success of creative teams is also determined by how their work is received by their peers. Having limited time and expertise, individuals use a variety of measures to identify which books to read, movies to watch, songs to listen, or sights to see. Yet, most metrics are subjective measures of quality that can have unknown biases. I develop a principled indicator that quantifies the long-term impact of scientific works. By virtue of its construction, my indicator is resistant to manipulation and rewards publication quality over quantity.

Acknowledgements

First, I would like to thank my advisor, Luís Amaral, for his guidance, ideas, and support over this journey. He took a chance on an unknown student after one awkward phone call and for that I am forever grateful.

I would also thank my committee members, Neda Bagheri, Linda Broadbelt, and Brian Uzzi, for their valuable feedback and advice.

The Amaral Lab is an authentic idea factory encouraging open and serious discussions on science, statistics, silly movies, and everything in between. Many thanks to each and every "Amaralian" that have shared my stay in the lab.

To produce great science you need great teams. Thanks to my collaborators, Jordi Duch, Marta Sales-Pardo, Filippo Radicchi, Haroldo Ribeiro, and Teresa Woodruff for their many insights and suggestions. I would especially like to thank Xiaohan Zeng for the long discussions and his constant optimism, and Murielle Dunand who thought studying gender biases in movies would make for a *short* summer project.

I could not have come to Northwestern without the financial support from the FCT-Portugal through PhD Studentship SFRH/BD/76115/2011.

Finally, I must thank my family for their unconditional love and support. Thanks to my parents, Margarida Pereira and Eliseo Moreira, for never second-guessing my decision to go to school in another continent. Thanks to my girlfriend, Emily Hanson, for all the cookies and pretzels, and for always being there for me. This is for all of you.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Humans are a cooperative species. Governments must engage in trade deals to provide their respective citizens with more choices of goods and exotic delicacies. International and private institutions regularly fund scientific collaborations involving researchers from different universities and countries [1]. Companies strategically share resources and partners to create surprising new innovations [2, 3]. Artists who work in close proximity act share ideas and act as each other's critic and fan, thereby improving the overall quality of their works [4]. As illustrated by these examples, we band together to find solutions to problems that no single individual could solve alone. Indeed, a "collective intelligence" can emerge for a group of individuals that is not simply the sum of individual intelligence [5].

Despite the advantages of working as a team [4, 6–9], personal and hierarchical differences between individuals and institutions can create conflicts, insecurity, and miscommunication that detract from the team's effectiveness [10–13]. The exact effect of individual characteristics such as age, gender, or location on team impact is an area of active research.

Gender is a particularly relevant characteristic. Even though the general population is nearly gender-balanced, this is not observed in most sectors of society. While some argue that some professions are more suited to a single gender, the fact is that gender diversity is increasingly regarded as a desired condition by many institutions and corporations. Indeed, a prolonged gender imbalance in a given occupation can turn to unconscious bias that, over time, will give rise to the unhealthy stereotype that only males (or females) are suited for that job [14].

In science, technology, engineering, and mathematical (STEM) disciplines, researchers have shown that females make for better collaborators than males [15–17] and that mixed-gender teams produce higher impact works than single-gender groups [18], others have reported that males publish more [19] and are more prolific collaborators than females [20, 21]. Given the current gender gap in STEM disciplines it is vital that we understand how researchers' gender facilitates or hinders the effectiveness of collaborations.

However, studying gender discrimination in science teams is complicated by claims that there are innate mathematical and logical ability differences between genders [22, 23], and that females choose to leave academia to raise children [24] or to pursue a scientific career in other industries [25]. One profession where no such arguments can be made is acting. Indeed, the movie industry gives the same accolades to female and male actors. Moreover, the fact that movie productions are usually just a few months long and that actors can go several years without appearing in a motion picture make acting more amenable to childcare than most other careers.

Yet, there is evidence for gender discrimination in the movie industry. While females are present in nearly all movies, action movies are typically associated with males, whereas romance movies are more closely identified with females [26, 27]. Furthermore, Hollywood has an insidious gender wage gap, as recently brought to light by some of the industry's most famous actors and actresses [28–30]. The origin of this gender discrimination and the effect of gender diversity in movie-making teams are still unsettled questions.

To fully determine the success of creative teams it is not enough to study their gender diversity. We must also analyze how the work produced by teams is perceived by their peers. Scientific collaborations create publications whose impact can be quantified using bibliometric indicators. Unfortunately, despite the rather large number of *ad-hoc* bibliometric indicators of scientific impact proposed in recent years [31–40], there have been surprisingly few attempts to develop a rigorous framework that reliably quantifies scientific impact [41–45]. Such a framework can be used to promote science of excellent quality, with the capability to promote innovation, economic growth, and social well-being.

In this dissertation, I present a quantitative, large-scale study of the effect of gender diversity in creative teams, coupled with a rigorous framework to quantify the impact of scientific works.

1.1. Gender disparities in scientific collaborations

Collaborations bring many benefits to all scientists involved. Studies show that collaborations can decrease experimental costs [3], increase researcher productivity [1, 6, 46] and creativity [2, 4]. Moreover, teams have a greater chance of producing publications with higher impact than individuals [8], especially if they constitute novel collaborations [7, 9].

Given that collaborations can deeply impact researchers' careers, it is vital to understand the individual factors that enable a collaboration to be successful such as researcher nationality [47], institute, [12], discipline [13], or gender [48]. The effect of this last factor is of particular interest. On one hand, some studies revealed that, compared to males, females have fewer single-author publications than males [19], prefer to work in less hierarchical structures [15], show less self-interest [17], and are more cooperative [16], suggesting females make for better collaborators. On the other hand, other researchers showed that males can be more productive than females [20, 49] and have more international collaborations [21].

However some of these apparently contradictory results rely on small samples or self-reported surveys and thus have small statistical power.

Furthermore, female researchers are at a disadvantage in nearly all science, technology, engineering, and math (STEM) disciplines. Females comprise only a small percentage of faculty members [50] in STEM and there is a growing gender gap with advancing levels of science specialization [51], the so-called “leaky pipeline” phenomenon. Several researchers also report that female faculty suffer systemic and selective pressures creating a “glass ceiling” that prevents career advancement [52–56] and that females are more risk-averse than males [57].

A proper study of gender diversity in scientific collaborations should take structural factors such as academic positions and publication volume into consideration. Indeed, after controlling for age, discipline, and career stage, Bozeman et al. find that females overall collaborate more than males after [58, 59]. Moreover, McDowell et al. find evidence for gender homophily in collaborations among economists [60], i.e., researchers prefer to collaborate with others of the same gender. Thus, the presence of gender homophily suggests that females have fewer opportunities for collaboration [61], which could help explain some of the apparently contradictory results on gender differences in collaborations. A systematic, large-scale study clarifying the role that gender diversity plays in scientific collaborations would go a long way towards understanding productivity differences between male and female researchers in STEM disciplines.

1.2. Gender discrimination in movie productions

Movies have the power to make us afraid, laugh, cry, think, and even angry. Some actors can obtain a high level of notoriety from their movies which enable them to get cult-like followings [62], dictate fashion trends [63], and even exert political influence [64]. On the whole, the movie industry has an enormous impact on the world economy. In 2015, 708 movies were released worldwide, which generated US$38 billion in revenue [65] and involved more than 600,000 direct jobs [66].

A movie can be viewed as a collaborative act between several actors, producers, directors, screenwriters, and other crew members. Therefore it is reasonable to assume that individual characteristics have an effect on the impact of movie-making teams, perhaps even more so than individual characteristics on scientific collaborations, since even so-called "one-actor" movies often require tens of supporting crew as well as a director and one or several producers.

In principle, gender should not play a role in the effectiveness of movie production teams: outstanding female and male actors are both similarly laudable, and the fact that, on average, actors participate in a single movie production per year precludes the need for female actors to go on maternity leave. Yet, examples of female discrimination were abundant throughout much of the 20^{th} century [26, 67]. Females actors suffer from age [68, 69] and salary [70] discrimination, and get less acting opportunities than their male counterparts [71, 72].

Several researchers have suggested that the emergence of the Hollywood "studio system" may have been at least partly responsible for the observed gender discrimination in the movie industry [69, 70, 73, 74]. In 1920, the five biggest studios in Hollywood (MGM, Paramount, Warner Bros., RKO, and Fox) banded together into a cartel that controlled every aspect of a motion picture, from casting of actors and hiring of directors and writers, all the way to distribution and exhibition of the final movie [75]. The few leaders of the production companies composing the studio system — white males such as Louis B. Mayer, David Sarnoff, David O. Selznick, or Jack Warner — essentially gained absolute control over the Hollywood movie industry.

The studio system started to crumble when, in 1944, actress Olivia de Havilland successfully sued Warner Bros. to end long-term contracts in Hollywood [76]. This decision gave actors greater creative freedom to chose their projects. The studio system was finally disbanded in 1948, after the U.S. Supreme court it to be in violation of anti-trust laws [77].

Adverse effects of the studio system's policies continued to be felt for years after its dissolution. Female screenwriters were present at the start of the movie industry, and even though they were the minority gender, the average female screenwriter had the same visibility as the average male screenwriter [73]. However, with the establishment of the studio system, female screenwriters were quickly pushed to the background. Only recently have female TV and movie screenwriters started to gain recognition again [78, 79].

Surprisingly, most studies performed so far on gender discrimination against actors, producers, directors, or screenwriters are either mostly qualitative, or consider only recently-released or highest-grossing movies. A comprehensive, large-scale analysis of historical patterns of female representation in the movie industry is still lacking. Such an analysis can yield valuable insights regarding the effect of gender diversity in movie productions.

1.3. Scientific impact of published research

The exponential growth of scientific literature in the past half century has all but strained researchers ability to keep up with recent developments. To choose what to browse, read or cite is now a very challenging task for researchers. Simultaneously, the scientific workforce also experienced a tremendous growth. In order to continue generating ever more specialized, high quality knowledge, universities, funding agencies, and reviewers need to be able to evaluate the creativity and productivity of researchers. Neither researchers nor evaluating entities have in-depth experts on all fields, therefore they need to rely on proxies or indicators of publication quality, researcher impact, etc.

Bibliometric indicators are measures that consider one or more of counts of scientific publications and citations received by them in the scientific literature, co-authorship and concentration within specific journals, journal prestige just to name a few [80–82]. The number of citations, in particular, represent a measure of the impact or influence of not only specific publications but scientific journals [83], individual researchers [31, 84], research groups [85], institutions [21, 86, 87] or even whole cities and nations [88–90].

Various bibliometric indicators have been proposed such as the notorious Journal Impact Factor [91] and the h-index [31] which measure the impact of scientific journals and individual researchers, respectfully. Yet, despite their growing numbers [32–40], for the most part, existing bibliometric indicators constitute simple heuristics of citation counts and thus can be biased by career stage or publication volume, or be susceptible to manipulation [80–82, 92–99]. To address these issues, many researchers sought to develop bibliometric indicators that are unbiased by collaboration contribution [100, 101], researchers' career stage [102], journal citation skewness [42, 103, 104], or field size [41, 105]. The increasing demand for the evaluation and accountability of science both from within the scientific community and the public [106–111] means we can no longer afford to rely on flawed indicators of performance.

Citation counts span many orders of magnitude, thus it is ill-advised to work with the raw number of citations directly when creating an indicator [112]. Furthermore, extensive research on the aging of scientific literature shows that publications' citation rates change over time and eventually reach a steady-state [113–118]. As a result of this process of accumulating citations, any set of publications can be characterized by a cumulative distribution of citations. This distribution represents the probability of a publication acquiring a given number of citation after a certain elapsed time period.

A lognormal distribution was one of the first proposed functional forms for the citation distribution [119]:

$$P(n) = \frac{1}{n\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln n - \mu)^2}{2\sigma^2}\right) \quad ,$$

where μ and σ represent, respectively, the mean and standard deviation of $\ln n$. Several researchers have since provided empirical evidence for the use of a lognormal model to study citation distributions [41, 42, 45, 120–122].

More recently, inspired by Burrell's idea of the existence of a latent variable that determines the number of citations receive by a publication [115, 123], Stringer et al. used a modified lognormal model to demonstrate that the distribution of the number n of citations to publications published in a given journal in a given year converges to a stationary discrete lognormal functional form after, on average, ten years [42, 124]. With their model, Stringer et al. can successfully quantify the long-term impact of publications published in a scientific journal. This suggests that the framework of the discrete lognormal may be used to develop an unbiased bibliometric indicator of scientific impact at several levels.

1.4. Objectives

The primary goal of my research is the quantification of the effect of gender diversity in creative teams. I first present a quantitative analysis of the origins of gender disparities in two distinct domains that are each of paramount importance to society as whole: scientific collaborations, the main drivers of knowledge creation worldwide, and movie-making teams, the creators one of the most popular forms of entertainment. I then focus on the quantification of the impact of the work produced by some of these teams. Namely, I present a framework to quantify the long-term impact of scientific publications.

In Chapter 2, I study gender diversity in scientific collaborations. Historically, female researchers have been at a disadvantage in STEM disciplines. Females have lower publication rates and shorter careers than males. These observed gender disparities make it difficult to interpret differences in collaborations patterns between male and female researchers. I perform a quantitative analysis of researcher collaborations that properly controls for these historic disadvantages suffered by females. I also analyze systemic differences both between and within several STEM disciplines.

Some researchers have posited that observed gender differences in science may be due to innate ability differences between genders, or females choosing to leave academia. For these reasons, in Chapter 3, I turn my studies to the acting career, as it is a profession with no innate differences between males and females but one where gender discrimination nevertheless still exists. I propose a possible cause for the low female representation among actors in the in the U.S. movie industry. I then find how the gender diversity of producers and directors influences the gender composition of actors in a movie production. I also investigate the role of genre and movie budget on female representation in the industry.

To determine the effect of individual characteristics on team dynamics, we need to quantify the impact of the output from those teams. Therefore , in Chapter 4, I design and rigorously validate a principled framework to measure the long-term impact of scientific publications grounded on the functional form of the discrete lognormal distribution. I use this framework to construct a bibliometric indicator to measure the scientific impact of the publications authored by a researcher and those associated with a given research institution.

CHAPTER 2

Gender differences in collaboration patterns of STEM researchers

This work was published with Xiaohan Zeng, Jordi Duch, Marta Sales-Pardo, Filippo Radicchi, Haroldo Ribeiro, Teresa Woodruff, and Luís Amaral as “Differences in Collaboration Patterns Across Discipline, Career Stage, and Gender” in the Meta-Research section of *PLoS Biology* [125].

2.1. Abstract

Collaboration plays an increasingly important role in promoting research productivity and impact. What remains unclear is whether female and male researchers in science, technology, engineering, and mathematical (STEM) disciplines differ in their collaboration propensity. Here, we report on an empirical analysis of the complete publication records of 3,980 faculty members in six STEM disciplines at select U.S. research universities. We find that female faculty have significantly fewer distinct co-authors over their careers than males, but that this difference can be fully accounted for by females’ lower publication rate and shorter career lengths. Next, we find that female scientists have a lower probability of repeating previous co-authors than males, an intriguing result because prior research shows that teams involving new collaborations produce work with higher impact. Finally, we find evidence for gender segregation in some sub-disciplines in molecular biology, in particular in genomics where we find female faculty to be clearly under-represented.

2.2. Author summary

Collaboration is increasingly important role in promoting research productivity and impact. What remains unclear is whether female and male researchers differ in their collaboration practices. In our study, we report on an empirical analysis of the complete publication records of 3,980 faculty members in six science, technology, engineering, and mathematical disciplines at select U.S. research universities. First we find that female faculty have significantly fewer distinct co-authors over their careers than males, but that this difference can be fully accounted for by females' lower publication rate and shorter career lengths. Next, we find that female scientists have a lower probability of repeating previous co-authors than males, an intriguing result because prior research shows that teams involving new collaborations produce work with higher impact. Finally, we find evidence for gender segregation in some sub-disciplines in molecular biology, in particular in genomics where we find female faculty to be clearly under-represented.

2.3. Introduction

It is widely acknowledged that collaboration is critical to the scientific enterprise [12, 13, 47, 126–129]. Although the motivations determining collaboration propensity is still the subject of much research, scientists benefit from collaboration both in terms of productivity and impact [2–4, 6, 46]. For example, Bordons et al. [1] showed that for biomedical research there is a positive correlation between productivity and collaboration at the author level, and Wuchty et al. [8] showed that teams produce publications with higher impact than individuals. Moreover, teams that include novel collaborations have a greater likelihood of producing higher impact work [7, 9].

Since research suggests that collaboration patterns affect a researcher's career performance, it is important to understand whether there are gender differences in collaboration patterns [14, 48]. Indeed, Kyvik and Teigen [19] reported that the productivity of both genders is positively correlated with the level of collaboration, and that females have fewer single-author works than males.

Prior research suggests that women tend to be more collaborative and less competitive than men in decision making, making them potentially better collaborators [15–17], but recent studies have reported contradicting results about which gender is more collaborative [20, 21, 58, 59, 130].

Because most STEM fields have much larger numbers of males than of females, homophily would suggest that female academics have fewer opportunities for collaboration [61]. McDowell et al. [60] find evidence of gender homophily in collaborator choice among a sample of economists and that females preferentially apply to larger departments to increase their chances of finding collaborators. Bozeman et al. not only find evidence of the same gender homophily [58] but also that, after controlling for gender disparities, females overall collaborate more than males [59].

To investigate the role of gender in collaborative behavior, we perform a large-scale empirical analysis on the publication records of faculty members for six STEM disciplines. Our analyses yield three main findings. First, female faculty have significantly fewer distinct co-authors than male faculty, but that this difference can be fully accounted for by the shorter career lengths of current female faculty and their lower publication rate. Second, female faculty tend to have a lower probability of repeating a collaboration, a strategy that has been shown to produce work of greater impact. Third, for the discipline of molecular biology, we find evidence for gender segregation in some sub-disciplines. In particular, we find that female faculty are clearly under-represented in genomics.

2.4. Data

We obtain complete faculty rosters, as of Fall 2010, for departments of chemical engineering, chemistry, ecology, materials science, molecular biology and psychology from several top research universities in the United States (US) (Tables B.1–B.3). We consider all active faculty members as of 2010, including tenure-track and research faculty, but exclude emeritus professors. We identify the researchers' gender from their departmental website photograph. If they have no photograph we use their given name to identify the gender (faculty with ambiguous names were excluded). We then obtain bibliometric data for 3,980 faculty members from Thomson Reuters' Web of Science (WoS) based on the biographical information listed on their websites and *curricula vitae*. See [50] for details on data acquisition and validation, and Table 2.1 for aggregate statistics.

Table 2.1. **Characteristics of the faculty cohorts in our study**.

Discipline	**Depts.**	**Faculty**			**Publications**		
		Female	**Male**	**Ratio**	**Female**	**Male**	**Ratio**
Chemical Engineering	31	98	567	1:5.8	6,392	66,328	1:10.4
Chemistry	35	198	1,020	1:5.2	13,790	137,723	1:10.0
Ecology	15	106	328	1:3.1	3,976	22,425	1:5.6
Materials Science	26	98	473	1:4.8	9,538	75,373	1:7.9
Molecular Biology	11	168	474	1:2.8	9,882	51,234	1:5.2
Psychology	10	171	279	1:1.6	7,143	20,976	1:2.9
Total	129	839	3,141	1:3.7	50,721	374,059	1:7.4

2.5. Results

2.5.1. Gender differences in number of collaborators

Since scientific publications are the direct product of scientific research and collaboration, the number of distinct co-authors a researcher has accrued throughout her career is a good proxy of how strongly she seeks collaborations. Because collaboration patterns may be discipline-specific, we examine each discipline separately [131]. Moreover, because collaboration patterns may depend on career stage, we also account for career stage in our analyses.

We focus on the number of *distinct* co-authors; that is, we count only once co-authors that appear multiple times in the publications of an individual. We do this because co-authoring publications with new collaborators more likely indicates the introduction of new expertise into the team and the expansion of one's professional network.

We calculate the distribution of total number of distinct co-authors over the career of the scientists in our database. Our raw results show that for all six disciplines, females on average have a significantly lower number of distinct co-authors over their careers than males (Fig. 2.1). However, in order to properly interpret these results, we must account for the fact that until 1980 there were hardly any female faculty, which implies that female faculty typically have shorter career length and thus are likely to have fewer publications than their male colleagues [50]. Moreover, because of the gender gap in the number of publications [50, 132], it is necessary to control for publication rate when comparing the number of co-authors of females and males. Thus, we test the null hypothesis that there is no gender difference in the number of distinct co-authors when controlling for the number of publications (see Appendix A.1). The confidence intervals constructed under this hypothesis show that once we account for the number of publications, the observed difference in the distribution of the number of distinct co-authors of female and male faculty is not statistically significant (Fig. 2.1).

2.5.2. Repeated co-authors and propensity to collaborate

The data from Fig. 2.1 shows that female and male faculty accrue an average number of new distinct co-authors per publication that is indistinguishable from the average for males. However, this observation does not imply that females and males accrue new collaborators in the same manner, or that they have the same propensity to collaborate.

2.5.2.1. Accruing new collaborators. Consider a publication of researcher i and n_c co-authors. The number n_n of distinct co-authors that i accrues can be expressed as

$$n_n = n_c \, (1 - f_r) \ , \tag{2.1}$$

where f_r is the fraction of repeated co-authors. Eq. (2.1) makes explicit that both team size (that is, n_c) and propensity to repeat collaborations affect the number of new distinct co-authors to be gained from each publication. We first investigate the effect of the repetition of co-authors on the gender disparity in the number of distinct co-authors. Researchers who frequently co-author with the same team will not accumulate co-authors as rapidly as those

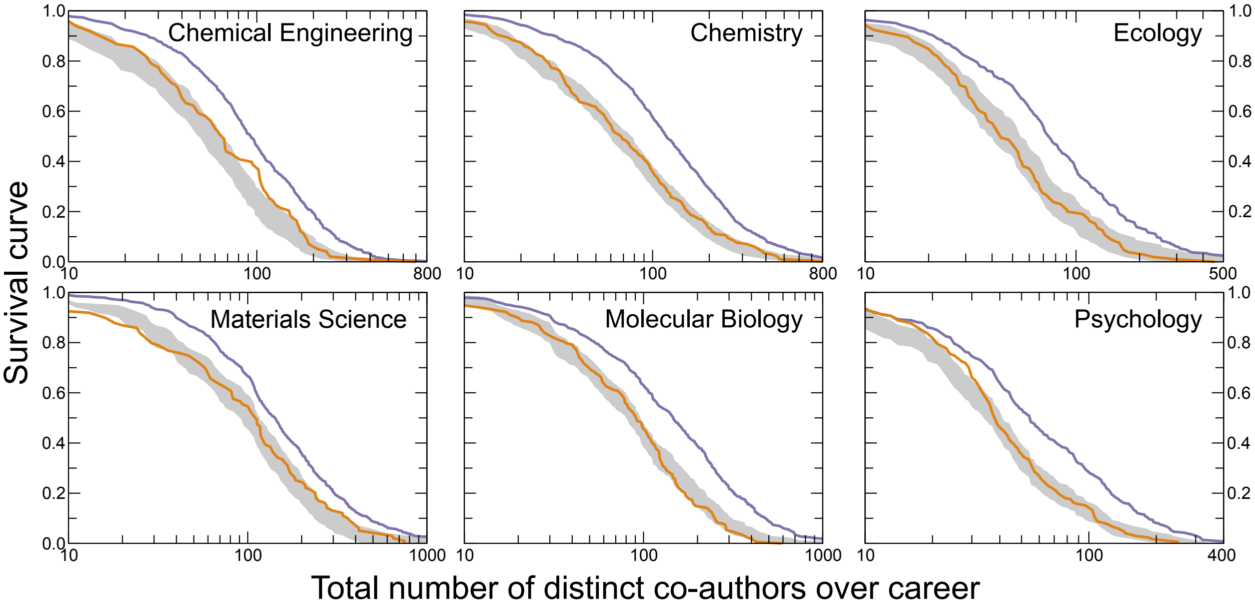

Figure 2.1. **Lower number of publications by female scientists results in lower total number of distinct co-authors**. Survival curve of the total number of co-authors over careers of females (orange) and males (purple). We test the null hypothesis that there is no gender difference in the total number of distinct co-authors for females and males with similar number of publications. The grey shaded region indicates the 95% confidence interval obtained under the null hypothesis. To construct the confidence interval, we generate samples of N_F males, where N_F is the number of females in our dataset. For a female with n_F publications, we select a male whose number of publications falls in the range of $[0.8\ n_F,\ 1.2\ n_F]$ (see Appendix A.1). Note that the curve for females falls inside the confidence interval, indicating that after correcting for number of publications, females and males have comparable numbers of distinct co-authors over their careers. The curve for males falls outside the confidence interval because some male researchers in the dataset have very large numbers of publications (see Fig. 7 of [50]).

who seek out new collaboration opportunities. To quantify the tendency to repeat previous co-authors, we calculate f_r for each author, and obtain the distribution of f_r for both genders for each discipline. We then test whether the two samples could have been drawn from the same distribution.

We show in Fig. 2.2 the probability distribution functions of f_r for females and males. The data show that females have an f_r approximately 20% smaller than males, indicating that female faculty repeat co-authors less frequently than male faculty. More frequent repetition of co-authors may also be an indicator that a few co-authors are responsible for most collaborations. We use the Gini coefficient [133] and the disparity index to quantify the degree of inequality in the distribution of collaboration frequencies, and find that females do tend to distribute their co-authoring opportunities more equally among their collaborators than males (Figs. C.1–C.3).[1]

2.5.2.2. Average team size. We next study the average number of co-authors per publication, n_c. Researchers who collaborate with larger teams have higher numbers of co-authors per publication. However, the number of co-authors changes as a function of the publication year and author's career stage (Fig. C.5). Since female faculty entered academia more recently and on average have shorter career lengths than male faculty [50], we need to account for these two factors when comparing team sizes. In Fig. 2.3 we show that, except for molecular biology, the two genders do not differ significantly in the number of co-authors per publication when their publication year and career stage are taken into consideration.

2.5.3. The case of molecular biology

Our findings for molecular biology are intriguing. While there are no significant differences during the first ten years, beyond ten years, publications authored by females in molecular biology have significantly lower number of co-authors per publication than those authored by males. To further detail this observation, we bin the publications authored by females according to the number of co-authors, after accounting for increases in team size over the period considered. Assuming that females do not prefer any particular team size, the

[1] Although the gender difference in the tendency to repeat co-authors is significant, our ability to establish its statistical significance on the total number of distinct co-authors is hampered by the heterogeneity in team size and number of publications (Fig. C.4).

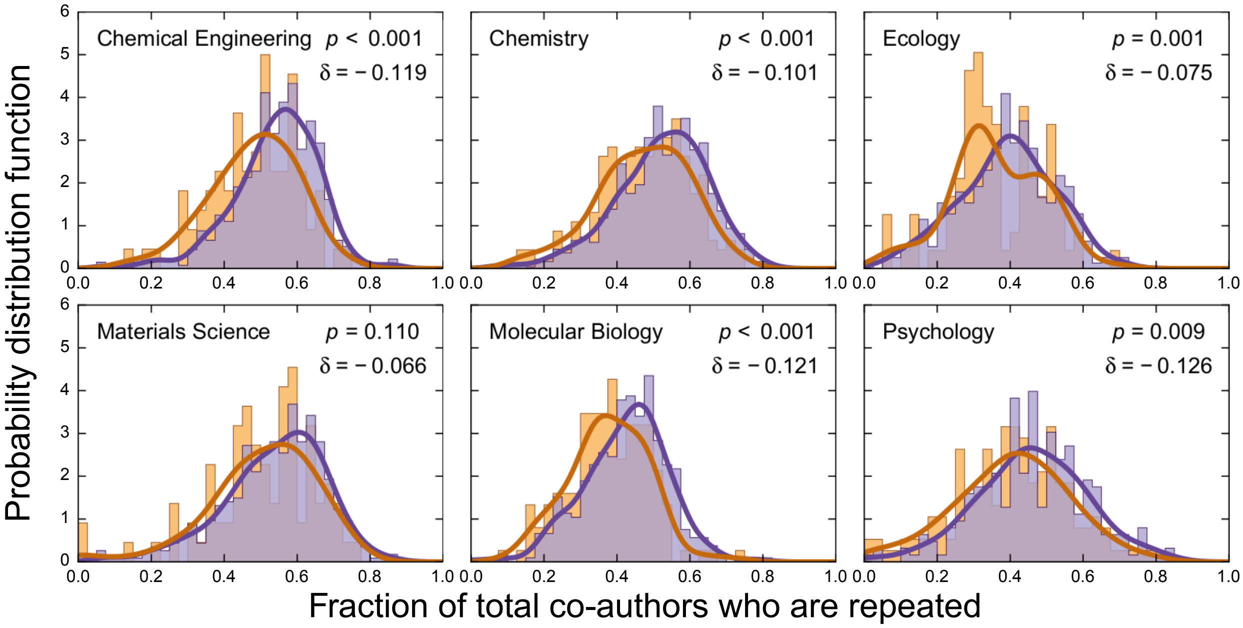

Figure 2.2. **Gender differences in the propensity to co-author with prior collaborators**. Probability distribution of the fraction of total coauthors who are repeated for all females (orange) and males (purple) in the dataset with at least 10 publications. We exclude single-author publications. Orange and purple lines are kernel density estimation of the distributions for females and males with bandwidth given by Scott's Rule [134]. We obtain p-values for the validity of the null hypothesis that the samples were drawn from the same distribution using the Kolmogorov-Smirnov test. For all disciplines, we find $\delta = 2(\bar{f}_{r,F} - \bar{f}_{r,M})/(\bar{f}_{r,F} + \bar{f}_{r,M}) < 0$, where $\bar{f}_{r,F}$ and $\bar{f}_{r,M}$ are the average f_r of the female and male faculty, respectively. Females have f_r smaller than those of males, suggesting that, except for materials science, female faculty have a lower propensity than male faculty to repeat collaborations.

fraction of publications by females in each bin should remain approximately constant. For each bin, we then calculate how much the observed number of publications by females deviate from the number expected from the null hypothesis using the hypergeometric distribution (see Appendix A.1). Figure C.6 demonstrates that female faculty in molecular biology departments have a distinct behavior from females in other disciplines: They consistently author significantly more publications than expected in teams smaller than average, and significantly fewer publications than expected in teams larger than average. We make this

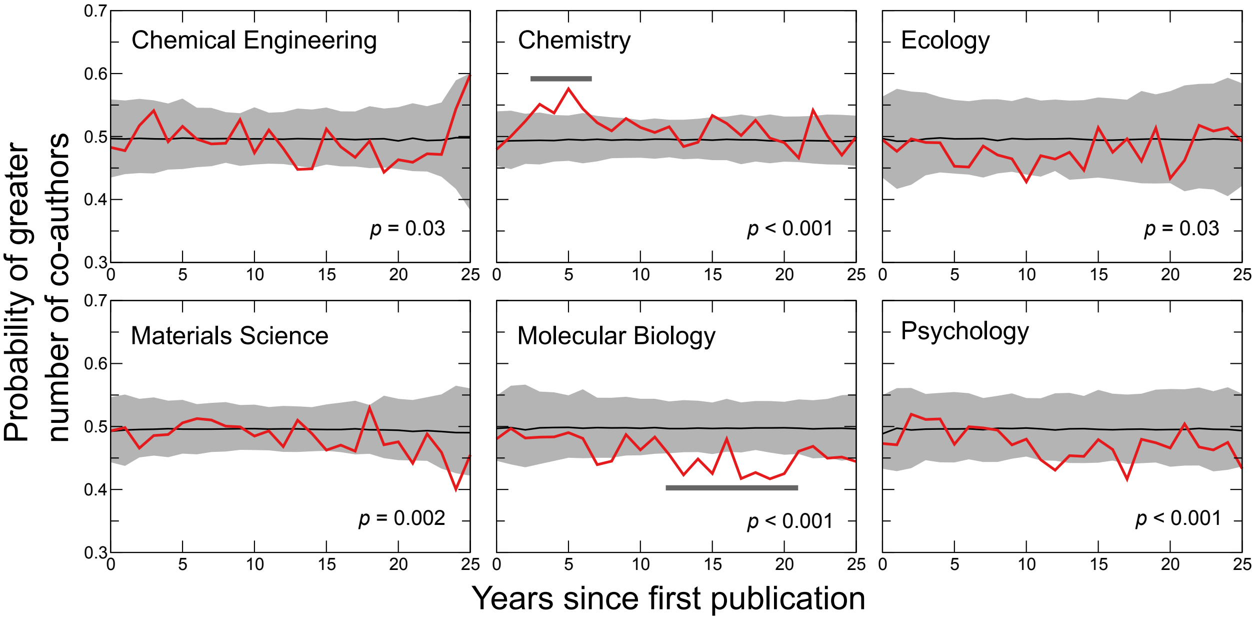

Figure 2.3. **Male and female faculty have similar number of co-authors per publication for five other disciplines, but not for molecular biology**. Probability of females having greater number of co-authors per publication in a given year of her career than a male peer at the same career stage (red lines). We use z-scores to account for the increasing size of research teams and the fluctuations over career stage (see Appendix A.1). We indicate the 99% confidence intervals by the grey areas, and the medians of the probabilities obtained from random ensembles by black lines. The p-values are obtained under under the null hypothesis that there is a 99% probability of any value being outside the confidence interval. Note that although the difference in the average size of teams appears to be statistically significant, it is not consistent along the career stage, except for chemistry for the first few years, and for molecular biology in later career stages (dark horizontal bars).

fact visually apparent by shading in grey regions where the observed value is significantly different from the null hypothesis.

