Practices, Networks and Success in Creative Careers: Study of Inequalities using Large-scale Digital Behavioural Data

Mohsen Jadidi

Institute for Web Science and Technologies
University of Koblenz–Landau
&
GESIS – Leibniz Institute for the Social Sciences
mohsen.jadidi@gmail.com

Vom Promotionsausschuss des Fachbereichs 4: Informatik der Universität Koblenz–Landau zur Verleihung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation.

PhD thesis at the University of Koblenz-Landau

Datum der wissenschaftlichen Aussprache:	
Vorsitz des Promotionsausschusses	Prof. Dr. Ralf Lämmel
Berichterstatter:	Jun.-Prof. Dr. Tobias Krämer
Berichterstatter:	Prof. Dr. Claudia Wagner

Abstract

In the last decade, policy-makers around the world have turned their attention toward the creative industry as the economic engine and significant driver of employments. Yet, the literature suggests that creative workers are one of the most vulnerable work-forces of today's economy. Because of the highly deregulated and highly individuated environment, failure or success are believed to be the byproduct of individual ability and commitment, rather than a structural or collective issue. This thesis taps into the temporal, spatial, and social resolution of digital behavioural data to show that there are indeed structural and historical issues that impact individuals' and groups' careers. To this end, this thesis offers a computational social science research framework that brings together the decades-long theoretical and empirical knowledge of inequality studies, and computational methods that deal with the complexity and scale of digital data. By taking music industry and science as use cases, this thesis starts off by proposing a novel gender detection method that exploits image search and face-detection methods. By analysing the collaboration patterns and citation networks of male and female computer scientists, it sheds lights on some of the historical biases and disadvantages that women face in their scientific career. In particular, the relation of scientific success and gender-specific collaboration patterns is assessed. To elaborate further on the temporal aspect of inequalities in scientific careers, this thesis compares the degree of vertical and horizontal inequalities among the cohorts of scientists that started their career at different point in time. Furthermore, the structural inequality in music industry is assessed by analyzing the social and cultural relations that breed from live performances and musics releases. The findings hint toward the importance of community belonging at different stages of artists' careers. This thesis also quantifies some of the underlying mechanisms and processes of inequality, such as the Matthew Effect and the Hipster Paradox, in creative careers. Finally, this thesis argues that online platforms such as Wikipedia could reflect and amplify the existing biases.

Zusammenfassung

Die Aufmerksamkeit politischer Entscheidungsträger weltweit richtet sich in den letzten 10 Jahren verstärkt auf die Kreativwirtschaft als signifikanter Wachstums- und Beschäftigungsmotor in Städten. Die Literatur zeigt jedoch, dass Kreativschaffende zu den gefährdetsten Arbeitskräften in der heutigen Wirtschaft gehören. Aufgrund des enorm deregulierten und stark individualisierten Umfelds werden Misserfolg oder Erfolg eher individuellen Fähigkeiten und Engagement zugeschrieben und strukturelle oder kollektive Aspekte vernachlässigt. Diese Arbeit widmet sich zeitlichen, räumlichen und sozialen Aspekten digitaler behavioraler Daten, um zu zeigen, dass es tatsächlich strukturelle und historische Faktoren gibt, die sich auf die Karrieren von Individuen und Gruppen auswirken. Zu diesem Zweck bietet die Arbeit einen computergestützten, sozialwissenschaftlichen Forschungsrahmen, der das theoretische und empirisches Wissen aus jahrelanger Forschung zu Ungleichheit mit computergestützten Methoden zum Umgang mit komplexen und umfangreichen digitalen Daten verbindet. Die Arbeit beginnt mit der Darlegung einer neuartigen Methode zur Geschlechtererkennung, welche sich Image Search und Gesichtserkennungsmethoden bedient. Die Analyse der kollaborativen Verhaltensweisen sowie der Zitationsnetzwerke männlicher und weiblicher Computerwissenschaftler*innen verdeutlicht einige der historischen Bias und Nachteile, welchen Frauen in ihren wissenschaftlichen Karrieren begegnen. Zur weiterführenden Elaboration der zeitlichen Aspekte von Ungleichheit, wird der Anteil vertikaler und horizontaler Ungleichheit in unterschiedlichen Kohorten von Wissenschaftler*innen untersucht, die ihre Karriere zu unterschiedlichen Zeitpunkten begonnen haben. Im Weiteren werden einige der zugrunde liegenden Mechanismen und Prozesse von Ungleichheit in kreativen Berufen analysiert, wie der Matthew-Effekt und das Hipster-Paradoxon. Schließlich zeigt diese Arbeit auf, dass Online-Plattformen wie Wikipedia bestehenden Bias reflektieren sowie verstärken können.

To my family.

Acknowledgments

Starting a PhD is a personal decision, finishing it is a collective effort. I have been privileged to have people around me who supported me in this journey, emotionally and intellectually. People who generously shared their knowledge, expertise and experiences. Those whose words and presence empowered me to take on challenges and go beyond my own limits.

First I would like to thank my supervisor, Claudia Wagner, who has been guiding me throughout my research. This work would have not been possible without her encouragement, valuable feedback, inspiring discussions, and patience. I would also like to thanks Tobias Krämer for his valuable support in the final stage of my PhD.

My sincerely gratitude goes to friends and co-authors who contributed directly in this research. I am particular thankful to Markus Strohmaier for his trust in accepting me to the program. Special thanks must go to Haiko Leitz for invaluable hours of discussions and brainstorming, and foremost introducing me to the fascinating world of sociology; Fariba Karimi from whom I learned greatly about complex network theory; Mattia Samory for his straightforward, precise and timely advice in challenging times.

I would also like to thank my friends and colleagues at GESIS, specially those in the computational social science department; David Brodesser and Arnim Bleier for helping me to stay motivated and keep a balanced head when my mind insisted otherwise; Maria Zens and Andreas Schmitz for their invaluable feedback on my research, and stimulating conversations about the "Field theory"; Fabian Flöck who helped me alot with bureaucratic matters when I first moved in Cologne; My fellow PhD students Lisette Espin, Julian Kohne, Lisa Posch, Anna Samoilenko, Indira sen and Olga Zagovora who never hesitated to offer their support. Special thanks must go to the Leibniz and GESIS PhD Networks for their seminars and workshops offerings.

My deepest gratitude go to my family for the unlimited love and unconditional support. Thanks for believing in me and supporting my decisions. Without you I could not be even close to where I am right now.

I feel extremely lucky to have a caring and supportive partner next to me, Lisa. Thanks for joining me in this crazy rollercoaster ride. Thanks

for all the feedback and discussions on my work along the way. This would not have been possible without your encouragement, unending patience, and love.

MOHSEN JADIDI

mohsen.jadidi@gmail.com

Cologne, Germany

EXPERIENCE (LAST 10 YEARS)

Co-Founder & Director
02.2019 - present

Timcheh e.V. , Cologne

- Post-PhD sabbatical project
- Initiated and lead a non-profit organization
- Planning & coordinating projects, and building collaborations among communities

Scientific Coordinator
05.2021 - 11.2021

Leibniz Institute for Social Sciences, Knowledge Exchange and Outreach Dep., Training Team

- Developed and planned the roadmap for the data science trainings of the department
- Planned and coordinated a 6-week online data science summer school with 200 participants

Research Associate
02.2015 - 11.2021

GESIS - Leibniz Institute for Social Sciences, Cologne
Computational Social Science Dep., Data Science Team

- Led and collaborated on researches that inform the EU policies on online platforms and social media
- Built and maintained infrastructure to collect digital behavioral data from online platforms
- Participated on writing funding applications
- Actively contributed to strategic development of department e.g., defining research direction and building collaboration with researchers and institutions

Guest Researcher
09.2018 - 01.2019

Telefónica R&D, Barcelona

- Conceptualized and Initiated a location intelligence app based on the use of company's internal data and social media data
- Carried out spatial data analysis to quantify the movement patterns of cellphone users in a city

Junior Data Scientist
10.2013 - 12.2014

GESIS - Leibniz Institute for Social Sciences, Cologne
Knowledge Discovery Team

- Developed a method to detect popular topics and their evolution on Twitter. The method also clusters Twitter users based on their interest to similar topics

EDUCATION

Ph.D., Computer Science
| 2022

University of Koblenz-Landau
Institute for Web Science and Technologies

Thesis: Practices, Networks and Success in Creative Careers: Study of Inequalities using Large-scale Digital Behavioural Data

M.Sc., Computer Science
| 2015

University of Bonn
Intelligent Analysis and Information Systems

B.Sc., Software Engineering
| 2011

Azad University of Tehran
Computer Engineering

SKILLS

Python, R, Jupyter, HTML, D3.js, SQL, MongoDB, Amazon EC2, Linux system administration (Ubuntu), Windows server administration

web crawling, research design, hypothesis testing, statistical analysis, data analysis, machine learning

scientific writing, grant writing

English (C2), German (B2), Farsi (Native)

VOLUNTEER EXPERIENCE

Event Staff
| 05.2019

https://think-about.io/

Event Staff
| 18.2016

https://www.icwsm.org

Web Developer
| 07.2017

https://www.ic2s2.org/

PUBLICATIONS

The Hipster Paradox in Electronic Dance Music: How Musicians Trade Mainstream Success Off Against Alternative Status. Mohsen Jadidi, Haiko Lietz, Mattia Samory and Claudia Wagner. The International AAAI Conference on Web and Social Media. 2022. Atlanta, Georgia, USA.

Gender Disparities in Science? Dropout, Productivity, Collaborations and Success of Male and Female Computer Scientists. Mohsen Jadidi, Fariba Karimi, Haiko Lietz and Claudia Wagner. The Journal of Advances in Complex Systems, 2017.

It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. International AAAI Conference on Web and Social Media. 2015. University of Oxford, Oxford, UK.

Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods. Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi and Markus Strohmaier. International World Wide Web Conference. 2016. Montreal, Canada.

(Don‘t) Mention the War: A Comparison of Wikipedia and Britannica Articles on National Histories. Anna Samoilenko, Florian Lemmerich, Maria Zens, Mohsen Jadidi, Mathieu Génois and Markus Strohmaier , International Conference on World Wide Web. 2018. Lyon, France.

TRAININGS

Foundations of User Experience (UX) Design | 2022
Coursera + Google

Essential Time- and Self-Management Skills | 2020
Dr. Anna Maria Beck, Cologne

Causal Analysis | 2017
GESIS Spring Seminar

Complex Networks: Theory, Methods, and Applications | 2017
Lake Como School of Advanced Studies

Lisbon Machine Learning School (LxMLS) | 2013
Instituto Superior Técnico

Contents

Chapter 1

Introduction

1.1 Motivation

Society is a stratification system based on the hierarchy of positions identified by power, status, and resources [Par40]. Individuals are positioned at various levels in the hierarchy based on their socioeconomic status, gender, race, and other personal and cultural characteristics. This differentiation system manifests itself as the unequal distribution of rewards and opportunities such as income and wealth, unequal access to education, cultural, and digital resources and different treatment by juridical systems. Such inequalities can have significant consequences on our society: poverty [CH04], crime [Bou09], conflicts and social unrest [Ste05], happiness and health in society [GP02, Lut05], and economic growth. The significance of this issue is also reflected in the United Nations's 2030 Sustainable Development Goals (SDG) in which addressing "gender equality" and "reduced inequality" among and within countries are among the top 17 priorities to achieve a better future.

Historically, inequality has been largely studied from an economic point of view. These studies are mainly concerned with the distribution of material entities, in particular income and wealth, over individuals and households. From this point of view, known as *inequality of outcome*, the condition of a just society is to have an equal outcome (i.e., income, wealth) for individuals [LPT08]. This view not only overlooks the diversity of preferences and tastes, but also denies the importance of individual responsibility and choice [Phi04]. Inequality is a complex issue that extends over other areas and levels of society. A fair society is based not only on the equal outcome for individuals but also on providing a fair condition for its members "... to lead the kind of lives they value–and have reason to value" [Sen01]. In this proposition, the goal should be to prevent and dispel *inequality of opportunities* in which specific individuals and groups face consistently inferior opportunities — economic, political, cultural, and social — than others. In

another word, individuals cannot hold responsible for circumstances beyond their control: their race, sex, urban, and rural location [RT15].

In her seminal work, Frances Stewart [Ste05] emphasizes the role of culture, and more specifically, cultural groups, in the unequal treatment of individuals and their access to resources. She identifies two types of inequalities. *Vertical Inequality* (VI) "which lines individuals or households up vertically and measures inequality over the range of individuals". From this perspective, inequalities among individuals arise from their differences in terms of human capital, skills, talent, social connections, and initial conditions of life [BT79]. *Horizontal Inequality* arises from the belonging of individuals to different social and cultural groups. She argues that the contemporary discourse about inequality has put the individuals on the center of analysis and neglects the critical "group" dimension of inequality. Group membership, she argues, is a primary need of human life that makes up for individuals' identity (or identities), as well as gender, age, ethnic, religious, racial, or regional affiliations. These identities shape individuals' behavior, their social interactions, and how they are perceived and treated by others. Therefore, horizontal or between-groups inequalities, may be the cause of prejudice, discrimination, marginalization, or other types of disadvantages, and must be accounted for.

One of the domains that horizontal and vertical inequalities are most prevalent in, and have destructive impacts on individuals' and groups' well-being is the career outcome. Jobs are more than just a source of income; they often become a core aspect of identity [DRB10], enabling the development of new skills, and the forging of enduring attachment [Hal02]. Additionally, career success such as high income or high status, is associated with a higher level of job satisfaction, and consequently, the well-being of individuals [DS04, Hal02]. A multitude of studies has investigated how variables such as socio-demographics (e.g., gender), human capital (e.g., education level), organizational sponsorship (e.g., mentorship), personality (e.g., cognitive ability), and personal relations [NESF05] are empirically related to subsequent career success. These studies highlight the role of individual investment in one's careers and the impact of structural and cultural conditions that help or hinder the access of individuals and groups to certain types of resources or "capitals". Inequality in access to resources or opportunities, leads to inequality of career outcome among individuals or groups.

For decades researchers have been using qualitative methods such as surveys and in-depth interviews to understand the conditions and dynamics of career developments. While they are useful devices to gain precise and valuable insight into individuals' attitudes, opinions, and personal history, they also pose critical limitations. The subjective nature of these methods, for example the reactivity of the researcher with participants (and vice versa), makes it difficult to maintain objectivity and avoid bias [Nor97, GL10]. Furthermore, the expensive and tedious data collection not only demand a

greater effort to obtain information, but can also call their practicality into question [SB07].

The digitalization of society and the emergence of computational social science [LPA+09] has opened up new perspectives and approaches to overrun these limitations. The online space has become a new dimension of economic, political, and social participation, which creates a new way of working, socializing, and knowledge consumption. Through interactions, people leave behind an unprecedented amount of digital footprints that can be used to trace individual-level behavior in offline and online spaces, in real-time, across cultures, and on a population scale [GM11]. These digital footprints are not only the product of *intentional* content production of users (i.e., active footprints), but may also be compiled and generated by other users or algorithms (i.e., passive footprints) [MLB18]. They enable us to go back in time to study the origins of phenomena [HBG04], measure things that we could not measure before [NSL+12], and revisit old theories to test their validity [STFMC11]. Now, for the first time, we have the ability to collect rich relational data of social interactions on a global scale. To obtain information not only about the structure of these relationships but also their content.

The online space has also become a space where new forms of inequality arise, and the existing ones reproduce. Attention is arguably the most valuable currency in the online space. Those who accumulate enough attention increase their chance to be seen and recognised for their work, and therefore their chance of success [CZW15]. Furthermore, online platforms have become the dominant source of information with the far-reaching potential to influence our individual and collective perception. The content bias in these platforms, such as under- or misrepresentation of certain groups, can have significant consequences. For example, they can produce new and/or amplify the existing stereotypes about certain demographics and impair their career development through the glass ceiling effect [CHOV01, RSZ14], or sway our collective memory toward certain narratives that ignore certain perspectives [FM12, GGMTY17].

This thesis taps into the potential of digital trace data and interdisciplinary research frameworks to improve our understanding of career inequalities in the creative industries. Creative industry is broadly defined as "those activities which have their origin in individual creativity, skill and talent and which have a potential for wealth and job creation through the generation and exploitation of intellectual property" [Cre01]. The function of the creative class is to create new ideas and technologies in different areas such as science and engineering, architecture and design, education, arts, music, and entertainment [Flo14b]. Their activities are associated with the economic boom of post-industrial cities, significant drivers of employment, and facilitators of urban regeneration. Policy makers around the world have turned toward the creative industry as the driver of innovations and catalyst

of economic growth. A recent report shows that the economic contribution of creative industries in Europa is greater than those of telecommunications, high technology, pharmaceuticals or the automotive industry. This includes a 10% increase in the share of employment since 2013 [1].

The "high bohemian" persona of creative culture [Flo14b], together with the idea that individual talent and creativity are the main assets and the driving force behind the creative industry have lead to an optimistic believe that it should be prone to discrimination and unfair treatment of individuals. After all, talent and creativity are "everyone's natural asset to exploit" [Ros09]. However, looking at the specific conditions of work and employment in the creative industry, it becomes obvious that the promise of "full opportunity and unfettered social mobility for all" [Flo14b] is far from a reality. Rather, creative workers have become one of the most vulnerable work-forces of the "new economy" – the economy that benefits from globalization of business and the revolution in information technology and communication technology (ICT) [Poh02]. Network-based recruitment, low or unwaged entry-level jobs, temporal contracts, high degree of geographical mobility and high market risk are just a few examples of working conditions that lead to exclusion and marginalization of groups and individuals [EW13]. For example, the reliance on relationships and recommendations to secure contracts implies that access to industry networks is crucial for survival [Wit01a]. At the same time, it is well known that access of minority groups to such networks is usually impeded by cultural, political and social barriers [Bou83, Ste05]. For example studies on film industry [BCR03, SCP17, Lut15, WLL19], music industry [GT21, Cit00, Mor19, BLM07], and academia [LNG+13, SKS05] show that minorities in these fields, such as women or people of color, suffer from different sorts of discrimination and prejudice over the coarse of their careers.

While people are "free" to fullfill their creative dreams, they have to "market" their products in a deregulated and highly individuated environment. The ideas of freedom, independence, and self-actualization lend themselves to the intensive practices of self-monitoring, self-marketing, and often self-exploiting. Individuals need to take the role of entrepreneurs and devise the right strategies to signal their talent and creativity, in order to obtain recognition, reputation, legitimacy and consequently to achieve success. At the same time, instead of operating within organisations with clear structure, pre-defined rules and expectations, "people become their own micro-structures [...] and do the work of the structures by themselves" [McR02]. What is considered as high value and legitimate is decided upon the evaluations, negotiations and interactions of actors involved. In other words, the underlying structure is emerged from the collective behaviour of participants. The power structure becomes hidden, and therefore it becomes more

[1] `https://www.rebuilding-europe.eu/`

difficult to identify and hold stack-holders responsible for their decisions and actions. Measuring and understanding the role and impacts of this informal, dynamic and often hidden structure demands analyzing a spectrum of practices that make up the creative careers. This thesis leverages the scale, granularity, and the depths of digital behavioural data, and the advances of computational methods to investigate this complex nature of creative careers.

Using the temporal, spatial, and social resolution of digital data, this thesis investigates the various aspects of career development and sheds light on the extent to which inequality manifests, evolves and impacts careers. Computational methods are used to collect, match and analyse large scale datasets from diverse online sources. Social science theories are used to inform different steps of this research, starting from data collection and research design to interpreting the results and discussing the findings. Notably, this thesis uses "network" as a common language between social sciences and computational sciences. While social scientists particularly study networks as phenomena ("structuralist interpretation") [Eri13], both groups use networks as an apparatus to assess and characterize natural or human phenomena. However, many traditional algorithms are often too complex and unable to deal with large networks. Computational scientists propose and use algorithms that are able to deal with the scale and granularity of digital behavioural data. The shift of the two disciplines toward each other could provide new opportunities to inquiry into social phenomena such as inequality at scale. Hence, this thesis borrows concepts from complex network theories and methods from social network analysis as intermediaries to connect theory and empiry.