2.5.3.1. Segregation among sub-disciplines. Although we restrict our analysis to researchers within the same discipline, academic disciplines such as molecular biology comprise several sub-disciplines. If females and males are segregated across sub-disciplines so that more males work in sub-disciplines with large teams, and more females in those with small teams, then this segregation could give rise to the gender gap in the average number of co-authors per publication.

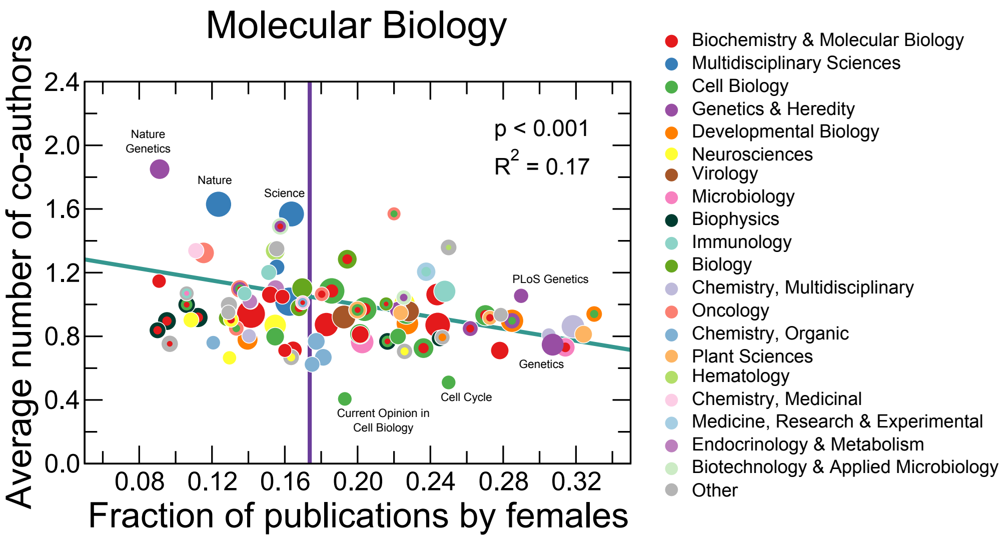

Figure 2.4. **Female faculty in molecular biology departments publish more in journals and sub-disciplines where typical team size is smaller**. We show correlation between the average number of co-authors corrected for the annual average and the fraction of publications authored by females, grouped by journal. We only consider publications authored after the tenth year mark in an author's career. We restricted the publication types to "article", "letter", and "note." The size of the circle is proportional to the logarithm of the number of publications in that journal or sub-discipline. We use journal category in the *ISI Journal Citation Report* as the sub-disciplines. Journals with multiple categories are plotted as concentric rings. The purple line indicates the total average fraction of publications by females for all the publications authored by faculty in molecular biology in our cohort, f_M (17.3%). The blue line is a weighted linear regression, in which we assign to each journal a weight equal to the number of publications. We only include data points within the range of $[0.5 f_M,\ 2 f_M]$.

We find that at journal level the average number of co-authors is strongly and significantly anti-correlated with the fraction of publications authored by females (Fig. 2.4). The strong and statistically significant anti-correlation indicates that females publish more in journals (and, presumably, sub-disciplines) where the typical team size is smaller, and less in those where the typical team size is larger (see Figs.C.7–C.11 for results for other disciplines).

The journal-level analysis strongly suggests the existence of gender segregation across sub-disciplines. However, many journals are multi-topic and even multidisciplinary, thus they may not accurately represent narrower research topics. To overcome this limitation of the journal-level analysis, we must determine the research topic of each publication at a finer scale. To this end, we use a highly accurate and reproducible topic classification algorithm to identify the topics of publications [135]. We identify a total of 69 topics using the titles and abstracts from the set of 61,116 publications by molecular biology faculty in our database. Table B.4 lists the identified topics and the most representative words and journals associated with them.

For the publications in each topic, we calculate the average team size and fraction of publications by females (Fig. 2.5). Using a 99% confidence region [136], we identify seven topics that are outliers; of those, two are in molecular biology (Table 2.2). All the outlier topics in chemistry and of the outlier topics in materials science actually have larger representations of publications by female faculty and larger team sizes. In contrast, the outlier topics in molecular biology have just larger team sizes. Looking at the representative journals for each of the outlier molecular biology topics, it becomes clear that topic 6 refers to genomics.

Genomics (topic B5) is particularly relevant when attempting to explain the smaller team sizes of female authored molecular biology papers. Genomics is unique because it has a very striking under-representation of females and markedly larger team sizes. Moreover, because it is a topic with a very large number of publications, it strongly affects the characteristics of the entire discipline. These results prompt the question of why females are under-represented in genomics. Table B.5 shows that 19 of the 20 most prolific researchers in our database working in genomics are male. A recent study suggests that the labs of prominent male researchers have lower than average fractions of female graduate students and postdocs [137]. Since the protégés of prominent scientists have such an important role in populating faculty positions in molecular biology, the under-representation of females in those labs propagates all the way to the level of tenured faculty.

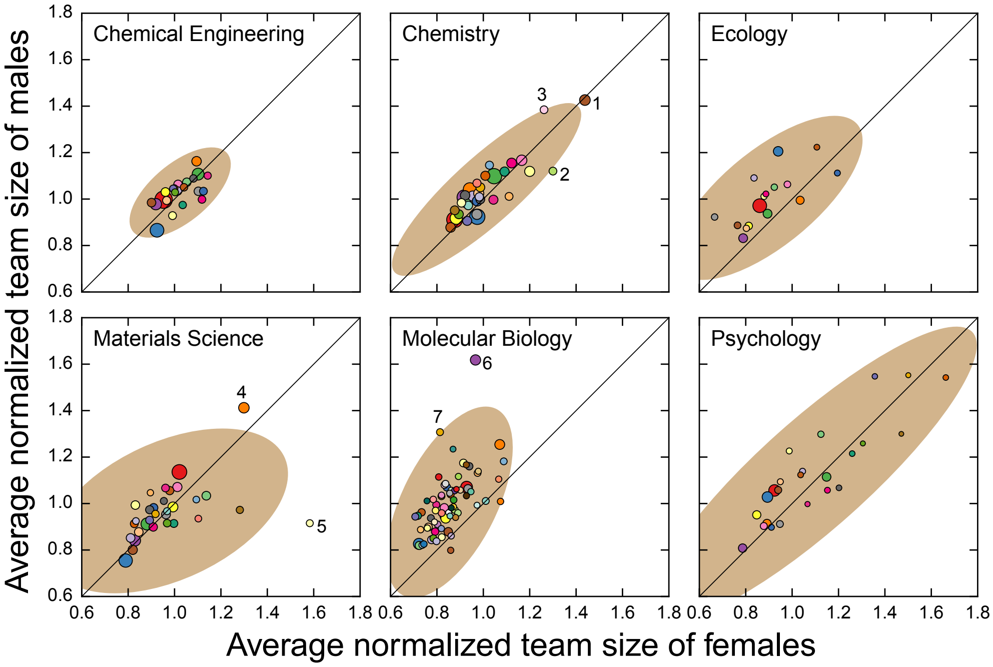

Figure 2.5. **Topic dependence of female representation in publications in the six disciplines**. We show the average number of co-authors corrected for the annual average for male faculty versus that for female faculty. Note for molecular biology most of the data points fall above the line $y = x$, indicating that for most topics females work in smaller teams than males. We label the seven topics which fall outside the 99% confidence region (brown ellipse) (see Table 2.2 for topic details).

In order to investigate the origins of the distinct characteristics of the outlier topics, we turn again to the lists of the scientists with the most publications in each topic (Tables B.5, B.6). We then repeat the analysis of Fig. 2.5 but excluding the publications of the 5 most prolific scientists for each outlier topic. Strikingly, we find that the characteristics of these topics revert to the mean for the entire discipline. That is, the gender of the most prolific authors determines the characteristics of the topic. We believe that this finding raises an important question: Why females have not been able to succeed in genomics in proportion to their numbers? No female in our dataset made it into the top 10 most prolific scientists in genomics, the first female appearing in 12^{th} place. If genomics was gender blind, and

Table 2.2. **Topics within considered disciplines that are outliers when considering the differences in average team size between male and female faculty in our database**. **Topic** represents the topic number identified by the topic classification algorithm and is field-specific [135]; **Outlier topic** represents the topic in Fig. 2.5.

Discipline	**Topic**	**Outlier topic**	**Representative journals**	**No. publs.**	**Norm. ratio by females**	**Mean norm. team size**
Chemistry	C4	1	Cancer Research, Bioconjugate Chemistry, Antimicrobial Agents and Chemotherapy	4,809	1.5	1.43
	C14	2	Nucleic Acids Research, Physical Review E, Genome Biology	1,354	1.3	1.37
	C18	3	Journal of Membrane Science, Radiochimica Acta, Journal of Natural Products	1,399	1.2	1.14
Materials Science	M0	4	Biomaterials, PNAS, Journal of Biological Chemistry	4,547	2.0	1.39
	M29	5	Organome tallics, Journal of Chemical Physics, Surface Science	1,742	1.0	0.99
Molecular Biology	B5	6	Nature Genetics, Genetics, Nucleic Acids Research	4,186	1.0	1.51
	B10	7	Molecular Biology and Evolution, Genetics, American Journal of Botany	899	1.1	1.22

considering that females comprise 26% of the biology researchers in our database, this would be an unlikely situation ($p \simeq 0.0095$).

2.6. Discussion

A number of recent studies support the hypothesis that there are gender differences in collaboration patterns [14, 48] and that collaboration has a significant impact on scientific productivity and impact [7, 8]. Evidence suggests that self-selection among female researchers due to greater career risks, and female scientists' decreased access to funding can, respectively, cause gender differences in publication rate and impact [50, 60].

Our present analysis conclusively shows that females do have fewer distinct co-authors over their careers, but that this gap can be accounted for by differences in number of publications. We also find evidence for the hypothesis that female scientists are more open to novel collaborations than their male counterparts, a behavior that was shown to correlate with producing work of greater impact [7].

It could be, however, that females have fewer distinct collaborators not purely because, as the females in our cohort they publish fewer publications, but because female scientists do not participate in research teams to the same extent as male scientists. We believe that this possibility is unlikely since there is strong evidence that females are generally more collaborative than males both in academic life [21, 59] and in other realms [15–17].

Concerning our finding that females appear to be more likely to engage new collaborators, it could be that females are simply more effective collaborators and are able to make the most of their lower representation in STEM disciplines. Wolley et al. showed that females typically have greater group intelligence than males [5] giving some credence to this hypothesis. An alternative explanation for the greater repetition of collaborations by males is unwarranted authorship in publications for the purpose of increasing one's publication counts. Anecdotal evidence suggests that, while the number of scientists pursuing such gaming of the system is small, they do tend to be male.

Lastly, our finding of female exclusion from genomics is of particular interest, especially because of what it may imply concerning the cultural milieu of this sub-discipline. The importance of culture on gender segregation is supported by recent studies showing the existence of gender stereotyping in physics and its negative consequences for females in that field [138, 139]. It is known that in some molecular biology sub-disciplines such as telomere research (topic B21) the participation of female scientists has been encouraged. Indeed, 6 of the 10 most prolific researchers in this topic are female (Table B.7). The top three researchers, Elizabeth Blackburn, Virginia Zakian, and Carol Greider conducted their doctoral research under the mentorship of Joseph Gall, who is known for having supported female scientists at a time when misogyny was widely accepted. The important role of prominent scientists in encouraging both males and females to pursue careers in research is also illustrated by William H Bragg's role in the recruitment of female scientists to crystallography. In contrast, the cultural milieu in institutions such as Genentech [140] likely had a chilling effect on female participation in genomics.

One caveat of our study is that it is limited by the fact that we are only able to track those scientists that persisted within academia. We believe it is important to also investigate to what extent our findings would still hold for scientists that were unable to remain in academic positions at top universities. In a perverse way, it could be that females' propensity to collaborate creates both better publications and a successful research program, and greater risk when the time comes for tenure decisions. Another caveat is that we are not able to identify which coauthors may be trainees (graduate students or post-docs), a situation that in many cases would be more representative of mentorship than of typical collaboration.

CHAPTER 3

Probable causes of gender discrimination in the U.S. movie industry

The work in this chapter is submitted for publication and was completed with contributions from Murielle Dunand, and Luís Amaral.

3.1. Abstract

Gender parity has been slowly but steadily increasing in many sectors of society. One sector where one would expect to see near gender parity is the movie industry, yet the numbers of females in most function of the U.S. movie industry remain surprisingly low. Here, we study the historical trends of female representation among actors, directors, and producers and attempt to gain insights into the causes of the lack of gender parity in the industry. We demonstrate that the advent of the studio system, a period where the "Big Five" Hollywood studios deliberately cooperated to control all aspects of the movie industry, had an extremely negative impact on female representation. Indeed, female representation among actors, directors, and producers dropped by more than half after the emergence of the studio system, to values so low that the gender imbalance is still observed presently. Moreover, we find that the gender diversity of a movie's producers influences both the gender of the director and the gender composition of the cast, and that female directors have a statistically significant preference for more gender-balanced casts. Additionally, we find that female directors are over-represented in two genres — Documentary and Romance — but under-represented in seven other genres. Lastly, we find that actress representation in higher

budget movies grew during the studio system, and that the increase in female representation in the 1960s was most evident in the lowest budget movies.

3.2. Introduction

Gender diversity is increasingly regarded as a desirable condition by educational, business, and governmental organizations. Recent research shows that more gender-balanced groups are better at complex decision-making [5] and females show less self-interest are are better at complex moral reasoning than males [17]. Indeed, the proportion of women faculty members in many STEM fields has been steadily increasing [50], as has the number of females in corporate suites and in political office [141, 142]. These trends are positive because the absence of women in leadership positions has a negative impact on women's aspirations and advancement and may perpetuate gender biases [71].

A factor muddying the discussion of the causes for lack of gender diversity is the argument that males may be better suited to some professions (i) because of greater physical strength, greater mathematical ability, or some other advantage; or (ii) because, unlike females, they do not have to interrupt their careers due to childbearing. However, there is one career for which neither of these arguments would 'hold much water' — acting. Indeed, unlike many other professions for which it is much easier to name prominent male exponents than female exponents, the same is not true for acting: Marlene Dietrich, Katharine Hepburn, and Meryl Streep are just as recognizable as Douglas Fairbanks, Humphrey Bogart, and Tom Hanks. Moreover, the fact that most actors participate in at most a single movie per year and can go several years without appearing in a motion picture, makes the career more flexible and amenable to actresses taking time away to care for young children.

Yet, there is evidence of significant gender discrimination against females in the U.S. movie industry [26, 67, 70]. Females not only are offered less roles than males in certain markets [71], they will also feature in fewer films if they have repeatedly co-starred with the same counterparts [72], or as they age [68, 69]. Indeed, age affects an actor's earnings

Table 3.1. **Coverage of gender information for the movies in our dataset**. See Appendix A.2 for details on gender assignment.

		Gender		
Role	**Coverage**	**Males**	**Females**	**Unknown**
Actors	98%	144,460	81,294	0
Directors	98%	6,281	543	71
Producers	94%	19,232	5,572	753

potential differently depending on gender. For females stars, movie earnings peak in the mid-thirties, whereas for males stars they do not peak until they reach fifty [70].

We believe that the study of the historical patterns of female representation among actors is likely to yield insights into the causes of gender discrimination without the confounding effect of *potentially* different innate abilities for the profession. Thus, we study here the temporal evolution of female representation in the cast of over 15 thousand U.S.-produced movies released between 1894 and 2011. We find that prior to the establishment of the Hollywood studio system (1920-1930) [75], female representation stood at nearly 30%, but that it had decreased by nearly a third by the late 1940s and that it would take another 15 years before it returned to pre-studio system levels. Below, we show that concentration of decision power among a small cadre of male executives predated the drop in female representation and that only the breakdown of the studio system let female representation rise again.

3.3. Background

The early movie industry was considerably more diverse in terms of gender and geography than it would become by the time the Great Depression arrived. Until the mid 1910s, France, Italy and the U.S. were all important movie production countries. Within the U.S., movies were being produced along the East Coast, from New York to Florida. However, within fifteen years, this situation would change dramatically. In the U.S., the attempt by Thomas Edison and the Motion Picture Patent Company (MPPC) to control movie production pushed many in the industry to relocate to California, and away from the legal reach of the MPPC (Fig. 3.1a). In Europe, the first World War greatly hindered the development of the industry. As a result, by the 1920s, Hollywood was the dominant player in the global movie industry both in terms of the number of movies being produced (Fig. 3.1a) and in terms of the profits captured [143].

Economic growth and co-location prompted industry consolidation and the emergence of the so-called "studio system". The "Big Five" studios (MGM, Paramount, Warner Bros., RKO, and Fox) formed a cartel that controlled every aspect of a motion picture, from the casting of actors, hiring of the director and the screenwriters, all the way to the distribution and exhibition of the final movie [75]. Through the studio system, a handful of individuals — men such as Louis B. Mayer, David Sarnoff, David O. Selznick, or Jack Warner — gained essentially absolute control over the industry.

3.4. Results

Using a dataset [146] comprising 15,425 U.S.-produced movies released between 1894 and 2011 (see Table 3.1 and Appendix A.2 for details), we find that the studio system had a similar impact on female representation among movie directors and producers (Figs. 3.1c, 3.1d). By the 1930s, female representation among producers and directors had dropped to less than half of the levels observed prior to 1920. At the level of producers, the drop was particularly severe for executive producers (Fig. C.13). This fact is particularly significant because producers' decisions are so impactful. They are responsible for overseeing a movie's

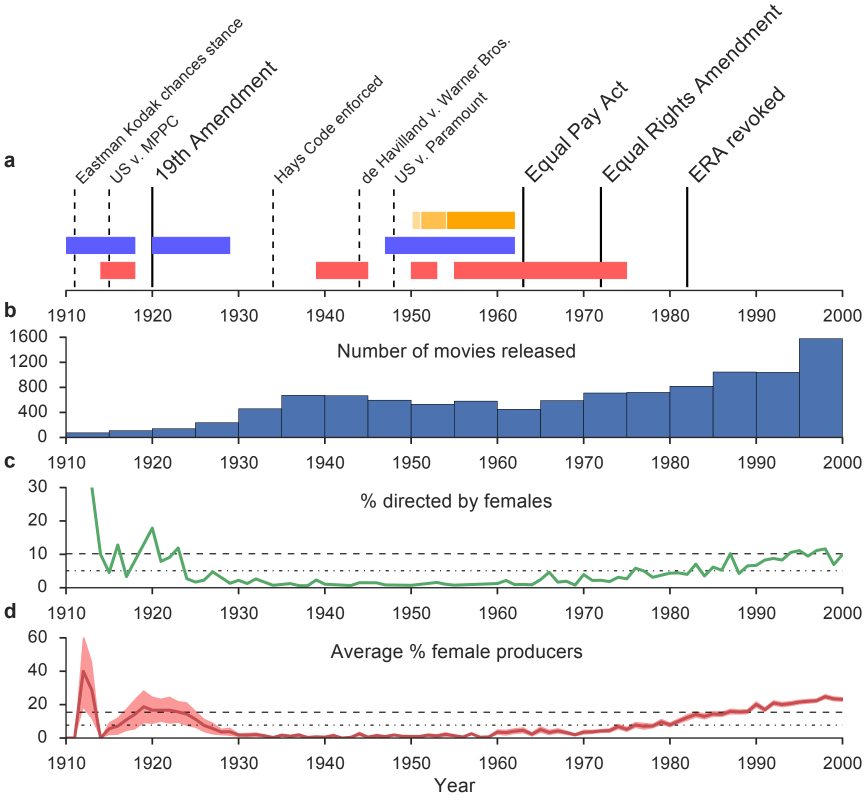

Figure 3.1. **Historical trends of gender imbalance in the U.S. movie industry**. (**a**) Timeline of 20th century events relevant to the U.S. movie industry. Orange shadings indicate the rates of TV adoption in U.S. households (from unsaturated to saturated: < 30%, < 60%, < 90%) [144]. Blue bars identify, chronologically, the duration of the MPPC control [143], consolidation of the studio system [75], and the blacklisting of industry participants [145]. Red bars represent major wars (chronologically, World War I, World War II, Korean war, and Vietnam war). (**b**) Number of U.S.-produced movies released annually and recorded in IMDb. (**c**) Percentage of movies directed by females as a function of release year. The data shows a U-shape. Remarkably, the percentage of movies directed by females in the early 1900s (dashed line, approximately 10%) was only reached again in 1994, having remained below half of that level for 59 years (dash-dotted line). (**d**) Percentage of female producers in movies (mean ± standard error). As for directors, the percentage of female producers for movies in the early 1900s (dashed line, approximately 15%) was only reached again in 1987, having remained below half of that level for 53 years (dash-dotted line).

finances, selecting and managing the cast and crew, and are involved in all movie-making facets, from conception to distribution [147, 148].

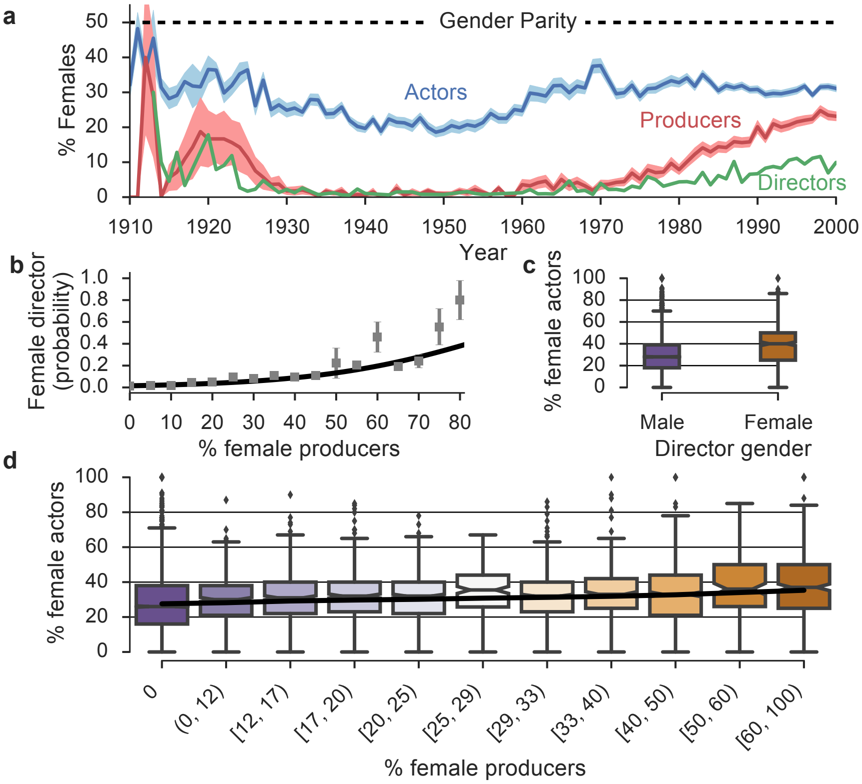

Figure 3.2. **Power and gender discrimination in the U.S. movie industry**. (**a**) Lack of gender parity is apparent for three of the most visible functions in the movie industry: producers, directors, and actors. Note the dramatic drop in female representation for these 3 functions starting in 1920. We hypothesize that the power structure within the movie industry contributes to gender discrimination. We test our hypothesis in the following panels using data from movies with a single director. (**b**) Logistic regression on the probability of a director being female as a function of the percentage of producers that are female. We find a significant correlation (pseudo-$R^2 = 0.11$, $\beta = 0.046 \pm 0.002$, $p < 0.001$), strongly suggesting that the gender of the producers contributes to explaining the gender of the director. (**c**) Impact of the gender of the director on the gender representation of actors. We find that, compared to male directors, female directors have a significant preference for a more gender-balanced cast (Mann-Whitney test, $U = 2.4 \times 10^6$, $p < 0.001$). (**d**) Linear regression on the percentage of actors that are female as a function of the percentage of producers that are female. We find a significant increase (slope $= 0.13 \pm 0.007$, $t = 19$, $p < 0.001$) in the percentage of female actors cast as the percentage of female producers grows (black line, representing bins left-edge values). Movies with no female producers are binned together ($N = 7974$); remaining movies ($N = 4936$) are divided into 10 equal-sized bins.

The establishment of the studio system also affected the gender diversity of a movie's cast. Figure 3.2 clearly demonstrates that the emergence of the studio system had a negative impact of female representation within casts. Between 1920 and 1940, we observe a reduction of nearly a third in the percentage of females cast for the typical movie.

The temporal evolution of female representation in three of the most visible functions in the movie industry displays the same overall "U-shape" (Fig. 3.2a). In the early years of the U.S. movie industry, female representation is high compared to mid-century levels, between 10% (for directors) and 33% (for actors). Interestingly, we find a recovery of female representation starting in the mid-1950s for actors (and in the late 1970s for producers and directors). Not coincidentally, the vise-like grip of the Big Five had started to ease just a few years earlier (Fig. 3.1a). First, Olivia de Havilland's 1944 legal victory against Warner Brothers Pictures [76], started to free actors from the endless contracts tying them to a studio. Then, in 1948, the U.S. Supreme Court ruled that the structure of the movie industry violated anti-trust laws [77].

These results mirror prior findings for screenwriters. Female screenwriters were highly visible at the start of the movie industry [73]. However, this visibility dramatically decreased with the establishment of the studio system. Only recently have female TV and movie screenwriters started to gain recognition again [78, 79].

An important difference to acting, however, is that the changes being brought by the studio system for screenwriters could be interpreted as supporting the hypothesis that males are innately better writers, and that the reduction in the representation of female screenwriters was due to increased competition for economically attractive positions. No such argument can be made about innate ability for acting. In the case of acting it is unlikely that competition among individuals with different innate abilities is the mechanism driving the historical patterns of female representation within movie casts.

3.4.1. Impact of gender of power brokers

Because of the decision-making power held by producers, we next investigate whether the gender diversity of the producer affects the gender of the director selected for a movie. To test this hypothesis we first perform a logistic regression on the probability of a director being female as a function of the percentage of producers that are female (Fig. 3.2b). We find a significant correlation (pseudo-$R^2 = 0.11$, $\beta = 0.046 \pm 0.002$, $p < 0.001$), strongly suggesting that the gender of the producers contributes to explaining why the overwhelming majority of directors are male (Fig. 3.2a. See also [149]).

To further test our hypothesis, we verify whether the gender of the director affects the gender representation of actors. Splitting movies according the gender of the director (Fig. 3.2c) reveals that female directors have a statistically significant preference for more gender-balanced casts (Mann-Whitney test, $U = 2.4 \times 10^6$, $p < 0.001$). As a final test, we perform a linear regression on the percentage of actors that are female as a function of the percentage of producers that are female (Fig. 3.2d). We find a significant increase (slope $= 0.13 \pm 0.007$, $t = 19$, $p < 0.001$) in the percentage of female actors cast as the percentage of female producers grows, indicating that the gender of the producers also contributes to explaining the gender composition of a movie's cast [147, 150].

3.4.2. Impact of movie genre

A second movie characteristic that will likely affect female representation is genre. Action, Adventure, or War are all genres typically associated with male characteristics, whereas

Romance may be more identifiable with females [26, 27]. Additionally, actors need to consider the genre(s) of the movies they participate in. Novice actors are more likely to be hired in the future if they restrict to the same genre, whereas more established actors have a higher chance to get hired if they diversify the genres of their work [151].

In order to investigate the role of genre on female representation, we group movies according to genre (Fig. 3.3). For clarity, we omit genres with fewer than 700 movies. Note that, in IMDb, movies are usually classified into multiple genres (median 2) which means movies sharing an "unpopular" genre may still be considered. To check for female discrimination we compare how many females actually directed movies in a given genre with what would be expected under a genre-unbiased null model (Fig. 3.3b). We observe that, while female directors are over-represented in Documentary and Romance movies, they are under-represented in seven of the fifteen most popular genres (Mystery, Sci-Fi, Horror, Adventure, Crime, Thriller, and Action). Notably, female directors do not appear to be over or under-represented in the two most common genres, Comedy and Drama. These results confirm the impact of gender preconceptions on hiring decisions. Consistent with all the findings reported, as female director representation decreases, so does the percentage of female actors (Fig. 3.3c).