1.2 Problem Statement, Objectives, and Approach

This section introduces the main problem statement of this thesis. Furthermore, the objectives and general approach to tackling the challenges of the main problem are presented.

Problem statement. The creative industry is one of the world's most rapidly growing economic sectors. This growth includes the number of people who seek to build a career within this industry. Yet, the existing production models and working conditions put certain groups and individuals in disadvantageous positions. While some enjoy excess access to social, economic, and cultural resources, others have to struggle to maintain their career. Previous studies suggest that it is the interdependence of social (i.e., social transactions), cultural (i.e., norms and values), and individual processes (e.g., creativity) that limit or empower personal and career development. For example, cultural hostility toward minority groups pushes them away from the preeminent social chambers that generate and control

a wealth of information, knowledge, and social supports. Despite considerable attention by policymakers and various scholars, our understanding of the extent and consequences of such inequality is still limited to specific demographics, time, dimensions, and processes. We still do not fully know nor understand the variety of ways in which individuals and groups develop their careers, what influences such decisions, and how different strategies lead to different outcomes. This is specially challenging in creative careers which lack the traditional formal organization structures.

Objectives. The main objectives of this thesis are to (i) make progress towards understanding the extent, dimension, and mechanisms of inequality in creative careers, (ii) collect large-scale, multi-faceted, and historical digital behavioral data that can support future research, and (iii) propose an interdisciplinary research framework that brings together theoretical insight from social sciences and methodological tools from computational sciences. Of particular interest is how to characterize individuals and groups' careers with respect to their access to available resources within a social system and the pattern of successful careers. Furthermore, this thesis aims to take a step toward understanding the ways online platforms can reproduce and enforce existing biases that could potentially harm specific demographics.

General approach. To reach these objectives, this thesis follows a data-driven approach guided by sociological theories. I collect and analyse data from a variety of online platforms to characterize creative careers and the socio-cultural systems in which they take place. I leverage the theoretical depth and analytical power of "networks" to draw a line between career practices, socio-cultural structure, opportunity, and success. Networks exist as a pattern of ties that capture the relationship between different entities. "Nodes" represent the individuals and "edges" their relationship with respect to a form of capital. Through interactions, individuals possess or exchange resources and form relations with one another. These relations build the structure of the social space in which individuals operate, and can be analyzed using social network analysis [BC11a]. Those who exhibit a similar configuration of ties are believed to occupy similar positions in the social or cultural structure of a system. Positions are identified by a combination of numbers of node-specific (or egocentric) properties. From this perspective, a career is defined as series of positions successively occupied in the successive state of a system. Moreover, I propose, measure and evaluate indices to measure success from an objective point of view. A form of success that is measurable and identifiable by third parties rather than individuals' evaluations of their own career. By measuring the position and success of individuals at each point in time, I construct longitudinal data that is used to identify the patterns of successful careers. Here, I look at success not only as an instance of vertical inequality that is shaped by individual practices, but also as a byproduct of other forms of inequality. For this purpose I use

fixed-effect and mixed-effect regression analysis that accounts for the dependency between observations. I perform a number of studies that describe the extent and evolution of inequality along multiple dimensions. This thesis mainly focuses on careers in music and academic disciplines.

1.3 Research Questions

Aligned with the objectives above, this thesis is built around the following three general research questions to investigate the *dimensions* (RQ1), *evolution* (RQ2) and *mechanisms* (RQ3) of inequality within creative careers:

RQ1 What are the dimensions of inequality in creative careers?

RQ2 How do they evolve over time?

RQ3 What are the underlying mechanisms and processes of inequalities in creative careers?

Each study touches upon these questions and provides a unique insight into the various dimensions of creative careers.

1.4 Main Publications

The core chapters of this thesis are based on results from the following publications:

- Article 1 [KWL+16a]: Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Strohmaier, Markus.Inferring gender from names on the web: A comparative evaluation of gender detection methods. *Proceedings of the 25th International Conference Companion on World Wide Web*. 2016. 10.1145/2872518.2889385.
- Article 2 [JKLW18b]: Mohsen Jadidi, Fariba Karimi, Haiko Lietz, and Claudia Wagner. Gender disparities in science? Dropout, productivity, collaborations and success of male and female computer scientists. *Advances in Complex Systems*. 2017. 10.1142/S02195259175001141.
- Article 3 [JKTWng]: Haiko Lietz, Mohsen Jadidi, Daniel Kostic, Milena Tsvetkova, and Claudia Wagner. The Matthew Effect in computer science: A career study of cohorts from 1970 to 2000. *Under review*. 2021.
- Article 4 [JLMW21]: Mohsen Jadidi, Haiko Lietz, Mattia Samory, and Claudia Wagner. The Hipster Paradox in Electronic Dance Music: How Musicians Trade Mainstream Success Off Against Alternative Status. *AAAI Conference on Web and Social Media*. 2021.

- Article 5 [WGJS15]. Claudia Wagner, David Garcia, Mohsen Jadidi, and, Markus Strohmaier, and Claudia Wagner. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *AAAI Conference on Web and Social Media*. 2015.

Additionally, the following publications contributed to formulating the basic ideas of this thesis.

- Article 6 [SLZ+18]: Anna Samoilenko, Florian Lemmerich, Maria Zens, Mohsen Jadidi, Mathieu Génois, and Markus Strohmaier. 2018. (Don't) Mention the War: A Comparison of Wikipedia and Britannica Articles on National Histories. In *Proceedings of the 27th International Conference Companion on World Wide Web*. 2018

1.5 Main Contributions

The main contributions of this thesis are summarized as follows:

1. First, this work introduces a Computational Social Science (CSS) framework that allows for theoretical and empirical investigation of inequality in creative careers 1.1

2. Second, this thesis empirically shows that online platforms are rich sources of relational data for investigating the opportunity structure within the creative industry.

3. Third, this thesis provides empirical evidence on dimensions and mechanisms of inequality and how they impact groups' and individuals' success.

4. Finally, this thesis offers a collection of novel datasets that support further research into the working mechanisms and conditions of creative careers.

Figure 1.1 shows the scheme of the CSS framework that constitutes the theoretical and methodological backbone of this thesis. It is divided into two main parts. The theory part is to leverage the decades-long theoretical and empirical knowledge of inequality studies for identification, conceptualization, and partially, the operationalization of research questions. It helps us to understand the context of the study, identify the essential concepts, and choose the correct measurement indicators. I start my inquiries with two questions that form the basis of previous studies, namely, *inequality among whom?* and *inequality of what?*. The first question concerns the subject of analysis by relying on two related concepts: 1) *Vertical Inequality* (VI) – inequality among individuals based on individuals' capacities and effort, and

2) *Horizontal Inequality* (HI) – inequality among culturally defined groups such as race and gender groups. The second question aims to investigate the dimensions and mechanisms of inequalities through two mutually inclusive perspectives: *inequality of outcome* and *inequality of opportunities*. While the first view quantifies inequality in terms of overall disparity in economic conditions (e.g., income, status), the latter takes a more profound and broader view on inequality by focusing on exogenous "circumstances" that shape individuals' opportunities to pursue their desired life plans.

The second part of the framework informs the choice of the methodological approach and data collection procedures. The *distributive approach* uses statistical methods to assess the distribution of resources over groups or individuals in a social system. The *relational approach* rests on the notion that social relations are the major determinant of life chances. Here, researchers use relational data, such as social interactions and affiliations, to measure the social positions and consequently the level of opportunity of groups and individuals in a socio-cultural system.

Moreover, the choice of data sources are divided into two groups. *Traditional data* sources have been used by social scientists and economists in the last decades - this includes survey data, or observation data from ethnographic studies and lab studies. *Digital behavioural data* has emerged as a result of technological advances and their widespread use. Each data source exhibits certain advantages and disadvantages that influence the scope of studies.

Finally, the framework also offers to extend our investigations to the *online space* as a new domain of social, cultural, and political interactions. Here, we can identify new instances and mechanisms of inequality and examine how the new socio-technical systems mirror or amplify the existing inequalities of the *offline space*.

1.6 Structure of this Thesis

The remainder of this thesis is organized as follows:

The Part I of this thesis starts by assessment of inequality in scientific careers. Here I focus on computer science discipline because of its collaborative nature, its long-standing issue of gender bias, its ongoing transformation, and a driver of the digital revolution.

Chapter 2 proposes a novel method that outperforms the existing gender detection method across heterogeneous sub-populations by augmenting traditional methods with face recognition techniques. I evaluate and compare frequently used name-based gender detection methods based on their overall accuracy and biases when used for names from different ethnic groups.

Being able to infer the gender of individuals, I investigate the extent,

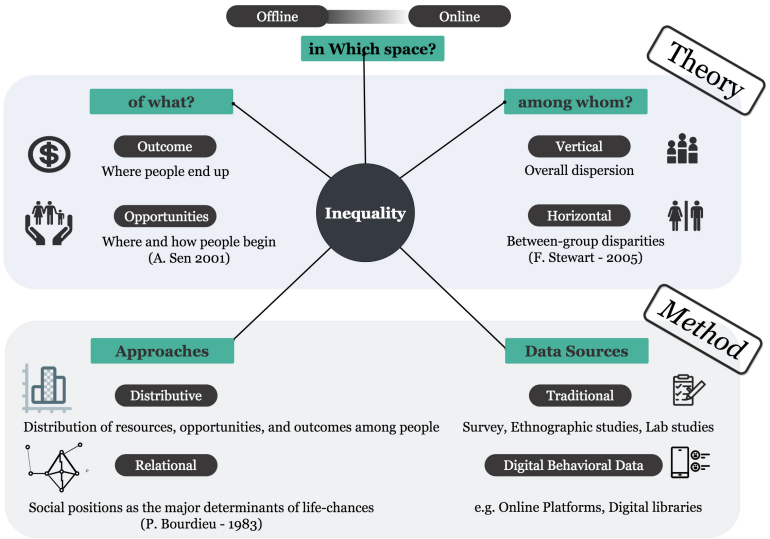


Figure 1.1: Scheme of the Computational Social Science framework to study inequality that is proposed by the author

dimensions, and evolution of HI and VI in the entire field of computer science. In **Chapter 3**, I identify multiple dimensions of the gender gap, show how they evolve, and discuss how they could potentially harm the career of women. By comparing the co-authorship network of male and female scientists, I quantify the structural disadvantage of women and identify which collaboration patterns are related to scientific success. Next, in **Chapter 4**, I examine different aspects of vertical (i.e., productivity and recognition) and horizontal (i.e., gender gap) inequalities across, and within cohorts of scientists that started their career in a particular year. I assess the predictive power, hence the impact, of numbers of meritocratic and non-meritocratic characteristics of scientists in their early-career, on their future success and dropout.

In Part II I shift my focus toward the music industry and in particular, the club culture of Electronic Dance Music (EDM). Similar to science, every musical genre and subculture may favour certain behaviour and practices based on its socio-cultural history and characteristics. Compared to many other music genres, EDM is a relatively young subculture that is witnessing a great cultural and social transformation in the last decade. Moving from the counter-culture movement to the center of mainstream culture, it has become one of the biggest pool of money and talents within music industry. **Chapter 5** follows a formal sociological approach based on bipartite networks to study one of the underlying mechanisms of success in this field – the hipster paradox – using digital traces of performing live and releasing music. I show different types of career trajectories of EDM artists, and quantify career patterns that are associated with success.

Finally, in **chapter 6** in part III I look at social media as a new space in which inequality could manifest itself. I take Wikipedia as use case and ask if successful men and women – those who are recognized for their achievements – receive equal treatment and attention by the Wikipedia community. Here I show that the gender gap in Wikipedia exists along multiple dimensions within six language editions.

1.1 provides an overview of the main chapters of this thesis. Each chapter is summarized by showing its relationship with the presented research questions, main publications, data, and methods utilized to achieve the described goals.

This thesis concludes with **Chapter 7** where I summarize the main results and contributions of each chapter. Moreover, I discuss the important implications of my findings on real-world applications. Finally, I provide an overview of limitations and future directions that can guide future researches.

Table 1.1: **Thesis Outline**: This table summarizes the main chapters of this thesis. Each chapter is based on a publication that answers a particular research question (RQ).

Chapter	Publication	RQ	Contributions	Space	Among whom?	of What?	Approach	Data
Chapter 2	[KWL+16a]		• offer a novel gender detection approach for a large and heterogeneous population	-	-	-	-	names and images of people
Chapter 3	[JKLW18b]	RQ1,2,3	• offer empirical evidence of multiple dimensions of VI & HI and their evolution in computer science • identify gender-specifics pattern of successful careers	Offline	HI: male and female scientists	• Outcome • Opportunity	• Distributive • Relational	historical records of • Publications • Citations
Chapter 4	[JKTWng]	RQ1,2,3	• comparison of HI & VI inequality among and within cohorts of computer scientists • identify the relation of early career achievements on future career outcome • offer empirical evidence for the Matthew Effect in computer science	Offline	HI: male and female scientists VI: career cohorts	• Outcome • Opportunity	• Distributive	historical records of • Publications • Citations
Chapter 5	[JLMW21]	RQ1,2,3	• propose a multifaceted characterization of musician's careers • offer empirical evidence of structural inequality in EDM club culture • identify patterns of successful careers	Offline	VI: EDM artists	• Outcome • Opportunity	• Relational	Historical records of • Live performances • Music releases
Chapter 6	[WGJS15]	RQ1,2	• empirical evidence of multiple dimensions of gender bias on Wikipedia along multiple languages	Online	HI: prominent men and women	• Outcome • Opportunity	• Distributive • Relational	• Wikipedia articles • Databases of prominent people in history

Part I

Inequality in Science

Introduction

Science constitutes the the core of creative industry [Flo14a]. It is a collective endeavor that employs human curiosity and creativity to produce objective knowledge about the world, drive innovation and growth, and ultimately increase the quality of our lives. However, issues such as gender and racial inequities have hindered scientific works for decades, in which certain ideas and perspectives dominate our scientific discourses and research agenda. This may result in sub-optimal solutions, innovations and technologies that fail to address the need of society at large. A healthier and more inclusive society demands a scientific system that values and practices equity. The first step to address this issue is to identify and acknowledge the existing, as well as the historical dimensions of inequality.

Science is a broad and complex field, in which every discipline might be characterized by certain history, culture, structure and production practices. However despite such differences, publications and citations are at the core of every scientific work. While publications are the primary form of production, citations offer a means for reputation and recognition building. By tapping into this information we can assess scientific systems and careers. This thesis takes computer science as the domain of study to explore and identify various domains of inequality in scientific careers. Computer science makes an interesting case because of its collaborative nature, its long-standing issue of gender discrimination, its ongoing transformation, and its role as the driver of the digital revolution. I start my analysis in chapter 2 by proposing a novel gender detection methods to infer the binary gender attribute of scientist using their names and pictures. Being able to infer the gender of individuals, I investigate the extent, dimensions, and evolution of horizontal and vertical inequalities in the entire field of computer science. In chapter 3, I identify multiple dimensions of the gender gap, show how they evolve, and discuss how they could potentially harm the career of women. Comparing the co-authorship networks of male and female scientists, I quantify the structural disadvantage of women and identify which collaboration patterns are related to scientific success. Finally, in chapter 4, I examine different aspects of vertical (i.e., productivity and recognition) and horizontal (i.e.,

gender gap) inequalities across, and within cohorts of scientists that started their career in a particular year. I assess the predictive power, hence the impact, of numbers of meritocratic and non-meritocratic characteristics of scientists in their early-career, on their future success and dropout.

H

Chapter 2

Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods

Abstract. Computational social scientists often harness the Web as a "societal observatory" where data about human social behavior is collected. This data enables novel investigations of psychological, anthropological and sociological research questions. However, in the absence of demographic information, such as gender, many relevant research questions cannot be addressed. To tackle this problem, researchers often rely on automated methods to infer gender from name information provided on the web. However, little is known about the accuracy of existing gender-detection methods and how biased they are against certain sub-populations. In this paper, we address this question by systematically comparing several gender detection methods on a random sample of scientists for whom we know their full name, their gender and the country of their workplace. We further suggest a novel method that employs web-based image retrieval and gender recognition in facial images in order to augment name-based approaches. Our findings show that the performance of name-based gender detection approaches can be biased towards countries of origin and such biases can be reduced by combining name-based an image-based gender detection methods.

2.1 Introduction

The Web enables studies of human social behavior on a very large scale. For many research questions, demographic information about individuals (such as age, gender or ethnic background) is highly beneficial but often

particularly difficult to obtain.

This has led previous research to employ different methods for inferring the gender of individuals from names. For example, in [GJT09] the authors determine the gender of individuals using the name repository from the US Social Security Administration and study the relationship between gender and job performance among brokerage firm. Mislove et al. [MJA+11] used the same name repository to infer the gender of Twitter users by mapping their self-reported names to the name database. In another study the authors aim to study gender disparities in science and infer the gender of scientists based on a similar approach [WJK+13]. Unfortunately, most previous work does not provide information on how accurate different gender detection methods are and/or how biased they are against certain sub-populations. Although crowd sourcing methods can be seen as an alternative for automated gender detection methods, they do not scale well and are expensive. In the absence of a full name, more sophisticated methods such as supervised machine learning models are used to harness the users content for detecting the gender (see e.g. [RYSG10]). Yet, separate models are needed for gender detection methods in each language community [CSR13].

In this paper we evaluate and compare frequently used name-based gender detection methods. We report overall accuracy and also bias, i.e., deviating accuracy for different demographic sub-populations (e.g. men, women and people living in different countries). Moreover, we propose novel methods that increase the accuracy of gender detection across heterogeneous sub-populations by augmenting traditional methods with face recognition techniques.

2.2 Data and Method

For our evaluation, we utilized ground truth data from a previous study on global gender disparities in science [LNG+13]. It consists of a manually labeled random sample of academics, their full names, institutions, countries, and their gender. The ground truth was created by inspecting CVs, pictures and institutional websites. After removing ambiguous and repetitive names, the final name list consist of 693 male names and 723 female names.

We then evaluate different name-based gender detection methods using the full names of our manually labeled scientists as input. Finally, we propose a new mixed method that combines name-based and image-based gender detection.

2.2.1 Gender detection methods

In the following, we review some prominent unsupervised approaches that only require a name or picture as an input. These approaches do not re-

quire training and are widely used in scientific research as mentioned in the introduction.

Security Administration's baby names data. The US Social Security Administration (SSA) covers registered baby names in the United States since 1880. Many gender detection tools such as the "gender" package in R[1] or the OpenGenderTracking[2] rely on this database.

IPUMS Census data. Integrated Public Use Microdata Series (IPUMS) census data consists of samples of the American population drawn from fifteen federal censuses from the American Community Surveys between 1850 to 2000. This database is also used in the "gender" package in R and other web-based name extraction packages[3].

Sexmachine. The list of 40,000 names is primarily collected by Jörg Michael. Because of its availability, several libraries in various languages (see for example C's (*gender.c*) or Python's *Sexmachine* library[4]) use this database. Given a name, Sexmachine makes a guess whether the name is male, mostly male, female, mostly female or unclear. The advantage of this name list is that it provides detailed information about how popular a first name is in a country and how strongly it is associated with a given gender. Therefore, it enables the disambiguation of names based on the country of origin. The list also provides information for a variety of countries including China and India.

Genderize. Apart from publicly available name data bases there are numerous commercial applications that incorporate various databases from online resources to assess gender. The problem with commercial applications is the difficulty to determine how the data is gathered and processed. Among commercial detection methods are *Facebook graph API*, *Gender API*, *Namsor* which is based on *Gender API* and *Genderize*. In this work, we analyzed the latter method. *Genderize* utilizes big datasets of information, from user profiles across major social networks and exposes this data through its API. The response includes a confidence value[5].

Face++. In addition to name-based gender detection methods, face recognition algorithms have become a popular tool for inferring the gender, e.g., for social media users. Among those, image-based application *Face++* seems to provide high performance [ZCY15]. This approach requires access to a picture of the person.

In order to derive the gender for a specific scientist, we propose to initially collect the first five *Google thumbnails* using the full name as search query term and then apply image-recognition on the search results. This approach

[1] `https://cran.r-project.org/web/packages/gender/`
[2] `http://opengendertracking.github.io/`
[3] `https://usa.ipums.org/usa-action/`
[4] `https://pypi.python.org/pypi/SexMachine/`
[5] `https://genderize.io/`

does not necessarily require that the collected pictures depict the scientist we originally searched for, but the idea is that we collect a sample of pictures that depict people who are named like the person we searched for. The advantage of using the full name as input is that for first names that are ambiguous or unisex, the combination of first and last name is often a better indicator of the gender associated with the certain culture.

A Novel Mixed Approach. In addition, we propose mixed methods that combine name-based detection methods with an image-based face recognition approach. We test two variations of this method. In method *Mixed1*, the best name-based approach, namely *Genderize*, is used first. For the remaining unidentified names, the image-based method *Face++*, is used. In method *Mixed2*, *Genderize* and *Face++* have equal weight. For the weighting, we do not use a binary decision for each method, but also take the reported confidence as a numeric value into account. In doing so, this method can handle ambiguous names more efficiently. Note that method *Mixed1* does not require retrieving pictures for the whole population and is therefore more efficient than method *Mixed2*.

Table 2.1: Per-class and overall precision and recall of various gender detection methods. The mixed approach outperforms all other methods by at least 9%.

	SSA	IPUMS	Sexmachine	Genderize	Face++	Mixed1	Mixed2
female precision	0.96	0.96	0.97	0.95	0.92	0.91	0.93
female recall	0.79	0.69	0.77	0.86	0.81	0.95	0.94
female F_1	0.86	0.80	0.85	0.90	0.86	0.93	0.93
male precision	0.98	0.92	0.98	0.98	0.86	0.96	0.98
male recall	0.70	0.68	0.72	0.77	0.85	0.89	0.88
male F_1	0.82	0.78	0.83	0.86	0.85	0.92	0.93
accuracy	0.75	0.68	0.74	0.82	0.83	0.92	0.91

2.3 Results and Discussion

The results displayed in Table 2.1 show that among individual methods, image-based *Face++* and *Genderize* perform relatively better than others. However, the overall best results are achieved by the mixed approaches, which outperform all others by at least 8% accuracy. Although all evaluated methods achieve high overall precision, recall rates vary. All gender detection methods show comparable results for both classes (male and female) and therefore no systematic gender-bias can be asserted.

By contrast, Table 2.2 indicates that the error rates strongly depend on the country of residence of an individual. While name-based approaches work quite well for western industrialized countries, their performance deteriorates for emerging nations such as China, South Korea or Brazil. Clearly,

Table 2.2: Accuracy of various gender detection methods for people from different countries. For most countries mixed approaches perform best.

	# instances	SSA	IPUMS	Sexmachine	Genderize	Face++	Mixed1	Mixed2
United States	419	0.82	0.76	0.84	0.83	0.91	0.91	0.90
China	113	0.20	0.11	0.67	0.28	0.65	0.50	0.56
United Kingdom	96	0.94	0.92	0.92	0.94	0.81	0.98	0.94
Germany	82	0.87	0.88	0.96	0.94	0.87	0.96	0.93
Italy	75	0.93	0.92	0.94	0.98	0.79	0.99	1
Canada	60	0.87	0.77	0.86	0.91	0.90	0.96	0.93
France	58	0.93	0.92	0.80	0.96	0.81	0.97	1
Japan	56	0.79	0.70	1	0.90	0.62	0.91	0.94
Brazil	44	0.29	0.29	0.15	0.44	0.81	0.90	0.93
Spain	39	0.96	0.92	0.92	1	0.92	1	1
Australia	31	0.89	0.89	0.90	0.86	0.86	0.94	0.93
India	29	0.67	0.17	0.71	0.78	0.83	0.83	0.93
South Korea	27	0.04	0.00	0.58	0.11	0.74	0.37	0.66
Switzerland	25	0.78	0.70	0.56	0.83	0.88	0.90	0.92
Turkey	21	0.43	0.14	0.79	0.81	0.86	1	1

popular names of these countries are not covered sufficiently in the databases at this point in time. For these countries, an image-based approach leads to substantially better results (e.g., for South Korea the accuracy of image-based approaches is at least 16% better than the best name-based method). The *Genderize* method that also harnesses social media performs poorly for China, presumably due to accessibility to the Chinese social networking websites. Our proposed mixed approaches outperform the existing methods for the majority of the countries.

Conclusions: Our results suggest that the performance of name-based gender detection approaches varies according to the country of origin and that performance for emerging nations is particularly weak. Significant enhancements can be achieved by combining name-based with image-based gender detection methods. In the future, our findings could be combined with machine learning approaches to develop better methods for assessing demographic attributes of users on the Web.

Chapter 3

Gender Disparities in Science? Dropout, Productivity, Collaborations and Success of Male and Female Computer Scientists

Abstract. Scientific collaborations shape novel ideas and new discoveries and help scientists to advance their scientific career through publishing high impact publications and grant proposals. Recent studies however show that gender inequality is still present in many scientific practices ranging from hiring to peer review processes and grant applications. While empirical findings highlight that collaborations impact success and gender inequality is present in science, we know little about gender-specific differences in collaboration patterns, how they change over time and how they impact scientific success. In this paper we close this gap by studying gender-differences in dropout rates, productivity and collaboration patterns of more than one million computer scientists over the course of 47 years. We investigate which collaboration patterns are related with scientific success and if these patterns are similar for male and female scientists. Our results highlight that while subtle gender disparities in dropout rates, productivity and collaboration patterns exist, successful male and female scientists reveal the same collaboration patterns: compared with scientists in the same career age, they tend to collaborate with more colleagues than other scientists, establish longer lasting and repetitive collaborations, bring people together that have not been collaborating before and collaborate more with other successful scientists.

3.1 Introduction

Collaboration is the core task of any scientific discourse. In the course of a collaboration, new ideas shape and eventually result in new discoveries and scientific publications [Moo04]. As a result, collaborations impact researchers' scientific career and academic success [PFP+14, Pet15, SPS+14a, SRNM+15]. For example, previous research has shown that the centrality of a scientist in a collaboration network is associated with his/her success [SPS+14a, SRNM+15] and co-authorship strength is related to high productivity and citations [Pet15, PFP+14].

At the same time gender inequality is still rife in science, for example, in hiring [MRDB+12], grant applications [LH08, vdLE15], peer reviews [MBD12, KGC14], earnings [Hol01, WC06], tenure [SKS05], satisfaction [Hol01], patenting [DMS06], productivity [WJK+13, DZSP+12], labor division in scientific collaborations [MLSS16], internationality of collaborations [LNG+13] and scientific success [LNG+13]. For example, a report from 2006 showed that only one quarter of full professors are female and they earn 80% of their male colleagues on average [WC06]. More recent research showed that women are more likely to take executive roles in collaborations [MLSS16], their collaborations are more domestically oriented and papers where women are the lead author (i.e. solo author, first author or last author) receive fewer citations [LNG+13].

While empirical findings highlight that collaborations impact success and that gender inequalities are present in science in various forms, little is known about gender-specific differences in collaboration patterns and how these differences may impact career success. A mentionable exception is a very recent study that investigated if female and male researchers in science, technology, engineering and mathematical (STEM) disciplines differ in their collaboration patterns [ZDSP+16]. While this work offers interesting insights into the average number of co-authors and strength of collaboration among male and female researchers across various disciplines, it does not analyze the temporal evolution of collaboration patterns across career ages and how different network features relate to scientific success.

In this work, we aim to close this gap by presenting an empirical study on the temporal collaboration network of researchers that contribute to the field of computer science and explore which patterns are related with scientific success (measured by number of citations and h-index). We analyze gender-differences in dropout rates, productivity and collaboration patterns and explore if successful male and female scientists show the same collaboration behavior. This study is conducted over time since the collaboration network as well as the success and productivity of scientists naturally change with career age. We use a panel regression method to explain the relation between the success of scientists at different career ages and various features that characterize their collaboration behavior in the past and their gender.

Our results show that (1) the dropout rate of women is consistently higher than the dropout rate of men, especially at the beginning of an academic career (40% of men and 47% of women stop publishing after the year in which they published their first publication); (2) the average productivity of men is higher and the productivity gap is increasing over time. However this difference can be explained by the higher number of senior male scientists. We do not find any significant differences in the average productivity of men and women within the same career age; (3) the overall gender homophily within the community has been increasing over the past few years. In particular the homophily among women is higher than it is among men when controlling for network topology and size. We only find small differences between the degree, k-core and clustering coefficient of men and women over time; (4) we find that the number of collaborators, the collaboration duration and strength, the success of collaborators and the ability to bring other scientists together are positively correlated with success. We do not find any gender-specific differences in how collaboration behaviour impacts scientific success.

3.2 Data

To construct a time-evolving collaboration network we use DBLP [Ley09], a comprehensive collection of computer science publications from major and minor journals and conferences. While DBLP offers name-disambiguation [RH10, Ley09, RWL+06], it does not provide information about citations. Therefore, we use publication titles to combine the DBLP dataset with the Aminer dataset [TZY+08] that contains all citation relations among papers in DBLP.

To infer the gender of authors we use a method that combines the result of a name-based (Genderize.io[1]) and an image-based (Face++ [2]) gender detection services. Previous research has shown that the accuracy of this method for most countries is above 90% (see *Mixed1* in Table 3.1) [KWL+16b]. Since we have a very low accuracy for Chinese and Korean names, we label their gender as unknown in order to reduce noise in our analysis. To detect Chinese names we compile a list of 202,045 unique names using "China Biographical Database Project (CBDB)"[3]. For compiling a list of Korean names with use Wikipedia as our data source. To do this, we extract the page titles of all the backlinks to the Wikipedia page "Korean names"[4]. The page titles include the names of prominent Korean figures (e.g,. singers) with a Wikipedia page that describe the origin of the

[1] `https://genderize.io/`
[2] `https://www.faceplusplus.com/`
[3] `http://projects.iq.harvard.edu/cbdb/home`
[4] `https://en.wikipedia.org/wiki/Korean_name`

Table 3.1: Accuracy (the proportion of true results among total number of cases) for various gender detection methods for scientists across different countries. For most countries the mixed approaches that combine image- and name-based gender detection perform best.

	# instances	SSA	IPUMS	Sexmachine	Genderize	Face++	Mixed1	Mixed2
United States	419	0.82	0.76	0.84	0.83	0.91	0.91	0.90
China	113	0.20	0.11	0.67	0.28	0.65	0.50	0.56
United Kingdom	96	0.94	0.92	0.92	0.94	0.81	0.98	0.94
Germany	82	0.87	0.88	0.96	0.94	0.87	0.96	0.93
Italy	75	0.93	0.92	0.94	0.98	0.79	0.99	1
Canada	60	0.87	0.77	0.86	0.91	0.90	0.96	0.93
France	58	0.93	0.92	0.80	0.96	0.81	0.97	1
Japan	56	0.79	0.70	1	0.90	0.62	0.91	0.94
Brazil	44	0.29	0.29	0.15	0.44	0.81	0.90	0.93
Spain	39	0.96	0.92	0.92	1	0.92	1	1
Australia	31	0.89	0.89	0.90	0.86	0.86	0.94	0.93
India	29	0.67	0.17	0.71	0.78	0.83	0.83	0.93
South Korea	27	0.04	0.00	0.58	0.11	0.74	0.37	0.66
Switzerland	25	0.78	0.70	0.56	0.83	0.88	0.90	0.92
Turkey	21	0.43	0.14	0.79	0.81	0.86	1	1

name of that person (e.g., Wikipedia page of a Korean singer and actor[5]). Using this method we compile a list of 6,451 unique Korean names.

Our data consists of 3,085,544 publications and 7,849,398 citations that have been created in the time span of 47 years, between 1970 and 2016. Among all publication, 717,471 papers (23%) receive at least one citation from other papers inside the DBLP corpus.

First, we build a collaboration network where each node represents an author and each edge a co-authorship relation. Each edge is labeled by one or multiple date(s) that correspond to the publication date(s) of papers. We later use this information to study the network evolution over time. The complete collaboration network consists of 1,634,682 nodes and 7,304,250 edges. 699,370 ($\approx$ %43) nodes were identified as men, 227,473 were identified as women ($\approx$ %14), and for 707,839 authors ($\approx$ %43) we could not infer their gender (e.g., Chinese or Korean names, name contains only initials).

We infer the *career ages* of scientists by comparing the first and last publication record inside the DBLP corpus. For example, a scientist who has only published papers in 1995, in 2000 and in 2005, has a career length of 11 years. In 1995 her career age is 1, in 2000 it is 6 and in 2005 it is 11.

Figure 3.1 (left) shows that the scientific community is growing rapidly in recent years and is becoming more gender balanced. The inset shows while in 1970 there were 16 times more men than women publishing in computer science venues, in 2015 we only find 3 times more men than women. Figure 3.1 (right) shows the proportion of men and women that are part of the

[5] `https://en.wikipedia.org/wiki/Ahn_Jae-wook`

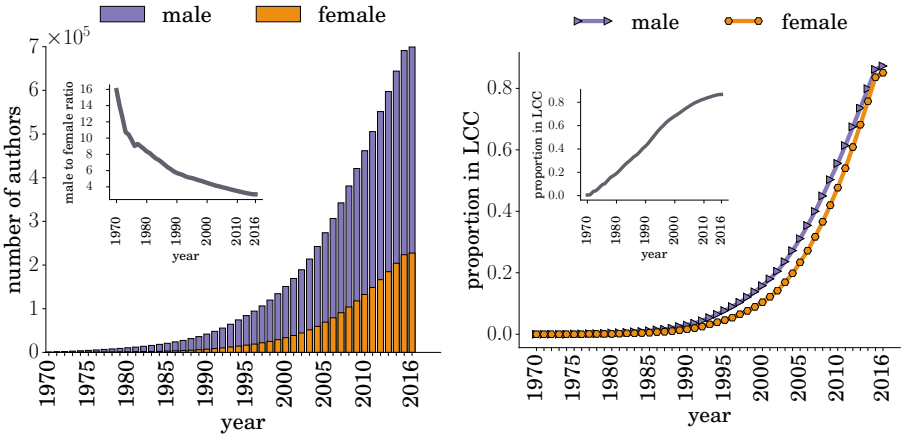


Figure 3.1: **Left: Presence of men and women in the community.** The main figure shows the total number of men and women in yearly snapshots of the network. The inset shows the relative size of men compared to women. Women are always underrepresented in the community. However the gap is decreasing over time. **Right: Growth of Largest Connected Component (LCC).** The main figure shows the proportion of men and women that belong to the LCC in yearly snapshots of the network. For example in the year 2000, around 20% of men and 10% of women were part of the LCC. There is always a higher proportion of men that belong to the LCC. The inset shows the overall proportion of authors in the LCC, including those labeled with unknown gender. Over time, the community is becoming more connected and the relative size of LCC is increasing.

Largest Connected Component (LCC). For example, in 2000 around 20% of men and 10% of women belonged to the LCC. The proportion increased to around 85% and 80% in 2015 for men and women, respectively. The plot suggests that the proportion of men in the LCC has always been higher than those of women. However, the gap is closing in recent years. Furthermore, the inset shows the proportion of nodes that belong to the LCC has been increasing regardless of gender which indicates that the community has become more connected over time.

3.3 Results

To investigate the evolution of gender disparities in the computer science community between 1970 and 2015, we compare (1) *dropouts* (number of male and female scientists that stop publishing), (2) *productivity* (number of publications per author), (3) *collaboration patterns* and (4) *scientific success* (number of citations and h-index) of male and female scientists.

3.3.1 Dropout

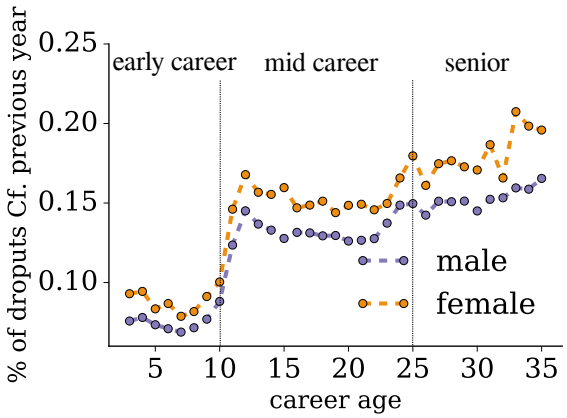


Figure 3.2: **Dropout rate:** Proportion of men and women at different career ages that permanently stop publishing. Most scientists (40% of men and 47% of women) drop out one year after their first publication (not shown). Of those that continue, 8% of men and 9% of women drop out after their second year (from here on shown). After the drastic dropout at the very beginning, the rate shows three phases. The first corresponds to early-career researchers (career age 2-10) for which we observe a dropout rate between 7% and 10% every year. In career ages 11 and 12, the rate jumps to 15% for men and 17% for women. In the second phase related to mid-career researchers (career age 11-25), the dropout rate fluctuates between 13% and 18%. The third phase corresponds to senior researchers with (career age above 25). They drop out at a rate of 14% to 21% (for career age above 35 fluctuations increase). Women consistently have higher rates (2 percentage points) across all career ages.

Leaky pipelines are frequently claimed to cause gender disparities in science. This metaphor implies that women drop out of academia at a higher rate as they advance in their career [Wic97, Pel96]. To compare the dropout rates of male and female scientists we first infer their career age based on their publications. We assume that a scientist who has not published any paper in 10 or more years has left academia, since staying in academia requires publishing. Scientists who died will also be counted as dropouts, but we do not expect that the proportion of men and women who die in the same career age is significantly different. Since our dropout definition requires to observe at least 10 years after each publication, we limit our dataset to scientists who published at least one publication before 2006. That means people who started their scientific career after 2006 are not included in our analysis. This leaves 326,329 men and 84,859 women for the dropout analysis.

Figure 3.2 shows the percentage of men and women who permanently

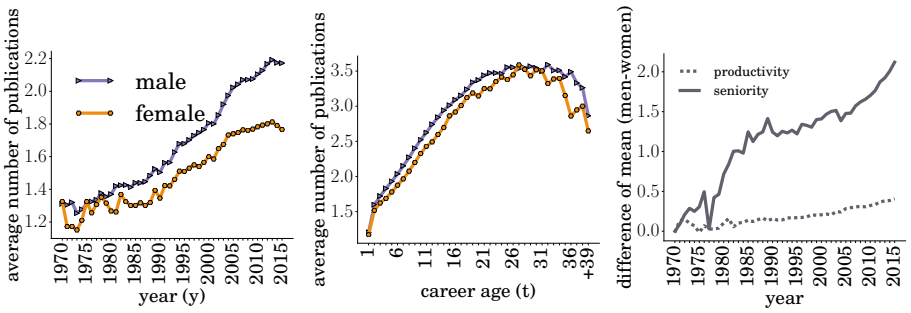


Figure 3.3: **Left: Productivity gap (calendar years).** Average productivity (number of publications) of men and women over calendar years. Although productivity increases for both sexes, men tend to be slightly more productive than women. In this analysis we neglect the year 2016 as it might be affected by censoring bias and missing publications. **Middle: Productivity gap (career ages).** Average productivity of men and women over career ages. Three phases can roughly be detected: (1) career age 1-20: increase of productivity; (2) career age 21-30: stable productivity, 3) career age 31 and on: decreases of productivity. The average productivity of men and women at the same stage of the career is very similar. **Right: Productivity gap vs. seniority gap.** Differences between the mean productivity of men and women (productivity gap) and the mean career ages of men and women (seniority gap) in the same calendar year. The Pearson correlation between the two differences is 0.86 with $p = 10^{-15}$.

dropped out of the academic pipeline at different stages in their academic career. The main message is that scientists tend to stay in the field if they manage to survive the first year in which they publish. 40% of the male and 47% of the female authors do not enter a second year (read caption for further details). For those who do survive, 32% of the men and 31% of the women stay for up to 10 years and become *early-career researchers*, 25% of the men and 20% of the women stay for up to 25 years and become *mid-career researchers*, and only 3% of the men and 2% of the women become *senior researchers* and stay 26 and more years in the field. This gender difference of careers entails a comparability issue we need to address in the remainder of the paper.