3.4.3. Impact of movie budget

As the Hollywood studio system was reeling from the lost legal battles of the 1940s, three major societal changes would force the industry to rethink its strategy. Television, which started entering U.S. households in the late 1940s, had reached over 75% households by the late 1950s [144]. Simultaneously, the Hollywood Blacklist interrupted the careers of screenwriters, actors, and directors with suspect political views [145]. A decade later, the

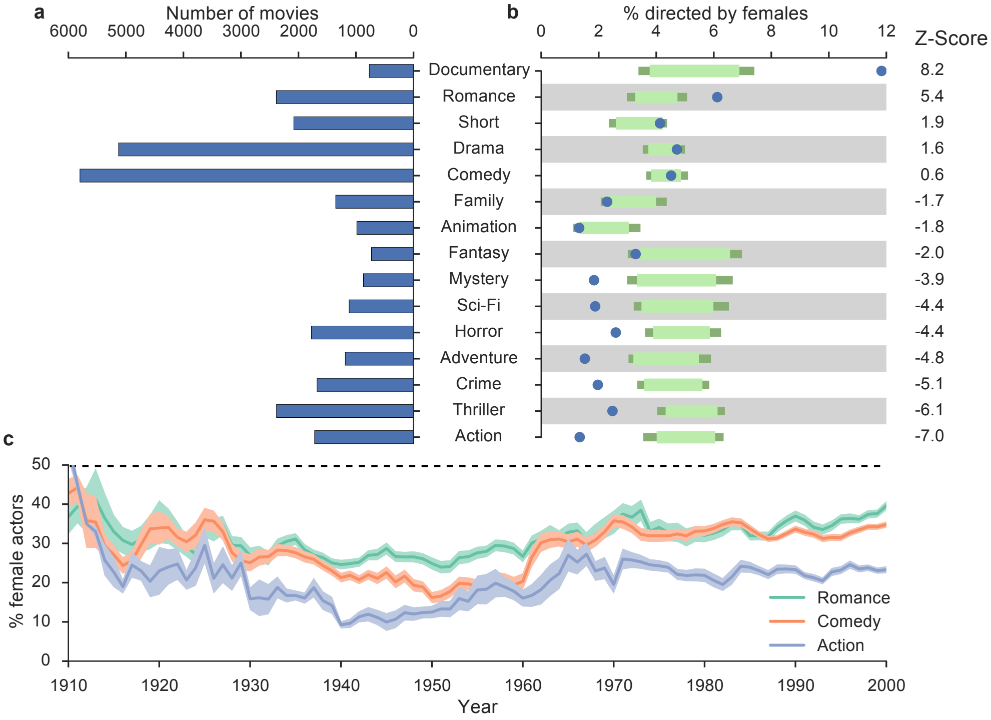

Figure 3.3. **Impact of genre on director gender**. (**a**) Number of movies classified into a given genre. Note that, in IMDb, movies are usually classified into multiple genres (median 2). We omit genres with fewer than 700 movies and consider only movies with a single director. (**b**) Female directors are over-represented (z-score > 3) in Documentary and Romance movies but under-represented (z-score < -3) in Mystery, Sci-Fi, Horror, Adventure, Crime, Thriller, and Action movies. Observed percentage of movies directed by females is indicated by the blue circles. We calculate 95% and 99% confidence intervals (light and dark green bars, respectively) by bootstrapping 1,000 samples the evolution of each genre under a binomial process for selecting a movie's director (see Appendix A.2 for simulation details). (**c**) Historical percentage of female actors (mean ± standard error) in movie genres with over-represented (Romance), typical (Comedy), and under-represented (Action) female directors. Note that, as female director representation decreases, so does the percentage of female actors. Data is smoothed over a 3-year rolling window. Black dashed line represents level of gender parity.

Vietnam War, the Civil Rights movement, and second-wave feminism forced new voices into the movie industry. As a reaction, the big Hollywood studios directed their focus towards big budget movies — blockbusters — that would have a better chance of bringing people to the theaters and achieve large profits [152, 153], and left small budget movies to independent studios [149, 154].

Prompted by these changes, we next investigate the impact of movie budget on female representation within movie casts. We have budget information for nearly 36% (5,476/15,425) of the movies in our dataset. We partition these movies by decade, and within each decade partition movies into deciles according to budget. In order to better visualize the impact of movie budget and time on female representation, we calculate deviations from the average female representation for all movies within the specific decade (Fig. 3.4). Along a column in Fig. 3.4, positive (negative) values indicate that movies within the budget decile have higher (lower) than average female representation.

In the 1910s, prior to the establishment of the studio system, there is no apparent pattern to the fluctuations in female representation according to movie budget. With the establishment of the studio system, however, we observe that higher than average female representation becomes concentrated in lower budget movies. Remarkably, during the 1930s, 1940s and 1950s, higher than average female representation shifts to increasingly higher budget movies (Fig. 3.4, leftmost green arrow).

We can understand this shift if we assume that female stars 'sold' movies just as well as male stars and that higher budget movies — which where expected to bring in greater revenues — would require greater gender balance of their casts. Indeed, while some studies have reported no impact of actors or directors on movie income [153, 155], others reported that well-known or recently successful actors, directors, and even producers positively impact movie revenues [156, 157].

In the 1960s, with the emergence of the independent studios and the greater power of male and female movie stars, we observe a dramatic change in the level of female representation as a function of movie budget. While in the previous decade, higher than average female representation was observed for movies with large budgets, during the 1960s higher than average female representation shifts to the movies with the lowest budgets. Note that this shift is accompanied by an overall increase in female representation (Fig. 3.2a). This means that females entering the industry are entering at the 'bottom'.

Strikingly, from the 1960s on, we find a steady increase in the budget size of the movies for which female representation is higher than the average (Fig. 3.4, rightmost green arrow). Again, this could be understood as the industry being unable to keep successful females stuck in the movies with the lowest budget.

3.5. Discussion

Our study suffers from two limitations. First, it does not capture changes in gender representation among starring actors, as the cast is not always shown in starring order on IMDb. Second, we do not know which percentage of female actors in the early 1900s may have been confined to minor roles or whether that percentage changed over time. Nonetheless, the U-shape of the time-series for all major movie industry functions (Fig. 3.2a) suggests that the studio system resulted in females being systematically excluded from most functions in the industry [73, 78, 79].

Our analysis supports the hypothesis that concentration of power in the hands of a few white males during the heyday of the studio system led to exclusion of other groups. The economic shift in favor of actors after the de Havilland decision [76] and the power that change brought to some actresses enabled them to later play roles as producers and directors leading to a virtuous cycle of increased female presence in the U.S. movie industry. Such a step could have had its own added benefits, as female directors have a slightly higher chance

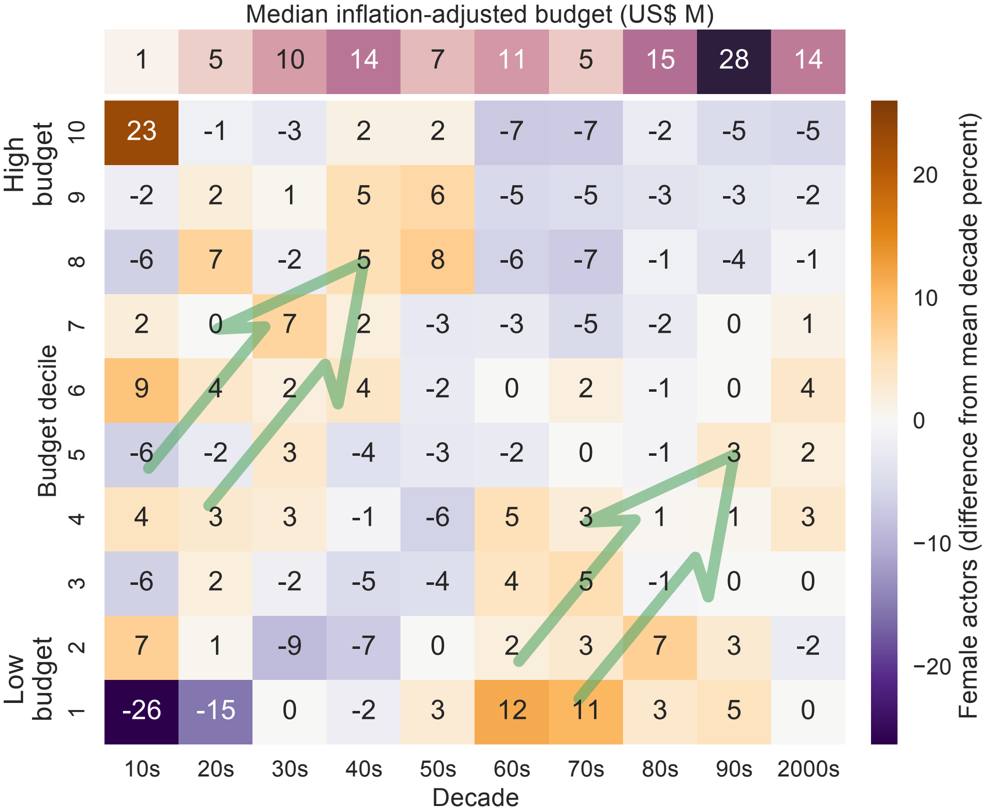

Figure 3.4. **Female cast participation as a function of time and movie budget**. In order to highlight possible dependency on budget, we calculate the percentage of female actors in the movies released each decade according to budget decile. We show the difference between the percentage for each cell and the mean percentage of female actors per decade. Top row shows the median inflation-adjusted movie budget (U.S. $millions) for each decade.

of directing award-winning movies [158], and collaborations between producers and female actors can have a positive impact on a movie's revenue [74].

Our results are consistent with the broader hypothesis that periods in which an industry grows in importance, with increasing financial rewards, and with greater consolidation may be particular susceptible to dramatic decreases in diversity. Thus, our study adds to the known examples of gender discrimination in such areas as computer science (despite the first programmers being female [159], the discipline became extremely popular among males with the advent of the home PC that was almost exclusively marketed to boys [160]) and medicine (by 1900, females struggled to be accepted in medical schools, yet in the previous century they performed almost all medical tasks without training [161]). This interpretation is also consistent with the increase in female representation in professions that lost prestige, such as teaching elementary school [162].

CHAPTER 4

A discrete lognormal model to quantify scientific impact

This work was published with Xiaohan Zeng, and Luís Amaral as "The Distribution of the Asymptotic Number of Citations to Sets of Publications by a Researcher or From an Academic Department Are Consistent With a Discrete Lognormal Model" in *PLoS ONE* [163].

4.1. Abstract

How to quantify the impact of a researcher's or an institution's body of work is a matter of increasing importance to scientists, funding agencies, and hiring committees. The use of bibliometric indicators, such as the h-index or the Journal Impact Factor, have become widespread despite their known limitations. We argue that most existing bibliometric indicators are inconsistent, biased, and, worst of all, susceptible to manipulation. Here, we pursue a principled approach to the development of an indicator to quantify the scientific impact of both individual researchers and research institutions grounded on the functional form of the distribution of the asymptotic number of citations. We validate our approach using the publication records of 1,283 researchers from seven scientific and engineering disciplines and the chemistry departments at the 106 U.S. research institutions classified as "very high research activity". Our approach has three distinct advantages. First, it accurately captures the overall scientific impact of researchers at all career stages, as measured by asymptotic citation counts. Second, unlike other measures, our indicator is resistant to

manipulation and rewards publication quality over quantity. Third, our approach captures the time-evolution of the scientific impact of research institutions.

4.2. Introduction

The explosive growth in the number of scientific journals and publications has outstripped researchers' ability to evaluate them [164]. To choose what to browse, read, or cite from a huge and growing collection of scientific literature is a challenging task for researchers in nearly all areas of Science and Technology. In order to search for worthwhile publications, researchers are thus relying more and more on heuristic proxies – such as author and journal reputations – that signal publication quality.

The introduction of the *Science Citation Index* (SCI) in 1963 [91] and the establishment of bibliographic databases spurred the development of bibliometric measures for quantifying the impact of individual researchers, journals, and institutions. Various bibliometric indicators have been proposed as measures of impact, including such notorious examples as the Journal Impact Factor and the h-index [31, 165]. However, several studies revealed that these measures can be inconsistent, biased, and, worst of all, susceptible to manipulation [80–82, 92–99]. For example, the limitations of the popular h-index include its dependence on discipline and on career length [166].

In recent years, researchers have proposed a veritable alphabet soup of "new" metrics – the g-index [167], the R-index [33], the ch-index [35], among others – most of which are *ad-hoc* heuristics, lacking insight about why or how scientific publications accumulate citations.

The onslaught of dubious indicators based on citation counts has spurred a backlash and the introduction of so-called "altmetric" indicators of scientific performance. These new indicators completely disregard citations, considering instead such quantities as number of article downloads or article views, and number of "shares" on diverse social platforms [168–170]. Unfortunately, new research is showing that altmetrics are likely to reflect popularity rather than impact, that they have incomplete coverage of the scientific disciplines [171, 172], and that they are *extremely susceptible to manipulation.* For example, inflating the findings of a publication in the abstract can lead to misleading press reports [173], and journals' electronic interfaces can be designed to inflate article views and/or downloads [174].

Citations are the currency of scientific research. In theory, they are used by researchers to recognize prior work that was crucial to the study being reported. However, citations are also used to make the research message more persuasive, to refute previous work, or to align with a given field [175]. To complicate matters further, the various scientific disciplines differ in their citation practices [176]. Yet, despite their limitations, citations from articles published in reputable journals remain the most significant quantity with which to build indicators of scientific impact [96].

It behooves us to develop a measure that is based on a thorough understanding of the citation accumulation process and also grounded on a rigorous statistical validation. Some researchers have taken some steps in this direction. Examples include the ranking of researchers using PageRank [43] or the beta distribution [44], and the re-scaling of citation distributions from different disciplines under a universal curve using the lognormal distribution [41].

One crucial aspect of the process of citation accumulation is that it takes a long time to reach a steady state [42]. This reality is often ignored in many analyses and thus confounds the interpretation of most measured values. Indeed, the lag between time of publication and perception of impact is becoming increasingly relevant. For example, faced with increasingly large pools of applicants, hiring committees need to be able to find the most qualified researchers for the position in an efficient and timely manner [177, 178]. To our knowledge, only a few attempts have been made in developing indicators that can predict future impact using citation measures [179, 180] and those have had limited success [181].

Here, we depart from previous efforts by developing a principled approach to the quantification of scientific impact. Specifically, we demonstrate that the distribution of the asymptotic number of accumulated citations to publications by a researcher or from a research institution is consistent with a discrete lognormal model [42, 124]. We validate our approach with two datasets acquired from Thomson Reuters' Web of Science (WoS):

- Manually disambiguated citation data pertaining to researchers at the top United States (U.S.) research institutions across seven disciplines [50]: chemical engineering, chemistry, ecology, industrial engineering, material science, molecular biology, and psychology;
- Citation data from the chemistry departments of 106 U.S. institutions classified as "very high research activity".

Significantly, our findings enable us to develop a measure of scientific impact with desirable properties.

4.3. Data

We perform our first set of analyses on the dataset described by Duch et al. [50]. This dataset contains the disambiguated publication records of 4,204 faculty members at some of the top U.S. research universities in seven scientific disciplines: chemical engineering, chemistry, ecology, industrial engineering, material science, molecular biology, and psychology (see [50] for details about data acquisition and validation). We consider here only 230,964 publications that were in press by the end of 2000. We do this so that every publication considered has had a time span of at least 10 years for accruing citations [124] (the researcher's publication dataset was gathered in 2010).

We perform our second set of analyses on the publication records of the chemistry departments at the top U.S. research institutions according to [182]. Using the publications' address fields, we identified 382,935 total publications from 106 chemistry departments that were in press by the end of 2009 (the department's publication dataset was gathered in 2014).

In our analyses we distinguish between "primary" publications, which report original research findings, and "secondary" publications, which analyze, promote or compile research published elsewhere. We identify as primary publications those classified by WoS as "Article", "Letter", or "Note" and identify all other publications types as secondary publications.

Moreover, to ensure that we have enough statistical power to determine the significance of the model fits, we restrict our analysis to researchers with at least 50 primary research publications. These restrictions reduce the size of the researchers dataset to 1,283 researchers and 148,878 publications. All 106 departments in our dataset have a total of more than 50 primary research publications.

4.4. The distribution of the asymptotic number of citations

Prior research suggests that a lognormal distribution can be used to approximate the steady-state citation profile of a researcher's aggregated publications [41, 121]. Stringer et al. demonstrated that the distribution of the number $n(t)$ of citations to publications published in a given journal in a given year converges to a stationary functional form after about ten years [42]. This result was interpreted as an indication that the publications published in a single journal have a characteristic citation propensity [123] which is captured by the distribution of the “ultimate” number of citations. Here, we investigate the asymptotic number of citations n_a to the publications of an individual researcher as well as the set of all researchers in a department at a research institution.

We hypothesize that n_a is a function of a latent variable ψ representing a publication's “citability” [115]. The citability ψ results from the interplay of several, possibly independent, variables such as timeliness of the work, originality of approach, strength of conclusion, reputation of authors and journals, and potential for generalization to other disciplines, just to name a few [119, 183]. In the simplest case, citability will be additive in all these variables, in which case the applicability of the central limit theorem implies that ψ will be a Gaussian variable, $\psi \in N(\mu_a, \sigma_a)$, where μ_a and σ_a are respectively the mean and standard deviation of the citability of the publications by researcher a. Therefore, the impact of a researcher's body of work is described by a distribution characterized by just two parameters, μ and σ. Similarly, because in the U.S. departments hire faculty based on their estimated quality, the researchers associated with a department will presumably be similar in stature or potential.

Unlike citations, which are observable and quantifiable, the variables contributing to ψ are neither easily observable nor easy to quantify. Moreover, mapping ψ into citations is not a trivial matter. Citation counts span many orders of magnitude, with the most highly cited publications having tens of thousands of citations [112]. Large-scale experiments on cultural markets indicate that social interactions often create a “rich get richer” dynamics, far distancing the quality of an underlying item from its impact [184]. Citation dynamics are no different. For example, Duch et al. recently showed that the h-index has a power-law dependence on the number of publications N_p of a researcher [50]. Here, we reduce the potential distortion of citation-accruing dynamics by focusing on the logarithm of n_a. In effect, we take n_a to be the result of a multiplicative process of the same variables determining

ψ. Thus, we can calculate the probability $p_{dln}(n_a)$ that a researcher or department will have a primary research publication with n_a citations, as an integral over ψ:

$$p_{dln}(n_a|\mu,\sigma) = \int\limits_{\log_{10}(n_a)}^{\log_{10}(n_a+1)} \frac{\mathrm{d}\psi}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\psi-\mu)^2}{2\sigma^2}\right) \ . \tag{4.1}$$

Most researchers also communicate their ideas to their peers via secondary publications such as conference proceedings which, in many disciplines, are mainly intended to promote related work published elsewhere. Some secondary publications will have significant timeliness, in particular review papers and editorial materials, and therefore will likely be cited too. Most of them, however, will not be cited at all. If accounting for secondary publications, Eq. (4.1) has to be generalized as:

$$P(n_a|\mu,\sigma,f_s,\boldsymbol{\theta}) = (1-f_s)p_{dln}(n_a|\mu,\sigma) + f_s\, p_s(n_a|\boldsymbol{\theta}) \ \ , \tag{4.2}$$

where f_s is the fraction of secondary publications in a body of work and $p_s(n_a|\boldsymbol{\theta})$ represents the probability distribution, characterized by parameters $\boldsymbol{\theta}$ and not necessarily lognormal, of n_a for secondary research publications. We found that in practice Eq. (4.2) can be well approximated by:

$$P(n_a|\mu',\sigma',f_s) = f_s\,\delta_{0,n_a} + (1-f_s)p_{dln}(n_a|\mu',\sigma') \ \ ,$$

where δ is the Kronecker delta. Surprisingly, we found that $\mu' \approx \mu$ and $\sigma' \approx \sigma$, suggesting that secondary publications have citation characteristics that are significantly different from those of primary publications.

4.5. Results

Fig. 4.1 shows the cumulative distribution of citations to primary research publications of two researchers in our database (see the Supporting Information of [163] for the results for all 1,283 researchers) and two chemistry departments. Using a χ^2 goodness-of-fit test with re-sampling [185], we find that we can reject the discrete lognormal model, Eq. (4.1), for only 2.88% of researchers and 1.13% or departments in our database. The results of our statistical analysis demonstrate that a discrete lognormal distribution with parameters μ and σ provides an accurate description of the distribution of the asymptotic number of citations for a researcher's body of work and for the publications from an academic department.

Fig. 4.2 displays the sample characteristics of the fitted parameters. The median value of $\hat{\mu}$ obtained for the different disciplines lies between 1.0 and 1.6. Using data reported in [176] we find a significant correlation ($\tau_{Kendall} = 0.62$, $p = 0.069$) between the median value of $\hat{\mu}$ for a discipline and the total number of citation to journals in that discipline (Fig. 4.3). This correlation suggests that $\hat{\mu}$ depends on the typical number of citations to publications within a discipline. This dependence on discipline size can in principle be corrected by a normalization factor [41, 98, 103].

We also plot the fraction of secondary publications, f_s, for all the researchers. We find that nearly a fourth of the publications of half of all researchers are secondary, but intra-discipline variation is high. Inter-discipline variability is also high: 17% of the publications of a typical researcher in chemistry are secondary, whereas 60% of the publications of a typical researcher in industrial engineering are secondary.

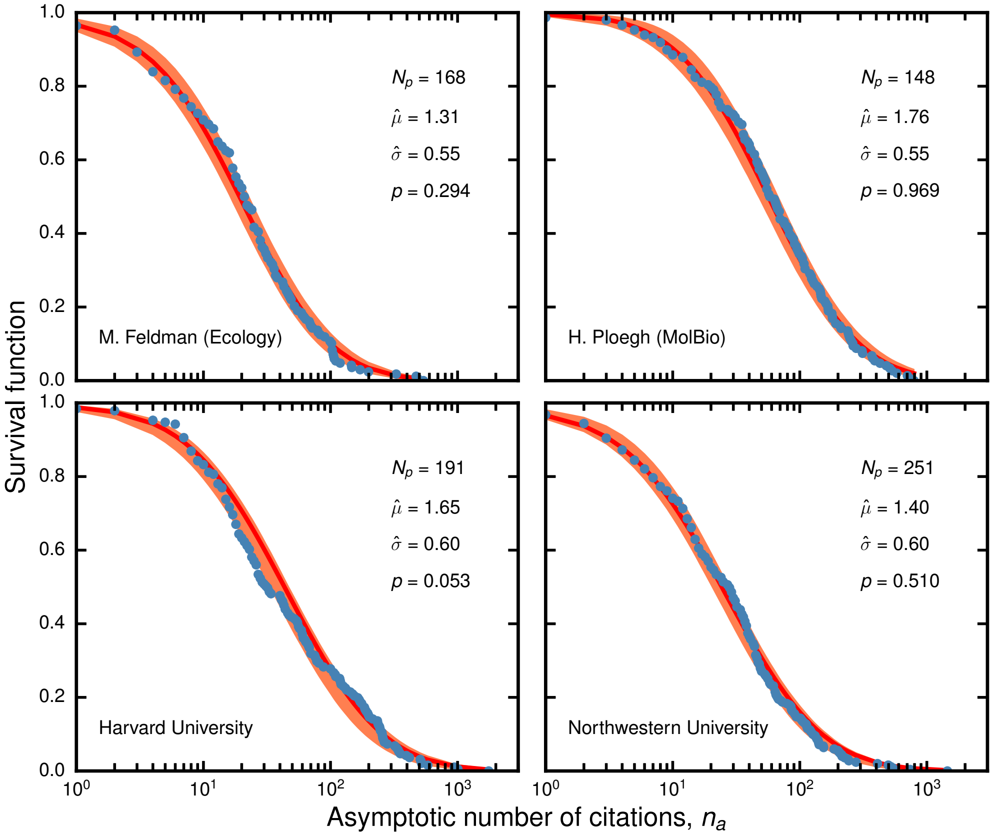

Figure 4.1. **Distribution of the asymptotic number of citations to publications for researchers and chemistry departments in our database**. We fit Eq. (4.1) to all citations accrued by 2010 to publications published by 2000 for two researchers (**top row**), and to all citations accrued by 2013 to publications published in 2000 for two chemistry departments (**bottom row**). The red line shows the maximum likelihood fit of Eq. (4.1) to the data (blue circles). The light red region represents the 95% confidence interval estimated using bootstrap (1000 generated samples per empirical data point). We also show the number of publications N_p in each set and the parameter values of the individual fits.

4.5.1. Reliability of estimation

We next investigate the dependence of the parameter estimates on number of publications, N_p, both at the individual level – testing the effect of sample size – and at the discipline level – testing overall dependence on N_p. To test for sample size dependence, we fit the model to subsets of a researcher's publication list. We find that estimates of σ are more sensitive to sample size than estimates of μ (Figs. C.15, C.16). However, this dependence becomes

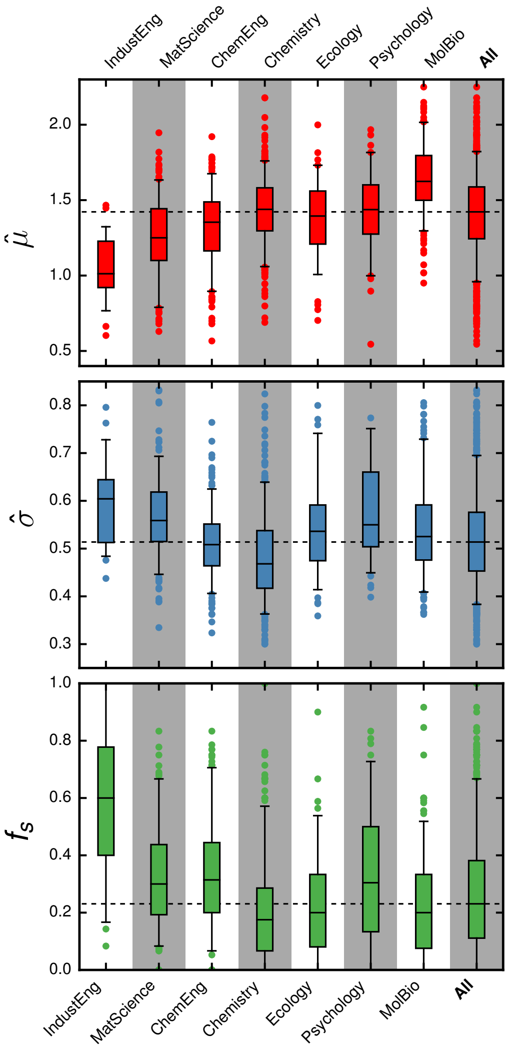

Figure 4.2. **Parameter statistics of all 1,283 researchers in the database grouped by discipline**. We show the maximum likelihood fitted model parameters (**top** and **center**) and the fraction of secondary publications (**bottom**). The black horizontal dashed line indicates the median of all researchers. For clarity, we do not show the values of $\hat{\sigma}$ for 9 researchers that are outliers.

rapidly negligible as the sample size approaches the minimum number of publications we required in creating our sample ($N_p \geq 50$).

Next, we test whether, at the discipline level, there is any dependence of $\hat{\mu}$ on N_p. We find no statistically significant correlation, except for a very weak dependence ($\mathrm{R}^2 \sim 0.035$,

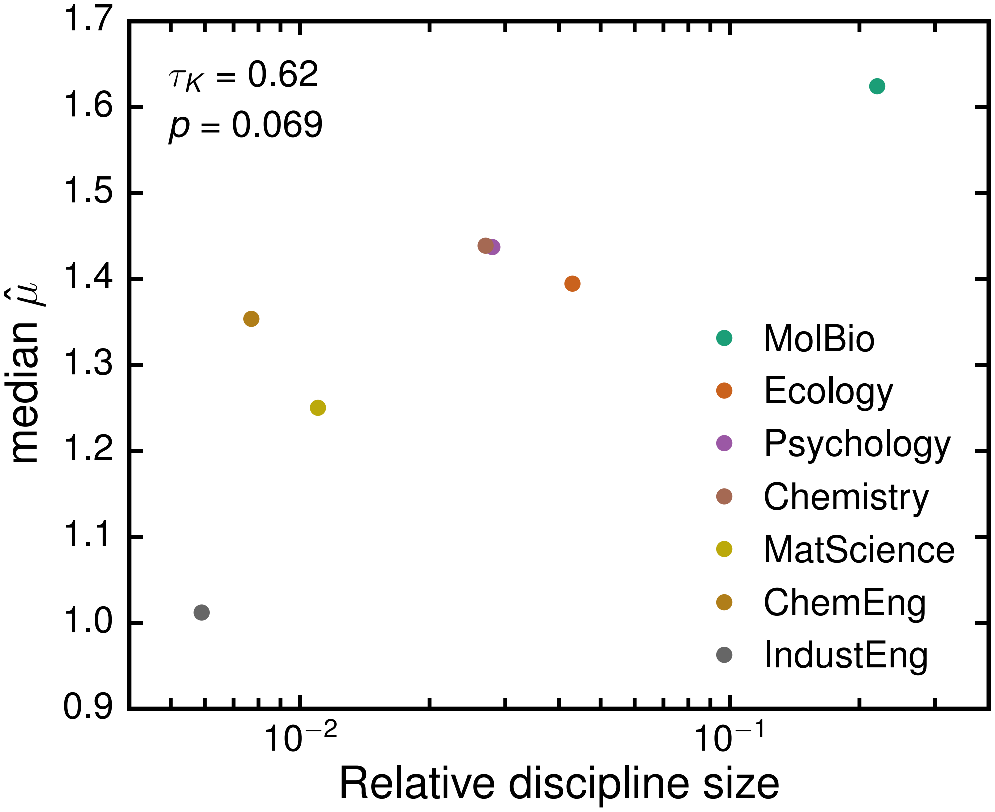

Figure 4.3. **Correlation between median $\hat{\mu}$ for a discipline and the discipline's relative size**. We use Rosvall et al. [176] reported values of the relative number of citations to publications in journals of several disciplines as a proxy for relative field size and compare them with the median value of $\hat{\mu}$ in each discipline. A Kendall rank-correlation test yields a $\tau_K = 0.62$ with $p = 0.069$. This correlation suggests that $\hat{\mu}$ depends on the typical number of citations of a discipline.

$p = 0.0052$) of $\hat{\sigma}$ on N_p for chemical engineering (Table B.8). This is in stark contrast with the h-index which exhibits a marked dependence on number of publications [166].

Then, we test for variation of the estimated parameter values along a researcher's career. To this end, we order each researcher's publication records chronologically and divide them into three sets with equal number of publications and fitted the model to each set of publications. Each set represents the citability of the publications authored at a particular career stage of a researcher. Time trends in the estimated values of μ would indicate that

Table 4.1. **Trends of $\hat{\mu}$ on career stage for the seven disciplines considered**. We divide each researcher's chronologically-ordered publication records into three sets with equal number of publications (start, middle, and end) and fit the model to each set of publications to obtain $\hat{\mu}^s$, $\hat{\mu}^m$, and $\hat{\mu}^e$. We then used ordinary-least-squares to perform a linear regression on the time dependence of $(\hat{\mu}^s, \hat{\mu}^m, \hat{\mu}^e)$. We then calculate the fraction of researchers whose μ exhibits a statistically significant dependence on career length, by performing a two-tailed significance test on the slope of the regression. We use a randomization test (1,000 samples), combined with a multiple hypothesis correction [186] (*false discovery rate* of 0.05) to calculate a p-value: for each researcher, we randomly re-order his or her publications, divide them into three sets with equal number of publications and fit the model to each set of publications, and calculate the new slope; we obtain a p-value by comparing the original slope of the fit with the distribution of the randomized slopes.