3.3.2 Productivity

Various explanations, from funding to family responsibilities and international collaboration, have been offered to solve the productivity puzzle discussed in the introduction. Our results show that the average productivity, regardless of gender, has been increasing over time and that gender differences prevail (cf. figure 3.3, left). On average, men tend to have higher

publication rates than women in all calendar years and the gap is widening after 2005.

We offer a solution to the productivity puzzle. The productivity gap almost vanishes when the average productivity of men and women in the same career age is compared (cf. figure 3.3, middle). Three phases of productivity become very similar for men and women: In their first two decades scientists tend to increase their productivity each year. In the following 10 years their average productivity is rather stable and scientists produce about 3.0 to 3.5 publications per year on average. Towards the end of long careers productivity drops again.

This result is in line with previous studies that found a similar pattern of productivity over the chronological age of scientists [Phe94, ARPS11a, VV09, Leh54]. However, the literature also reports different productivity trajectories for scientists of different citation impact [SWD+16a] and for researchers in different disciplines [BD77, KT96]. Recent research also highlights that while the aggregated pattern of productivity is surprisingly similar for researchers that are placed in institutions of different prestige rank, high diversity can be observed in the production trajectories of individual scientists [WMCL16].

Comparing scientists only for similar career ages amounts to controlling for seniority. Figure 3.3 (right) shows that the *productivity gap*, measured as the difference between the mean productivity of men and women in the same year, is paralleled by a *seniority gap*, measured as the difference between the mean career age of men and women in the same year. They not only increase over time but are strongly and significantly correlated (Pearson correlation coefficient 0.86, $p = 10^{-15}$). This suggests the simple explanation that men are more productive on average because they have a larger fraction of senior authors.

3.3.3 Collaboration patterns

Previous studies have either focused on a specific country (e.g., Zeng et al. [ZDSP+16] focus on the US) or ignored the time dimension (e.g., West et al. [WJK+13] ignore the career age of men and women when analyzing the average authorship-position on papers).

Here we investigate *how collaboration patterns and the network positions of male and female researchers change over time in an entire scientific field, computer science.*

For structural analyses and later regressions analyses of gender and success we operationalize several concepts of network embeddedness. Node degree, the number of co-authors, is a measure of the *size* of a researcher's ego network. Three measures offer insights into ego network properties. *Cohesion* is the extent to which a network has evolved into a hierarchical structure of increasingly dense cores embedding into each other. Since the

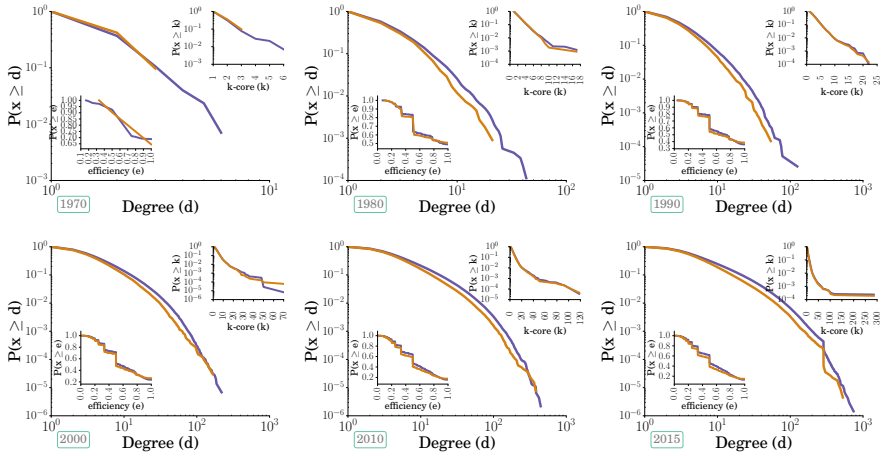


Figure 3.4: **Evolution of degree, *k*-core and efficiency distributions over 6 decades**: Main figures show the degree distributions of male and female scientists. The top-right and bottom-left insets show the k-core and efficiency distributions, respectively. Each plot refers to one specific year and describes the structure of the network including all collaborations that occurred between the beginning of 1970 and the end of the given year. As the cumulative network grows, the distributions grow fatter tails. In the beginning (1970 and 1980), women tended to collaborate with fewer researchers (lower degree) and with researchers that were themselves less well connected (lower k-core) than men. Women also tend to collaborate slightly more with colleagues that also collaborate with each other (lower efficiency).

best operationalization is costly to implement [MW03a] we use the k-core metric instead where k is an ego's maximum number of co-authors that have at least k neighbors themselves [B.S83].

Neither degree nor k-core tell if ego networks contain structural holes. Both the absence and the presence of such voids of connectivity are indispensible for the functioning of social networks. Closure, the absence of structural holes, is needed for trustful coordination while the presence of structural holes is accompanied by possibilities of brokerage, the reaping of advantages from tapping different pockets of information at multiple sides of the structural hole [Bur05]. We operationalize *closure* through the clustering coefficient, the density of an ego network excluding ego [WS98], and *brokerage* using Burt's efficiency, the normalized number of co-authors minus their average degree within the ego network, excluding ties to ego [Bur95].

To also capture the dynamics of structural order and disorder – or closure and brokerage – we introduce two measures relating to team assembly [GUSA05a]. *Collaboration strength* is the median number of publications of ego's collaborations, and *collaboration duration* is the median maximum publication year difference of ego's collaborations. If those scores are low,

collaborations are less trustful, and brokerage is more pronounced.

Structural gender disparities. Figure 3.4 depicts the growth of distributions of degree, k-core (top-right inset) and efficiency (bottom-left inset) for six points in cumulative time, distinguished by men and women. The tails of the degree and k-core distributions reveal that collaboration at the macro level has been increasing over decades, regardless of gender. We also observe that, in earlier years, men have slightly broader degree and k-core distributions compared to women. As the total network grows and the number of women increases, women emerge with ego networks that are as sizable and cohesive as those of men. With respect to efficiency, men tend to have slightly higher probabilities to act as bridges across structural holes. This is an intriguing result since previous work has shown that brokers tend to be more influential [UBMK12, Bur04b, Bur95].

Table 3.2: **Cliff's d-test to measure the distance between distributions.** Each value shows the d-statistic comparing degree, k-core and efficiency distributions for men and women for networks cumulated up to the given year (cf. figure 3.4). Positive (negative) values indicate whether the distribution of men (women) is dominant. The value of d ranges from -1 (when every observation for women are greater that those of men) to 1 (when every observation for men are greater that those of women). The differences between the distributions are significant but small for all years except the earlier ones when the network itself was small. In all significant cases, the distribution for men is dominant. *Note:* $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

	1970	1980	1990	2000	2010	2015
degree	0.000	0.007***	0.023***	0.064***	0.097***	0.069***
k-core	0.000	0.007***	0.023***	0.063***	0.089***	0.051***
efficiency	-0.075	-0.028	0.002	0.027***	0.061***	0.074***

To quantify the comparison of these distributions for men and women, we use Cliff's d-test that measures the extent to which one distribution is statistically dominant over the other one [Cli93]. Table 3.2 gives the d-statistics for degree, k-core and efficiency for six points in cumulative time. We observe small but significant differences between the distributions. In all significant cases, the distribution for men is the dominant distribution – i.e., men have larger and more cohesive networks, and they are more likely to be positioned at structural holes.

To quantify the change inherent to these distributions, we study the mean of the log-transformed values and look at the men-to-women ratio over cumulative time. Figure 3.5 shows that men tend to have larger and more cohesive networks at any time, though the gaps are decreasing. Regarding brokerage, the gender gap closes until 1983, in 1989 men have higher

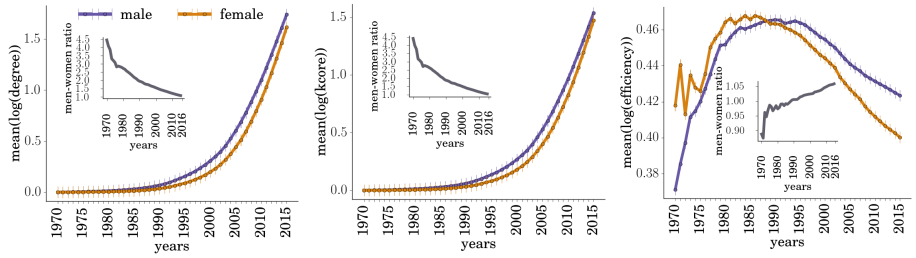


Figure 3.5: **Changes of degree (left), *k*-core (middle) and efficiency (right).** The main figures show the changes in means of log-transformed values over cumulative time. The insets show corresponding men-to-women ratios (ratios above (below) one indicate higher mean log-efficiency for men (women)). For degree and k-core men tend to have higher values, but the gap is decreasing over time. The gender gap in efficiency shows three phases: In the first phase (1970–1982) women are stronger brokers than men (ratios are below 1). In the second phase (1983-1993) the average log-efficiencies are not distinguishable. In the third phase (1994-2015) men are stronger brokers.

log-efficiency for the first time, and by 1994 men are significantly stronger brokers on average.

Collaboration patterns across career ages. Although the results so far indicate that gender-specific differences in collaboration practices exist, other confounding factors, such as the career-age distribution of men and women or the computer-science specialties in which men and women are unequally embedded, may explain our results. To address this problem to some extent, we use multiple logistic regression models in which we use a single collaboration concept as the independent variable and gender as the dependent variable.

Diagnosing the relationship between position and gender requires accounting for dynamic effects. To explore the temporal stability of the bivariate relationships, we fit several models for increasing time periods (e.g., the model for the year 2000 is based on the cumulative collaboration network of all publications that have been published before or in 2000). To establish temporal comparability, we only study authors which are active in the final year of each period (e.g., the model for the year 2000 is based on those authors in the cumulative collaboration network which had published in 2000).

This reduces the sample size to the one given in the last column of figure 3.3.

To further control for the career age of researchers, we replace a raw feature score s by its corresponding career-age z-score separately for each period. For example, for each scientist i in a specific year, we measure how

much her feature score at career age τ, $s_i(\tau)$, deviates (in terms of standard deviation) from the average degree of scientists at the same career age:

$$z_i(\tau) = \frac{s_i(\tau) - \langle s(\tau) \rangle}{\sigma[s(\tau)]} \tag{3.1}$$

Table 3.3 shows the odds ratio and z-statistics for each regression.

Table 3.3: **Association between collaboration features and gender.** Each model assesses the relationship between different collaboration features and gender ($male = 0$, $female = 1$) while controlling for the career age of scientists. Each cell gives the odds ratio from a logistic regression model that only uses a single collaboration feature to explain the gender of scientists in the collaboration network at the end of the given year. No significant effects are observed for early periods. For periods up to more recent years, nodes with higher clustering coefficient, lower efficiency and lower collaboration duration are more likely to correspond to female scientists. Degree and k-core are significant but exhibit effect sizes close to 1. We do not find any significant gender difference with respect to collaboration strength. *Note:* $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Period (1970–)	Clustering coefficient	Efficiency	k-Core	Degree	Collaboration duration	Collaboration strength	Sample size (female ratio)
1970	0.0 (-0.0)	2.158 (0.627)	0.727 (-0.713)	0.667 (-0.967)	- [1]	-	228 (0.05)
1980	1.669 (1.809)	0.81 (-0.754)	1.04 (1.414)	0.94 (-2.489)*	0.922 (-2.18)*	1.069 (0.726)	2,145 (0.09)
1990	1.916 (7.455)***	0.502 (-6.368)***	1.024 (2.602)**	0.969 (-5.798)***	0.951 (-4.737)***	1.008 (0.202)	11,104 (0.13)
2000	1.412 (10.069)***	0.67 (-8.071)***	0.999 (-0.438)*	0.984 (-11.833)***	0.951 (-11.013)***	1.0 (-0.02)	46,486 (0.16)
2010	1.649 (32.17)***	0.444 (-31.304)***	1.0 (-0.582)	0.99 (-25.03)***	0.945 (-24.985)***	1.0112 (1.438)	147,163 (0.2)
2015	1.818 (43.272)***	0.414 (-41.075)***	0.998 (-5.806)***	0.992 (-31.56)***	0.941 (-34.689)***	0.993 (-0.937)	192,687 (0.21)

Before 1990 no significant effects can be observed. For periods up to more recent years we find that scientists whose ego networks are more closed, contain fewer structural holes and are more short-lived are more likely to be female. This statistical analysis confirms our earlier results that men and women do differ structurally, particularly regarding brokerage and closure, starting in the 90s. The finding that women, on average, embed into networks with shorter collaboration duration may be interpreted to be in line with results by Zeng et al. [ZDSP$^+$16] who found that women have a lower probability of repeating previous collaborations than men. It should be noted that in all cases the coefficient of determination is close to zero, i.e. each feature alone can only explain a small proportion of variance in the response variable.

Mixing of men and women. *Homophily*, the tendency to associate with similar others, is one of the fundamental factors that shape social ties [Moo01, KW09]. Homophilic behaviour combined with group size differences can limit minorities to stretch their overall degree [KGW$^+$17]. Con-

[1] The z-score could not be computed because the corresponding value for all authors is equal to 1 and therefore standard deviation is equal to zero.

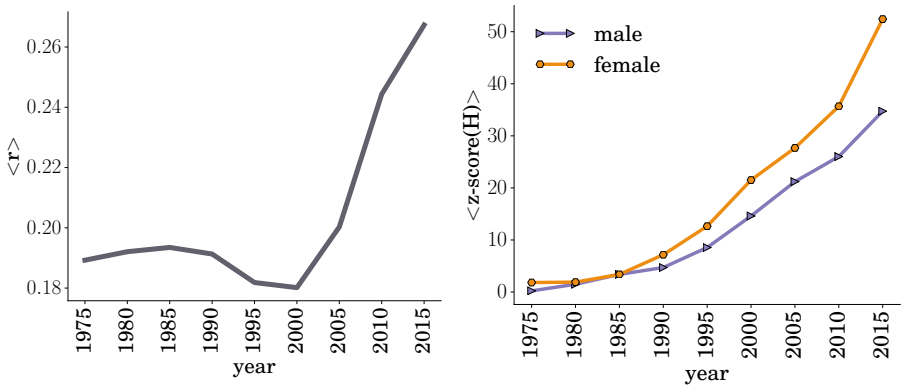


Figure 3.6: **Gender assortativity and homophily.** (Left) Newman gender assortativity r computed for annual snapshots of the collaboration network. Gender assortativity is stable until about 2000 and subsequently increases. (Right) z-Score of homophily computed using equation 3.2 for annual snapshots and 100 instances of a corresponding null model (i.e. a network in which we reshuffle the links but keep the degree intact). z-Scores indicate the deviation (in terms of standard deviation) from the homophily we would expect in a randomized network. They are computed separately for men and women. Homophily increases monotonically, women are more homophilic than men and the gap widens. All curves are smoothed using a 5-year moving average.

sequently, it can impact the opportunities afforded to minorities to access novel ideas and information. Since we are interested in observing how homophily is changing over time, we analyze the collaborative behaviour of scientists within each year separately rather than looking at the accumulated collaboration network for each year.

To diagnose global changes of homophily, we use Newman's assortativity measure r that captures the extent to which collaborative ties exist across gender ($r < 0$) and among the same gender ($r > 0$) compared to what we would expect from the node's degree [New03]. Figure 3.6 (left) suggests that assortativity was relatively stable in the past but started to increase in 2000.

The increasing trend in gender assortativity requires a detailed analysis to uncover whether the increase is mainly produced by the behaviour of one group or both groups. To assess the homophily for each gender separately, we look at the proportion of links between women (H_f) and men (H_m):

$$\begin{aligned} H_m &= \frac{E_{m,m}}{E_{m,m} + E_{f,m}} \\ H_f &= \frac{E_{f,f}}{E_{f,f} + E_{f,m}} \end{aligned} \tag{3.2}$$

Here $E_{m,m}$ refers to male-to-male edges, $E_{f,f}$ to female-to-female edges, and

$E_{f,m}$ to female-to-male edges. For example, $H_f = 1$ means that women only collaborate with women.

To assess the significance of the observed mixing pattern, we compare the observation to null models in which we keep the network size and the degree of the nodes intact and reshuffle the edges. Using this model we generate 100 synthetic networks for each yearly snapshot of our empirically observed co-authorship network. The synthetic networks represent random baselines that are expected if men and women are gender-blind during co-author selection. As a last step, we compute the mean and standard deviation of male and female homophily and the report the corresponding z-score.

Figure 3.6 (right) shows how many standard deviations the empirical homophily deviates from the expectation if the interactions would not be impacted by gender. We again see that the homophilic behaviour of men and women is increasing over time. However, the homophilic behaviour of women exceeds the expectation more than those of men.

Note that our baseline model assumes that every computer scientist can in theory collaborate with any other computer scientist. In reality subfields and specialties constrain who could collaborate with whom. If women are a minority that focuses on selected topical areas (e.g., Human Computer Interaction), then we would observe higher homophily for women than expected from our baseline model, assuming that collaborations within subfields are more likely than across subfields. That means, while our work shows that women tend to collaborate more with other women than expected, we do not answer the question why this is happening. Gender is one possible explanation, but also the gender composition of certain subfields will play a role. Therefore, whether the observed homophily is the result of authors' choices (choice homophily) or emergent structures (induced homophily) requires a deeper investigation that we leave for future works. [KW09, ST11]

3.3.4 Success

Here, we aim to understand the relationship between collaboration patterns, gender and scientific success. Specifically, we seek to answer *which collaboration patterns are related with scientific success and if these patterns are similar for male and female scientists.* To quantify scientific success, our dependent variable, we use two common measures: *citation impact*, the raw number of citations an author has accumulated up to a given year, and the h-index, the number of an author's publications that have accumulated at least h citations [Hir05b]. While the number of citations can be driven by single high-impact papers, the h-index combines the assessment of both quantity (number of papers) and quality (number of citations). A scientist needs to produce a high number of high quality papers in order to obtain a high h-index.

We create two different regression models that describe the relationship

between the collaborative behaviour of scientists and their success.

The first model (*ego model*) relies on the ego-centric properties of a node defined in the previous subsection. Because of a high correlation between degree and k-core (Pearson correlation of 0.75 with $p < 0.001$), we do not use k-core in our model to avoid multicollinearity. The second model (*1-hop model*) extends the ego model by including information about a node's median neighbourhood structure.

Moreover, the academic system naturally changes over time (e.g., with respect to size, number of relevant venues, publication and citation practices). Therefore comparing scientists that started their career in different decades may confound our results. To control for this effect, we add the starting decade of an author's career to our model. To study the effect of gender in collaboration and on success, we include gender as an interaction term in our models.

Table 3.4: **Sample size for regression of success.** Beside the number of authors we also list the number of observations since we have multiple observations per author (one for each year in which they were active).

	Men	Women	Total
Number of authors	72,076	13,746	85,822
Number of observations	734,474	131,194	865,668

The population of scientists is restricted to those with careers of at least 10 years and at least 5 publications. This way we focus only on people who have decided to pursue an academic career. For each scientist we record her collaborative features for all stages of her academic career, i.e. our panel data consists of multiple observations (at least 5) for each author, one for each career age. Furthermore, we ignore the first 5 career ages to give authors enough time to accumulate citations. Table 3.4 shows the size of our panel.

To account for within-subject correlation and unbalanced observations for subjects (e.g., missing observations), we use the General Estimation Equation (GEE) regression model [LZ86] with an *exchangeable correlation structure*. This structure meets our cumulative research design by assuming that the correlations between features for the same author at different career ages are stationary.. We fit the GEE model with a Gaussian distribution and the identity link function to the data. To assess the goodness of the fit we use the marginal R^2 which is an extension of R^2 statistics for GEE models [Zhe00]. Similar to R^2, marginal R^2 can be interpreted as the proportion of variance in the response variable explained by the fitted model.

We consider a scientist as successful if she has a higher citation impact or h-index than an average scientists in the same career age. Therefore we again use equation 3.1 to compute the age-specific z-scores for the number

of citations and the h-index. Since the z-scores of our dependent variables are skewed, we use of log of the z-scores instead. The independent variables are transformed into z-scores but not logged. Therefore, the coefficients quantify the association between above-average collaboration features and success.

Table 3.5: **GEE model for citation impact.** Odd ratios of coefficients are given for the number of citations as the dependent variable. Values in brackets give z-statistics for the coefficients. The ego model shows that degree, collaboration duration and collaboration strength are sizeably and positively related to scientific success. Efficiency has a small positive but significant effect. The 1-hop neighbourhood model confirms these observations and finds that median number of citations as well as career age of alters significantly add to ego's success while clustering coefficient of alters has a negative effect. There is no gender effect. *Note:* $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

	ego model	ego model +interactions	1-hop model	1-hop model +interactions
Intercept	1.682(125.699)***	1.681(125.098)***	1.700(133.419)***	1.700(132.788)***
gender(reference=men)[women]		1.004(1.518)		1.002(0.783)
clustering	0.999(-1.234)	0.999(-1.213)	0.999(-2.098)	0.999(-1.769)
clustering*gender		1.000(0.211)		1.000(-0.297)
degree	1.128(69.524)***	1.127(63.104)***	1.122(58.912)***	1.120(53.123)***
degree*gender		1.007(1.364)		1.008(1.659)
efficiency	1.004(7.467)***	1.004(6.88)***	1.005(7.811)***	1.005(7.95)***
efficiency*gender		0.999(-0.499)		0.998(-1.263)
median collaboration duration	1.021(57.046)***	1.021(52.313)***	1.015(34.412)***	1.015(31.314)***
median collaboration duration*gender		1.000(0.048)		1.000(-0.177)
median collaboration strength	1.007(10.833)***	1.007(9.732)***	1.005(8.445)***	1.005(8.039)***
median collaboration strength*gender		0.998(-1.092)		0.999(-1.051)
neighbours median age			1.004(5.473)***	1.004(4.872)***
neighbours median age*gender				1.000(-0.007)
neighbours median clustering			0.998(-5.666)***	0.997(-5.403)***
neighbours median clustering*gender				1.001(0.894)
neighbours median degree			1.005(1.613)	1.006(1.71)
neighbours median degree*gender				0.993(-1.701)
neighbours median n citations			1.036(6.376)***	1.036(5.452)***
neighbours median n citations*gender				1.000(0.036)
Marginal R^2	0.217	0.217	0.296	0.296

Tables 3.5 and 3.6 report odd ratios and size effects for the number of citations and the h-index, respectively, as proxies for success. All four models (the ego and 1-hop models for citation impact and h-index) agree that embedding into large enduring networks with some repetition of collaborations is the primary explanation of academic success. Structural closure is a significant predictor in the h-index models. Brokerage, however, the tapping of

Table 3.6: **GEE model for *h*-index.** Odd ratios of coefficients are given for the h-index as the dependent variable. Values in brackets give z-statistics for the coefficients. The ego model shows that degree, collaboration duration, efficiency and collaboration strength are sizeably related to scientific success. Clustering coefficient has a small but significant effect. The 1-hop neighbourhood model confirms these observations and finds that the median career age sizeably and the number of citations as well as the degree of alters significantly add to ego's success. There is no gender effect. *Note:* $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

	ego model	ego model +interactions	1-hop model	1-hop model +interactions
Intercept	1.971(100.784)***	1.968(100.382)***	2.009(107.311)***	2.008(107.022)***
gender(reference=men)[women]		1.016(5.396)***		1.012(4.485)***
clustering	1.005(7.335)***	1.005(6.676)***	1.005(6.718)***	1.005(6.215)***
clustering*gender		1.000(0.026)		0.999(-0.404)
degree	1.172(96.373)***	1.172(87.289)***	1.157(78.669)***	1.156(71.131)***
degree*gender		1.002(0.428)		1.005(0.993)
efficiency	1.017(23.733)***	1.016(21.355)***	1.021(26.531)***	1.021(25.288)***
efficiency*gender		0.999(-0.331)		0.997(-1.326)
median collaboration duration	1.032(66.816)***	1.032(61.709)***	1.022(41.479)***	1.022(38.399)***
median collaboration duration*gender		0.998(-1.832)		0.998(-1.423)
median collaboration strength	1.013(10.364)***	1.013(8.727)***	1.009(8.303)***	1.009(7.256)***
median collaboration strength*gender		0.998(-0.965)		0.998(-0.788)
neighbours median age			1.016(16.516)***	1.016(15.386)***
neighbours median age*gender				0.995(-2.137)
neighbours median clustering			0.999(-1.441)	0.999(-1.437)
neighbours median clustering*gender				1.000(0.087)
neighbours median degree			1.018(6.28)***	1.019(5.883)***
neighbours median degree*gender				0.991(-2.134)
neighbours median n citations			1.035(6.441)***	1.034(5.501)***
neighbours median n citations*gender				1.004(0.513)
Marginal R^2	0.220	0.220	0.298	0.298

various information resources, is also a significant predictor of success, even a strong one when success is measured through the h-index. Interestingly, in the latter case, closure also turns significant. Much in line with the existing literature [GUSA05a, PBV07] this means that trustful relations are not an option but a requirement for authors and fields to thrive. Successful scientists keep reproducing a large network of core collaborators while simultaneously adding new collaborators from a variety of social circles. While long-lasting research partnerships can lead to collaborations that increase success through increased productivity [Pet15], new collaborators and brokerage can increase visibility within the community and make a researcher more influential [UBMK12, Bur04b].

In addition to the effects of ego-centric features, the 1-hop models demonstrate that collaborating with successful and senior scientists is beneficial for a researcher, especially in the h-index models. In that case, when success is also assessed in terms of productivity, collaborating with highly-connected scientists is also beneficial. This is probably the effect that teams of junior and senior researchers can produce outputs of quantity and quality [WJU07a]. Given that, in the ego model, creating trustful relations (enduring and strong) is a stronger predictor of success than brokerage, one may expect that collaborations with strong brokers may be beneficial. The observation that ties to co-authors with highly closed networks have a negative effect on success may be considered as supporting evidence for this conjecture.

Finally, our analysis shows that no significant gender-specific differences exist in how collaboration patterns impact success, since no interactions between gender and collaboration patterns can be found. This is evidence that *successful male and female scientists exhibit the same collaborative behaviour* and that *no differences exist in which collaboration patterns may explain the success of men and women in computer science.*

While the same collaboration patterns explain the success of male and female scientists, our previous analysis (see table 3.3) revealed that men and women do embed into significantly different ego networks. Networks of female researchers are significantly smaller, more closed, more devoid of structural holes and – on the median – more short-lived, while men take roles as explorers of large spaces who maintain trustful relations on the long run. Male collaborative behaviour is the one associated with success in academia. This suggests that women are on average less likely to adapt the collaborative behaviour that is related to success. However, those women who do become successful computer scientists show the same collaboration patterns as their successful male colleagues.

Interestingly, gender has a minuscule but significant effect on the h-index but not on the number of citations in the 1-hop models (tables 3.5 and 3.6). Also note that the regression models only explain 20–30% of the variance, i.e. our purely structural approach misses central aspects of the research

practice.

3.4 Discussion

Gender gap in academia, especially in STEM fields, has been a great concern over the past decades and many studies have tried to quantify the extent to which gender inequalities are present in science.

In this study we extend previous work by looking at the collaboration behaviour of scientists over a period of 47 years, from 1970 to 2016. In particular, we focus on one field, computer science, with a collaborative community [WJU07a] and a wide gender gap [LCMF15].

To assess the gender gap, we first compare the *productivity* of men and women in terms of number of publications. In line with previous studies [CZ84a, WJK+13, BA03, WW97, DZSP+12, Pro08, Sta04] our findings confirm that a productivity gap exists and our work also indicates that the gap is widening over the past few years. However, we also show that the productivity gap between male and female computer scientists can in part be explained by the distribution of the career ages of male and female scientist. Across all calendar years more senior male than senior female scientists were active in computer science and the average productivity of senior scientists is naturally higher than those of junior scientists.

The lower number of senior female scientists can in part be explained by the lower number of women in computer science in the past (e.g., in 1970, 16 times more men than women published at computer science venues). But also the dropout rate of women is consistently 2% higher than the dropout rate of men which may also contribute to the seniority gap between men and women.

Surprisingly, we observe an increasing trend in gender homophily over the years and in particular among women. The high homophily among women is alarming since recent research highlights the importance of diverse interactions in receiving visibility [KGW+17] and accessing novel ideas and information [DG11].

The evolution of the collaboration network suggests that in earlier years men tended to have broader degree (more collaborators) and k-core (more well-connected collaborators) compared to women but in recent years the gap is closing. Yet, men tend to have a slightly broader degree and k-core distribution than women. Additionally, women show slightly higher clustering compared to men suggesting that women are more involved in triadic relations. This indicates that women are less likely to collaborate with colleagues that are not connected among themselves and consequently women are less likely to bridge structural holes in the collaboration network.

As also suggested by previous work [UBMK12, Bur04b], we find that scientists who function as bridges tend to be more successful. A scientist with

a low clustering coefficient brings other scientists together who would probably not collaborate otherwise. Our success analysis reveals that successful scientists tend to collaborate with more colleagues than other scientists in the same career age, they establish long lasting and repetitive collaborations (i.e. they form strong ties), they bring people together that have not been collaborating before and they collaborate with other successful scientists. In part our results support findings from previous research that showed that publications co-authored with super ties (i.e. long lasting, repetitive collaborations) positively impact long-term citations [Pet15].

Interestingly, we do not find any gender differences with respect to how collaboration behavior impacts scientific success. This suggests that successful male and female scientists reveal similar collaboration behavior and success is not directly impacted by gender when controlling for collaboration patterns and scientific age. Also hiring outcomes (which are a another measure of success) are not directly effected by gender after controlling for scholarly productivity and relative prestige between hiring and placing institution [WLC16]. Interestingly, Zeng et al. found that female scientists have a lower probability of repeating previous collaborations compared to men [ZDSP+16]. This is an intriguing result since our work shows that repetitive and long lasting collaborations have a positive impact on success.

Our work does not allow to answer the causal question if certain collaboration strategies (e.g. repetitive collaborations or bringing people from different communities together) lead to success or if the observed patterns are a consequence of success. For example, successful people may be involved in repetitive collaborations because others want to collaborate with them again. It is very likely that these relationships are not unidirectional causal, but mediated by an unobserved variable, the skills and knowledge of a scientist.

Although our statistical models controlled for different factors such as career age, our work is limited to characteristics that are measurable and observable in our data. The main contribution of this work is a large scale temporal and gender-sensitive analysis of productivity, dropouts, collaboration patterns and success of computer scientists over the course of 47 years. We hope our results shed light into the understanding of collaboration patterns that are related with scientific success and gender differences in scientific collaborations, productivity and career paths. In future it would be interesting to extend this analysis to more academic fields and also explore disparities across ethnic groups.

Chapter 4

The Matthew Effect in computer science: A career study of cohorts from 1970 to 2000

Abstract. Inequality prevails in science. Vertical inequality signals who or what belongs to the core of a field. Few are successful, most perish quickly. Also, decades after the "productivity puzzle" has been identified, horizontal inequality among women and men is still puzzling. But the literature has identified the Matthew Effect as a central mechanism in explaining both inequalities and academic success. Using large-scale bibliographic data and following a computational approach, we study the evolution of inequality for cohorts from 1970 to 2000 in the whole field of computer science as it becomes a "big science." We find that vertical inequality in productivity increases over a scholar's career but is historically invariant. Vertical inequality in recognition is larger but stable across cohorts and careers. Gender inequality prevails regarding productivity, but there is no evidence for differences in recognition. The Matthew Effect is shown to accumulate productivity (publications) and recognition (citations) advantages and to become stronger over the decades. Predicting total-career outcomes from early-career achievements and endowments, we identify and discuss two paths to recognition-based success. We obtain our results by integrating various modeling steps and incorporating methods and epistemological strategies from the formal sciences and machine learning into computational sociology.

4.1 Introduction

Half a century ago, Price diagnosed that the science system exibits an "essential, built-in undemocracy", meaning that academic achievements like productivity and recognition are strongly concentrated among a very limited number of persons or organizations [Pri63]. He observed inequality in the form of productivity and recognition distributions and found this pattern to be stable as science grows, perpetuating a system where a "few giants" coexist with a "mass of pygmies" [Pri63, p.53]. The broadness of these distributions has been found to be a universal property of the science system [Lot26, Bra34, ACORC11, RCC14]. Inequality can be quantified via aggregate statistics like the Gini coefficient [All78]. Early work on chemistry cohorts had found inequality in productivity (publications) and recognition (citations) to increase as a cohort ages [ALK82]. Using full-scale bibliographic databases, recognition inequality has been found to decrease [LGA09, PP14, PPPF18] over time as the academic system is transitioning from "little science" to "big science" – from a scholar-centered to a globalized, interdisciplinary, team- and project-based mode of knowledge production [Gib94].

From a field [Bou88] or network [Whi70] theoretic perspective, such *vertical inequality* (i.e., inequality among individuals) is functional. In science, the broad "power distributions" [Fla17] of productivity and recognition resemble the hierarchies or "pecking orders" [Cha80] of academic fields in which authors (or ideas) take positions throughout their careers. By signaling who or what belongs to the core of an academic formation, inequality in positions reduces the information observers must process and creates meaning horizons for future transactions [Fuc01, WOSMP04].

Vertical inequality tends to be considered fair if it is merit-based [SSB17]. Scientific practice is considered fair if the merit of a knowledge claim is not based on ascribed characteristics like the gender, age, or ethnicity of the person that makes it [Mer73, Col79]. Inequality among persons belonging to different groups constitutes *horizontal inequality* and is dysfunctional [Ste05]. Gender has a strong influence on inequality patterns because, as a habituated principle of distinction, it structures life in fields [BW92]. The "productivity puzzle" concerns the observation of horizontal inequality in productivity: In the little science days, women produced about half as much as men [CZ84b, CS91], particularly over the first decade of their career [RH79, Lon92].

This observation from the early days of science studies is put into perspective by large-scale analyses of the science system. Each year, women are 20 percent more likely to drop out of science than men [HGSB20]. In computer science, women do publish less than men per year on average for the first several years of employment [WLC16]. But productivity inequality almost vanishes when women and men are compared for the same career

ages, stressing the explanatory power of dropouts and the importance of life course approaches [JKLW18a, HGSB20, AL20]. However, even when the survival bias is removed, women have fewer publications than men when they become a professor [LS16, ARPS11b]. Women are less likely to take prestigious author positions on publications [WJK+13, HSFH18], yet they are more likely to perform better in the job market [WLC16]. Horizontal inequality in recognition has also been reported [CZ84b, LPKL12, LNG+13]. While [RH79] concluded that there is evidence for discrimination, gender disparities can also be the unintended outcome of a plethora of contributing factors [CS91, CW11, WLC16, AL20, HJCL20].

Regarding the explanation of horizontal and vertical inequality, the literature has identified an endogenous process of reproduction as the main mechanism that operates behind multiple interacting factors: the Matthew Effect [DE06, Per14, BVR18]. In his explanations of advancement in academic careers, [Mer68, Mer88] referred to the *Matthew Effect* (ME) as a *cumulative advantage* process according to which "initial comparative advantages of trained capacity, structural location, and available resources make for successive increments of advantage such that the gaps between the haves and the have-nots in science ... widen until dampened by countervailing processes." [Mer88, p. 606] The larger the ME (i.e., the more reward or recognition is a function that attributes positive returns to individual status), the more "the rich get richer rendering the poor relatively poorer" [Pag15, p. 34].

In this paper, we take a computational approach to the problem of inequalities in academia and their origins. Using bibliographic data on the whole field of computer science, we define cohorts from 1970 to 2000 and study the careers of authors over 15 years. We find that vertical inequality in productivity is slightly increasing over the course of academic careers, inequality in recognition is larger but stable, and these trends are invariant as the field matures as a big science. Horizontal gender inequality exists, but recognition inequality finds an explanation in productivity inequality. Over the decades, we diagnose the emergence of an imperative to "publish or perish." We identify the operation of the ME as a mechanism that accumulates productivity and recognition advantages but not simply explains the observed patterns of inequality. Studying the effect of early-career author achievements and capitals on dropout and total career success, we identify two different paths to success – the "one hit" and the "steady" path – that should be evaluated in future work.

Methodologically, we arrive at these insights by leveraging the potential of systematically combining abductive, inductive, and deductive inference strategies [BT21]. Our methods include pattern description, scaling and autocorrelation analysis, and explanatory prediction with cross-validated regression analysis. We map our epistemological strategy to the integrative modeling framework recently proposed [HWA+21] and hope that our

work exemplifies the benefits of using large-scale behavioral data and methods from the formal sciences and machine learning for a more cumulative sociological knowledge production.

In the next section, we distill from the literature a middle-range evolutionary theory of careers in competitive fields that has the ME at its center and guides our analysis. Then, we present our integrative research design, discuss our results in detail, and conclude our work. For readability, materials and methods are placed at the end.

4.2 The Matthew Effect in the center of theory

Social scientific thought regarding the ME has been heavily influenced by Merton [DE06, Mer68]. For example, regarding the emergence of vertical inequality, getting a more (less) prestigious job entails an increase (decrease) in productivity [AL90]. As a consequence, the ME makes it increasingly difficult for an individual to stay in academia [CC73]. If positive feedback does not set in early in a career, the respective scholar requires a motivation to be productive for the love of the work – the "sacred spark" (ibid., p. 114–5) – or some amount of tenacity [Hub02].

To this argument, computational approaches have added a more formal and general perspective. Most directly, the ME has been demonstrated in quantifications of the extent to which past achievement (collaborators or citations) predicts current achievement [Per14, RPP18]. Scaling laws provide other vivid evidence. Career reinforcement via the ME shows as increasing returns of the average number of citations per paper as an author becomes more productive [CBvLvR09]. For highly-cited authors, staying in academia twice as long means being up to 2.8 times more productive and being up to eight times more recognized. At the same time, the ME is found to operate via author prestige below a citation threshold and via publication visibility above that threshold [PFP$^+$14]. A "barrier" that young scientists must overcome to excel has also been found analytically as part of an explanation for broad career durations [PJYS11].

Returning to the example of the role of departmental prestige in careers, it was shown that prestige operates and reproduces in networks. As a scholar climbs up the career ladder, she advances into the core of a field and becomes part of a reproductive vortex that makes it increasingly hard to *not* benefit from collective dynamics [Bur04a, CAL15, WMLC19]. Cores harbor the few positions that strongly influence how a field reproduces [Fuc01]. [PP12] introduce the concept of autocatalytic feedback to model these dynamics.

Another telling example relates to the question if there is a breakthrough moment at some point in a career. For a long time, the observation that the ME also takes the form of a cumulative disadvantage had sustained the hypothesis that success either comes early or not at all [ZM72]. Surprisingly,

though most computer scientists are most productive in their fifth year (after hiring), there is a huge variance in productivity career patterns [WMCL17]. And success can come at any time in a career, but it depends on persistence, ability to excel, and, last but not least, luck [SWD+16b].

Studies predicting the success of scholars or publications have found that current productivity and recognition [AAK12, PPP+13, Maz12, DJC15], combined with an intrinsic "fitness" of individuals to reproduce [WSB13], and mediated through networks [SPS+14b] are positively correlated with future success. The observation that the early career of a scientist is predictive of her or his later success and gains in predictive power diminish as more career ages are used for prediction provides further evidence for the ME [Maz12, PPP+13, WSB13].