Discipline	**Upward trend in $\hat{\mu}$**	**Downward trend in $\hat{\mu}$**
ChemEng	12%	19%
Chemistry	26%	6%
Ecology	8%	8%
IndustEng	0%	33%
MatScience	10%	11%
MolBio	5%	8%
Psychology	0%	0%
All	**16%**	**9%**

the citability of a researcher's work changes over time. We find such a change for 25% of all researchers. For over 64% of those researchers whose citability changes of over time we find that $\hat{\mu}$ increases (Table 4.1).

In general, a department has many more publications than any single researcher. Thus, we are able to apply the model from Eq. (4.1) to each year's worth of departmental publications. This fine temporal resolution enables us to investigate whether there is any time-dependence in the citability of the publications from a department. Figure 4.4 shows the time-evolution of $\hat{\mu}$ for the chemistry departments at four typical research institutions. We see that both $\hat{\mu}$ (circles) and $\hat{\sigma}$ (vertical bars) remain remarkably stable over the period considered.

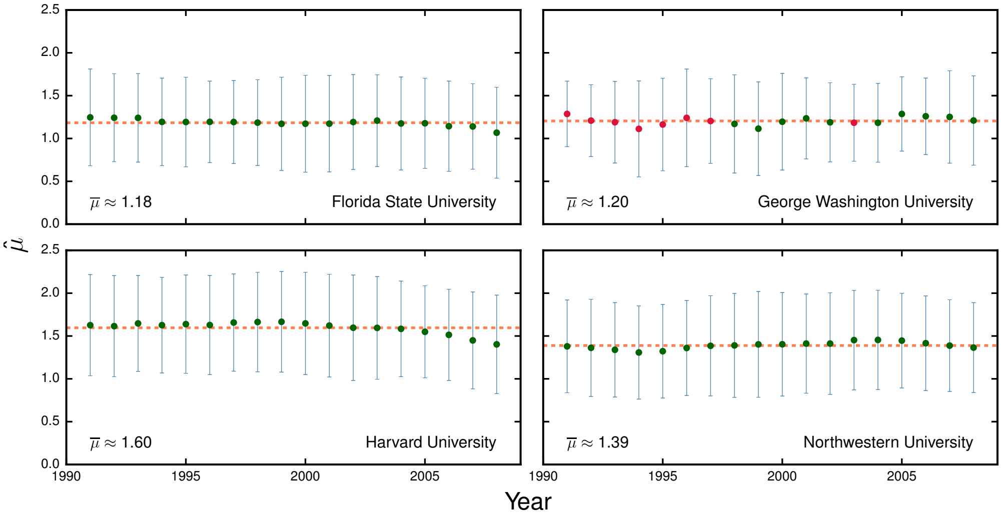

Figure 4.4. **Time-evolution of departments $\hat{\mu}$.** Each circle and bar represent, respectively, the $\hat{\mu}$ and $\hat{\sigma}$ for a given year of publications. We estimate the parameters in Eq. (4.1) for sets of departmental publications using a "sliding window" of 3 years. Fits for which we cannot reject the hypothesis that the data is consistent with a discrete lognormal distribution are colored green. We also show each department's average value of $\hat{\mu}$ over the period considered (orange dashed lines).

4.5.2. Development of an indicator

In the following, we compare the effectiveness of μ as an impact indicator with that of other indicators. First, we test the extent to which the value of μ_i for a given researcher is correlated with the values of other indicators for the same researcher. In order to provide an understanding of how the number of publications N_p influences the values of other metrics, we generate thousands of synthetic samples of n_a for different values of N_p and μ_i, and a fixed value of σ for each discipline. We find that μ is tightly correlated with several other measures, especially with the median number of citations (Fig. 4.5). Indeed $\hat{\mu}$ can be estimated from the median number of citations:

$$\hat{\mu} \cong \log_{10}[\text{median}(n_a)] \quad , \tag{4.3}$$

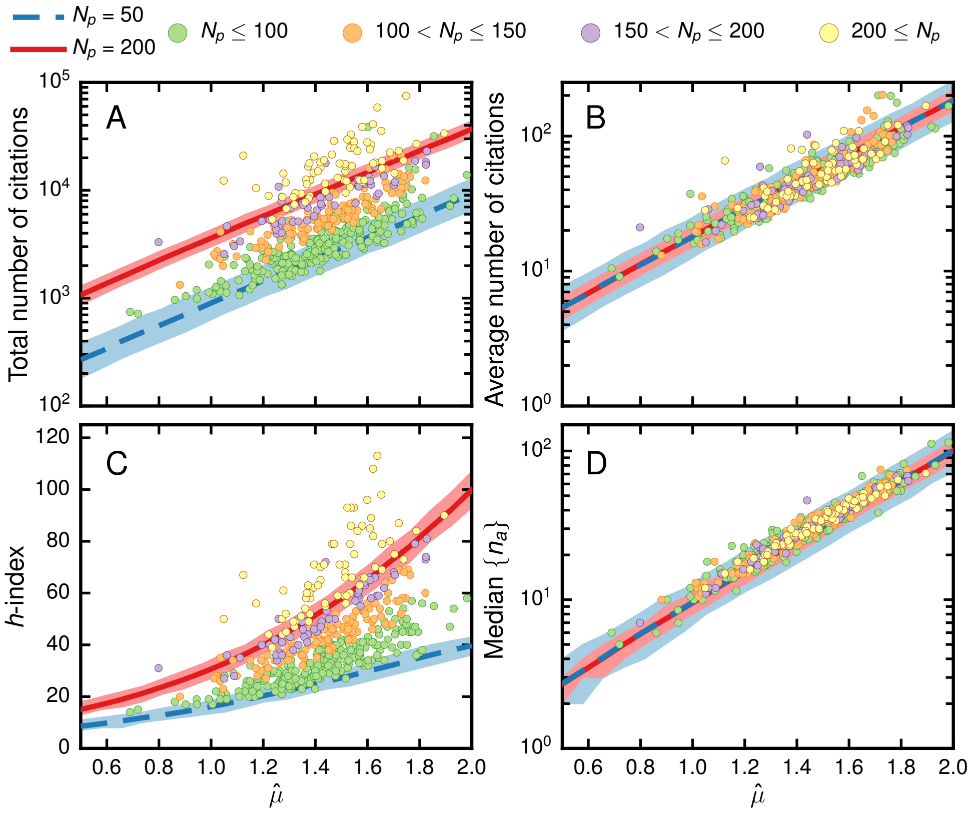

Figure 4.5. **Dependence of popular impact metrics on the values of $\hat{\mu}$ and number of publications N_p for researchers in chemistry**. We generate 1000 synthetic datasets for each of 20 values of $\hat{\mu}$ from 0.5 to 2.0, inclusive, and for $N_p = 50$ (blue) and $N_p = 200$ (red). We use the average $\hat{\sigma}$ of all researchers in chemistry. For each pair of values of $\hat{\mu}$ and N_p we calculated the average value and 95% confidence interval. The colored circles indicate the observed values of the corresponding metrics for chemistry, which have been grouped according to their number of publications N_p. Values for 22 researchers fall outside of the figures' limits: 3 in A, 7 in B, 4 in C, 3 in D. (A) The total number of citations depends dramatically on N_p, which in turn depends strongly on career length, and can be influenced by just a few highly cited publications. (B) The average number of citations is less susceptible to changes in N_p but can still be influenced by a small number of highly cited publications. (C) The h-index, like the total number of publications, is strongly dependent on N_p. (D) The median number of citations to publications, like the average, is not very dependent on N_p, and can capture most of the observed behavior.

This close relation between mean and logarithm of the median further supports our hypothesis of a lognormal distribution for the asymptotic number of citations to primary publications by a researcher.

An important factor to consider when designing a bibliometric indicator is its susceptibility to manipulation. Both the number of publications and total or average number of citations are easily manipulated, especially with the ongoing proliferation of journals of dubious reputation [187, 188]. Indeed, the h-index was introduced as a metric that resists manipulation. However, it is a straightforward exercise to show that one could achieve $h \propto \sqrt{N_p}$ exclusively through self-citations. Indeed, because the h-index does not account for the effect of self-citations, it is rather susceptible to manipulation, especially by researchers with low values of h [189, 190].

In order to determine the true susceptibility of the h-index to manipulation, we devise a method to raise a researcher's h-index using the least possible number of self-citations (see Appendix A.3 for details). Our results suggest that increasing the h-index by a small amount is no hard feat for researchers with the ability to quickly produce new articles (Fig. 4.6, left).

Our proposed indicator, μ, is far more difficult to manipulate. Because it has a more complex dependence on the number of citations than the h-index, to increase μ in an efficient manner we use a process whereby we attempt to increase the median number of citations of a researcher's work (see Appendix A.3 for details). Specifically, we manipulated μ for all the researchers by increasing their median number of citations. Remarkably, to increase μ by a certain factor one needs at least 10 times more self-citations than one would need in order to increase the h-index by the same factor (Fig. 4.6, right).

While a difference of 2 to 3 orders of magnitude in number of required self-citations may seem surprising for a measure so correlated with citation numbers (Fig. 4.5), the fact that $\hat{\mu}$ is actually dependent on the citations to half of all primary publications by a researcher (Eq. (4.3)) makes $\hat{\mu}$ less susceptible than the h-index to manipulation of citation counts from a small number of publications. This view is also supported by the fact that increasing

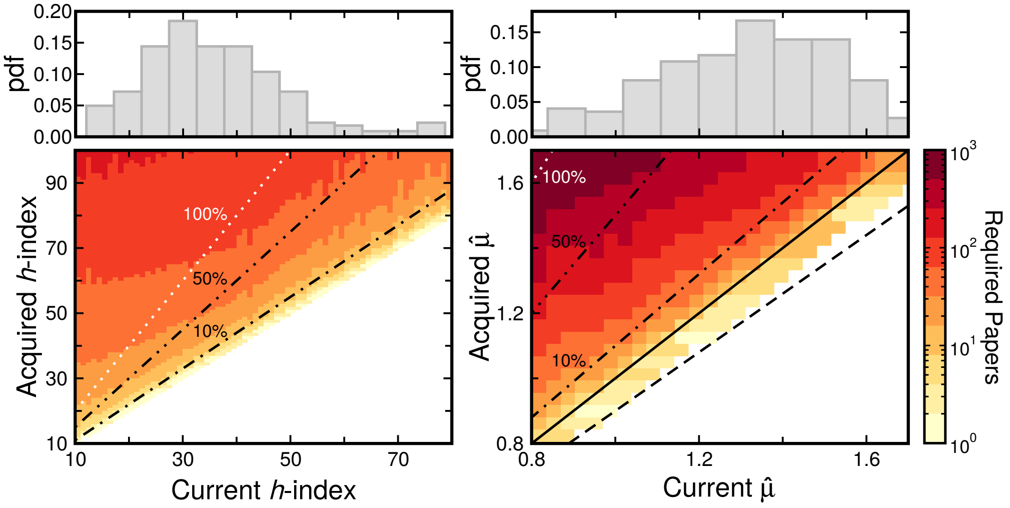

Figure 4.6. **Comparison of the susceptibility of h-index (left) and μ (right) to manipulation**. **Bottom panel**: For each researcher in the database, we add publications with self-citations until we reach the desired value of index (see main text for details). The dashed black, dotted-dashed black and dotted white lines indicate the number of publications required to increase the index value by 10%, 50% and 100%, respectively. The solid diagonal black line indicates when the current value of $\hat{\mu}$ is equal to the manipulated $\hat{\mu}$. The dark blue vertical line represents the average value of the indicator amongst all researchers in our database. **Top panel**: Distributions of current h-index (left) and $\hat{\mu}$ (right) for all researchers in the database.

citations may actually decrease $\hat{\mu}$, as we may be adding them to a publication that would not be expected to receive that number of citations given the lognormal model. As a result, manipulation of scientific performance would be very difficult if using a μ-based index.

4.5.3. Comparison of parameter statistics

Finally we estimate the parameters in Eg. (4.1) for chemistry journals and compare $\hat{\mu}$ of chemistry departments and journals in selected years, and all chemistry researchers in our database (Fig. 4.7. See Fig. C.18 for $\hat{\sigma}$ and f_s comparison). In order to make sense of this comparison, we must note a few aspects about the data. The researchers in the database were affiliated with the top 30 chemistry departments in the U.S., whereas the set of chemistry

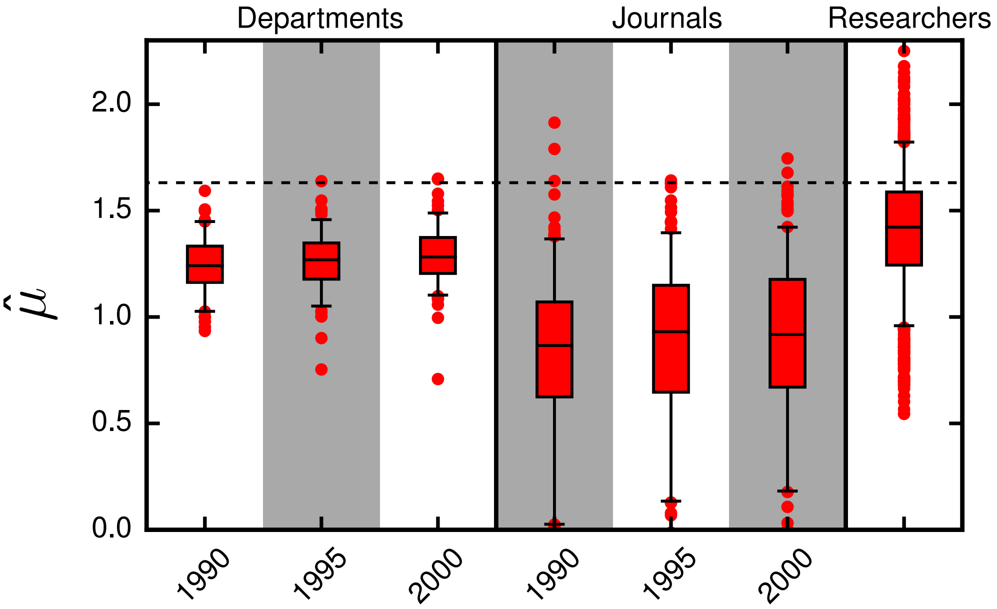

Figure 4.7. **Comparison of $\hat{\mu}$ across departments, journals, and researchers**. We show the maximum likelihood fitted $\hat{\mu}$ for chemistry departments and chemistry journals in select years, and for all chemistry researchers in our database. The black horizontal dashed lines mark the value of the corresponding parameter for the *Journal of the American Chemical Society* in 1995. For clarity, we do not show $\hat{\mu}$ for 23 journals that are outliers.

departments covers all the chemistry departments from very high research activity universities. Thus, it is natural that the typical $\hat{\mu}$ of researchers is higher than that of departments. Not surprisingly, we find that $\hat{\mu}$ is typically the lowest for journals.

4.6. Discussion

The ever-growing size of the scientific literature precludes researchers from following all developments from even a single sub-field. Therefore researchers need proxies of quality in order to identify which publications to browse, read, and cite. Three main heuristics are familiar to most researchers: institutional reputation, journal reputation, and author reputation.

Author reputation has the greatest limitations. Researchers are not likely to be known outside their (sub-)field and young researchers will not even be known outside their labs. Similarly, if we exclude a few journals with multidisciplinary reputations (Nature, Science, PNAS, NEJM), the reputation of a scientific journal is unlikely to extend outside its field. Institutional reputations are the most likely to be known broadly. Cambridge, Harvard, Oxford, and Stanford are widely recognized. However, one could argue that institutional reputation is not a particularly useful heuristic for finding quality publications within a specific research field.

Our results show that the expected citability of scientific publications published by (i) the researchers in a department, (ii) a given scientific journal, or (iii) a single researcher can be set on the single scale defined by μ. Thus, for a researcher whose publications are characterized by a very high μ, authorship of a publication may give a stronger quality signal about the publication than the journal in which the study is being published. Conversely, for an unknown researcher the strongest quality signal is likely to be the journal where the research is being published or the institution the researcher is affiliated with. Our results thus provide strong evidence for the validity of the heuristics used by most researchers and clarify the conditions under which they are appropriate.

CHAPTER 5

Conclusion

In this dissertation, I have presented a rigorous analysis of gender disparities in creative teams. I first analyzed the differences in collaboration patterns between male and female STEM researchers. Then, I studied the origins of gender discrimination in the U.S. movie industry. I have also presented a framework to quantify scientific impact of individual researchers and academic institutions.

My work is noteworthy in that all results are derived from rigorous statistical analysis of large-scale datasets. I used the Web of Science database of scientific publications when quantifying the effect of gender diversity in scientific collaborations and the impact of scientific publications; and the Internet Movie Database when quantifying the effect of gender diversity in movie productions.

In Chapter 2, I studied gender differences in scientific collaborations. I first proved that, even though female researchers have less distinct collaborators, this is only due to the fact that females publish less than males and have shorter career lengths. I then showed that, despite these disadvantages, females actually have a higher propensity to engage in novel collaborations, suggesting their work to be of higher impact than that of males. Finally, I presented evidence of female exclusion from genomics, a sub-disciplines of molecular biology.

In Chapter 3, I present evidence for how females have been discriminated against in the U.S. movie industry. Namely, I demonstrated that during the years of the Hollywood studio system, female representation among actors, directors, and producers dropped by more than half. This under-representation may be at least partially responsible for today's observed gender imbalance in the movie industry as I also found that the gender diversity of a movie's producers influences both the gender of the director and the gender composition of the cast, and that female directors have a statistically significant preference for more gender-balanced casts. Additionally, I showed that female directors are over-represented in Documentary and Romance, and under-represented in seven other genres Mystery, Sci-Fi, Horror, Adventure, Crime, Thriller, and Action. Finally, I found that higher than average female representation became concentrated in higher budget movies during the studio system, but in the 1960s higher than average female representation shifted to the movies with the lowest budgets.

In Chapter 4, I put forth the notion that scientific publications have a latent "citability" that can be estimated using the asymptotic number of citations. Specifically, I determined that the asymptotic number of citations n_a to sets of publications by a researcher or associated with an academic department can be described by a discrete lognormal distribution. I performed a principled statistical analysis of the properties of this distribution and showed that the mean citability, μ, can be used as an unbiased bibliometric indicator of scientific impact for researchers, departments, and journals. Furthermore, μ is resistant to manipulation, unlike other popular indicators such as the h-index, and can be well approximated by the median of the logarithm of n_a.

5.1. Societal implications of gender biases in creative teams

Collaborations decrease many barriers towards producing works of high impact which benefits all creators involved in the process. I have determined that gender has a profound effect in creative teams. While other researchers have also reported gender effects in teams, my research has several distinctive features. First, I leveraged the power of "Big Data" in order to avoid many sampling biases of small datasets that can lead to inaccurate conclusions. Second, I controlled for some inherent complexities in my systems of study — scientific collaborations and movie productions — that can make it difficult to draw the correct inferences, regardless of dataset size. For instance, a direct analysis of co-authorship patterns in collaborations would lead to the erroneous conclusion that male researchers collaborate more than females; and only by correcting for differences in publication volume and shorter career lengths can we uncover the true relation (Figs. 2.1, 2.3). Similarly, only by accounting for the fact that there are very few female movie directors can we show the rich genre differences across movie directors (Fig. 3.3c). My findings illustrate the need to always consider the context of where the data collected when performing any analysis.

While my study on scientific collaborations is limited to U.S. faculty members in seven distinct STEM disciplines, it could easily be extended to all scientific disciplines where collaborations are the norm. Thus we could precisely determine how field-dependent the gender effects in collaborations are. Large gender differences across sub-disciplines of the same large discipline could indicate the presence of strong gender discrimination, such as the case of molecular biology (Fig. 2.4). Conversely, academic practices in disciplines showing very minimal gender differences in collaborations warrant a deeper look as they may be promoting female participation in science.

Many researchers investigate the factors contributing to the the gradual loss of female representation along the academic career path — the “leaky pipeline”. The findings from studies can be combined with the present work to create guidelines or policies that ensure proper institutional support for both genders. For instance, highlighting the academic achievements of female researchers and creating inclusive environments for female postdoctoral students and faculty members could foster an increase in female representation in science. Indeed, evidence indicates that increasing the visibility of female leaders in careers of low female representation, such as business and politics, has a positive contribution to female advancement and can decrease gender biases [71, 141, 142].

Conversely, my study of gender representation in the U.S. movie industry clearly shows how the lack of female role models can have a strong negative impact on gender diversity: while the early U.S. movie industry had a relatively high gender diversity — females composed about one third of the cast in the typical at the time — once the studio system was established, many females either left the industry or were forced to leave, especially those working behind the camera (Fig. 3.2a). My results suggest that the accumulation of power in the hands of the few white male leaders of the big Hollywood studios lead to females being excluded from the industry. Furthermore, this negative influence can carry a strong inertia, as evidenced by the fact that, after the studio system was dissolved, it took decades for female representation to recover to pre-studio levels.

If there is power consolidation in sectors experiencing big growth, we should create incentives for teams at the top to remain as diverse as possible so as to avoid instituting biases — not just gender-related — that can take years to dispel. The discipline of computer science provides another illustrative example of this phenomenon. Programming pioneers such as Ada and Grace Hopper certainly made the nascent field appealing to females [159]. Then, with the creation of the PC, there was the opportunity to make the field accessible to the general public. Unfortunately, the product was almost exclusively marketed towards young males, which lead to a surge of male interest in the field [160]. Despite educational reforms and focus groups aimed at increasing female interest in the discipline, the stereotypical computer programmer is still overwhelmingly a young male.

Finally, it is worth noting that the factors I identified as possible causes of gender biases may also explain under-representation of other minorities in creative teams. For example, given the appropriate datasets, my analyses can be adapted to study the effect of racial or ethnic diversity in movie success, or to understand how cultural diversity of ideas affect the impact of a scientific publication.

5.2. Guidelines to quantify the impact of creative works

The goal of science is to accurately quantify and measure natural phenomena. For this reason alone we should move away from using heuristics and *ad-hoc* measures if we want to measure the impact of science itself in a rigorous way. Upon closer look, the disadvantages of bibliometric indicators such us the h-index and Journal Impact Factor far outweigh their touted ease-of-use and simple interpretation. For instance, the fact that the h-index increases monotonically over time makes it unsuitable to compare researchers at different career stages, as it penalizes younger researchers with fewer publications. Moreover, because of its dependence on publication volume, the h-index can be boosted if researchers spread their results over many publications, in effect encouraging quantity over quality of scientific research. This incentive system has consequences for hiring committees and funding agencies that may use bibliometric indicators as a first screen of their potentially hundreds of applicants. The solution to this problem is not to propose corrections to the h-index or more-complex *ad-hoc* indicators but instead use a principled, data-driven approach to quantify scientific impact.

In contrast to most existing heuristic bibliometric indicators, my proposed framework is grounded on the latent citability of a publication and as such can be used to systematically characterize the impact of any set of publications, be it those authored by a researcher, associated with a department, or published in a journal. Being able to place work by all these three entities in the same scale — the expected citability — is especially relevant when there is uncertainty about a researcher's impact. It may be difficult to directly assess the impact of a young researcher with few publications or an unknown researcher from a different discipline, but the expected citability of the journals where they have published work or their institutional address will provide a good indication of expected researcher impact. Conversely, work by researchers with high expected citability may be noteworthy even if it is published in unknown journals. Thus, my framework will enable hiring committees and funding agencies to speed up their evaluating process while simultaneously be confident that they are making sound decisions.

To create a general framework of impact of creative works, it is not enough to have a solid mathematical foundation. Any proposed model must be carefully validated against a representative dataset. For the case of scientific impact, I validated the citability framework against hundreds of thousands of publications across different scientific disciplines. With large datasets from various domains becoming more accessible than ever before, my approach can be applied to not only quantify scientific impact in other disciplines but the impact of most creative works in areas outside of science.

One such domain is the movie industry. In the U.S., movies that are deemed "culturally, historically, or aesthetically significant" to the country are preserved in the National Film Registry (NFR) [191]. For U.S. productions, induction into the NFR is perhaps the closest indicator of latent movie impact. Indeed, Wasserman et al. recently showed that the *long-gap citations*, i.e., the number of times that a movie is referenced in other movies that are 25+ years younger is a good predictor of induction into the NFR [192]. Thus these researchers demonstrated that *long-gap citations* constitute a quantitative, principled indicator of movie significance, much like the expected citability for publication impact.

References

[1] Maria Bordons, Isabel Gómez, M. Teresa Fernández, M. Angeles Zulueta, and Aida Méndez. Local, Domestic and International Scientific Collaboration in Biomedical Research. *Scientometrics*, 37(2):279–295, 1996.

[2] Gautam Ahuja. Collaboration networks, structural holes, and innovation: A longitudinal study. *Adm. Sci. Q.*, 45(3):425–455, 2000.

[3] Jeffrey H. Dyer. Effective interfirm collaboration: how firms minimize transaction costs and maximize transaction value. *Strateg. Manag. J.*, 18(7):535–556, 2002.

[4] Brian Uzzi and Jarrett Spiro. Collaboration and creativity: The small world problem. *Am. J. Sociol.*, 111(2):447–504, 2005.

[5] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(2010):686–688, 2010.

[6] Rebecca Gajda. Utilizing Collaboration Theory to Evaluate Strategic Alliances. *Am. J. Eval.*, 25(1):65–77, 2004.

[7] Roger Guimerà, Brian Uzzi, Jarrett Spiro, and Luís A. Nunes Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702, 2005.

[8] Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. The Increasing Dominance of Teams in Production of Knowledge. *Science*, 316(5827):1036–1039, 2007.

[9] Jon R. Katzenback and Douglas K. Smith. *The Wisdom of Teams.* New York: Harper Business, 2008.

[10] Amy Edmonson. Psychological Safety and Learning Behavior in Work Teams. *Adm. Sci. Q.*, 44(2):350–383, 1999.

[11] Karen A. Jehn, Gregory B. Northcraft, and Margaret A. Neale. Why Differences Make a Difference: A Field Study of Diversity, Conflict, and Performance in Workgroups. *Adm. Sci. Q.*, 44(4):741–763, 1999.

[12] Jonathon N. Cummings and Sara Kiesler. Collaborative Research Across Disciplinary and Organizational Boundaries. *Soc. Stud. Sci.*, 35(5):703–722, 2005.

[13] Benjamin F. Jones, Stefan Wuchty, and Brian Uzzi. Multi-university research teams: Shifting impact, geography, and stratification in science. *Science*, 322(5905):1259–1262, 2008.

[14] Editorial. Science for all. *Nature*, 495:5, 2013.

[15] Jennifer L. Berdahl and Cameron Anderson. Men, Women, and Leadership Centralization in Groups Over Time. *Gr. Dyn. Theory, Res. Pract.*, 9(1):45–57, 2005.

[16] Rolf Kümmerli, Caroline Colliard, Nicolas Fiechter, Blaise Petitpierre, Flavien Russier, and Laurent Keller. Human cooperation in social dilemmas: comparing the Snowdrift game with the Prisoner's Dilemma. *Proc. R. Soc. London B Biol. Sci.*, 274(1628): 2965–2970, 2007.

[17] Chris Bart and Gregory McQueen. Why women make better directors. *Int. J. Bus. Gov. Ethics*, 8(1):93–99, 2013.

[18] Lesley G. Campbell, Siya Mehtani, Mary E. Dozier, and Janice Rinehart. Gender-Heterogeneous Working Groups Produce Higher Quality Science. *PLoS One*, 8(10): 1–6, 2013.

[19] Svein Kyvik and Mari Teigen. Child Care, Research Collaboration, and Gender Differences in Scientific Productivity. *Sci. Technol. Hum. Values*, 21(1):54–71, 1996.

[20] Sooho Lee and Barry Bozeman. The Impact of Research Collaboration on Scientific Productivity. *Soc. Stud. Sci.*, 35(5):673–702, 2005.

[21] Giovanni Abramo, Ciriaco Andrea D'Angelo, and Gianluca Murgia. Gender differences in research collaboration. *J. Informetr.*, 7(4):811–822, 2013.

[22] Daniel Voyer, Susan Voyer, and M. P. Bryden. Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. *Psychol. Bull.*, 117 (2):250–270, 1995.

[23] Janet S. Hyde, Sara M. Lindberg, Marcia C. Linn, Amy B. Ellis, and Caroline C. Williams. Gender Similarities Characterize Math Performance. *Science*, 321(5888): 494–495, 2008.

[24] Stephen J. Ceci and Wendy M. Williams. Understanding current causes of women's underrepresentation in science. *Proc. Natl. Acad. Sci.*, 108(8):3157–3162, 2011.

[25] Henry Etzkowitz and Marina Ranga. Gender dynamics in science and technology: From the "leaky pipeline" to the "vanish box". *Brussels Econ. Rev.*, 54(2-3):131–147, 2011.

[26] Stacy L. Smith, Marc Choueiti, and Katherine Pieper. Gender Inequality in Popular Films: Examining On Screen Portrayals and Behind-the-Scenes Employment Patterns in Motion Pictures Released between 2007–2013. Technical report, USC Annenberg, 2014.

[27] Peter Wühr, Benjamin P. Lange, and Sascha Schwarz. Tears or Fears? Comparing Gender Stereotypes about Movie Preferences to Actual Preferences. *Front. Psychol.*, 8(March):428, 2017.

[28] Alicia Adamczyk. Why You Should Care About the Hollywood Wage Gap, 2016. URL `http://time.com/money/4207416/hollywood-wage-gap`. Accessed June 30, 2017.

[29] Erica Gonzales. Captain America Has a Lot to Say About the Gender Wage Gap, 2016. URL `http://www.harpersbazaar.com/culture/film-tv/a15627/captain-america-chris-evans-gender-wage-gap/`. Accessed June 30, 2017.

[30] Mary Sollosi. 13 Stars Who Spoke Out on the Gender Pay Gap, 2017. URL `http://ew.com/movies/stars-gender-pay-gap/natalie-portman-no-strings-attached`. Accessed June 30, 2017.

[31] Jorge E. Hirsch. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci.*, 102(46):16569–16572, 2005.

[32] Leo Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[33] BiHui Jin, LiMing Liang, Ronald Rousseau, and Leo Egghe. The R- and AR-indices: Complementing the h-index. *Chinese Sci. Bull.*, 52(6):855–863, 2007.

[34] S. Alonso, F. J. Cabrerizo, E. Herrera-Viedma, and F. Herrera. Hg-Index: a New Index To Characterize the Scientific Output of Researchers Based on the H- and G-Indices. *Scientometrics*, 82(2):391–400, 2009.

[35] Fiorenzo Franceschini, Domenico Maisano, Anna Perotti, and Andrea Proto. Analysis of the ch-index: an indicator to evaluate the diffusion of scientific research output by citers. *Scientometrics*, 85(1):203–217, 2010.

[36] D. Gnana Bharathi. Evaluation and ranking of researchers - bh index. *PLoS One*, 8 (12):e82050, 2013.

[37] Tommaso Lando and Lucio Bertoli-Barsotti. A new bibliometric index based on the shape of the citation distribution. *PLoS One*, 9(12):e115962, 2014.