The literature thus portrays the ME as a feedback mechanism that generates vertical inequality. This process is central to a middle-range evolutionary *theory of careers* in competitive fields that is taking shape at the intersection of the social and computational sciences. It is a *field theory* [Bou88] because the academic fields, as settings that delimit agents' social positions and interactions, are the loci that harbor the ME [WOSMP04]. Emerging from collective action, field structure acts as a memory in which advantages accumulate and lead to institutionalization [PP14, Fla17, PPPF18]. This field-endogenous feedback process operates behind (i.e., it reinforces or impedes) life-course variables like creativity, self-perceptions, dispositions, access to resources, and environmental conditions [CC73, CS91, PP12]. *Competition* for ideas, positions, and funds results. Careers are tournament-like endeavors [Sø86] with the goal to improve one's rank in the academic "pecking order" [Cha80]. Ranks translate to positions in networks, and upward or downward mobility resembles approaching or withdrawing from network cores [Bur04a, CAL15]. Only few make it up those "chains of opportunity", for most the way is down [Whi70]. As an *evolutionary theory*, it looks for path dependence and the long-term consequences of initial conditions [CS91, Wra11]. Small differences in ability, persistence, or luck accumulate and lock a career into an upward or downward path [PRSP12, WMLC19]. Since this is a collective phenomenon, good ideas can fail if they are put forth at the "wrong time" [New09, BJS11], but if the time is "right," success breeds success in an avalanche-like way [MEH+11]. Finally, it is a *middle-range theory* that is capable of explaining life courses and the emergence of inequality in general [Dan87, O'R96, FS09].

This theory prepares the ground for understanding horizontal (gender) inequality as co-generated by the ME [LF95, DE06]. Some or many of the career variables exemplified above are likely to be gender-correlated and thus generate outcome differences as they interact with the ME [XS98, CS91]. For example, absence from the job market (e.g., because of motherhood) leads to disadvantages which accumulate [Col79, DE06]. And women's disadvantages grow early in a career [RH79, Lon92].

4.3 Research design

We identify two research gaps. The first relates to cohort design and data availability. Older analyses tend to have sound cohort designs but are often restricted in the amount of data (number and size of cohorts) that were studied. For example, [ZM72] only analyze one cohort, while [ALK82] analyze three cohorts. More recent computational analyses tend to study large amounts of data but are often restricted regarding cohort design. For example, [PPP+13] aggregated scientists that started their career in the same decade. [PFP+14] group people into one cohort that published their first paper in a competitive journal within the same 15 years. These cohorts are heterogeneous with respect to career age and do not include unsuccessful scientists and early career researchers.

The second research gap relates to the field's transformation from little science to big science. This forces one to be explicit about system growth, that is, growth of the computer science discipline. [Pri63] argued that recruiting more people into science implies that less talented people will enter. [ZM72] hypothesized that this leads to larger differences between the most and the least talented one, suggesting that inequality should be higher in more recent cohorts than in older ones. Addressing this is important as winners in today's science have relatively more to gain than those in the past [Xie14]. But much of the research spawned by Merton happened before science changed its face [Mer68].

Our work is an attempt at an integrated modeling approach to vertical and horizontal inequality in an academic field. By "integrated" we mean that we are interested in both explaining and predicting inequality [HWA+21]. We study 15-year careers in the entire computer science discipline for cohorts from 1970 to 2000. Starting with descriptive modeling, we explore the evolution of vertical and horizontal inequality regarding productivity and recognition. In an explanatory modeling step, we then mount the evidence that the ME is the underlying causal mechanism that generates the patterns of vertical inequality we observe. In a predictive modeling step, we delineate the early career of computer scientists and inquire how accurately it predicts total-career achievements. Finally, we integrate the insights from the previous modeling steps. That is, we identify the meritocratic and non-meritocratic career factors that predict whether an author drops out of the field after the early career stage or how successful she or he will become by the end of a 15-year career. Explanations of vertical and horizontal inequality then derive from the causal assumption that the ME accumulates advantages and, in the process, interacts with these career factors.

We study the field of computer science. This makes for an interesting case because the field is relatively young, in ongoing transformation, and a driver of the digital revolution. Last but not least, there are potentially large gender disparities since only one out of five computer scientists is

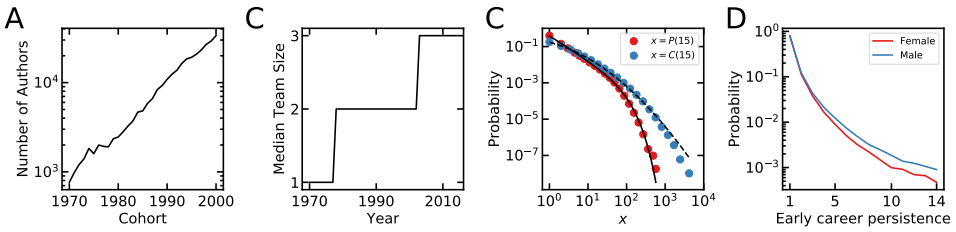


Figure 4.1: Context for the field of computer science. (A) The size of cohorts increases exponentially with time. (B) The median team size, measured from the number of authors per paper, increases over time. (C) Distributions of productivity (cumulative number of papers P per author at career age 15) and recognition (cumulative number of citations C per author at career age 15) are broad. The lines are best fits to the data: a truncated power law ($P(15)$) and a stretched exponential ($C(15)$). (D) The number of authors decreases with the number of years during which they publish persistently after the beginning of their careers (early career persistence). Female scientists show equal persistence in early career but after 4 years they are less likely to persist. Our method allows us to infer a binary gender attribute (see "Materials and methods").

female [LKW+19]. As main data source, we use DBLP, a comprehensive collection of computer science papers that were published in major and minor computer science outlets [Ley09]. We study cohorts from 1970 to 2000, where an author belongs to a cohort if they have published their first paper in the given year. For each cohort, we study careers over 15 years, including the start year. We measure productivity in terms of the number of publications since publications are the vehicles of academic communication [Mer68] and recognition in terms of the numbers of citation and h-index, two widely used measures of scientific impact [Mer88]. For details of our methods, we refer to the "Materials and methods" section at the end of the paper. Selected results obtained from the DBLP dataset [WLC16, JKLW18a] have been reported above.

Our cut of the DBLP dataset consists of 2.5 million publications from 1970 to 2014 that are authored by 1.4 million authors. Of those, about 300,000 authors started their career between 1970 and 2000 and are counted as cohort members. There are 7.9 million citations among publications which we use for all recognition analyses. Figures 4.1A and 4.1B show that cohorts grow exponentially with time and that the field is becoming a team science in the process (both signatures of a transformation into a big science). Vertical inequality at the most aggregate level (all publications and citations accumulated over an author's 15 year career, aggregated for all cohorts) is depicted via broad probability distributions. The citation distribution is broader than the productivity distribution, i.e., inequality in recognition

is higher (figure 4.1C). Correspondingly, the Gini coefficient, our measure of vertical inequality, is larger for recognition (0.83) than for productivity (0.68). This is not surprising since authors are physically constrained about the number of projects they can work on during any year but there are no such restrictions when it comes to the number of citations their work receives. The last plot shows early career *persistence*, that is, the number of career years during which an author publishes consecutively from the beginning of their career. Most authors persist for only one year before they become inactive (for at least a year) or drop out of computer science. Long persistence is decreasingly likely, especially for female scientists (4.1D).

4.4 Results

4.4.1 Vertical inequality over time

In the first, *descriptive* modeling step, we explore the evolution of vertical and horizontal inequality regarding productivity and recognition. If the ME is in place, how would inequality change over time? Intuitively, one might assume that inequality should increase if the rich get richer and that an increase in productivity inequality should directly translate to an increase in recognition inequality. This is what [ALK82] expect and find in their study of the aforementioned chemistry cohorts from the 50s and 60s. But they also find that the way of counting publications and citations – window vs. cumulative counting – is decisive. They find stable recognition inequality for cumulative counting. Increases are found only for window counting. Therefore, we discuss results regarding the evolution of vertical inequality and the comparison among cohorts using both 3-year window counting and cumulative counting of publications and citations. Our measure of vertical inequality is the Gini coefficient.

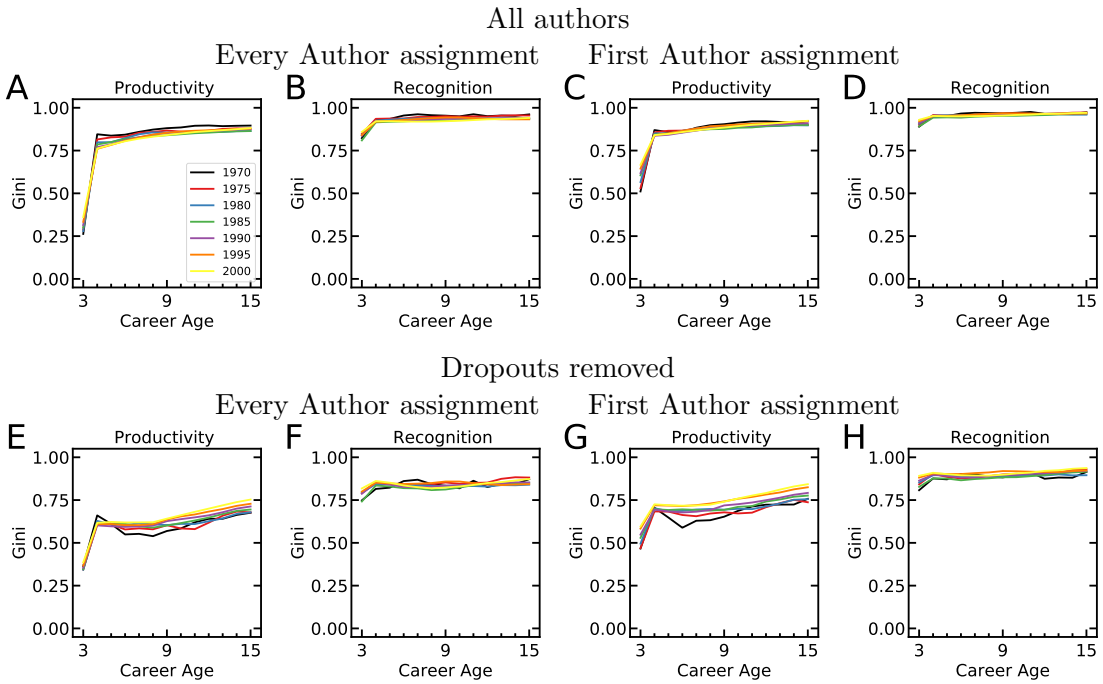


Figure 4.2: Inequality over career ages (window counting). Vertical inequality in productivity and recognition as a function of career ages, depicted for seven cohorts between 1970 and 2000. We count publications and citations in 3-year publication windows (given career age plus previous two career ages, $p_{3\mathrm{yr}}(t)$ and $c_{3\mathrm{yr}}(t)$, defined in "Materials and methods: Vertical inequality"). (First two columns) Assigning publications to all authors. (Last two columns) Assigning publications only to first authors. (Second row) Authors are filtered that have not published for ten consecutive years (most likely left academia). Inequality in recognition is always larger and more stable over the course of a career than inequality in productivity.

For window counting, we find that productivity inequality is slightly increasing over career years (figure 4.2A) while recognition inequality is larger but mostly stable (4.2B). We study several modifications to validate this finding. The sudden increase at career year 4 that can be seen in almost all subfigures is due to the fact that the career of an author starts with the first publication (i.e., in career year 1 every author has at least one publication while even those authors that eventually become highly cited may still have zero citations). As we saw in figure 4.1D, many authors drop out of academia early on, but their publication and citation counts influence the Gini coefficients. We introduce the convention that a *dropout* is present if an author is absent for at least ten consecutive years. When we remove dropouts[1] (figure 4.2E) then author careers are more comparable and productivity inequality drops, but the increasing trend remains. When these filters are applied to measuring recognition, the inequality level also drops but the stable trend does not change (figures 4.2F and J).

In computer science, the order of authors is typically important. The first author usually did the most valued part of the work. Hence, in our analysis, attributing publications only to first authors serves the purpose of studying scholars of heightened importance.[2] When we apply the first-author filter, we find marginally higher levels of inequality for productivity and recognition but again the same trend over career ages. In appendix 4.A, we report Gini trends for cumulative counting. Coefficients are systematically lower, there is an alleged strong increase of productivity inequality (figure 4.A.1A), and even a decrease in recognition inequality (figure 4.A.1B). But the modifications show that these are counting effects from carrying along authors that left academia in their early careers. All in all, no contradicting evidence is found. That means, our main result is robust: Productivity inequality is slightly increasing over career ages while recognition inequality is larger but mostly stable.

[1] Results are qualitatively similar for absences of five and ten consecutive years.

[2] In our cut of the DBLP dataset, 69% of all publications have author lists that are not alphabetically sorted. Since an author ranking by importance can be alphabetic by chance, the fraction where the author ranking is indicative of importance will be even higher.

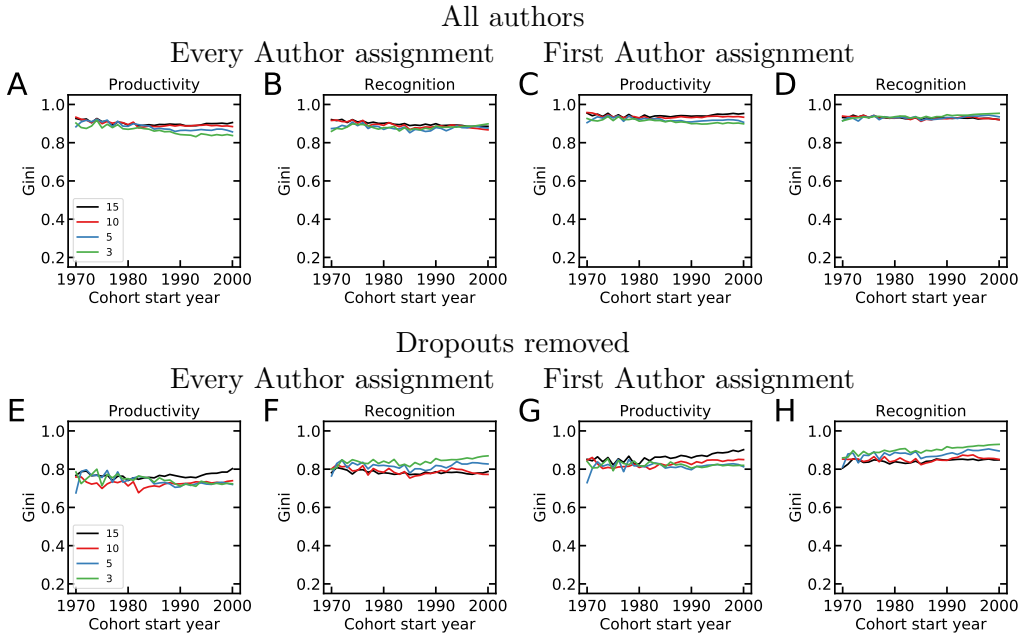


Figure 4.3: Inequality over cohorts (window counting). Vertical inequality in productivity and recognition as a function of cohort start year, depicted for career ages 3, 5, 10, and 15. We count the number of publications authored in a career age and the number of citations received in a career age by all publications authored until and in that career age ($p(t)$ and $c(t)$, defined in "Materials and methods: Vertical inequality"). (First two columns) Assigning publications to all authors. (Last two columns) Assigning publications only to first authors. (Second row) Authors are filtered that have not published for ten consecutive years (most likely left academia). Inequality is surprisingly stable over cohorts though they vary in size and the field has evolved over 45 years.

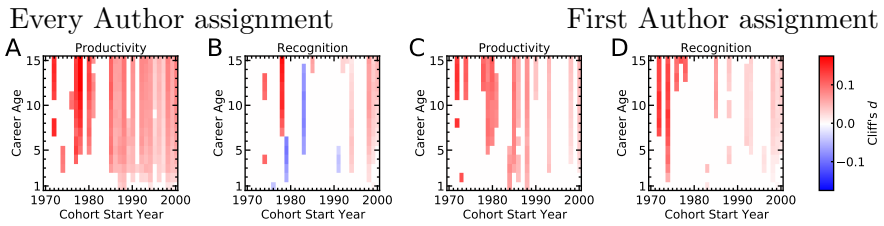


Figure 4.4: Gender differences. Horizontal inequality for productivity and recognition as a function of cohort start year and career ages. We compare the cumulative publications distribution $P_{\text{gender}}(t)$ and cumulative citations distribution $C_{\text{gender}}(t)$ of male and female scientists in the same cohort at the same career age t and test differences between these distributions. Color marks the effect size (Cliff's d). Positive values (red) indicate that men dominate women, while negative values (blue) reveal that women dominate men. Effects are only shown if they are significant ($p \leq 0.05$) according to a Mann–Whitney U test. Details in "Materials and methods: Horizontal inequality." Publications are assigned to all authors (A, B) or first authors only (C, D). In general, effects decrease with cohort start year and increase with career age.

Now turning to the historical analysis over cohorts, we address Zuckerman's ([ZM72]) hypothesis that recruiting more people into science will lead to larger differences between the most and the least talented one. Our results do not support this hypothesis since we do not see an increase in productivity and recognition inequality over cohorts (figure 4.3). As before, removing dropouts and restricting the analysis to authors with early-career persistence reduced inequality levels but does not alter interpretation.

4.4.2 Horizontal inequality over time

Increasing vertical inequality in science is not necessarily problematic if the evaluation is based solely on merit rather than on functionally irrelevant factors such as gender, race, nationality, age, or class. Due to its societal importance, we focus on gender inequality. Figure 4.4 shows a systematic comparison of the cumulative productivity and recognition distributions of male and female computer scientists, for every author (A and B) and first author (C and D) assignment. Positive values (red) indicate that the distribution of men is dominant, that is, men are more productive or recognized. Negative values (blue) reveal that the distribution of women are dominant.

There is a general pattern: horizontal inequality seems to accumulate and is more prevalent in the later career stages of scientists. If there are differences in productivity, it is always men publishing more. This gender productivity gap exists in almost all cohorts (figure 4.4A). For horizontal inequality in recognition, the picture is less clear. Female or male dominance

both exists sporadically in cohorts. In four cohorts, women are statistically more likely to have more citations than men; for the 1982 cohort even for ten consecutive career years. There is no cohort in which horizontal inequality shifts sign, that means, it is always one cohort's gender that is dominant (figure 4.4B). In total, gender inequality is more pronounced for productivity than for citations. For cumulative numbers of publications, 55% of 465 cohort-age pair differences are statistically significant; for cumulative numbers of citations 19% are significant. That means, the productivity gap does not automatically translate to a recognition gap. However, when there is a recognition gap it can be explained by a productivity gap: significant differences in citation are strongly correlated with differences in publications ($r = 0.91, p \leq 0.001$). As [AL20] found, the productivity gap is the puzzle to solve.

Limiting authors to first authors is one step towards solving this puzzle. When we restrict authors to "important" ones, the magnitude of the gender gap becomes smaller (28% significant cohort-career year pairs). This is particularly the case for the more recent cohorts. This suggests that in "big" team-based computer science, male scientists have boosted their productivity more via collaborations than female scientists, since larger differences in productivity between male and female scientists diminish when only first-author contributions are counted. Applying the first author filter makes the recognition gap a purely male phenomenon but also a phenomenon of the 70s (figure 4.4D). More pronounced gender roles likely contributed to a recognition gap which still finds explanation in a productivity gap (significant differences in citation are still strongly correlated with differences in publications, $r = 0.88, p \leq 0.001$). In sum, horizontal inequality exists. While it appears to be diminishing on the timescale of cohorts, there must be a mechanism that causally explains it on the career timescale.

4.4.3 Matthew Effect underlies careers

In the *explanatory* modeling step that now follows, we inquire if the ME can generate the patterns of vertical inequality we observe. The ME states that present achievement (productivity or recognition) causally depends on past achievement. This feedback process makes it easier to produce large numbers once they have been produced and amounts to an advantage that can accumulate over time. Our guiding theory describes this feedback process as a vortex, an autocatalytic mechanism that fuels itself [PP12]. This process is causal in the sense that emergent distributions of achievement influence future distributions [Fla17]. When the ME is fully operational – formally: when it is linear – it generates power law distributions which signal the absence of a characteristic scale [AB02]. In our case of computer science, productivity and recognition distributions are broad but not pure power laws. Distributions for individual cohorts are much like the (trun-

cated power law and stretched exponential) distributions that we measure when all cohorts are lumped together (figure 4.1C). These deviations can result from a quenched (sublinear) ME and other mechanisms and factors that interact with the ME but also from sampling and finite size effects intrinsic to the DBLP database.