[38] Fang Xu, Wenbin Liu, and Ronald Rousseau. Introducing sub-impact factor (SIF-) sequences and an aggregated SIF-indicator for journal ranking. *Scientometrics*, 102 (2):1577–1593, 2015.

[39] Marco Frittelli, Loriano Mancini, and Ilaria Peri. Scientific Research Measures. *J. Am. Soc. Inf. Sci. Technol.*, 67(12):3051–3063, 2016.

[40] Bruce Ian Hutchins, Xin Yuan, James M. Anderson, and George M. Santangelo. Relative Citation Ratio (RCR): A new metric that uses citation rates to measure influence at the article level. *PLOS Biology*, 14(9):1–25, 2016.

[41] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proc. Natl. Acad. Sci.*, 105(45):17268–17272, 2008.

[42] Michael J. Stringer, Marta Sales-Pardo, and Luís A. Nunes Amaral. Effectiveness of Journal Ranking Schemes as a Tool for Locating Information. *PLoS One*, 3(2):e1683, 2008.

[43] Filippo Radicchi, Santo Fortunato, Benjamin Markines, and Alessandro Vespignani. Diffusion of scientific credits and the ranking of scientists. *Phys. Rev. E*, 80(5):056103, 2009.

[44] Alexander M. Petersen, H. Eugene Stanley, and Sauro Succi. Statistical regularities in the rank-citation profile of scientists. *Sci. Rep.*, 1:181, 2011.

[45] Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying Long-Term Scientific Impact. *Science*, 342(6154):127–132, 2013.

[46] Donna J. Wood and Barbara Gray. Toward a Comprehensive Theory of Collaboration. *J. Appl. Behav. Sci.*, 27(2):139–162, 1991.

[47] András Schubert and Wolfgang Glänzel. Cross-national preference in co-authorship, references and citations. *Scientometrics*, 69(2):409–428, 2006.

[48] Jevin D. West, Jennifer Jacquet, Molly M. King, Shelley J. Correll, and Carl T. Bergstrom. The Role of Gender in Scholarly Authorship. *PLoS One*, 8(7):e66212, 2013.

[49] Richard A. Wanner, Lionel S. Lewis, and David I. Gregorio. Research Productivity in Academia: A Comparative Study of the Sciences, Social Sciences and Humanities. *Sociol. Educ.*, 54(4):238, 1981.

[50] Jordi Duch, Xiao Han T. Zeng, Marta Sales-Pardo, Filippo Radicchi, Shayna Otis, Teresa K. Woodruff, and Luís A. Nunes Amaral. The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLoS One*, 7(12):e51332, 2012.

[51] Jennifer Leadley. Women in U.S. Academic Medicine Statistics and Benchmarking Report. Technical report, Association of American Medical Colleges, 2009.

[52] Robert J. Menges and William H. Exum. Barriers to the Progress of Women and Minority Faculty. *J. Higher Educ.*, 54(2):123–144, 1983.

[53] Janis E. Jacobs. Twenty-five years of research on gender and ethnic differences in math and science career choices: what have we learned? *New Dir. Child Adolesc. Dev.*, 2005 (110):85–94, 2005.

[54] Molly Carnes, Claudia Morrissey, and Stacie E. Geller. Women's health and women's leadership in academic medicine: Hitting the same glass ceiling? *J. Women's Heal.*, 17(9):1453–1462, 2008.

[55] Nicholas H. Wolfinger. Problems in the Pipeline: Gender, Marriage, and Fertility in the Ivory Tower. *J. Higher Educ.*, 79(4):388–405, 2008.

[56] Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. Science faculty's subtle gender biases favor male students. *Proc. Natl. Acad. Sci.*, 109(41):16474–16479, 2012.

[57] Christine R. Harris, Michael Jenkins, and Dale Glaser. Gender differences in risk assessment: Why do women take fewer risks than men. *Judgm. Decis. Mak.*, 1(1): 48–63, 2006.

[58] Barry Bozeman and Elizabeth Corley. Scientists' collaboration strategies: implications for scientific and technical human capital. *Res. Policy*, 33(4):599–616, 2004.

[59] Barry Bozeman and Monica Gaughan. How do men and women differ in research collaborations? An analysis of the collaborative motives and strategies of academic researchers. *Res. Policy*, 40(10):1393–1402, 2011.

[60] John M. McDowell and Janet Kiholm Smith. The effect of gender-sorting on propensity to coauthor: Implications for academic promotion. *Econ. Inq.*, 30(1):68–82, 1992.

[61] Nadine V. Kegen. Science Networks in Cutting-edge Research Institutions: Gender Homophily and Embeddedness in Formal and Informal Networks. *Procedia - Soc. Behav. Sci.*, 79:62–81, 2013.

[62] Kate Egan and Sarah Thomas. Introduction: Star-Making, Cult-Making and Forms of Authenticity. In Kate Egan and Sarah Thomas, editors, *Cult Film Stardom Offbeat Attract. Process. Cultification*, pages 1–17. Palgrave Macmillan UK, London, 2013.

[63] Sarah Berry. *Screen Style: Fashion and Femininity in 1930s Hollywood.* Univ Of Minnesota Press, 2002.

[64] Anthony J. Nownes. An Experimental Investigation of the Effects of Celebrity Support for Political Parties in the United States. *Am. Polit. Res.*, 40(3):476–500, 2012.

[65] Motion Picture Association of America. 2015 Theatrical Market Statistics. Technical report, Motion Picture Association of America, 2015.

[66] Motion Picture Association of America. Creating Jobs, Trading Around the World. Technical report, Motion Picture Association of America, 2017.

[67] Stacy L. Smith, Katherine Pieper, and Marc Choueiti. Inclusion in the Director's Chair? Gender, Race, & Age of Film Directors Across 1,000 Films from 2007-2016. Technical Report February, USC Annenberg, 2017.

[68] Doris G. Bazzini, William D. McIntosh, Stephen M. Smith, Sabrina Cook, and Caleigh Harris. The Aging Woman in Popular Film: Underrepresented, Unattractive, Unfriendly, and Unintelligent. *Sex Roles*, 36(7-8):531–543, 1997.

[69] Anne E. Lincoln and Michael Patrick Allen. Double jeopardy in Hollywood: Age and gender in the careers of film actors, 1926-1999. *Sociol. Forum*, 19(4):611–631, 2004.

[70] Irene E. De Pater, Timothy A. Judge, and Brent A. Scott. Age, Gender, and Compensation: A Study of Hollywood Movie Stars. *J. Manag. Inq.*, 23(4):407–420, 2014.

[71] Deborah Dean. No human resource is an island: Gendered, racialized access to work as a performer. *Gender, Work Organ.*, 15(2):161–181, 2008.

[72] Mark Lutter. Is There a Closure Penalty? Cohesive Network Structures, Diversity, and Gender Inequalities in Career Advancement. *MPIfG Discuss. Pap. 13/9*, 2013.

[73] Laurel Smith-Doerr. Flexible Organizations, Innovation and Gender Equality: Writing for the US Film Industry, 190727. *Ind. Innov.*, 17(1):5–22, 2010.

[74] Vishal Narayan and Vrinda Kadiyali. Repeated Interactions and Improved Outcomes: An Empirical Analysis of Movie Production in the United States. *Manage. Sci.*, 62(2): 591–607, 2016.

[75] Marshall Deutelbaum. *The Genius of the System: Hollywood Filmmaking in the Studio Era.* JSTOR, 1989.

[76] De Haviland v. Warner Bros. Pictures, 67 Cal.App.2d 225, 1944.

[77] United States v. Paramount Pictures, Inc., 334 U.S. 131, 1948.

[78] Denise D. Bielby and William T. Bielby. Women and Men in Film: Gender Inequality Among Writers in a Culture Industry. *Gend. Soc.*, 10(3):248–270, 1996.

[79] Denise D. Bielby. Gender inequality in culture industries: Women and men writers in film and television. *Sociol. Trav.*, 51(2):237–252, 2009.

[80] F. Narin and Kimberly S. Hamilton. Bibliometric performance measures. *Scientometrics*, 36(3):293–310, 1996.

[81] Christine L. Borgman and Jonathan Furner. Scholarly communication and bibliometrics. *Annu. Rev. Inf. Sci. Technol.*, 36(1):2–72, 2005.

[82] Peter Vinkler. Characterization of the impact of sets of scientific papers: The Garfield (impact) factor. *J. Am. Soc. Inf. Sci. Technol.*, 55(5):431–435, 2004.

[83] Eugene Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(60): 471–479, 1972.

[84] Jevin D. West, Theodore C. Bergstrom, and Carl T. Bergstrom. The Eigenfactor Metrics TM : A Network Approach to Assessing Scholarly Journals. *Coll. Res. Libr.*, 71(3):236–244, 2010.

[85] Anthony F. J. Van Raan. Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67(3):491–502, 2006.

[86] Alain Molinari and Jean-Francois Molinari. Mathematical aspects of a new criterion for ranking scientific institutions based on the h-index. *Scientometrics*, 75(2):339–356, 2008.

[87] Lutz Bornmann, Felix de Moya Anegón, and Loet Leydesdorff. The new Excellence Indicator in the World Report of the SCImago Institutions Rankings 2011. *J. Informetr.*, 6(2):333–335, 2012.

[88] David A. King. The scientific impact of nations. *Nature*, 430(6997):311–316, 2004.

[89] Amin Mazloumian, Dirk Helbing, Sergi Lozano, Robert P. Light, and Katy Börner. Global multi-level analysis of the 'scientific food web'. *Sci. Rep.*, 3:1167, 2013.

[90] Qian Zhang, Nicola Perra, Bruno Goncalves, Fabio Ciulla, and Alessandro Vespignani. Characterizing scientific production and consumption in Physics. *Sci. Rep.*, 3:1–9,

2013.

[91] Eugene Garfield and Irving H. Sher. *Genetics Citation Index*. Institute for Scientific Information, Philadelphia, 1963.

[92] Michael H. MacRoberts and Barbara R MacRoberts. Problems of citation analysis: A critical review. *J. Am. Soc. Inf. Sci.*, 40(5):342–349, 1989.

[93] Jonathan R. Cole. A Short History of the Use of Citations as a Measure of the Impact of Scientific and Scholarly Work. In *web Knowl. A Festschrift Honor Eugene Garf.*, chapter 14, pages 281–300. Information Today, 2000.

[94] Wolfgang Glänzel and Henk F. Moed. Journal impact measures in bibliometric research. *Scientometrics*, 53(2):171–193, 2002.

[95] Lutz Bornmann and Hans-Dieter Daniel. What do we know about the h index? *J. Am. Soc. Inf. Sci. Technol.*, 58(9):1381–1385, 2007.

[96] Lutz Bornmann and Hans-Dieter Daniel. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80, 2008.

[97] S. Alonso, F. J. Cabrerizo, E. Herrera-Viedma, and F. Herrera. h-Index: A review focused in its variants, computation and standardization for different scientific fields. *J. Informetr.*, 3(4):273–289, 2009.

[98] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81(2):591–646, 2009.

[99] Allen W. Wilhite and Eric A. Fong. Coercive Citation in Academic Publishing. *Science*, 335(6068):542–543, 2012.

[100] Jorge E. Hirsch. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 85(3):741–754, 2010.

[101] Jonathan Stallings, Eric Vance, Jiansheng Yang, Michael W. Vannier, Jimin Liang, Liaojun Pang, Liang Dai, Ivan Ye, and Ge Wang. Determining scientific impact using a collaboration index. *Proc. Natl. Acad. Sci.*, 110(24):9680–5, 2013.

[102] Alexander M. Petersen, Massimo Riccaboni, H. Eugene Stanley, and Fabio Pammolli. Persistence and uncertainty in the academic career. *Proc. Natl. Acad. Sci.*, 109(14): 5213–18, 2012.

[103] Alexander M. Petersen, Fengzhong Wang, and H. Eugene Stanley. Methods for measuring the citations and productivity of scientists across time and discipline. *Phys.*

Rev. E, 81(3):036114, 2010.

[104] Zu-Guo Yang and Chun-Ting Zhang. A proposal for a novel impact factor as an alternative to the JCR impact factor. *Sci. Rep.*, 3:1–5, 2013.

[105] Jasleen Kaur, Filippo Radicchi, and Filippo Menczer. Universality of scholarly impact metrics. *J. Informetr.*, 7(4):924–932, 2013.

[106] Blaise Cronin and Kara Overfelt. Citation-based auditing of academic performance. *J. Am. Soc. Inf. Sci.*, 45(2):61–72, 1994.

[107] Ronald N. Kostoff. The Handbook of Research Impact Assessment (7th Ed.). Technical report, Office of Naval Research, Arlington VA, 1997.

[108] Peter Weingart. Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, 62(1):117–131, 2005.

[109] Colin Macilwain. What science is really worth. *Nature*, 465(June):682–684, 2010.

[110] Julia Lane and Stefano Bertuzzi. Measuring the results of science investments. *Science*, 331:678–680, 2011.

[111] Jean-Michel Fortin and David J. Currie. Big Science vs. Little Science: How Scientific Impact Scales with Funding. *PLoS One*, 8(6):e65263, 2013.

[112] Richard Van Noorden, Brendan Maher, and Regina Nuzzo. The top 100 papers. *Nature*, 514(7524):550–553, 2014.

[113] Leo Egghe. A heuristic study of the first-citation distribution. *Scientometrics*, 48(3): 345–359, 2000.

[114] Thijs Pollman. Forgetting and the ageing of scientific publications. *Scientometrics*, 47 (1):43–54, 2000.

[115] Quentin L. Burrell. Stochastic modelling of the first-citation distribution. *Scientometrics*, 52(1):3–12, 2001.

[116] Dag W. Aksnes. Characteristics of highly cited papers. *Res. Eval.*, 12(3):159–170, 2003.

[117] Hamid Bouabid. Revisiting citation aging: a model for citation distribution and life-cycle prediction. *Scientometrics*, 88(1):199–211, 2011.

[118] Alexander M. Petersen, Santo Fortunato, Raj K. Pan, Kimmo Kaski, Orion Penner, Armando Rungi, Massimo Riccaboni, H. Eugene Stanley, and Fabio Pammolli. Reputation and impact in academic careers. *Proc. Natl. Acad. Sci.*, 111(43):15316–15321, 2014.

[119] William Shockley. On the Statistics of Individual Variations of Productivity in Research Laboratories. *Proc. IRE*, 45(3):279–290, 1957.

[120] Leo Egghe. A noninformetric analysis of the relationship between citation age and journal productivity. *J. Am. Soc. Inf. Sci.*, 52(5):371–377, 2001.

[121] Sidney Redner. Citation statistics from 110 years of physical review. *Phys. Today*, 58 (6):49–54, 2005.

[122] Filippo Radicchi and Claudio Castellano. A Reverse Engineering Approach to the Suppression of Citation Biases Reveals Universal Properties of Citation Distributions. *PLoS One*, 7(3):e33833, 2012.

[123] Quentin L. Burrell. Predicting future citation behavior. *J. Am. Soc. Inf. Sci. Technol.*, 54(5):372–378, 2003.

[124] Michael J. Stringer, Marta Sales-Pardo, and Luís A. Nunes Amaral. Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. *J. Am. Soc. Inf. Sci. Technol.*, 61(7): 1377–1385, 2010.

[125] Xiao Han T. Zeng, Jordi Duch, Marta Sales-Pardo, João A. G. Moreira, Filippo Radicchi, Haroldo V. Ribeiro, Teresa K. Woodruff, and Luís A. Nunes Amaral. Differences in Collaboration Patterns across Discipline, Career Stage, and Gender. *PLoS Biol.*, 14 (11):1–19, 2016.

[126] John M. Levine and Richard L. Moreland. Collaboration: The Social Context of Theory Development. *Personal. Soc. Psychol. Rev.*, 8(2):164–172, 2004.

[127] Daniel Stokols, Kara L. Hall, Brandie K. Taylor, and Richard P. Moser. The Science of Team Science. Overview of the Field and Introduction to the Supplement. *Am. J. Prev. Med.*, 35(2 SUPPL.):S77–89, 2008.

[128] Holly J. Falk-Krzesinski, Katy Börner, Noshir Contractor, Stephen M. Fiore, Kara L. Hall, Joann Keyton, Bonnie Spring, Daniel Stokols, William Trochim, and Brian Uzzi. Advancing the Science of Team Science. *Clin. Transl. Sci.*, 3(5):263–266, 2010.

[129] Staša Milojević. Principles of scientific research team formation and evolution. *Proc. Natl. Acad. Sci.*, 111(11):3984–3989, 2014.

[130] Jonathan R. Cole and Harriet Zuckerman. The Productivity Puzzle: Persistence and change in patterns of publication of men and women scientists. In M. L.; Maehr and M. W. Steinkamp, editors, *Adv. Motiv. Achiev.*, pages 217–258. JAI Press, 1984.

[131] Giovanni Abramo, Ciriaco Andrea D'Angelo, and Gianluca Murgia. The collaboration behaviors of scientists in Italy: A field level analysis. *J. Informetr.*, 7(2):442–454, 2013.

[132] Yu Xie and Kimberlee A. Shauman. Sex Differences in Research Productivity: New Evidence about an Old Puzzle. *Am. Sociol. Rev.*, 63(6):847, 1998.

[133] Lidia Ceriani and Paolo Verme. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *J. Econ. Inequal.*, 10(3):421–443, 2011.

[134] David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons, New York, 1992.

[135] Andrea Lancichinetti, M. Irmak Sirer, Jane X. Wang, Daniel E. Acuna, Konrad Körding, and Luís A. Nunes Amaral. High-Reproducibility and High-Accuracy Method for Automated Topic Classification. *Phys. Rev. X*, 5(1):011007, 2015.

[136] Norman Richard Draper and Harry Smith. *Applied Regression Analysis.* John Wiley & Sons, Inc., Hoboken, NJ, USA, 2nd edition, 1998.

[137] Jason M. Sheltzer and Joan C. Smith. Elite male faculty in the life sciences employ fewer women. *Proc. Natl. Acad. Sci.*, 111(28):10107–10112, 2014.

[138] Ramón S. Barthelemy and Melinda Mccormick. Gender Discrimination in Physics and Astronomy : Graduate Student Experiences of Sexism and Gender Microaggressions. *Phys. Rev. Phys. Educ. Res.*, 020119(12):1–28, 2016.

[139] Allison J. Gonsalves, Anna Danielsson, and Helena Pattersson. Masculinities and experimental practices in physics: The view from three case studies. *Phys. Rev. Phys. Educ. Res.*, 12(2):020120, 2016.

[140] G. Kirk Raab. *CEO at Genentech, 1990-1995 : Oral History Transcript.* Bancroft Library, University of California, Berkeley, 2003.

[141] Judith Lynne Zaichkowsky. Women in the board room: one can make a difference. *Int. J. Bus. Gov. Ethics*, 9(1):91, 2014.

[142] Farida Jalalzai. *Shattered, Cracked, or Firmly Intact? Women and the Executive Glass Ceiling Worldwide.* Oxford University Press, 2013.

[143] Allen J. Scott. *On Hollywood: The Place, the Industry.* Princeton University Press, 2005.

[144] Cobbett Steinberg. *TV Facts*. Facts on File, Inc, 1980.

[145] Paul Buhle and David Wagner. *Hide in Plain Sight: The Hollywood Blacklistees in Film and Television, 1950-2002*. St. Martin's Griffin, 2004.

[146] João A. G. Moreira, Murielle L. Dunand, and Luís A. Nunes Amaral. U.S. movies with gender-disambiguated actors, directors, and producers. figshare. doi:10.6084/m9.figshare.4967876.v1, 2017. URL `https://doi.org/10.6084/m9.figshare.4967876.v1`.

[147] Martha M. Lauzen and David M. Dozier. The Role of Women on Screen and behind the Scenes in the Television and Film Industries: Review of a Program of Research. *J. Commun. Inq.*, 23(4):355–373, 1999.

[148] Gino Cattani, Simone Ferriani, Marcello M. Mariani, and Stefano Mengoli. Tackling the "Galácticos" effect: Team familiarity and the performance of star-studded projects. *Ind. Corp. Chang.*, 22(6):1629–1662, 2013.

[149] Irena Grugulis and Dimitrinka Stoyanova. Social Capital and Networks in Film and TV: Jobs for the Boys? *Organ. Stud.*, 33(10):1311–1331, 2012.

[150] Stacy L. Smith, Marc Choueiti, Katherine Pieper, Ariana Case, and Artur Tofan. Inclusion or Invisibility? Comprehensive Annenberg Report on Diversity in Entertainment. Technical report, USC Annenberg, 2016.

[151] Ezra W. Zuckerman, Tai-Young Kim, Kalinda Ukanwa, and James von Rittmann. Robust Identities or Nonentities? Typecasting in the Feature-Film Labor Market. *Am. J. Sociol.*, 108(5):1018–1073, 2003.

[152] David A. Garvin. Blockbusters: The economics of mass entertainment. *J. Cult. Econ.*, 5(1):1–20, 1981.

[153] S. Abraham Ravid. Information, Blockbusters, and Stars: A Study of the Film Industry. *J. Bus.*, 72(4):463–492, 1999.

[154] Robert R. Faulkner and Andy B. Anderson. Short-Term Projects and Emergent Careers: Evidence from Hollywood. *Am. J. Sociol.*, 92(4):879–909, 1987.

[155] Andrew Ainslie, Xavier Drèze, and Fred Zufryden. Modeling Movie Life Cycles and Market Share. *Mark. Sci.*, 24(3):508–517, 2005.

[156] Anita Elberse. The Power of Stars: Do Star Actors Drive the Success of Movies? *J. Mark.*, 71(4):102–120, 2007.

[157] Allègre L. Hadida. Commercial success and artistic recognition of motion picture projects. *J. Cult. Econ.*, 34(1):45–80, 2010.

[158] Mark Lutter. Creative Success and Network Embeddedness: Explaining Critical Recognition of Film Directors in Hollywood, 1900–2010. *MPIfG Discuss. Pap. 14/11*, 2014.

[159] Laura Sydell. The Forgotten Female Programmers Who Created Modern Tech, 2014. URL `http://www.npr.org/sections/alltechconsidered/2014/10/06/345799830/the-forgotten-female-programmers-who-created-modern-tech`. Accessed April 4, 2017.

[160] Jane Margolis. *Unlocking the Clubhouse: Women in Computing*. MIT Press, 2003.

[161] Regina Markell Morantz-Sanchez. *Sympathy and Science: Women Physicians in American Medicine*. Oxford University Press, 1985.

[162] Elizabeth Boyle. The Feminization of Teaching in America, 2004. URL `https://stuff.mit.edu/afs/athena.mit.edu/org/w/wgs/prize/eb04.html`. Accessed May 5, 2017.

[163] João A. G. Moreira, Xiao Han T. Zeng, and Luís A. Nunes Amaral. The distribution of the asymptotic number of citations to sets of publications by a researcher or from an academic department are consistent with a discrete lognormal model. *PLoS One*, 10(11):1–17, 2015.

[164] David Nicholas, Anthony Watkinson, Rachel Volentine, Suzie Allard, Kenneth Levine, Carol Tenopir, and Eti Herman. Trust and Authority in Scholarly Communications in the Light of the Digital Transition: setting the scene for a major study. *Learn. Publ.*, 27(2):121–134, 2014.

[165] Eugene Garfield. The history and meaning of the journal impact factor. *J. Am. Med. Assoc.*, 295(1):90–93, 2006.

[166] Leo Egghe. Dynamich-index: The Hirsch index in function of time. *J. Am. Soc. Inf. Sci. Technol.*, 58(3):452–454, 2007.

[167] Leo Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[168] Laura Bonetta. Should You Be Tweeting? *Cell*, 139(3):452–453, 2009.

[169] Sibele Fausto, Fabio A. Machado, Luiz Fernando J. Bento, Atila Iamarino, Tatiana R. Nahas, and David S. Munger. Research Blogging: Indexing and Registering the Change in Science 2.0. *PLoS One*, 7(12):e50109, 2012.

[170] Roberta Kwok. Research impact: Altmetrics make their mark. *Nature*, 500(7463): 491–493, 2013.

[171] Stefanie Haustein and Tobias Siebenlist. Applying social bookmarking data to evaluate journal usage. *J. Informetr.*, 5(3):446–457, 2011.

[172] Jason Priem, Heather A. Piwowar, and Bradley M. Hemminger. Altmetrics in the wild: Using social media to explore scholarly impact. *arXiv12034745v1 csDL 20 Mar 2012*, 1203.4745:1–23, 2012.

[173] Amélie Yavchitz, Isabelle Boutron, Aida Bafeta, Ibrahim Marroun, Pierre Charles, Jean Mantz, and Philippe Ravaud. Misrepresentation of Randomized Controlled Trials in Press Releases and News Coverage: A Cohort Study. *PLoS Med.*, 9(9):e1001308, 2012.

[174] Philip M. Davis and Jason S. Price. eJournal interface can influence usage statistics: Implications for libraries, publishers, and Project COUNTER. *J. Am. Soc. Inf. Sci. Technol.*, 57(9):1243–1248, 2006.

[175] Terrence A. Brooks. Evidence of complex citer motivations. *J. Am. Soc. Inf. Sci.*, 37 (1):34–36, 1986.

[176] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci.*, 105(4):1118–1123, 2008.

[177] Sune Lehmann, Andrew D. Jackson, and Benny E. Lautrup. Measures for measures. *Nature*, 444(7122):1003–1004, 2006.

[178] Alison Abbott, David Cyranoski, Nicola Jones, Brendan Maher, Quirin Schiermeier, and Richard Van Noorden. Metrics: Do metrics matter? *Nature*, 465(7300):860–862, 2010.

[179] Daniel E. Acuna, Stefano Allesina, and Konrad P. Kording. Future impact: Predicting scientific success. *Nature*, 489(7415):201–202, 2012.

[180] Amin Mazloumian. Predicting Scholars' Scientific Impact. *PLoS One*, 7(11):e49246, 2012.

[181] Orion Penner, Raj K. Pan, Alexander M. Petersen, Kimmo Kaski, and Santo Fortunato. On the Predictability of Future Impact in Science. *Sci. Rep.*, 3:3052, 2013.

[182] List of research universities in the United States, 2015. URL `https://en.wikipedia.org/w/index.php?title=List_of_research_universities_in_the_United_States&direction=next&oldid=659424554`. Accessed June 30, 2017.

[183] Adrian Letchford, Helen Susannah Moat, and Tobias Preis. The advantage of short paper titles. *R. Soc. Open Sci.*, 2(8):150266, 2015.

[184] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, 311:854–856, 2006.

[185] Bryan F. J. Manly. *Randomization, bootstrap and Monte Carlo methods in biology*. Chapman and Hall/CRC, 3rd edition, 2006.

[186] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, 57(1):289–300, 1995.

[187] John Bohannon. Who's Afraid of Peer Review? *Science*, 342(6154):60–65, 2013.

[188] Declan Butler. Investigating journals: The dark side of publishing. *Nature*, 495(7442): 433–435, 2013.

[189] Michael Schreiber. Self-citation corrections for the Hirsch index. *Europhys. Lett.*, 78 (3):30002, 2007.

[190] Leif Engqvist and Joachim G. Frommen. The h-index and self-citations. *Trends Ecol. Evol.*, 23(5):250–252, 2008.

[191] Library of Congress. National Film Preservation Board, 2017. URL `https://www.loc.gov/programs/national-film-preservation-board/film-registry/`. Accessed June 30, 2017.

[192] Max Wasserman, Xiao Han T. Zeng, and Luís A. Nunes Amaral. Cross-evaluation of metrics to estimate the significance of creative works. *Proc. Natl. Acad. Sci.*, 112(5): 1281–1286, 2015.

[193] Mark E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proc. Natl. Acad. Sci.*, 101(Supplement 1):5200–5205, 2004.

[194] Gender Guesser: Guess gender from first name in Python 2 and 3 (Version 0.4.0), 2016. URL `https://github.com/lead-ratings/gender-guesser`.

[195] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

[196] U.S. News & World Report: Best Colleges Rankings 2010 Edition, 2011. URL `https://web.archive.org/web/20100512221918/http://colleges.usnews.`

rankingsandreviews.com/best-colleges/national-universities-rankings. Accessed July 12, 2017.

[197] U.S. News & World Report: Best Graduate Schools in Chemical Engineering 2011 Edition, 2011. URL https://web.archive.org/web/20100914231426/http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-engineering-schools/chemical-engineering. Accessed July 12, 2017.

[198] U.S. News & World Report: Best Graduate Schools in Chemistry 2010 Edition, 2011. URL https://web.archive.org/web/20091001135906/http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-chemistry-schools/rankings. Accessed July 12, 2017.

[199] U.S. News & World Report: Best Graduate Schools in Ecology 2010 Edition, 2011. URL https://web.archive.org/web/20090428035058/http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-biological-sciences-programs/ecology. Accessed July 12, 2017.

[200] U.S. News & World Report: Best Graduate Schools in Material Engineering 2011 Edition, 2011. URL https://web.archive.org/web/20100915141756/http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-engineering-schools/material-engineering. Accessed July 12, 2017.

[201] U.S. News & World Report: Best Graduate Schools in Molecular Biology 2010 Edition, 2011. URL https://web.archive.org/web/20090428035103/http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-biological-sciences-programs/molecular-biology. Accessed July 12, 2017.

[202] U.S. News & World Report: Best Graduate Schools in Psychology 2010 Edition, 2011. URL https://web.archive.org/web/20100515154410/http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-psychology-schools/rankings. Accessed July 12, 2017.

APPENDIX A

Methods

A.1. Gender differences in collaboration patterns of STEM researchers

A.1.1. Co-author names matching

To calculate the number of distinct co-authors for a researcher, we used the following procedure. For each researcher, we maintain a set of standardized co-author names. For each co-author name, we convert the name to a string of last name and first name initials. For example, a co-author named "Jane Linda Smith" will be converted to "Smith JL". For each publication, we standardize the names of the co-authors, and add them to the set. We finally count the number of elements in the set.

Note that using this procedure, we treat "Jane Linda Smith" and "Jane Lily Smith" as the same name, because they are both converted to the string "Smith JL". Also, we treat "Jane Linda Smith" and "Jane Smith" as different names, since the former is converted to "Smith JL", while the latter is converted to "Smith J". In reality, for a single author's co-authors, the probability for either case to happen is very small, hence the error rate of our procedure is very low.

A.1.2. Confidence interval for the survival curve of total number of distinct co-authors

We use matched sampling to obtain the confidence interval for the survival curve of total number of distinct co-authors. We consider the null hypothesis that there is no difference in the total number of co-authors between females and the males with similar number of publications. To construct the confidence interval, we generate samples of N_F males, where N_F the number of females in our dataset. For a female with n_F publications, we select a male whose number of publications falls in the range of $[0.8\ n_F,\ 1.2\ n_F]$, a range small enough to produce good matches but large enough that there is at least one match. We then compute the survival curve for the obtained sample of male authors. We obtain the confidence interval by repeating this procedure 1,000 times.

The procedure is similar for the null hypothesis that there is no difference in the total number of co-authors between females and the males with equal number of publications, except that the sample of males consists of males who have the same number of publications as the females.