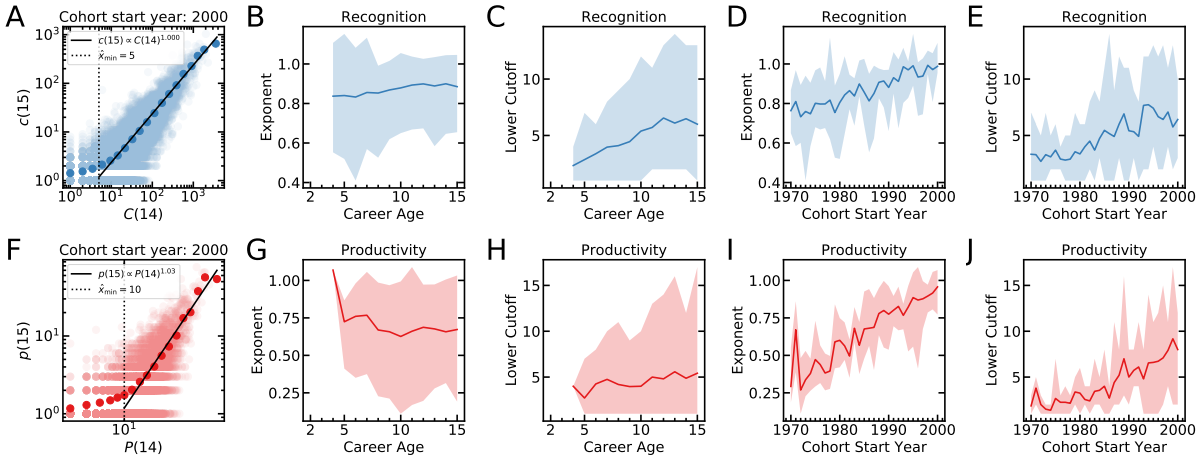


Figure 4.5: Matthew Effect. (First column: A, F) Measurement of the strength of a cohort's reproductive feedback as the exponent that relates an author's number of citations received, or papers produced, in a career age (y-axis) to the respective cumulative numbers in the previous career age (x-axis), shown for the 2000 cohort and the last career age. Exponents show as slopes of the continuous lines. Dotted lines indicate that feedback fully unfolds only above a lower cutoff. (B, G) For an average cohort, potential individual advantages from feedback are constant along the career path, for both recognition and productivity. (C, H) For an average cohort, the number of publications and citations required to take advantage from feedback increases along the career path. (D, I) For an average career age, potential individual advantages from feedback increase historically, but more so for productivity. (E, J) For an average career age, the numbers of citations and publications required to take advantage from feedback increase historically. (All columns but the first) Shaded areas are bounded by minima and maxima, lines show means.

We quantify the strength of the reproductive feedback of the field – the strength of the vortex – that a cohort experiences in a career age by regressing the number of publications or citations in a career age on the corresponding cumulative number in the previous career age (details in "Materials and methods: Reproductive feedback"). We interpret two parameters. The exponent of the scaling relationship quantifies the strength of reproductive feedback. An exponent that is larger than zero over time is indicative of a cumulative advantage. The lower cutoff states at and above which number of publications or citations the advantage accruing from past selection fully unfolds. It resembles the boundary to the basin of attraction of the feedback dynamics: once an author crosses it, she or he gets attracted by the reproductive vortex and advantages can accumulate. Examples of the fitting procedure are depicted in figures 4.5A and F. They show that scaling relationships are plausible fits to the data.

Our results show that the ME is present for productivity and recognition since all exponents are larger than zero. For an average cohort, the strength of the ME is stable over an author's career, allowing for a constant cumulative advantage. This holds true for recognition and productivity as there are no discernible trends in figures 4.5B and G. To enter the productivity basin of attraction (i.e., to actually reap benefits), an author must produce a certain number of publications that is constant over career ages (no trend in figure 4.5H). That means, regarding productivity, it is equally possible for an early- or late-career author to benefit from autocatalytic feedback. On average, all they need to do is produce a cohort-specific number of publications. However, getting one's publications cited becomes increasingly difficult as careers progress since the lower cutoff increases with career age (figure 4.5C). Regarding recognition, moving early is advantageous.

While the strength of the ME is stable over a computer scientists career, it does increase at the historical timescale of cohorts. Nowadays, the ME is fully operational for both recognition and productivity (the exponents in figures 4.5D and I are ≈ 1 for the 2000 start year). But the practices of authorship and citation started at different levels. Whereas the 1970 cohort already experienced a strong effect from past citations (exponent ≈ 0.8), the effect of the past number of publications started off weak (≈ 0.3). In other words, while getting cited has long been endowed with a strong reinforcement effect, becoming productive constantly became so over three decades. At the same time, the lower cutoff for reinforcement to set in has been growing historically for both recognition (figure 4.5E) and productivity (4.5J). In a field becoming a big science, this is likely how limited resources are distributed among an increasing number of scholars. For the authorship practice, this mechanism has a name: "publish or perish" [Gar96].

The ME underlies career dynamics in computer science. While we demonstrate that cumulative advantage is an active mechanism, it can only explain the most general finding regarding vertical inequality: A ME that is per-

sistently stronger for recognition than for productivity generates a level of vertical inequality that is persistently higher. However, reproductive feedback alone is not capable of explaining the inequality patterns we observe. Increases in vertical inequality would have to correlate with increases in cumulative advantage [ALK82]. That we do not find such a correlation is likely because inequality dynamics are highly complex due to many interacting factors.In principle, any author can benefit from the feedback dynamics of the ME. That we observe vertical inequality is due to the fact that authors benefit from it to a different extent. Crossing or not crossing the boundary to the basin of attraction is one explanation. Crossing the boundary sooner than later is another explanation that points to the importance of the early career. Our theory of careers states that horizontal inequality results from gender differences in the early career that are then amplified by the ME. We continue our modeling flow by taking into focus a variety of factors that have been shown to influence academic careers. In particular, we turn towards the importance of these factors in the early career in shaping the total-career outcome.

4.4.4 Prediction of dropout and success

Delimiting the early career

We study the effect of nine factors that describe either the early-career achievements of authors or their social, symbolic, and cultural capital in the early career. The constructs are fully described and operationalized in the section "Materials and methods: Independent variables" and summarized

Table 4.1: Independent variables used in prediction models. These variables characterize authors in their early career ages $[1, t_e]$. The variables are used to predict the success of authors and whether they drop out of computer science for ten consecutive years. Details are given in the section "Materials and methods". The end of the early career is chosen from a success prediction to be $t_e = 3$.

Variable	Description
Baseline	
Start year	Year in which cohort members started publishing
Achievement	
Productivity	Cumulative number of publications authored in the early career
Productivity (1st author)	Cumulative number of publications authored in the early career as a first author
Recognition	Cumulative number of citations received in the early career
Gender	
Male	Dummy
Female	Dummy
Undetected	Dummy
Social capital	
Social support	Number of distinct co-authors in the early career
Team size	Median number of authors of all publications produced in the early career
Senior support	Largest h-index of all co-authors in the early career
Symbolic capital	
Top venue	Smallest $h5$-index-based quartile rank of all journals and conference proceedings an author has published in in the early career
Cultural capital	
Ability	Number of citations that the publications produced in the early career accumulate until the end of the career
Ability (1st author)	Number of citations that the publications produced in the early career as a first author accumulate until the end of the career

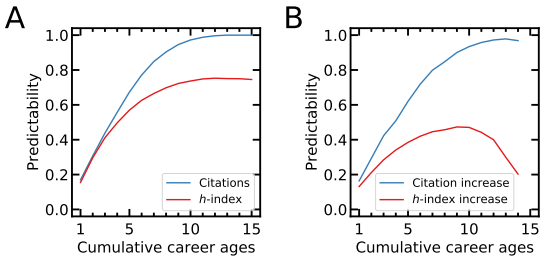


Figure 4.6: Predictability of success: Predictability is measured as the adjusted coefficient of determination R^2. Values are averages over all 31 cohorts. (A) Prediction of the cumulative number of citations $C(15)$ and the h-index $h(15)$ at the end of the career, using all independent variables in table 4.1 for varying time windows. $C(15)$ can be perfectly predicted from 15 career ages due to autocorrelation with the recognition variable. After the first 3 years, success after 15 years is predictable to 40%. (B) Predicting the citation and h-index increases $C^+(15)$ and $h^+(15)$ removes the autocorrelation. After a certain career age, predictability decreases because the predicted increases diminish. Predicting a zero increase is trivial.

in table 4.1. We control for the cohort start year (i.e., we are interested in changes that occur on the historical time scale). Obviously, to assess potential gender differences, we also include the gender of authors.

The question is: Which career ages resemble the early career? To identify the last career age t_e of the early career, we perform a first regression analysis. This third modeling step is purely *predictive*, that is we are not yet interested in the effects of the career factors just introduced. For now, we are only interested how predictable author success in career age 15 is as we use all factors in a black box and use an increasingly long early career $[1, t_e]$ for prediction. As success measures we use the cumulative number of citations and the h-index, a metric that integrates productivity and recognition. The dependent variables as well as the prediction model are also described in the "Materials and methods" section.

Figure 4.6A reports these predictions for an average cohort. Consistent with the literature [Maz12, PPP+13, WSB13], the concavity of both curves indicates that increases in predictability diminish as more career years are used for prediction. This means that those that eventually became successful did, on average, set the seeds for their success in their early careers. Therefore, the curves are indirect evidence for the ME as the mechanism that amplifies initial differences. The citation success and h-index at career age 15 are predictable to 44% and 41%, respectively, after three years. It is, therefore, true that, on average, success comes early. But this does not exclude that it may also come later [ZM72]. We chose $t_e = 3$. Note that, to determine the $[1, 3]$ early career interval, we made use of the autocor-

relation intrinsic to the prediction because of the independent recognition variable which becomes a perfect predictor for an early career of 15 years. In later success predictions, we predict the increases in citation scores and h-indices to remove any direct autocorrelations. Figure 4.6B shows that, when increases are predicted, the early career is slightly less predictive on average (curves are less steep) and there is now a strong limit of 47% to the predictability of h-index increases.

Dropout prediction

We proceed with the final step of *integrative* modeling: causal explanations of dropout and future success in out-of-sample predictions. These explanations are causal since we have assured that the ME is a safe assumption as it theoretically interacts with early-career factors. And they are predictive

Table 4.2: Dropout prediction. Each column corresponds to a separate logistic regression model that aims to predict whether (1) or not (0) an author dropped out of computer science (described in section "Materials and methods: Prediction models"). Dropout is predicted from the achievements and types of capital accumulated in the early career of the first three career ages. Cohort start year and gender are controlled for. Coefficients are reported as means (with standard deviations in brackets) from 10-fold cross validation. Goodness-of-fit measures (F1 and average precision) are also means across all folds.

3.5pt

	Model 1 *Baseline*	**Model 2** *Achievement*	**Model 3** *Gender*	**Model 4** *Social Capital*	**Model 5** *Symbolic Capital*	**Model 6** *Cultural Capital*
Start year	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
Productivity		-0.57(0.00)	-0.58(0.00)	-0.55(0.00)	-0.54(0.00)	-0.53(0.00)
Productivity (1st)		-0.28(0.00)	-0.27(0.01)	-0.22(0.01)	-0.21(0.00)	-0.20(0.00)
Recognition		-0.00(0.00)	-0.00(0.00)	0.00(0.00)	0.01(0.00)	0.07(0.00)
Female			0.05(0.01)	0.06(0.00)	0.07(0.00)	0.06(0.00)
Male			-0.06(0.00)	-0.07(0.00)	-0.08(0.00)	-0.08(0.00)
Undetected			0.02(0.00)	0.02(0.00)	0.02(0.00)	0.02(0.00)
Social support				-0.08(0.00)	-0.08(0.00)	-0.08(0.00)
Senior support				-0.03(0.00)	-0.01(0.00)	-0.00(0.00)
Median team size				0.22(0.00)	0.22(0.00)	0.22(0.00)
Top source					-0.31(0.01)	-0.29(0.01)
Ability						-0.02(0.00)
Ability (1st)						-0.02(0.00)
F1	0.44	0.68	0.68	0.68	0.68	0.68
Average precision	0.58	0.74	0.74	0.76	0.76	0.76

since we evaluate the learned coefficients on unseen data. Table 4.2 shows the results of a logistic regression model that uses dropout as a binary dependent variable. The most important factor for predicting dropout is early career productivity. Scientists that publish much in their first three career ages, not necessarily as a first author, are less likely to drop out. This is not surprising, given that dropout is defined as the absence of publications for ten consecutive career ages. While predictive accuracy hardly improves as more factors are added, other career factors do exhibit interpretable correlations. Publishing in a top source is a strong predictor of not dropping out. Having a publication in a top journal or conference proceedings is a symbolic capital that likely breeds further productivity. Social capital is a multi-faceted construct category. On the one hand, co-authoring publications with many others is positively correlated with dropout. This tells us that being one among many is not automatically an achievement; writing as a first author is. On the other hand, authors that stay in computer science have larger social support groups. Women are more likely to drop out than an average computer scientist. Having early senior support (a co-author with a high h-index) only makes an author slightly less likely to drop out.

Remarkably, dropout is not associated with the early-career recognition of early-career publications. Having many early citations does not make it more likely for an author to stay in computer science. We also experiment with a variable intended to measure an author's individual ability to excel, operationalized as the total-career recognition of early-career publications. As we see in all regressions (tables 4.2-4.4), this individual ability is correlated with early-career recognition. We see this because the latter becomes slightly predictive of dropping out (due to Elastic Net regularization) when ability is added in model 6. Interestingly, the cohort has no effect. Whatever the cause, the dropout-related patterns we observe are historically invariant on average.

Success prediction

Next, we study which factors can explain and predict how success in terms of recognition increases, on average, after the first three career ages (table 4.3). The dependent variable is the increase in citations until career age 15. The cohort has a small positive effect on success. This is probably an effect of the exponential growth of the field: As more publications are produced and reference lists become longer, more citations are made and accumulated [PPPF18]. This is controlled for in the following considerations. Early career productivity is a requirement for recognition, and its effect is at par with that of recognition. This resembles our earlier observation that total success is well predictable from early success (figure 4.6). Other than for dropout, early senior support is an important factor for a successful career. But similar to dropout prediction, publishing in large teams exhibits a negative

effect on citation success, and social support has a weakly positive effect.

Until and in model 5, being female is a negative predictor for an increase in citations. Similarly, publishing in a top source is negatively correlated with success. However, both effects vanish when the ability variable is added in model 6. Adding it also entails a huge increase in predictability. Ability, the total-career recognition of early-career publications, strongly explains citation increase. This is an interesting effect because, while ability is strongly correlated with recognition (as discussed above), it is not autocorrelated with the dependent variable (the total-career recognition of mid- and late-career publications). However, the effect is much smaller for ability as a first author. This suggests that what we construe as cultural capital is not an intrinsic individual attribute but an ability to excel in a team of authors. The explanation for the negative association of success with being

Table 4.3: Success prediction (citation increase). Each column corresponds to a separate linear regression model that aims to predict $C_i^+(15)$, the increase in citations an author gains after the early career of the first three career ages (described in section "Materials and methods: Prediction models"). Citation increase is predicted from the achievements and capitals accumulated in the early career. Cohort start year and gender are controlled for. Coefficients are reported as means (with standard deviations in brackets) from 10-fold cross validation. Goodness-of-fit measures (mean squared error and adjusted R^2) are also means across all folds.

3.5pt

	Model 1 *Baseline*	**Model 2** *Achievements*	**Model 3** *Gender*	**Model 4** *Social Capital*	**Model 5** *Symbolic Capital*	**Model 6** *Cultural Capital*
Start year	0.08(0.00)	0.06(0.00)	0.06(0.00)	0.05(0.00)	0.05(0.00)	0.03(0.00)
Productivity		1.01(0.01)	1.01(0.01)	0.88(0.01)	0.89(0.01)	0.68(0.01)
Productivity (1st)		0.45(0.01)	0.45(0.01)	0.45(0.01)	0.45(0.01)	0.36(0.01)
Recognition		1.01(0.01)	1.00(0.02)	0.91(0.01)	0.91(0.02)	-0.20(0.01)
Female			-0.09(0.01)	-0.10(0.01)	-0.10(0.01)	0.00(0.00)
Male			0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
Undetected			0.07(0.01)	0.04(0.01)	0.04(0.01)	0.00(0.00)
Social support				0.05(0.01)	0.05(0.01)	0.01(0.01)
Senior support				0.70(0.01)	0.70(0.01)	0.50(0.01)
Median team size				-0.09(0.01)	-0.09(0.01)	-0.02(0.01)
Top source					-0.09(0.01)	0.00(0.00)
Ability						0.58(0.01)
Ability (1st)						0.04(0.00)
Intercept	-158.94	-127.67	-128.13	-94.94	-95.42	-50.4
Mean squared error	40.17	31.97	31.97	31.54	31.55	22.94
Adjusted R^2	0.01	0.21	0.21	0.22	0.22	0.43

female in models 3-5 is, thus, not that women are less able to excel but less able to reap benefits from working in teams.

This prediction of late-career success is clearly influenced by our choice to count recognition and success in terms of total citations. Therefore, we put it into perspective by another set of models, shown in table 4.4. The independent variables stay the same, but now the dependent variable is the increase in h-index until career age 15. The h-index is a measure that compounds productivity and recognition: the maximum number h of publications that each have at least h citations [Hir05a]. Essentially, this indicator is an index of persistent success. As such, it quantifies another kind of career outcome. In these models, the effect of start year almost vanishes, that is, the dependent variable is less sensitive to field growth. Productivity matters to a lesser extent than in citation increase prediction. Most notably, authors whose early work is highly cited hardly benefit from this recognition in the long run. These insights are both trivial and revealing. They are trivial because the compound h-index requires high productivity to be high. But they are also revealing because recognition is not a predictor

Table 4.4: Success prediction (h-index increase). Each column corresponds to a separate linear regression model that aims to predict $h_i^+(15)$, the increase in h-index an author gains after the early career of the first three career ages (described in section "Materials and methods: Prediction models"). Independent variables and model description as in table 4.3.

3.5pt

	Model 1 *Baseline*	**Model 2** *Achievements*	**Model 3** *Gender*	**Model 4** *Social Capital*	**Model 5** *Symbolic Capital*	**Model 6** *Cultural Capital*
Start year	0.03(0.00)	0.02(0.00)	0.02(0.00)	0.02(0.00)	0.01(0.00)	0.01(0.00)
Productivity		0.43(0.00)	0.43(0.00)	0.38(0.00)	0.36(0.00)	0.34(0.00)
Productivity (1st)		0.15(0.00)	0.14(0.00)	0.14(0.00)	0.14(0.00)	0.13(0.00)
Recognition		0.03(0.00)	0.03(0.00)	0.00(0.00)	-0.01(0.00)	-0.12(0.00)
Female			-0.00(0.01)	-0.04(0.00)	-0.02(0.00)	-0.00(0.00)
Male			0.00(0.00)	0.00(0.00)	0.01(0.00)	0.01(0.00)
Undetected			0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
Social support				0.04(0.00)	0.04(0.00)	0.04(0.00)
Senior support				0.23(0.00)	0.22(0.00)	0.2(0.00)
Median team size				-0.11(0.00)	-0.09(0.00)	-0.09(0.00)
Top source					0.36(0.00)	0.34(0.00)
Ability						0.06(0.00)
Ability (1st)						0.01(0.00)
Intercept	-56.42	-41.84	-41.63	-31.1	-29.27	-24.59
Mean squared error	2.39	1.81	1.81	1.75	1.73	1.64
Adjusted R^2	0.01	0.25	0.25	0.28	0.29	0.32

anymore. It means that persistently producing publications is another path to success. The effect sizes in the social capital category are comparable in magnitude to those in the citation increase predictions (i.e., senior support and social support increase and large author teams decrease the likelihood of success).