A.1.3. Measuring gender difference in the distribution of collaboration opportunities

We use two methods, the Gini coefficient and the disparity index, to measure how homogeneously each author distributes all her/his collaboration opportunities among her/his co-authors. A high Gini coefficient or disparity index indicates inhomogeneity of collaboration frequency distribution, where the author collaborates highly frequently with only a small portion of her/his co-authors, but only a few times with each of the remaining majority. Thus, this author has a high propensity to concentrate her/his collaboration opportunities on a few co-authors. A low Gini coefficient or disparity index indicates that the author collaborates with each of her/his co-authors about equally frequently.

Gini coefficient. Consider author a with n_c co-authors. For each co-author c_i of a, we count the times of collaboration between a and c_i, y_i. That is, the number of publications a

has co-authored with c_i. We next arrange y_i in non-decreasing order, where $y_i \leq y_{i+1}$. The Gini coefficient of author a is calculated as

$$G(a) = \frac{2\sum_{i=1}^{n_c} i y_i}{n_c \sum_{i=1}^{n_c} y_i} - \frac{n_c + 1}{n_c} \quad . \tag{A.1}$$

Disparity index. We first calculate the weight of collaboration (link) between a and c_i as given by Newman [193],

$$w_{ac_i} = \sum_{j=1}^{k_{c_i}} \frac{1}{l_j - 1} \quad , \tag{A.2}$$

where k_{c_i} is the number of publications authored by a and c_i together, and l_j is the number of co-authors in publication j. Then we calculate for a the summation of the weights of collaboration (strength),

$$s_a = \sum_{i=1}^{n_c} w_{ac_i} \quad . \tag{A.3}$$

Finally, the disparity index is calculated as

$$\Upsilon(a) = \sum_{i=1}^{n_c} \left(\frac{w_{ac_i}}{s_a} \right)^2 n_c \quad . \tag{A.4}$$

We obtain the sample of Gini coefficients for female authors, $\{G_F\}$, and that for male authors, $\{G_M\}$. We then can obtain the significance of the difference between the two samples, by performing a Kolmogorov-Smirnov test on the cumulative distribution function curves of the two samples. We perform the same hypothesis test for $\{\Upsilon_F\}$ and $\{\Upsilon_M\}$.

A.1.4. Simulating total number of distinct co-authors

We simulate the process of accumulating distinct co-authors and then calculate the total number of distinct co-authors. For each author, we calculate the fraction of repeated co-authors, f_r. We then generate a list of publications, and record the number of collaborations with each distinct co-author. For each co-author in each publication, we decide if this co-author is a previous co-author with probability f_r. If yes, we choose a previous co-author with a probability proportional to the times of collaboration with that co-author, and increase the times of collaboration with that co-author by one. Otherwise, we add a new co-author to the list of co-authors. We do not use equal probability when choosing a previous co-author because this would lead to larger number of distinct co-authors than observed.

Initially, we assign to each author 100 publications, in each of which the author has 5 co-authors. The results show that, for most disciplines, females have significantly more distinct co-authors ($p < 0.0006$, Fig. C.4a). This is expected since females repeat co-authors less than males do. We next introduce the observed heterogeneity in the team size, by keeping the number of publications at 100 while using team sizes sampled from the author's publications. Figure C.4b shows that in this case the gender difference is no longer significant. Finally, we introduce the heterogeneity in the number of publications, by using the actual number of publications and the number of co-authors in each publication (Fig. C.4c). Now, females have significantly fewer number of distinct co-authors for most disciplines. These results clearly expose the origins of the results presented in Fig. 2.1 where by controlling for number of publications alone we observed no statistical significant difference between males and females in the number of distinct co-authors.

A.1.5. Confidence interval for the probability of greater number of co-authors per publication

We consider the probability that publications authored by female authors in our cohort have a larger number of co-authors than publications authored by male authors in our cohort as a function of the career stage of the authors. Since not all the publications are published at the same career stages of the authors, and the size of science teams is increasing with time, we do not consider raw numbers of co-authors but instead standard scores relative to career stages.

Let $n_i(y)$ denote the number of co-authors of publication i from discipline j in year y, and let $N_j(y)$ denote the total number of publications published in year y. We calculate the standard score of publication i in year y as

$$z_i(y) = \frac{n_i(y) - \mu_j(y)}{\sigma_j(y)} \quad , \tag{A.5}$$

where $\mu_j(y)$ is the average number of co-authors per publication from discipline j published in year y

$$\mu_j(y) = \frac{\sum\limits_k n_k(y)}{N_j(y)} \quad , \tag{A.6}$$

$\sigma_j(y)$ is the standard deviation of the number of co-authors per publication published in year y

$$\sigma_j(y) = \sqrt{\frac{1}{N_j(y)} \sum_k \left[n_k(y) - \mu_j(y)\right]^2} \quad . \tag{A.7}$$

We finally consider $z_i^c(s)$, the standard score of publication i as a function of the career stage $s = y - y_i$, where y_i is the year of the first publication of i's author. We then calculate for each career stage s the quantity $P\left[z_F^c(s) > z_M^c(s)\right]$, representing the probability that a publication authored by a female author has a standard score higher than that of a publication authored by a male author at the same stage of the career as the female author. We also compute the confidence intervals for these probability values, in the null hypothesis that there is no gender difference in the standard scores:

$$H_0 : z_F(t) = z_M(t). \tag{A.8}$$

We generate the confidence interval valid under this hypothesis using a re-sampling method: The populations of females and males are fixed, the values of all standard scores are also fixed, but values of the standard score are randomly reassigned among publications (this is the same as randomly reassigning the genders to authors). For each random configuration, we compute again the probability $P\left[z_F^c(s) > z_M^c(s)\right]$ and obtain the confidence interval by repeating this procedure 1,000 times.

A.1.6. Statistical significance of the number of publications with a given team size

To measure the extent to which females have different team sizes than expected, we use the hypergeometric distribution as the null model. We first account for the increasing trend in the team size over years (Fig. C.5). For publication i with n_i co-authors from discipline j in year y, we calculate the corrected team size, $\nu_i(y)$, by dividing n_i by the average number of co-authors for all the publications published in year y, $\mu_j(y)$,

$$\nu_i(y) = \frac{n_i(y)}{\mu_j(y)} \ , \ \mu_j(y) = \frac{\sum\limits_k n_k(y)}{N_j(y)} \ , \tag{A.9}$$

where $N_j(y)$ is the total number of publications published in year y. We then bin the publications according to $\nu(y)$.

For the discipline being considered, suppose there are N publications in total, of which N_F are authored by females. Consider a bin b in which there are N_b publications. If the females collaborate with teams of different sizes with equal probability, then the expected number of publications by females in b is

$$N^e_{F,b} = N_b \frac{N_F}{N} \quad . \tag{A.10}$$

Suppose that of the N_b publications in bin b, $N^o_{F,b}$ are authored by females. The probability of observing $N^o_{F,b}$ publications by females given by the hypergeometric distribution is then

$$P(X = N^o_{F,b}) = \frac{\binom{N_F}{N^o_{F,b}}\binom{N-N_F}{N_b-N^o_{F,b}}}{\binom{N}{N_b}} \quad . \tag{A.11}$$

The p-value of observing $N^o_{F,b}$ is then $P(X \leq N^o_{F,b})$. In Fig. C.6 we plot $\log \frac{N^o_{F,b}}{N^e_{F,b}}$ for each bin, and shade the regions where the p-value is significant. We use the Bonferroni correction in which the false discovery rate (FDR) is set to be 0.01. We reject the null model if p-value $< \frac{0.01}{m}$, where m is the number of bins and thus the number of hypotheses.

A.2. Probable causes of gender discrimination in the U.S. movie industry

A.2.1. Assigning gender to individuals

The gender of actors is explicitly mentioned in their individual biographical pages, thus we are able to fully determine their gender. For producers and directors that do not also have acting credits, we use indirect methods to assign a gender. If present, we parse the individual's biographical text for gender-specific pronouns (he/his/him/himself, or she/her/hers/herself). If the number of (male-) female-specific pronouns exceeds that of (female-) male-specific ones, we assume the individual is a (male) female. If the previous attempt is inconclusive, we use the Python package *gender-guesser* (version 0.4.0) [194] to “guess” the gender based on the first name of the individual. The output of *gender-guesser* is one of “female”, “mostly female”, “androgynous”, “unknown”, “mostly male”, or “male”. We only assign a gender if the guess is either “male” or “female”. If we still have not been able to assign a gender, we try to find a photograph of the individual. If all attempts fail, we mark the individual’s gender as “unknown”.

A.2.2. Null model for assigning gender to movie directors

We build a null model for gender assignment that preserves the number and genres of the movies produced each year. We consider only movies directed by a single director. We extract the number of movies of each genre released each year. For each year y in the period 1910–2000, we assign a director gender to each of N_y movies released that year while keeping track of each movie’s genre. The gender is female with probability p_y^d (equal to the fraction of active female directors in year y). After repeating this procedure for every year, we record the total fraction f_G of movies in genre G directed by females. For each genre G, we bootstrap the evolution the number of movies directed by females using 1,000 samples, and extract the 95% and 99% confidence interval bounds from the bootstrap samples.

A.2.3. Confidence interval for the probability of selecting a female actor

For any given year y, we assume the gender breakdown of the cast a_{i_y} for movie i_y to be the result of a binomial process $B(a_{i_y}, p_y^a)$ where an actor is female with probability p_y^a. Then, movie i_y has f_{i_y} female and $(a_{i_y} - f_{i_y})$ male actors. If we assume that each movie’s casting process within year y is an independent stochastic process, we can take the total actors $A_y = \sum_i a_{i_y}$ and the total female actors $F_y = \sum_i f_{i_y}$, and estimate $\widehat{p_y^a}$ from the observed fraction of female actors in all movies in a given year. Therefore, we calculate a confidence interval for the binomial proportion p_y^a using the Clopper-Pearson method [195] where F_y is the number of successes of $B(A_y, \widehat{p_y^a})$.

While the IMDb data violates the independence assumption, the error will be quite small because there are many more actors than those that can be cast within a single movie. Indeed, less than 12% of actors ever acted in more than 1 movie in a single year.

A.2.4. Data Availability

The movies, actors, directors, and producers datasets analyzed in Chapter 3 are available in *figshare* at doi.org/10.6084/m9.figshare.4967876.v1 [146].

A.3. A discrete lognormal model to quantify scientific impact

A.3.1. Model Fitting and Hypothesis Testing

We estimate the discrete lognormal model parameters of Eq. (4.1) for all 1,283 researchers in our database using a maximum likelihood estimator [124]. We then test the goodness of the fit, at an individual level using the χ^2 statistical test. We bin the empirical data in such a way that there are at least 5 expected observation per bin. To assess significance we calculate the χ_o^2 statistic for each researcher and then, for each of them, re-sample their citation records using bootstrap (1,000 samples) and calculate a new value of the statistics χ_i^2 ($i = 1$, . . . , 1,000). We then extract a p-value by comparing the observed statistic χ_o^2 with the re-sampled χ^2 distribution. Finally we use a multiple hypothesis correction [186], with a *false discovery rate* of 0.05, when comparing the model fits with the null hypothesis.

A.3.2. Generation of Theoretical Performance Indicators

For each discipline we take the average value of $\hat{\sigma}$ and 20 equally spaced values of μ between 0.5 and 2.0. We then generate 1,000 datasets of 50 and 200 publications by random sampling from Eq. (4.1). We then fit the model individually to these 2,000 synthetic datasets and extracted the h-index, average number of citations, total number of citations and median number of citations to publications with at least one citation. Finally, for each value of μ, we calculate the average and the 95% confidence interval of all the indicators.

A.3.3. Manipulation Procedure for h-index

We try to increase the h-index of a researcher by self-citations alone, i.e., we assume the researcher does not receive citations from other sources during this procedure. The procedure works by adding only the minimum required citations to those publications that would cause the h-index to increase. Consider researcher John Doe who has 3 publications with $\{n_a\} =$ (2,2,5). Doe's h is 2. Assuming those publications don't get cited by other researchers during this time period, to increase h by 1, Doe needs to publish only one additional publication with two self-citations; to increase h by 2 he must instead produce five publications with a total of eight self-citations, four of which to one of the additional five publications. We execute this procedure for all researchers in the database until they reached a h-index of 100.

A.3.4. Manipulation Procedure for μ

The manipulation of μ is based on Eq. (4.3). We try to change a researcher's μ by increasing the median number of citations to publications which have at least one citation already. We consider only self-citations originating from secondary publications, i.e., publications that will not get cited. For a given corpus of publications we first define a target increase in median, x and then calculate the number of self-citations needed to increase the current median by x citations and the corresponding number of secondary publications. We then take the initial corpus of publications and attempt to increase the median citation by $x + 1$. We repeat this procedure until we reach an increase in median citation of 2000.

APPENDIX B

Supplementary Tables

Table B.1. **US News & World Report 2010 Best Colleges [196], and Chemical Engineering [197] and Chemistry [198] specialty Graduate School Rankings**.

		Chemical Engn.		Chemistry	
Rank	University	Rank	Researchers	Rank	Researchers
1	Harvard University	—	—	5	24
1	Princeton University	6	22	16	20
4	Massachusetts Inst of Technology	1	39	1	32
4	Stanford University	5	20	1	23
4	Univ Pennsylvania	16	23	20	34
4	California Inst of Technology	3	10	1	26
10	Duke University	—	—	43	27
12	Northwestern University	16	17	9	30
14	Johns Hopkins University	23	13	28	21
15	Cornell University	13	17	9	26
17	Rice University	23	18	28	22
17	Emory University	—	—	36	20
20	Univ Notre Dame	30	20	62	35
21	Univ California, Berkeley	2	18	1	58
22	Carnegie Mellon University	16	23	50	28
24	Univ California, Los Angeles	23	13	12	54
27	Univ Michigan	13	23	16	48
35	Georgia Inst of Technology	11	38	26	42
39	Univ Wisconsin at Madison	6	19	7	46
39	Univ Illinois at Urbana-Champaign	11	20	7	44
42	Univ California, Santa Barbara	9	19	26	40
42	Rensselaer Polytechnic Inst	27	17	74	21
42	Univ California, Davis	30	27	34	38
42	Univ Washington	21	17	28	35
47	Univ Florida	23	22	36	44
47	Univ Texas at Austin	6	23	12	46
47	Pennsylvania State University	21	22	16	36
53	Ohio State University	27	17	28	41
56	Boston University	—	—	62	22
61	Univ Minnesota at Minneapolis St. Paul	3	33	22	42
61	Purdue University	15	27	22	55
68	Univ Delaware	10	24	62	32
77	Univ Colorado	19	22	28	51
88	North Carolina State University	20	24	50	28
106	Univ Massachusetts Amherst	30	18	50	27

Table B.2. **US News & World Report 2010 Best Colleges [196], and Ecology [199] and Materials Science [200] specialty Graduate School Rankings**.

		Ecology		Materials Science	
Rank	University	Rank	Researchers	Rank	Researchers
1	Harvard University	2	11	—	—
1	Princeton University	8	16	—	—
4	Massachusetts Inst of Technology	—	—	1	25
4	Stanford University	7	18	5	14
4	Univ Pennsylvania	—	—	13	20
4	California Inst of Technology	—	—	12	11
8	Univ Chicago	1	20	—	—
10	Duke University	5	31	—	24
12	Northwestern University	—	—	3	35
14	Johns Hopkins University	—	—	22	12
15	Cornell University	6	25	8	17
17	Rice University	—	13	37	15
21	Univ California, Berkeley	2	43	5	25
22	Carnegie Mellon University	—	—	13	19
24	Univ California, Los Angeles	—	—	22	19
27	Univ Michigan	—	51	7	30
35	Georgia Inst of Technology	—	—	8	39
39	Univ Wisconsin at Madison	—	31	16	20
39	Univ Illinois at Urbana-Champaign	—	68	2	23
42	Univ California, Santa Barbara	—	—	4	34
42	Rensselaer Polytechnic Inst	—	—	19	11
42	Univ Washington	—	6	26	15
47	Univ Florida	—	—	8	40
47	Univ Texas at Austin	8	16	—	—
47	Pennsylvania State University	—	58	8	24
53	Ohio State University	—	—	15	24
56	Boston University	—	—	—	31
58	Univ Georgia	10	24	—	—
61	Purdue University	—	—	16	15
68	Univ Delaware	—	—	45	14
88	North Carolina State University	—	—	31	16

Table B.3. **US News & World Report 2010 Best Colleges [196], and Molecular Biology [201] and Psychology [202] specialty Graduate School Rankings**. Univ California, San Francisco offers only graduate-level courses and thus is not part of the Best Colleges Rankings. We still include it since it is very highly ranked in the specialty of Molecular Biology.

		Ecology		Materials Science	
Rank	University	Rank	Researchers	Rank	Researchers
1	Harvard University	1	44	3	26
1	Princeton University	8	54	8	30
3	Yale University	8	38	3	25
4	Massachusetts Inst of Technology	2	67	—	—
4	Stanford University	3	24	1	32
4	California Inst of Technology	6	15	—	—
2	Washington University in St. Louis	—	155	—	—
14	Johns Hopkins University	8	32	—	—
11	Univ California, Berkeley	5	41	1	31
24	Univ California, Los Angeles	—	—	3	67
27	Univ Michigan	—	—	3	111
39	Univ Wisconsin at Madison	—	—	8	35
39	Univ Illinois at Urbana-Champaign	—	—	7	53
47	Univ Texas at Austin	—	123	—	—
61	Univ Minnesota at Minneapolis St. Paul	—	—	8	40
–	Univ California, San Francisco	4	50	—	—

Table B.4. **Research topics in molecular biology**. We show for each topic the list of most representative words and journals. The topic numbers and words are given by the topic classifying method [135], and the journals are those in which the number of publications is significantly more than expected to occur by chance if drawn from a hypergeometric distribution.

Topic	Representative words	Representative journals
B0	cell express activ signal develop	Development, Molecular and Cellular Biology, Cancer Research, Genes & Development, Journal of Immunology
B1	patient increas studi p signific	Journal of Clinical Investigation, Cancer Research, Circulation, Diabetes, Investigative Ophthalmology & Visual Science
B2	protein structur bind dna site	Biochemistry, Nucleic Acids Research, Molecular Cell, Journal of The American Chemical Society, EMBO Journal
B3	use model method data protein	Biophysical Journal, Nucleic Acids Research, Journal of The American Chemical Society, Physical Review Letters, Physical Review B
B4	channel receptor neuron cell activ	Journal of Neuroscience, Neuron, Journal of Neurophysiology, Nature Neuroscience, Journal of Physiology-London
B5	gene mutat sequenc genom chromosom	Nature Genetics, Genetics, Nucleic Acids Research, Genome Research, American Journal of Human Genetics

Continued on next page

Table B.4. Continued from previous page

Topic	Representative words	Representative journals
B6	virus infect cell viral protein	Journal of Virology, Virology, Journal of Immunology, Journal of Experimental Medicine, Nature Medicine
B7	protein membran cell transport vesicl	Journal of Cell Biology, Molecular Biology of The Cell, EMBO Journal, Journal of Cell Science, American Journal of Physiology
B8	male femal behavior sex receptor	Endocrinology, Development, Hormones and Behavior, Developmental Biology, Journal of Comparative Neurology
B9	cell microtubul protein spindl mitot	Journal of Cell Biology, Molecular Biology of The Cell, Current Biology, Genes & Development, Molecular and Cellular Biology
B10	speci sequenc phylogenet group data	Molecular Biology and Evolution, Genetics, American Journal of Botany, Systematic Botany, Molecular Phylogenetics and Evolution
B11	actin cell protein filament myosin	Journal of Cell Biology, Journal of Cell Science, Molecular Biology of The Cell, Current Biology, Neuron

Continued on next page

Table B.4. Continued from previous page

Topic	Representative words	Representative journals
B12	gene methyl histon cell dna	Molecular and Cellular Biology, Genes & Development, Molecular Cell, Genetics, Nature Genetics
B13	protein degrad ubiquitin substrat cell	Molecular Cell, Molecular and Cellular Biology, EMBO Journal, Genes & Development, Journal of Virology
B14	plant express gene protein cell	Plant Cell, Plant Journal, Plant Physiology, Plant Molecular Biology, Molecular Plant-Microbe Interactions
B15	gene v iron bacteria express	Journal of Bacteriology, Molecular Microbiology, Infection and Immunity, Applied and Environmental Microbiology, Biotechnology and Bioengineering
B16	c elsevi right reserv all	Developmental Biology, Bioorganic & Medicinal Chemistry Letters, Tetrahedron Letters, FEBS Letters, Biochemical and Biophysical Research Communications

Continued on next page

Table B.4. Continued from previous page

Topic	Representative words	Representative journals
B17	oxid no activ nitric heme	Biochemistry, Investigative Ophthalmology & Visual Science, Biochemical and Biophysical Research Communications, Archives of Biochemistry and Biophysics, Free Radical Biology and Medicine
B18	beta radic 2 dot center	Biochemistry, Journal of the American Chemical Society, Physical Review B, Physical Review Letters, Inorganic Chemistry
B19	light protein gene express arabidopsi	Plant Cell, Plant Physiology, Plant Journal, Genetics, Planta
B20	proteas inhibitor parasit activ cystein	Biochemistry, Journal of Medicinal Chemistry, Chemistry & Biology, Molecular and Biochemical Parasitology, Bioorganic & Medicinal Chemistry Letters
B21	telomer telomeras rna dna cell	Molecular and Cellular Biology, Genes & Development, Nucleic Acids Research, Molecular Cell, RNA-A Publication of the RNA Society

Continued on next page

Table B.4. Continued from previous page

Topic	Representative words	Representative journals
B22	class cell peptid t molecul	Journal of Immunology, Journal of Experimental Medicine, European Journal of Immunology, Immunity, International Immunology
B23	receptor activ thrombin bind platelet	Blood, Journal of Clinical Investigation, Biochemical Journal, Journal of Pharmacology and Experimental Therapeutics, Molecular Endocrinology
B24	charg lipid cation nmr concentr	Biochemistry, Journal of The American Chemical Society, Biophysical Journal, Langmuir, Journal of General Physiology
B25	cell oxidas neutrophil activ mice	Blood, Journal of Immunology, Journal of Experimental Medicine, Journal of Leukocyte Biology, Calcified Tissue International
B26	tumor dna skin mice adduct	Cancer Research, Molecular Carcinogenesis, Biochemistry, Carcinogenesis, Chemical Research In Toxicology
B27	phage gene genom rate cell	Journal of Bacteriology, Evolution, Molecular Biology and Evolution, Virology, RNA-A Publication of The RNA Society

Continued on next page

Table B.4. Continued from previous page

Topic	Representative words	Representative journals
B28	shock heart heat ventricular cardiac	American Journal of Physiology-Heart and Circulatory Physiology, Circulation Research, Circulation, Journal of Cardiovascular Electrophysiology, Heart Rhythm
B29	anion complex 1 angstrom 2	Journal of The American Chemical Society, Biochemistry, Chemical Communications, Journal of Organic Chemistry, Inorganic Chemistry
B30	mice cholesterol receptor apo cell	Journal of Lipid Research, Journal of Clinical Investigation, Circulation Research, Arteriosclerosis Thrombosis and Vascular Biology, Lipids
B31	zone fluid soil site depth	Development, Journal of Neuroscience, American Journal of Pathology, Geology, Journal of Geophysical Research-Planets
B32	coli e assembl pilus bladder	Infection and Immunity, Journal of Bacteriology, Molecular Microbiology, EMBO Journal, Organic & Biomolecular Chemistry
B33	ant coloni popul speci albican	Genetics, PLOS Biology, Evolution, Molecular Ecology, Insectes Sociaux

Continued on next page

Table B.4. Continued from previous page

Topic	Representative words	Representative journals
B34	surfac cell antibodi use film	Physical Review B, Applied Physics Letters, Langmuir, Journal of Physical Chemistry B, Nature Biotechnology
B35	cell protein signal kinas chemotaxi	Journal of Bacteriology, Molecular Microbiology, Biophysical Journal, Planta, Microbial Ecology
B36	spd lung macrophag cell protein	Journal of Immunology, American Journal of Respiratory Cell and Molecular Biology, Journal of Clinical Investigation, American Journal of Physiology-Lung Cellular and Molecular Physiology, Infection and Immunity
B37	protein aggreg diseas beta prion	Protein Science, Human Molecular Genetics, Annals of Neurology, ACS Chemical Biology, Archives of Neurology
B38	ligand bind structur kringl acid	Biochemistry, Journal of The American Chemical Society, Journal of Biomolecular NMR, Protein Engineering, Proteins-Structure Function and Genetics

Continued on next page

Table B.4. Continued from previous page

Topic	Representative words	Representative journals
B39	resist cell drug efflux mutant	Journal of Bacteriology, Antimicrobial Agents and Chemotherapy, Molecular Microbiology, American Journal of Physiology, Organic & Biomolecular Chemistry
B40	gene cell gut c human	Molecular Microbiology, Eukaryotic Cell, American Journal of Physiology, Journal of Nutrition, Cell Host & Microbe
B41	matrix fiber cell type tissu	Journal of Cell Biology, Journal of Clinical Investigation, Journal of Cell Science, Journal of The Acoustical Society of America, American Journal of Respiratory Cell and Molecular Biology
B42	l cell monocytogen host intracellular	Journal of Bacteriology, Infection and Immunity, Molecular Microbiology, Journal of Immunology, Journal of Experimental Medicine
B43	domain bind type vwf platelet	Blood, Journal of Clinical Investigation, Thrombosis and Haemostasis, Journal of Thrombosis and Haemostasis, Human Gene Therapy

Continued on next page

Table B.4. Continued from previous page

Topic	Representative words	Representative journals
B44	mrna rna express intestin protein	RNA-A Publication of The RNA Society, Journal of Lipid Research, Endocrinology, American Journal of Physiology, Biochemical and Biophysical Research Communications
B45	dna recombin protein meiotic chromosom	Genes & Development, Molecular and Cellular Biology, Genetics, Development, Molecular Cell
B46	express gene cell develop hoxa10	Cancer Research, Development, Developmental Biology, Molecular Endocrinology, Endocrinology
B47	receptor bind cell protein ligand	Bioconjugate Chemistry, Biochemical Journal, Journal of Neurochemistry, Journal of Medicinal Chemistry, Experimental Cell Research
B48	activ insulin acid islet increas	Biochemistry, Diabetes, Circulation Research, Journal of Clinical Investigation, Biochemical Journal
B49	subunit alpha protein gamma beta	Analytical Biochemistry, Archives of Biochemistry and Biophysics, Applied Microbiology and Biotechnology, Molecular Plant-Microbe Interactions, Journal of Phycology

Continued on next page

Table B.4. Continued from previous page

Topic	Representative words	Representative journals
B50	beta alpha termin subunit lh	Molecular Endocrinology, Endocrinology, Molecular and Cellular Endocrinology, Clinical Orthopaedics and Related Research, Bio-Technology
B51	reaction synthesi acid group use	Journal of The American Chemical Society, Biochemistry, Organic Letters, Tetrahedron Letters, Journal of Organic Chemistry
B52	mice diseas cell bone normal	Journal of Clinical Investigation, Blood, Investigative Ophthalmology & Visual Science, Molecular Therapy, Journal of Bone and Mineral Research
B53	protein activ cell kinas inositol	Molecular and Cellular Biology, Biochemical and Biophysical Research Communications, Biochemical Journal, Biotechniques, American Journal of Physiology-Endocrinology and Metabolism
B54	transcript activ promot bind protein	Molecular and Cellular Biology, Genes & Development, Molecular Cell, EMBO Journal, Journal of Virology

Continued on next page

Table B.4. Continued from previous page

Topic	Representative words	Representative journals
B55	cell protein assembl flagellar cilia	Journal of Cell Biology, Development, Cell Motility and The Cytoskeleton, Current Biology, Genetics
B56	cell oscil neuron period cycl	Journal of Neuroscience, Neuron, Journal of Neurophysiology, Nature Neuroscience, PLOS One
B57	reductas degrad protein cell j	Archives of Biochemistry and Biophysics, Circulation, American Journal of Medical Genetics, Calcified Tissue International, Protein Expression and Purification
B58	mitochondri cell protein death mitochondria	Molecular and Cellular Biology, Journal of Clinical Investigation, Circulation Research, Archives of Biochemistry and Biophysics, Current Genetics
B59	replic cell gene dna protein	Journal of Bacteriology, Molecular Microbiology, Genes & Development, Molecular Cell, Genetics
B60	toxin alpha nakatpas express cell	Developmental Biology, FEBS Letters, Memorias Do Instituto Oswaldo Cruz, Insect Biochemistry and Molecular Biology, European Journal of Biochemistry

Continued on next page

Table B.4. Continued from previous page

Topic	Representative words	Representative journals
B61	gene express protein cell 1433	Plant Physiology, Plant Molecular Biology, Plant Cell, Plant Journal, Maydica
B62	neuron express gene olfactori drosophila	Development, Neuron, Journal of Neuroscience, Genes & Development, Developmental Biology
B63	m tuberculosi infect immun secret	Infection and Immunity, Molecular Microbiology, Journal of Experimental Medicine, PLOS Pathogens, Structure
B64	protein coli respons gene stress	Journal of Bacteriology, Genes & Development, Molecular Microbiology, Molecular Cell, Molecular Biology of The Cell
B65	gene element boundari express domain	Development, Genes & Development, Molecular and Cellular Biology, Genetics, Nucleic Acids Research
B66	protein activ kinas signal inhibit	Chemistry & Biology, Cell Calcium, Science Signaling, Journal of Experimental Biology, Mutation Research
B67	activ kinas enzym acid phosphoryl	Biochemistry, Journal of Bacteriology, Biotechnology and Bioengineering, Applied Microbiology and Biotechnology, Archives of Biochemistry and Biophysics

Continued on next page

Table B.4. Continued from previous page

Topic	Representative words	Representative journals
B68	respons call pattern select differ	Journal of Neuroscience, Journal of Neurophysiology, Journal of The Acoustical Society of America, Journal of Molecular Evolution, Hearing Research

Table B.5. **The 20 most prolific scientists in our dataset publishing in topic B5 identified as genomics (outlier topic 6 in Table 2.2).**

Name	Publications in topic	Total publications	Gender
Lander ES	196	334	M
Vogelstein B	187	448	M
Chakravarti A	99	277	M
Boeke JD	86	220	M
Housman DE	85	213	M
Wilson RK	84	125	M
Botstein D	80	391	M
Kazazian HH	79	320	M
Permutt MA	77	204	M
Page DC	70	177	M
Kruglyak L	67	116	M
Walbot V	62	188	F
Zack DJ	61	149	M
Feinberg AP	60	149	M
Tilghman SM	52	131	F
Nathans J	51	144	M
Silver LM	43	150	M
Germino GG	39	86	M
Burge CB	33	51	M
Landweber LF	30	90	F

Table B.6. **The 20 most prolific scientists in our dataset publishing in topic B10 (outlier topic 7 in Table 2.2)**.

Name	**Publications in topic**	**Total publications**	**Gender**
Jansen RK	114	148	M
Hillis DM	49	143	M
Gutell RR	48	96	M
Andolfatto P	22	37	M
Warnow T	21	49	F
Lander ES	16	334	M
Wilson RK	14	125	M
Garcia BA	14	91	M
Hoekstra HE	13	38	F
Landweber LF	12	90	F
Shankland M	12	42	M
Irish VF	11	54	F
Dellaporta SL	10	51	M
Barrick JE	10	43	M
Silver LM	9	150	M
Matz MV	9	34	M
Gordon JI	8	396	M
Deng XW	8	200	M
Weissman JS	8	196	M
Bartel DP	8	121	M

Table B.7. **The 20 most prolific scientists in our dataset publishing in topic B21 identified as telomere research**.

Name	Publications in topic	Total publications	Gender
Blackburn EH	89	177	F
Zakian VA	59	109	F
Greider CW	58	86	F
Collins K	42	64	F
Campbell JL	27	117	F
Pardue ML	24	117	F
Weinberg RA	23	346	M
Boeke JD	20	220	M
Lambowitz AM	19	174	M
Bartel DP	19	121	M
Hanawalt PC	15	262	M
Sharp PA	11	396	M
Doudna JA	11	115	F
Altman S	10	170	M
Hemann MT	10	25	M
Kazazian HH	9	320	M
Bustamante C	9	205	M
Landweber LF	9	90	F
Paull TT	9	37	F
Vogelstein B	8	448	M

Table B.8. **Individual lognormal parameters show no dependence on N_p.** For each researcher within each of the seven disciplines we perform least-squares linear regression between the lognormal parameters $\hat{\mu}$ and $\hat{\sigma}$, and $\log_{10}(N_p)$. We used a permutation test to calculate the p-values: for each set of pairs, $(\hat{\mu}, N_p)$ and $(\hat{\sigma}, N_p)$, we performed 10,000 random swaps of all N_p and subsequent regression; we obtained a p-value by comparing the original slope of the fit with the distribution of the permuted slopes. $^*p < 0.05/7 \sim 0.0074$.

Parameter	**Discipline**	**slope (m)**	**intercept (b)**	**R^2**	$\boldsymbol{p}$
$\hat{\mu} = mN_p + b$	ChemEng	0.051	1.218	0.00187	0.5240
	Chemistry	0.073	1.286	0.00668	0.0744
	Ecology	-0.150	1.658	0.00844	0.4502
	IndustEng	0.379	0.348	0.04305	0.3166
	MatScience	0.106	1.043	0.01114	0.1278
	MolBio	0.156	1.332	0.01909	0.0410
	Psychology	0.104	1.229	0.00496	0.5732
$\hat{\sigma} = mN_p + b$	ChemEng	0.067	0.379	0.03542	0.0052*
	Chemistry	0.031	0.418	0.00650	0.0862
	Ecology	0.059	0.434	0.00806	0.4524
	IndustEng	-0.094	0.764	0.01657	0.5380
	MatScience	0.033	0.502	0.00921	0.1598
	MolBio	0.064	0.415	0.01688	0.0592
	Psychology	-0.092	0.761	0.02026	0.2302

Table B.9. **Individual discipline statistics of the lognormal model parameters**.

Parameter	**Discipline**	**Mean**	**Std Dev**	**Min**	**Median**	**Max**
$\hat{\mu}$	ChemEng	1.354	0.256	0.566	1.320	1.921
	Chemistry	1.439	0.225	0.689	1.437	2.179
	Ecology	1.395	0.308	0.702	1.375	1.999
	IndustEng	1.012	0.313	0.603	1.046	1.466
	MatScience	1.250	0.266	0.629	1.254	1.947
	MolBio	1.624	0.253	0.950	1.641	2.250
	Psychology	1.437	0.318	0.545	1.427	1.967
$\hat{\sigma}$	ChemEng	0.508	0.080	0.323	0.513	0.764
	Chemistry	0.468	0.095	0.300	0.482	0.956
	Ecology	0.536	0.124	0.359	0.546	0.896
	IndustEng	0.604	0.150	0.438	0.591	0.796
	MatScience	0.559	0.095	0.335	0.568	0.969
	MolBio	0.525	0.105	0.362	0.541	1.006
	Psychology	0.550	0.137	0.398	0.585	0.954

APPENDIX C

Supplementary Figures

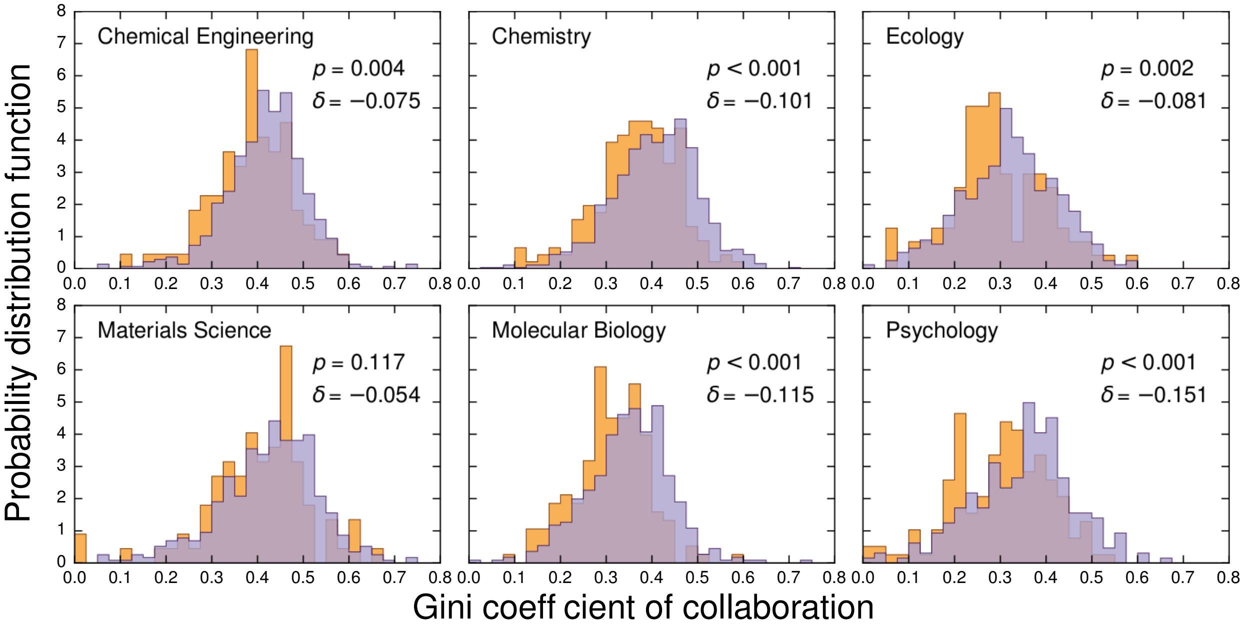

Figure C.1. **Gender differences in the propensity to repeat previous collaboration measured using the Gini coefficient**. Distribution of the Gini coefficient of collaboration heterogeneity [133] for females (orange) and males (purple) in the dataset with at least 10 publications. We exclude single-author publications. We obtain p-values for the validity of the null hypothesis that the samples were drawn from the same distribution using the Kolmogorov-Smirnov test. For all disciplines, we find $\delta = 2(\bar{G}_F - \bar{G}_M)/(\bar{G}_F + \bar{G}_M) < 0$, where $\bar{G}_F$ and $\bar{G}_M$ are the average Gini coefficient of the female and male faculty, respectively. Females have Gini coefficients smaller than those of males, suggesting that female faculty have a lower propensity than male faculty to repeat collaborations.

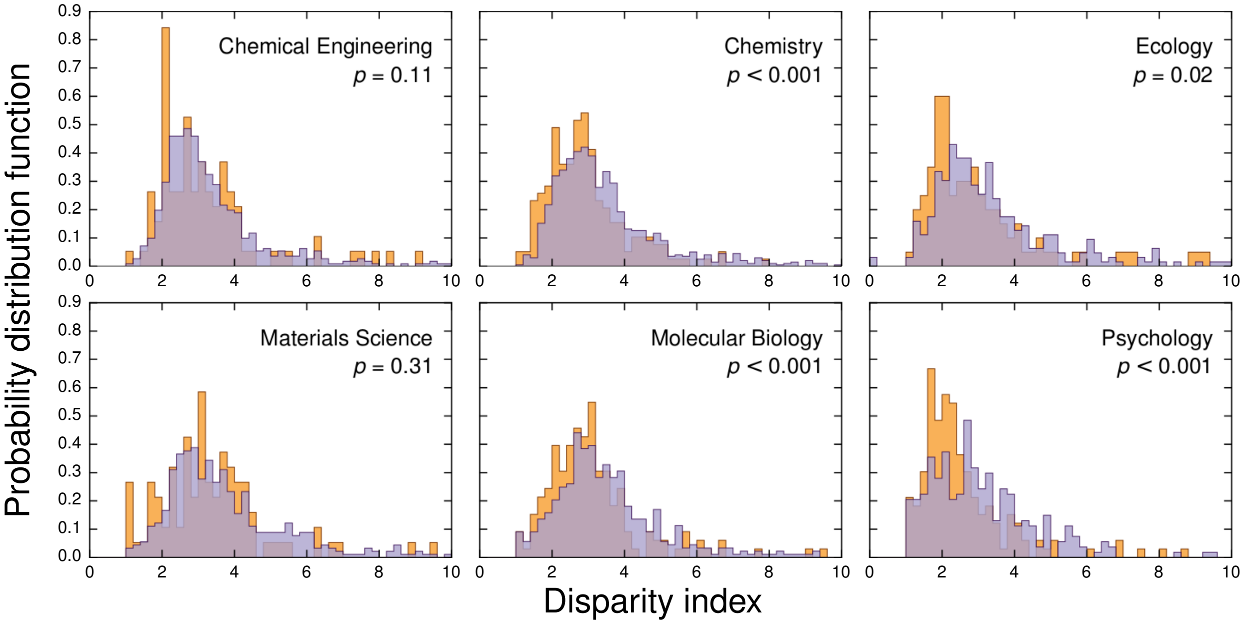

Figure C.2. **Gender difference in the propensity to repeat previous co-authors measured using the disparity index**. Distribution of the disparity index measuring the repetition of co-authors of females (orange) and males (purple). The p-values indicate the significance of the gender difference, obtained with Kolmogorov-Smirnov test. The result is in good agreement with that obtained using the Gini coefficient in Fig. C.1.

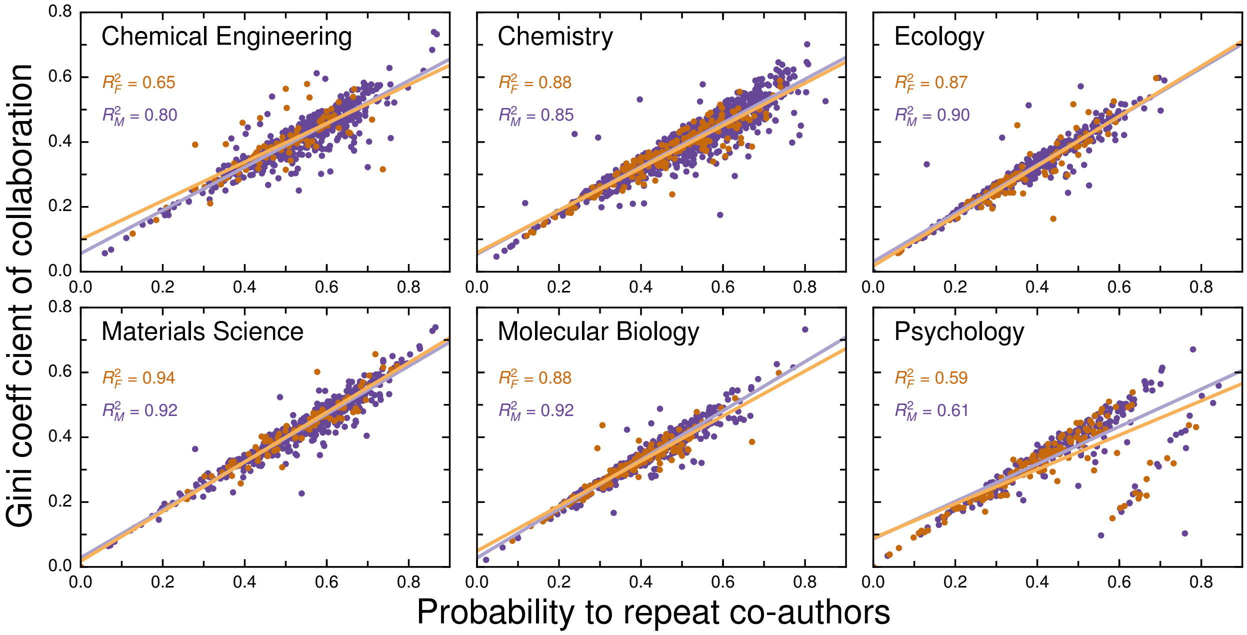

Figure C.3. **Correlation between Gini coefficient and probability to repeat previous co-authors**. Orange (female) and purple (male) lines are linear fits to data, and R_F^2 and R_M^2 are the corresponding coefficient of determination.

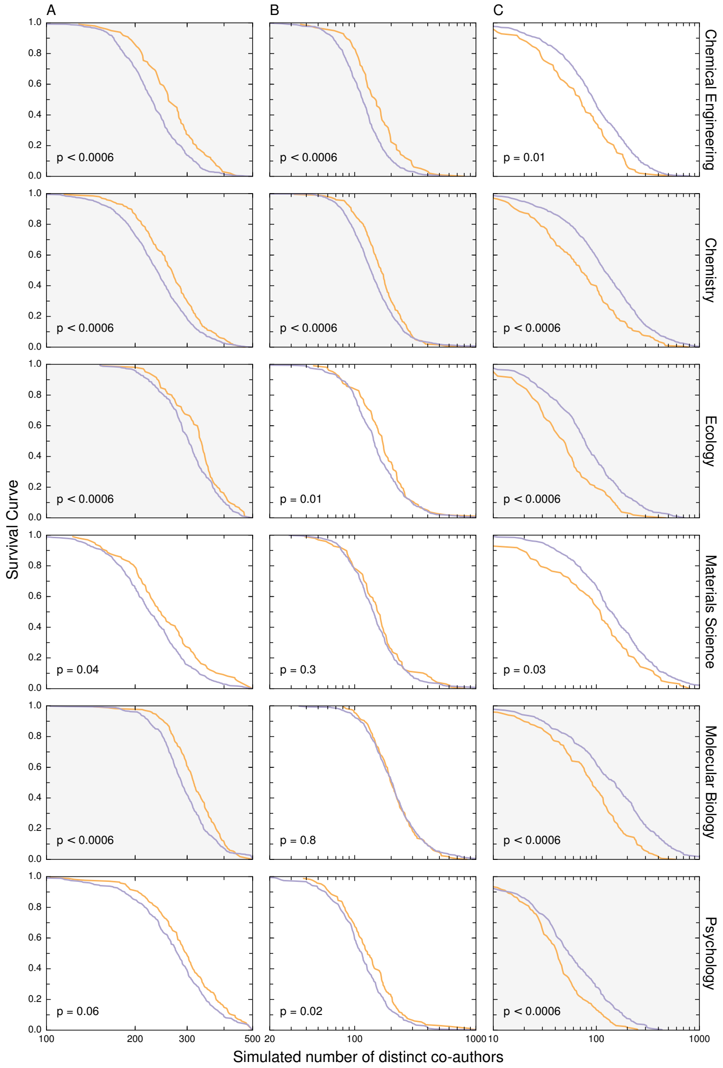

Figure C.4. **Heterogeneity in the number of publications and team size masks the effect of gender difference in the propensity to repeat co-authors**. Survival curves of the simulated total number of distinct co-authors with fixed number of publications and team size (**A**), fixed number of publications and team sizes sampled from real data (**B**), and both number of publications and team sizes from real data (**C**) for female (orange) and male (purple) faculty in all departments (see Appendix A.1 for details). We obtained p-values for the validity of the null hypothesis that the samples were drawn from the same distribution using the Kolmogorov-Smirnov test. Statistical significant results with $p < 0.01/18 \approx 0.0006$ (Bonferroni correction for multiple hypothesis) are shaded grey. When using fixed number of publications and team size, females have significantly more distinct co-authors. However, the gender difference disappears for most disciplines when using fixed number of publications but real team sizes. When we also use number of publications from the real data, females have significantly fewer distinct co-authors, consistent with Fig. 2.1.

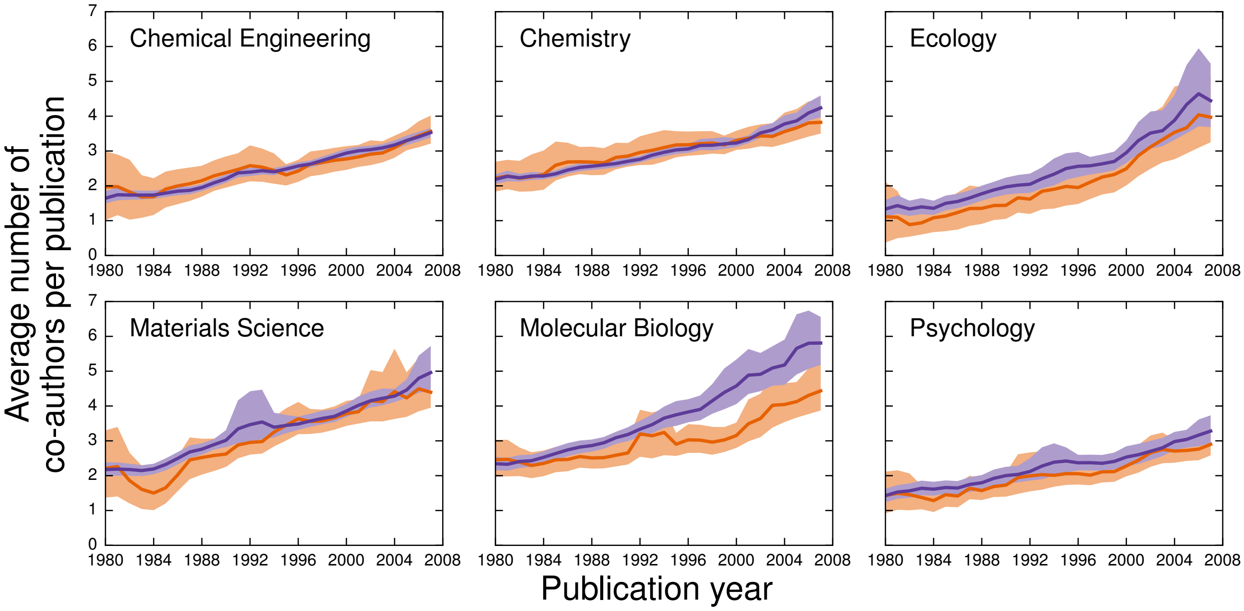

Figure C.5. **Growth of average number of co-authors during considered period**. Average number of co-authors per publication for females (orange) and males (purple) as a function of publication year. The data are smoothed using a moving averaging method with window size 3. The shaded region indicates the 99% confidence interval obtained with bootstrapping.

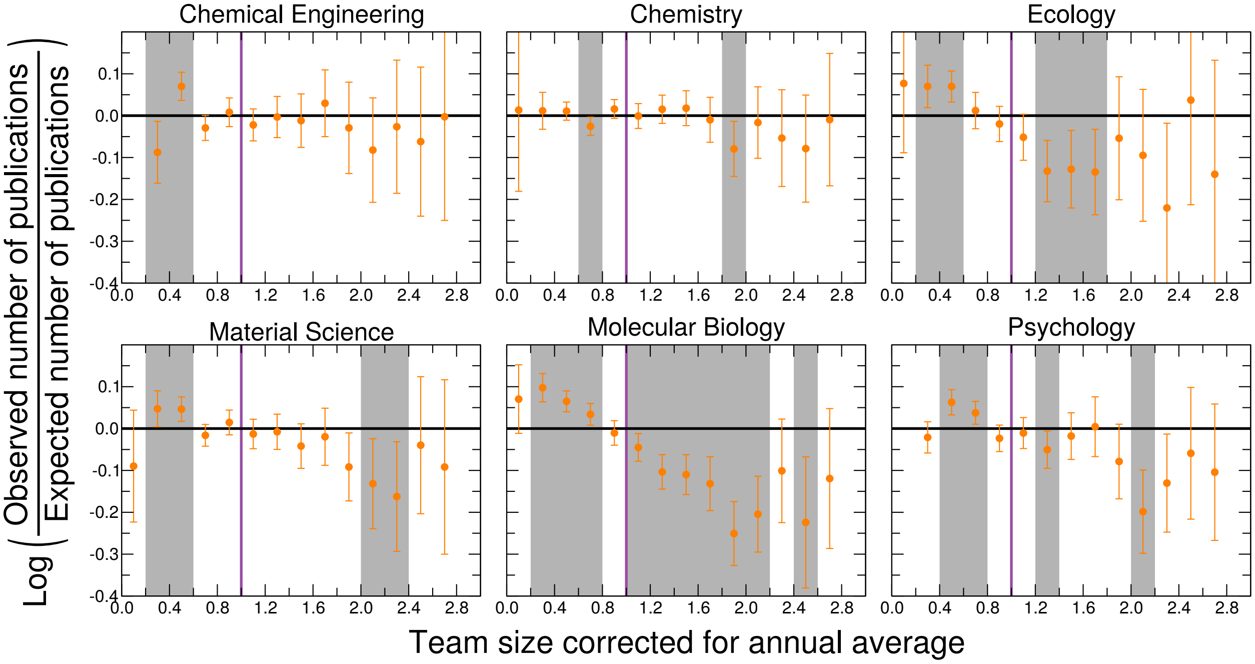

Figure C.6. **In molecular biology departments, female faculty work in smaller teams than male faculty**. Logarithm of the ratio of observed number of publications authored by females over that expected from a hypergeometric distribution (orange circles). The publications are binned by the number of co-authors corrected for the annual average with a bin size of 0.2. The shaded areas indicate that the observed number is significantly different from expected by the model, using the Bonferroni correction by treating each bin as an independent hypothesis test (see Appendix A.1 for details). The error bars indicate thrice the standard deviation. The black line indicates the ratio of 1.0, and the purple line indicates the average corrected team size. Note that for molecular biology, females have more publications than expected with smaller teams (corrected team size < 1.0) and fewer publications than expected with larger teams (corrected team size > 1.0).

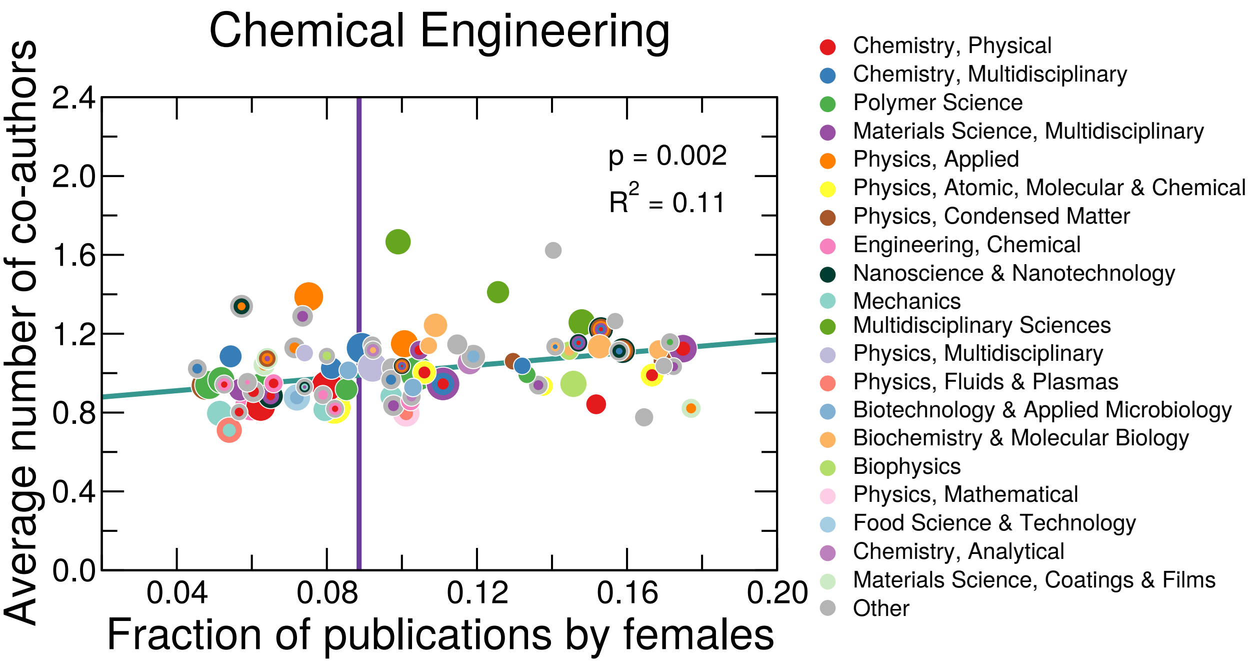

Figure C.7. **Correlation between the average number of co-authors corrected for the annual average versus the fraction of publications authored by female faculty in chemical engineering departments.** Publications are grouped by journal. We restricted the publication types to “article”, “letter”, and “note”. The size of the circle is proportional to the logarithm of the number of publications in that journal or sub-discipline. We use journal category in the *ISI Journal Citation Report* as the sub-disciplines. Journals with multiple categories are plotted as concentric rings. The purple line indicates the total average fraction of publications by females for all the publications authored by faculty in chemical engineering in our cohort, f_M. The blue line is a weighted linear regression, in which we assign to each journal a weight equal to the number of publications. We only include data points within the range of $[0.5f_M,\ 2f_M]$.

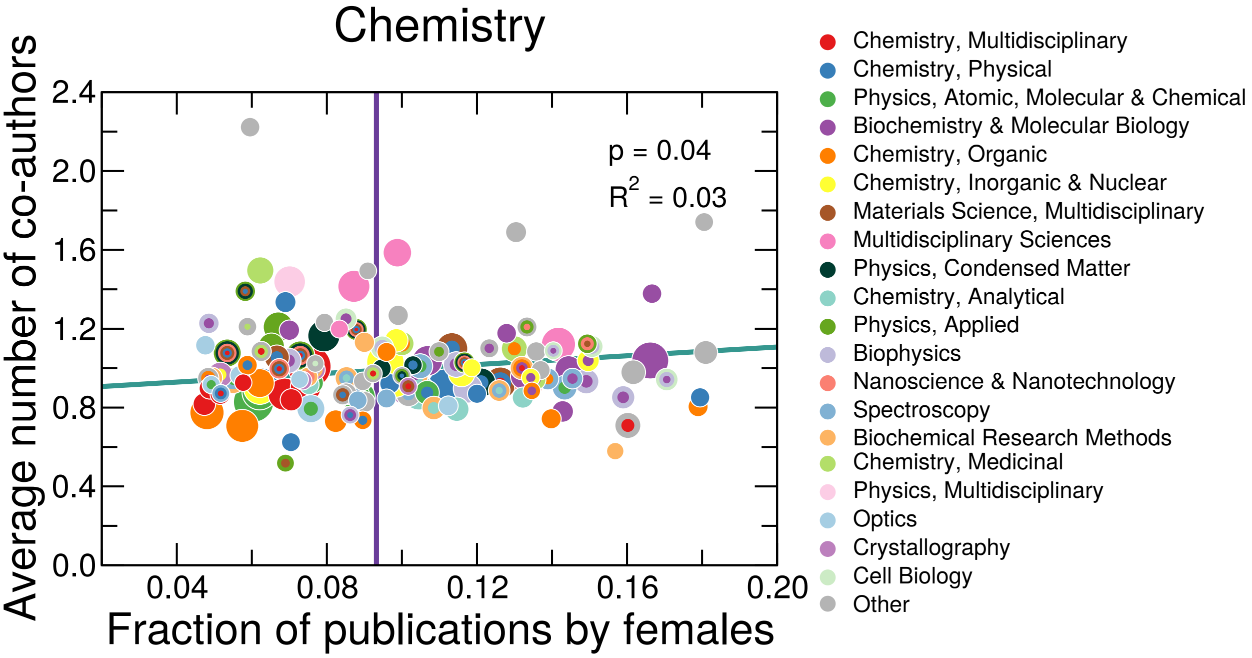

Figure C.8. **Correlation between the average number of co-authors corrected for the annual average versus the fraction of publications authored by female faculty in chemistry departments.** See the caption of Fig. C.7 for details.

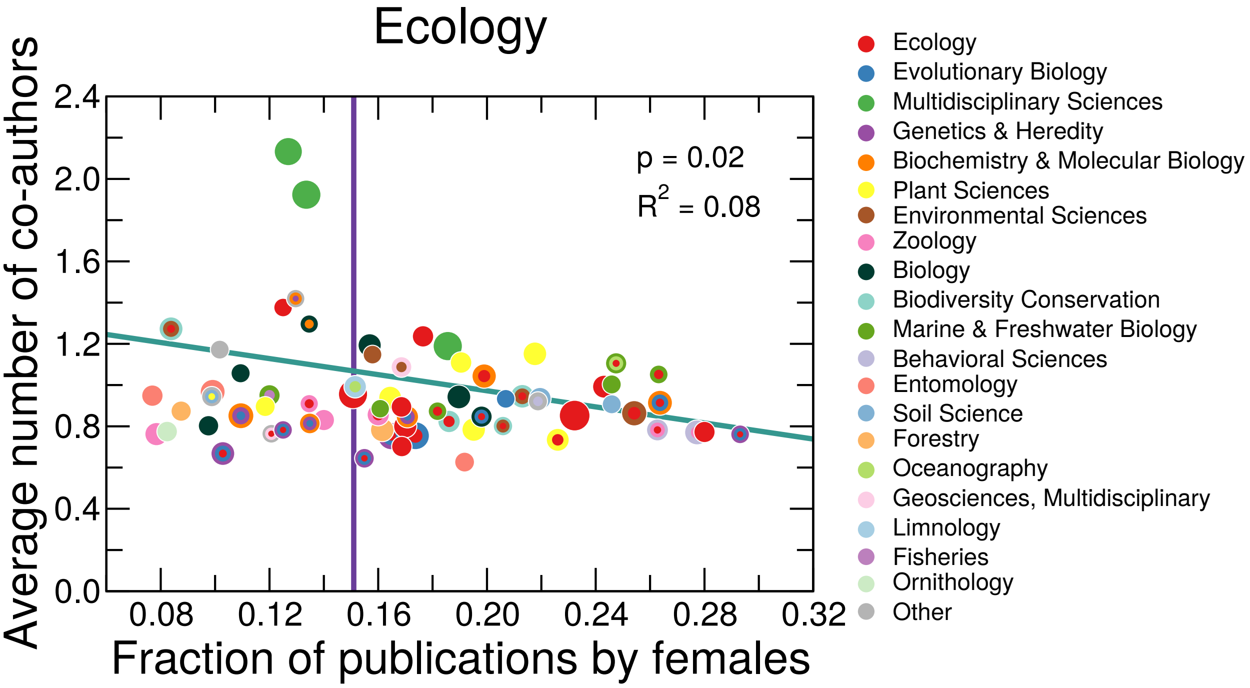

Figure C.9. **Correlation between the average number of co-authors corrected for the annual average versus the fraction of publications authored by female faculty in ecology departments.** See the caption of Fig. C.7 for details.

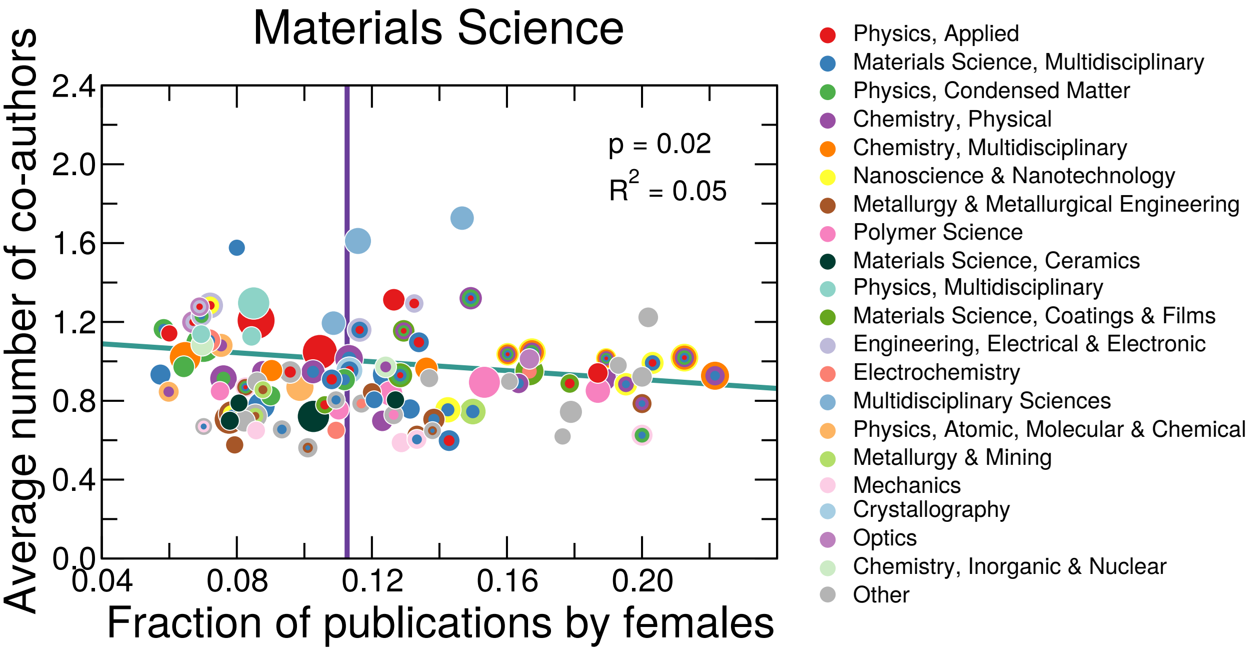

Figure C.10. **Correlation between the average number of co-authors corrected for the annual average versus the fraction of publications authored by female faculty in materials science departments.** See the caption of Fig. C.7 for details.

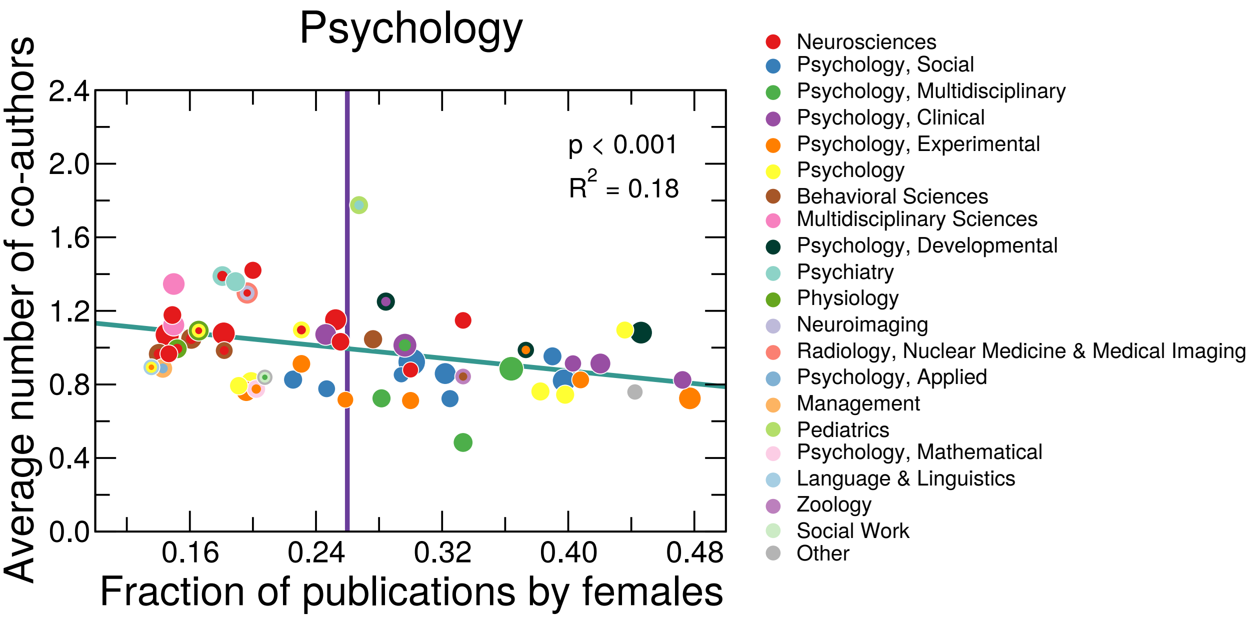

Figure C.11. **Correlation between the average number of co-authors corrected for the annual average versus the fraction of publications authored by female faculty in psychology departments.** See the caption of Fig. C.7 for details.

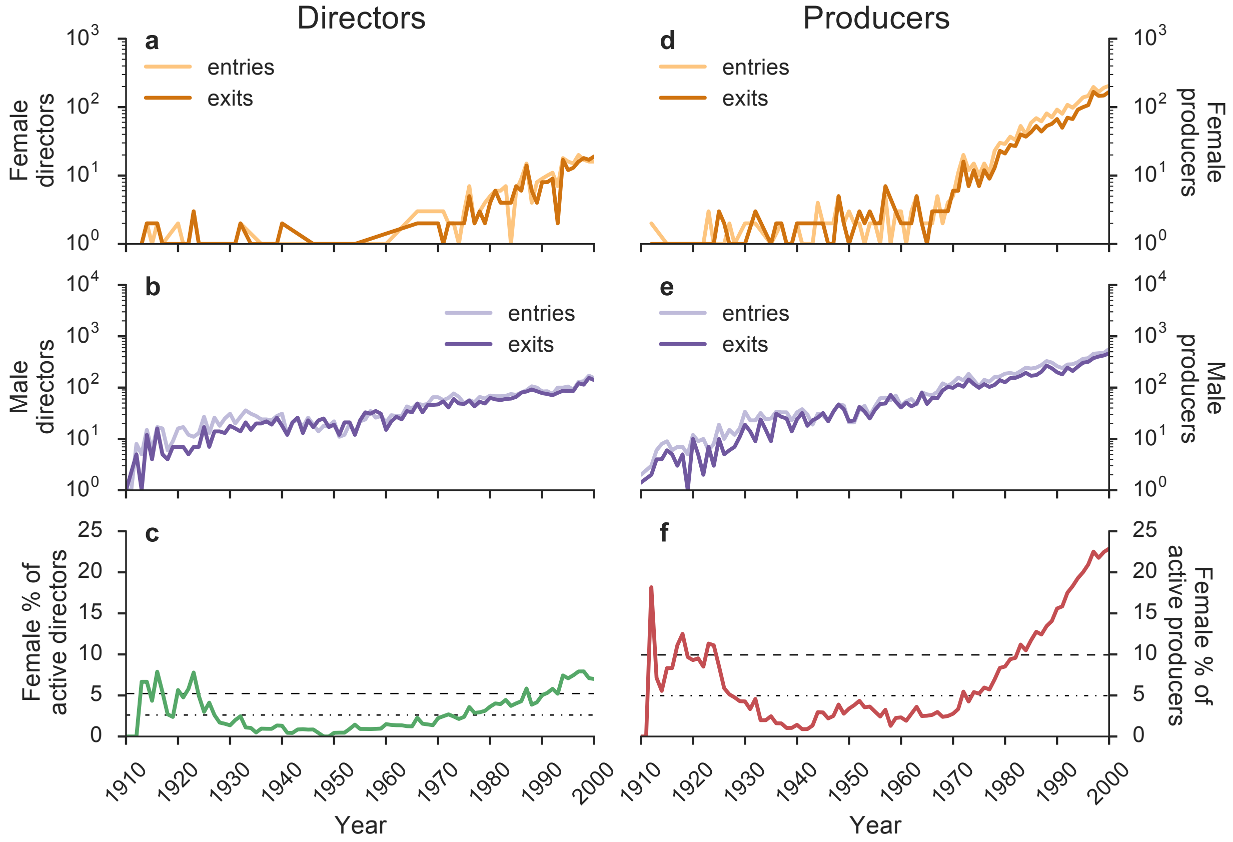

Figure C.12. **Evolution of female representation as directors and producers**. Number of (**a**) females and (**b**) males directing a movie for the first time (entry) or for the last time (exit) for U.S.-produced movies. For females, entries almost exactly balance exits. For males, between 1920 and 1940, the number of entries systematically exceeds the number of exits. (**c**) Percentage of females among active movie directors. The more equitable condition of the early 1900s (dashed line, approximately 5%) was only reached again in 1991, having remained below half of that level for 49 years (dash-dotted line). Number of (**d**) females and (**e**) males producing a movie for the first time (entry) or for the last time (exit) for U.S.-produced movies. For females, entries almost exactly balance exits until 1975, at which point entries exceed exits. For males, between 1920 and 1940, the number of entries systematically exceeds the number of exits. (**f**) Percentage of females among active movie producers. The more equitable condition of the early 1900s (dashed line, approximately 10%) was only reached again in 1983, having remained below half of that level for 44 years (dash-dotted line).

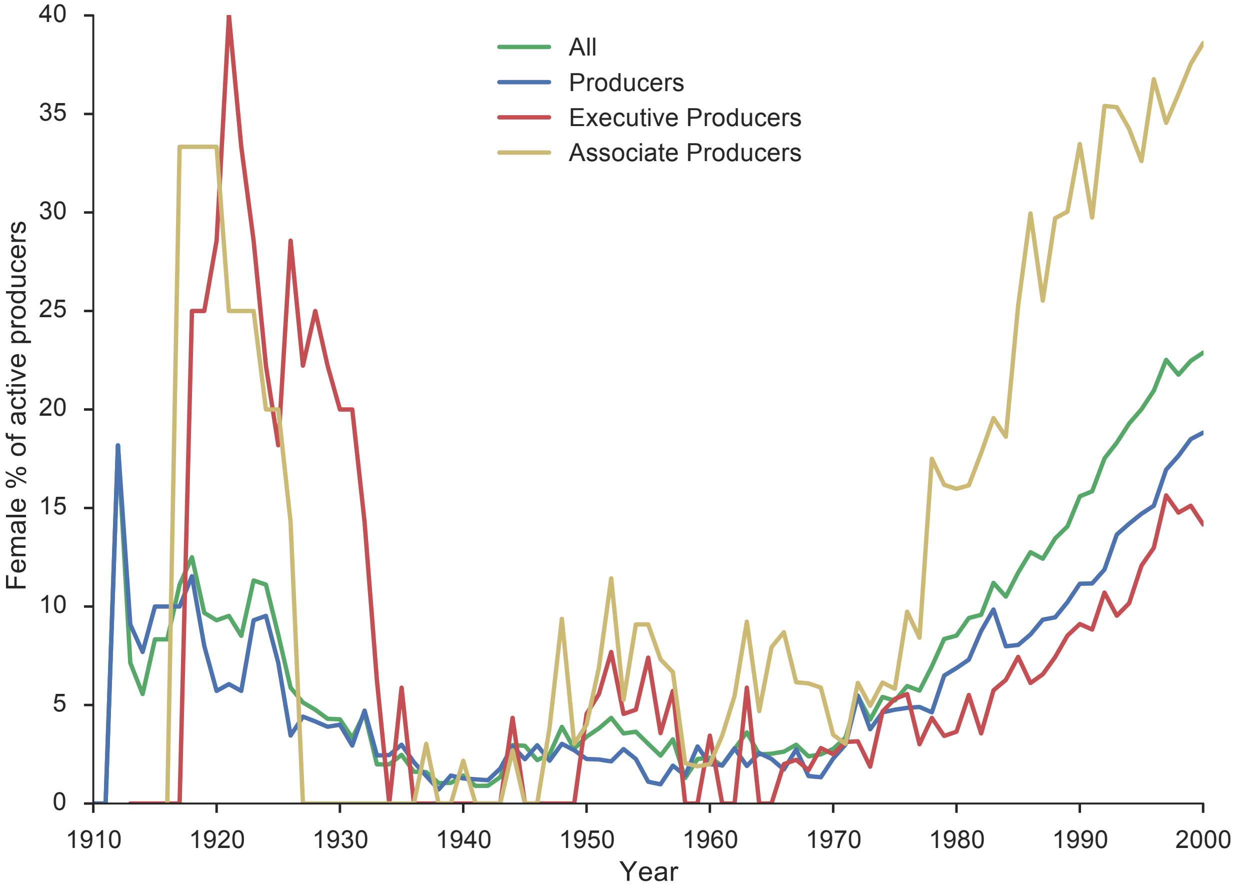

Figure C.13. **Historical trends of female representation in different producer roles**. Percentage of females among different active movie producer roles over time.

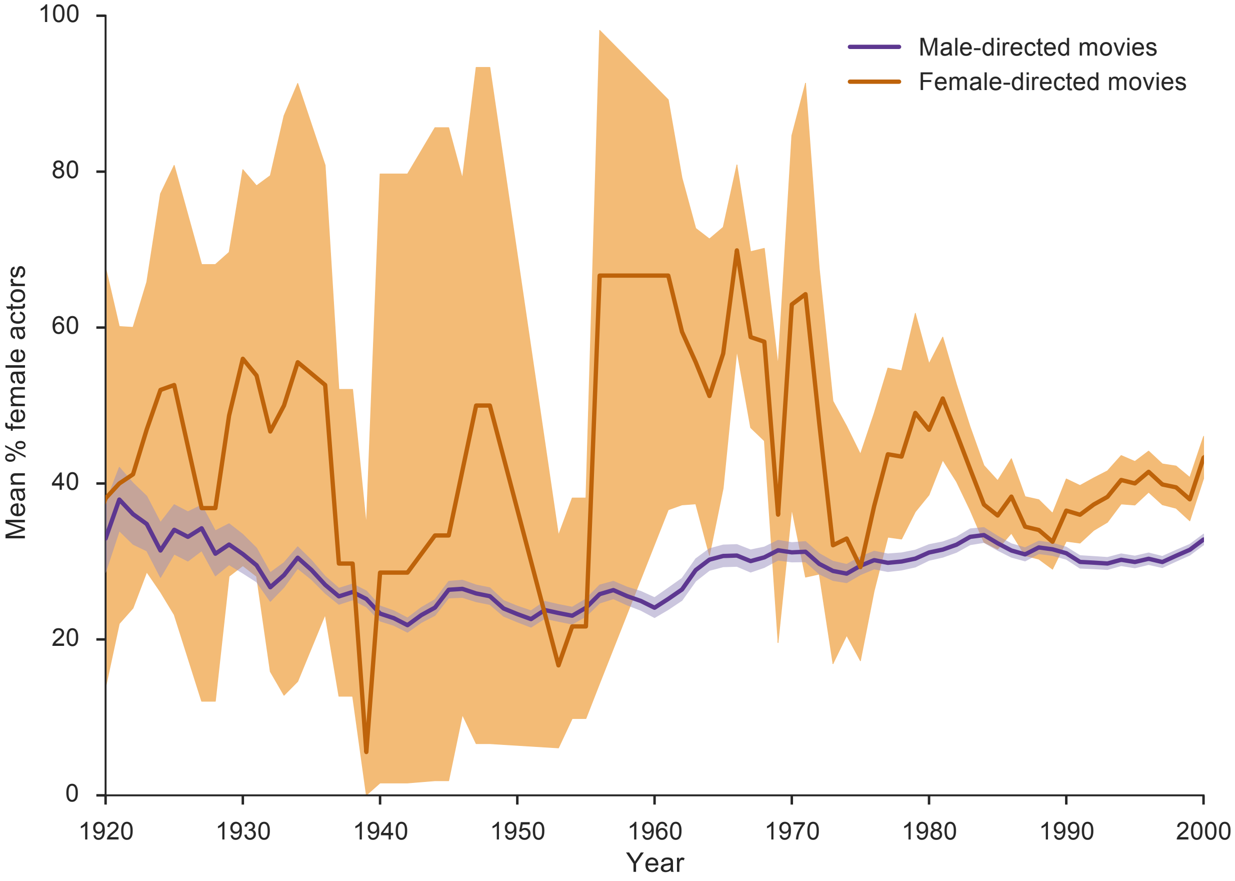

Figure C.14. **Actor preferences according to director's gender**. Historical trend of the mean percentage of female actors cast in movies directed by a female (orange) or a male (purple). Female directors have a significantly higher preference for female actors than male directors (Mann-Whitney test, $U = 2.6 \times 10^6$, $p < 0.001$). Shaded regions represent the 99% confidence bands calculated using the Clopper-Pearson [195] method under a binomial process for selecting a movie's cast (see Appendix A.2 for details).

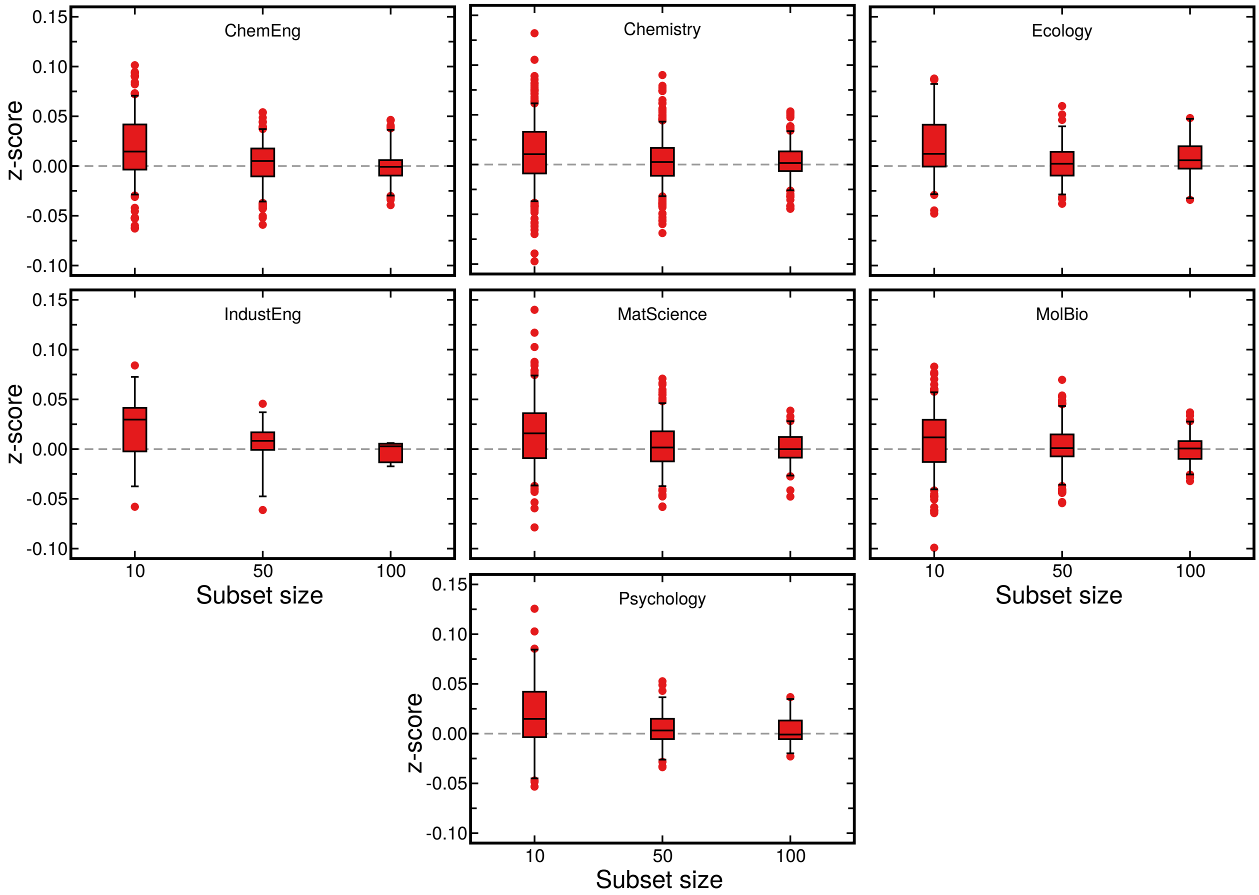

Figure C.15. **Dependence of $\hat{\mu}$ on number of publications at the individual level**. We fit the model to 1,000 randomized subsets of each researcher's publication list and compare the $\hat{\mu}$ obtained from fitting each subset of 10, 50, and 100 publications with the $\hat{\mu}$ associated with the complete publication list. Then, for each researcher and subset size, we calculate a z-score using the mean and standard deviation of the "sub-$\hat{\mu}$". For $N_p \geq 50$, the dependence on sample size is negligible for most researchers. Researchers with $N_p < 100$ are omitted from the calculation on the subset of size 100.

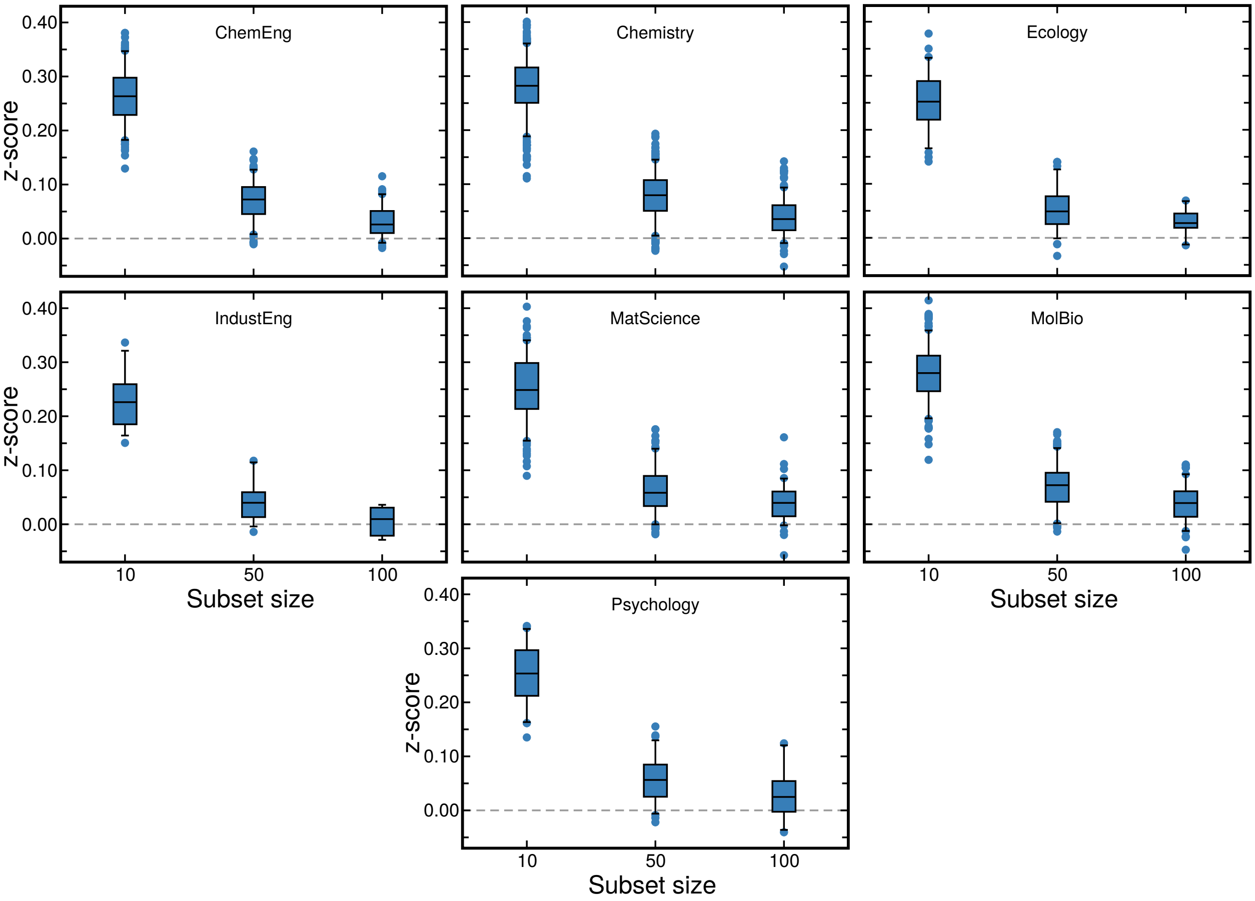

Figure C.16. **Dependence of $\hat{\sigma}$ estimates on number of publications at the individual level**. We use the same procedure as in Fig. C.15, except here we show the results for the dependence of $\hat{\sigma}$ on sample size. Estimates of $\hat{\sigma}$ are more dependent of sample size than $\hat{\mu}$. However, as in the case of $\hat{\mu}$, the dependence of $\hat{\sigma}$ on sample size decays rapidly with increasing sample size. Researchers with $N_p < 100$ are omitted from the calculation on the subset of size 100.

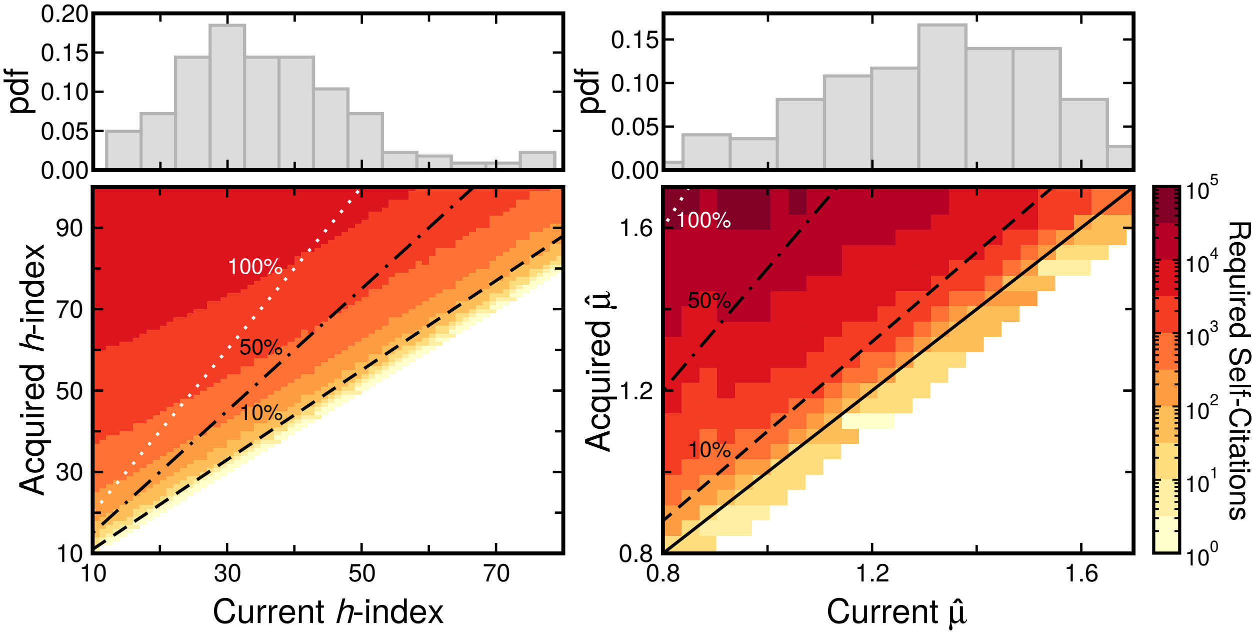

Figure C.17. **Susceptibility of impact measures to manipulation**. We used the same procedure as in Fig. 4.6, except here we show the required number of publications with self-citations that researchers need to publish in order to increase their indicators. Other details are the same as in Fig. 4.6.

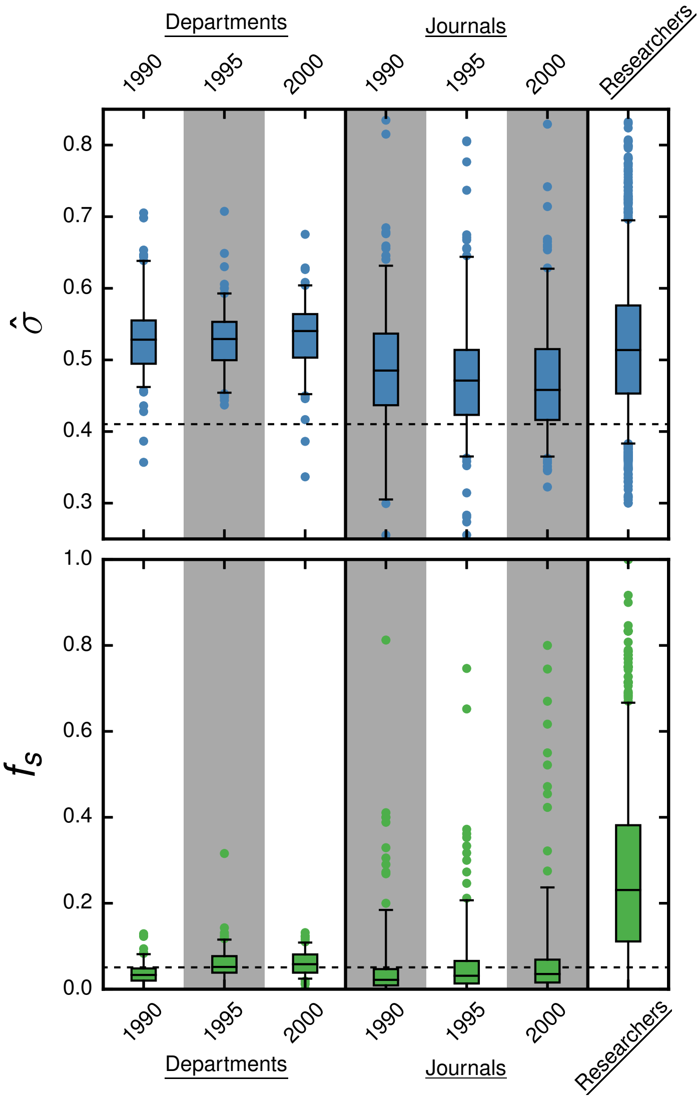

Figure C.18. **Comparison of $\hat{\sigma}$ and f_s across departments, journals, and researchers**. We show the maximum likelihood fitted $\hat{\sigma}$ (**top**) and the fraction of secondary publications (**bottom**) for chemistry departments and chemistry journals in select years, and for all chemistry researchers in our database. The black horizontal dashed lines mark the value of the corresponding parameter for the *Journal of the American Chemical Society* in 1995. For clarity, we do not show $\hat{\sigma}$ for 19 journals and 9 researchers that are outliers.