Gender is not a predictor of persistent success. This suggests that the gender effect in the citation-based predictions (table 4.3) is in fact a consequence of women being less productive due to dropping out. The h-index controls for persistence and creates gender comparability. Publishing in a top source is now very beneficial to success. This effect is robust to adding the ability variable, but adding the latter again reduces the effect of early-career recognition on success. These observations are compatible with the explanation that publishing in top sources helps breed the persistent kind of success (measured with the h-index) while potential "one hit" success (measured with total citations) needs not be published in top journals or proceedings. The small effect size of individual ability adds to this interpretation: Persistence to publish suffices to be successful.

4.5 Discussion

4.5.1 Summary and conclusion

We have been interested in inequalities in computer science, their changes over author careers as well as over the field's transformation from a little to a big science, and their origins. We found that vertical inequality in recognition is generally larger than inequality in productivity. The latter increases during the careers of an average cohort but, contrary to what has been previously suggested [ALK82], does not translate to an increase in recognition inequality: The inequality patterns of computer science cohorts from 1970 to 2000 are different than those of chemistry cohorts in the 60s and 70s. Since computer science exhibits an exponential influx of personnel, we have also checked if it leads to an increase of vertical inequality on the historical time scale of cohorts [ZM72]. We did not find such an effect: The inequality patterns are the same since the 70s. The productivity puzzle [CZ84b] shows as horizontal gender inequality. We found that men produce more publications than women, particularly towards the end of the career, though this phenomenon was more pronounced in the past. Compared to the literature on little science, this is a clear continuity [RH79, CZ84b, CS91, Lon92]. Regarding reports of horizontal inequality in recognition [CZ84b, LPKL12, LNG+13], we found that men having more publications does not automatically entail having more citations, but more citations find their explanation in more publications.

To guide our interpretations, we connect various theoretical and empirical strands to a middle-range evolutionary theory of careers in competitive

fields. According to this, fields contain cores that harbor the few positions that are contested by those wanting to make a career. The Matthew Effect (ME) is central to field dynamics: As a field autocatalytically reproduces, its core behaves much like a vortex that accumulates a scholar's advantages once she or he enters the vortex's basin of attraction [Fuc01, PP12]. Consistent with the general magnitudes of vertical inequality, we found that the ME is stronger in terms of citations than publications. For both categories, it is stable over an author's career. Regarding recognition, it becomes harder to benefit from it the more a career progresses (i.e., there is an early-citation advantage). Cumulative advantages from past achievements increase historically to unity, but regarding productivity, the effect started off weakly, indicating the rise of the imperative to publish or perish.

The early-citation advantage, the increasing necessity for persistence in publishing, and the fact that the first three career ages are already quite predictive of total-career success support our theory. It further states that the origins of inequalities are also to be found in the ME's interactions with intrinsic and behavioral author characteristics. Inquiring about the importance of early career achievements and capitals in shaping total-career outcomes, we found that early productivity is the one best predictor for not dropping out of computer science and for recognition-based success. While staying in the field is a condition for success, our regression models also show that success and having successful co-authors are not conditions for staying in the field.

Author behavior and their social capital consequences exhibit consistent effects: Authors with a large social support network are more likely to stay or to be successful. But authors that are part of large co-author collectives are more likely to drop out and to be unsuccessful, potentially because more of the same in terms of team structure hinders creativity [US05b, GUSA05b]. Women are more likely to drop out of the field. Part of the explanation may be – besides factors we could not measure bibliographically (e.g., becoming a mother) – that female computer scientists embed into collaboration networks that are smaller and more cohesive than the male counterparts, as established in previous work [Jad17]. If the female type of embedding is a disadvantage, as our results suggest, the latter would accumulate due to the ME. If men manage to inflate their publication counts more than women due to having more social capital [WLC16], this can also explain another part of the productivity puzzle: that the productivity gap between women and men is smaller when only publications authored as a first author are counted.

Our results so far underline the explanatory importance of productivity [HGSB20, AL20]. Yet, there is no simple productivity-based explanation of what makes a successful career in computer science [WMCL17]. Our empirical results lead us to propose that there are (at least) two paths to a successful career. The first path mainly operates on recognition. Success

shows as a high number of citations. Since this can be achieved with a single publication, we call it the *one hit* path to success. The earlier an author gets cited, the more she or he can benefit from the ME. But recognition later in the career also helps, it is just more difficult to benefit from reproductive field dynamics. On this path, it shows not to be important to publish in top journals or conference proceedings. Theoretically, after the ME has started operating on an author's prestige, it switches to operate on publication visibility once an author comes to be known for a particular idea or contribution to computer science [PFP+14]. When this happens, the ME can spill over to other publications [MEH+11]. The second path operates on the recognition of productivity. Success shows as a high h-index. Since this requires a fair amount of publications that each are fairly cited, we call it the *steady* path to success. High citation is not required, but publishing in top sources is beneficial – probably because the ME mainly operates on author prestige [PFP+14]. The first path is likely more subject to ability and luck [SWD+16b], the second path to the determination to publish research [CC73]. On both paths, it is beneficial to have senior co-authors. Both paths, but particularly the one hit path, benefit from the field's increasing citation potential that goes along with computer science's transformation into a big science [PPPF18].

Inequality, we conclude, is not simply explained by the ME. It is a complex phenomenon that arises from many interacting factors, social capital from network embedding being one of them. But the ME is the central mechanism that governs which author (and idea) gets to take a core position and, therefore, to influence the fate of the field. On average, core authors are productive, and persistent productivity is one path to success; following a high-citation strategy is another one. Between 1970 to 2000, when computer science became a team-based big science, vertical inequality remained constant and the female productivity gap narrowed, but the ME became stronger, up to the point where it is now common to publish or perish. Finally, our research design prohibits conclusions about gender discrimination since we do not know why authors drop out of the field or do not enter it in the first place. That said, all gender effects that we detected in operating computer science can be explained by women dropping out earlier, being less persistently productive, and having less social capital that could accumulate and breed further achievements.

4.5.2 Methodological considerations

We close our paper with a few methodological considerations. They are concerned with the hope that "big" behavioral data combined with powerful statistical and machine learning ways to crunch it allows us to build a more cumulative and more formal way of social scientific knowledge production [Wat17]. The first set of considerations relates to data. Inquiries

into inequality and the ME date back to the 70s when it was only possible to study small cohorts. Much of this research was done using one of two carefully constructed bibliographic chemistry cohort datasets [ALK82]. To the contrary, we use a large-scale dataset on the complete trails of computer science. The use of bibliographic traces has allowed us to reconstruct and study scholarly careers in historical comparison. While formal communication just resembles the observable part of careers, it is undeniably an important part since careers are subject to collective field dynamics that work on what is observable.

The ability to model processes with behavioral data comes at the cost of much less being able to model the individual. Gender is an example. We contribute insights into the social construction of binary gender. In particular, we show how a structural mechanism that accumulates achievements and is agnostic to individual traits – the ME – generates gender disparities simply because women and men differ in how they embed into collaboration networks. Augmenting behavior with data on cognitive states (e.g., whether computer scientists dropped out on free terms, because of structural constraints or even discrimination) would allow for deeper insights into the origins of inequality.

Operationalizing success via the number of citations is straightforward because it very well captures that success is a collective phenomenon. On the other hand, citation scores are not unobtrusive measures anymore. Citations have become a currency in science, scholars try to improve their scores, and the databases we use for research are also used to compute a scholar's market value. This adds a dysfunctional dimension to the ME that cannot be disentangled from the mechanism that generates functional inequality [Xie14, CLS17]. What is more, deriving a dependent variable from the citation practice limited us in constructing independent variables from that practice. As a result, our attempt to use the total-career recognition of early-career publications as a proxy for an author's intrinsic ability to excel more or less turned out to be a – however meaningful – control in our prediction models.

The second set of considerations relates to methods. We have followed the integrative modeling approach [HWA+21] and found that the combination of descriptive, explanatory, and predictive modeling came about naturally. With large-scale behavioral data, exploratory description is necessary because existing knowledge may not translate into meaningful research questions or hypotheses for testing: Past small-scale studies may not have captured new phenomena, and new large-scale studies may not generalize due to preprocessing decisions and design choices. To still let the literature guide our modeling, we distilled the theory of careers from a broad set of studies. Since this theory states that inequality breeds further inequality, the descriptive modeling of figures 4.2, 4.3, 4.A.1, and 4.A.2 is an act of *abductive* inference [BT21]: Assuming the ME as a reproductive mechanism,

traces of behavior (i.e., the data) are generated by patterns of inequality.

Downstream, these descriptions informed two acts of *inductive* reasoning. In the minimal explanatory modeling of figure 4.5, we mounted the evidence for the ME by inferring its parameters from past achievement as cause and present achievement as effect. In the second act of induction, the predictive modeling of figure 4.6, we were not interested in parameter estimates but in pure predictive accuracy. This way we delimited the early career as the first three years of a computer scientist's career.

All these steps serve to prepare the integrative modeling step. We have realized this step, which could be called explanatory prediction, by employing interpretable regression models with cross validation. Obviously, we could not use a statistical model that improves predictive accuracy at the expense of interpretability [Mol19] because the model coefficients tell which career factors theoretically interact with the ME in shaping career outcomes. Generalized linear models are still a good choice. In interpreting them, critically reflecting on results and performing multiple checks and modeling alternatives proved to be essential. For example, figure 4.4 and table 4.3, when taken at face value, suggest that women are recognized to a lesser extent than men, but in both cases we could explain the effect by the survivor bias introduced by women dropping out earlier or easier than men. Cross validation turns linear models into prediction models because the goodness of fit is estimated out of sample. Epistemologically, our prediction consists of a first inductive inference procedure in which coefficients are learned (from the data divided into ten "folds") and a second *deductive* inference procedure in which these coefficients are used to predict the dependent variables in other folds. Cross validation also entails a shift from significance scores to effect sizes and their robustness across folds. In the face of large sample sizes and biases in the data (e.g., from gender inference), this helps guard against false certainties.

4.6 Materials and methods

Data: We use DBLP [Ley09, The17], a comprehensive collection of computer science publications from major and minor journals and conference proceedings. From this dump, we remove *arXiv* preprints. The coverage of DBLP ranges from 55% in the 80s to over 85% in 2011 [WLC16]. Our dataset consists of 2.5 million publications from 1970 to 2014 that are authored by 1.4 million authors. Of those, 292.443 started their career between 1970 and 2000. We have added citations among publications by combining DBLP with the AMiner dataset [WZZT19, AMi17] via publication titles and year (reference removed). There are 7.9 million citations among publications. Author names in DBLP are disambiguated [RH10].

To infer the gender of authors, we have used a method that combines

the results of name-based (genderize.io) and image-based (Face++) gender detection services. The accuracy of this method is above 90% for most nationalities. Since the accuracy is very low accuracy for Chinese and Korean names, we label their gender as unknown in order to reduce noise in our analysis (reference removed). Since authors are free to chose the name under which they publish, the inferred variable is a true, socially constructed gender attribute.

Cohorts and career ages: Our main units of analysis are cohorts of computer scientists from 1970 to 2000. We consider a career to begin with an author's first publication in the database. Since DBLP covers publication years back to 1960, this ensures that authors of the earliest cohort have been at least absent for ten years. Imbalances in coverage over publication years cause earlier cohorts to be less homogeneous as we tend to miss more first publications. Given start years, we follow cohort members over career ages $t \in [1, 15]$.

Publication and citation counts: Our unit of observation is the individual author i in a cohort. For each author and career age, we measure the number of publications $p_i(t)$ authored in a career age, the cumulative number of publications $P_i(t)$ authored until and in a career age, the number of citations $c_i(t)$ received by $P_i(t)$ in a career age, and the cumulative number of citations $C_i(t)$ received by $P_i(t)$ until and in a career age. Citations are always counted coming from the whole field of computer science, not just from the same cohort.

Vertical inequality: To quantify vertical inequality we use the Gini coefficient $G(t)$ of the publication and citation distributions of authors in the same cohort at the same career age:

$$G(t) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i(t) - x_j(t)|}{2n \sum_{i=1}^{n} x_i(t)} \tag{4.1}$$

The numerator is the absolute difference of all pairs (i, j) of authors in a cohort. x is a placeholder for publication or citation counts. In figures 4.2 and 4.3, we use backward-looking 3-year windows, that is, to quantify inequality in productivity, $x(t) = p_{3\text{yr}}(t) = \sum_{\tau=0}^{2} p(t-\tau)$, and, to quantify inequality in recognition, $x(t) = c_{3\text{yr}}(t) = \sum_{\upsilon=0}^{2} \sum_{\tau=0}^{\upsilon} c_{t-\upsilon}(t-\tau)$, where $t \geq 3$ and the index $t - \upsilon$ of c defines the career age for the publications of which citations are counted. In appendix 4.A, we use cumulative counting, that is, to quantify inequality in cumulative productivity, $x(t) = P(t)$, and, to quantify inequality in cumulative recognition, $x(t) = C(t)$. A Gini coefficient of zero expresses perfect equality, where all authors in one cohort have produced equal amount of papers or received equal amount of citations. A Gini of one indicates maximal inequality among authors.

Horizontal inequality: To quantify horizontal inequality, we look at the differences between the cumulative distributions of productivity $x(t) = P(t)$ and recognition $x(t) = C(t)$ of male and female scientists in the same career age. For both $x(t)$, we rank all observations ascendingly (with adjusted ranks for ties) and perform the Mann–Whitney U test,

$$U(t) = R_m(t) - \frac{n_m(t)(n_m(t)+1)}{2}, \tag{4.2}$$

where $R_m(t)$ is the sum of the ranks and $n_m(t)$ is the number of male scientists. The U test allows us to assess the statistical significance of the difference between the distributions of male and female scientists [MW47]. To quantify the size of the difference, we compute Cliff's d,

$$d(t) = \frac{2U(t)}{n_m(t)n_f(t)} - 1, \tag{4.3}$$

where $n_f(t)$ is the number of female scientists [Cli93]. The value of d ranges from -1 (when every observation for women are greater that those of men) to 1 (when every observation for men are greater that those of women). For example, if $d = 0.8$ for the cumulative publication distribution, a randomly picked man has a 80% chance to have more publications than a randomly chosen woman. If $d = -0.8$ then a randomly picked woman has a 80% chance to have more publications than a randomly picked man.

Reproductive feedback: We quantify the ME as the extent to which authors reproduce their individual productivity and recognition over time via positive feedback. For each cohort and career age, we diagnose to what extent scholars author new publications or receive new citations in a career age proportional to their productivity or recognition in the previous career age. This relationship is quantified by the scaling law

$$x(t) \propto x(t-1)^{\beta(t)}, x(t-1) \geq x_{\min}(t-1), \tag{4.4}$$

where the exponent β and the lower cutoff $x_{\min}$ are the model parameters. If the scaling law is a plausible fit and the estimated exponent $\hat{\beta} > 0$, past productivity or recognition is advantageous to, because correlated with, present productivity or recognition. If this advantage accumulates over subsequent career ages, we speak of the ME that is then quantified by the sequence of $\hat{\beta}$s. To quantify the ME in productivity, we predict the number of publications $x(t) = p(t)$ by the cumulative number of publications $x(t-1) = P(t-1)$, and, to quantify the ME in recognition, we predict the number of citations $x(t) = c(t)$ by the cumulative number of citations $x(t-1) = C(t-1)$ [JNB03]. Predicting by the number of publications $p(t-1)$ and citations $c(t-1)$ yields less variance in $x(t-1)$, shorter time series, and marginally smaller exponents, but similar trends.

In figure 4.5A, we demonstrate the fitting procedure for the cohort start year 2000, career age 15, and the citation practice. The pale points are the observations for authors with $c_i(15) \geq 1$ and $C_i(14) \geq 1$. The full points result from putting these observations into 20 bins of exponentially increasing size. The model is fitted to the binned data using the method of ordinary least squares, and the coefficient of determination R^2 quantifies how well the model fits the corresponding unbinned data. The lower cutoff is estimated by choosing $x_{\min}$ such that $R^2(x_{\min})$ has its first maximum. This is a simple heuristic that, in our particular application scenario, underestimates both model parameters but mitigates statistical errors on the scaling exponent as well as biases from finite-size effects.

Independent variables: There are substantive and methodological reasons to not mix data from different cohorts. Substantively, we are interested in changes that may have occurred as computer science became a big science. Methodologically, it prevents to account for variations in the production and recognition functions of authors across career ages [PPP+13]. To account for this, the *baseline* category of independent variables contains a *start year* (cohort) control variable.

We are interested in the factors that affect an author's career. According to the Matthew Effect, advantages accumulate over time. The earlier in a career an advantage sets in, the more it can accumulate. Hence, all our independent variables are computed for the early career $[1, t_e]$. The value of t_e is determined in section 4.4.4. We refer to the field theory of [Bou88] when using the social, symbolic, or cultural capital concepts. The first construct category contains the early-career *achievements* of authors. *Productivity* is the cumulative number of publications $P_i(t_e)$. *Productivity (1st author)* $P_{i(1st)}(t_e)$ is the number of publications written as a first author. *Recognition* is the cumulative number of citations $C_i(t_e)$. While productivity is a performance measure, recognition is a real success measure in the symbolic capital sense of recognized cultural capital. To test for a gender effect, we include a *gender* category. Since gender could not be detected for all authors, we use *male*, *female*, and *undetected* as dummy variables.

Careers are affected by being able to reap benefits from embedding into social networks. Hence, our third construct category is *social capital*. *Social support* is the size of the social support network, measured in terms of the number of distinct co-authors in the early career. The transformation from little science to big science is marked by the emergence of team science [WJU07b]. Therefore, we study the effect of *team size*, defined as the median number of authors of all publications produced in the early career. *Senior support* quantifies the extent to which an author enjoys mentorship from a senior scientist. Our proxy is the largest h-index [Hir05a] of all co-authors j in the social support network: $\max(h_j(y))$. $h_j(y)$ is the maximum

cumulative number of publications h that each have accumulated at least h citations until y, where y is the year in which author i is in career age t_e.

We also expect *symbolic capital*, a reputation for academic worthiness, to influence career paths. One way to quantify it is to use the reputation of the sources (journals and conference proceedings) an author publishes in. We operationalize symbolic capital based on the $h5$-index [Goo20] of sources. Corresponding to the definition above, $h5_s(y)$ is the maximum cumulative number of publications $h5$ published in source s in the years $[y-4, y]$ that have accumulated at least $h5$ citations in those years. The binary *top source* variable is then 1 if an author has at least one publication in a source that belongs to the top 75% of the distribution in a given year.

Finally, career paths are influenced by an author's individual *ability* to excel [SWD+16b]. Such a measure is supposed to capture an intrinsic author property that, given that we measure it via behavioral traces, can only be a rough proxy. We reason that such an ability should be detectable early on and operationalize it as the number of citations $C_i^{\text{abil}}(15)$ that the publications produced in the early career accumulate until and in career age 15. As with productivity, we also use an *ability (1st author)* variable $C_{i(1\text{st})}^{\text{abil}}(15)$. Since an author's ability depends on the *cultural capital* invested in her or his life, we label the construct category accordingly.

All independent variables are standardized by subtracting the median and dividing the result by the range between the 1st and 3rd quartile.

Dependent variables: Authors can leave academia for a certain number of years in a row. We label each author in our corpus as a *dropout* if she or he has not published for ten consecutive years in the first 15 career ages. 59% of the authors are labeled as dropouts. This label is used as a binary variable in dropout predictions. Citation-based measures are commonly used to quantify the success of authors. Our first type of measure is $C_i(15)$ as defined above: the cumulative number of *citations* received by all publications published until and in career age 15. However, this measure is autocorrelated with the independent predictor $C_i(t_e)$. This autocorrelation inflates the coefficient of determination. Hence, we also define a *citation increase* variable that is not autocorrelated with a predictor [PPP+13]. It is defined as $C_i^+(15) = C_i(15) - C_i(t_e)$. Our second type of success measure is the *h-index* $h_i(15)$, as defined above, of all publications produced in the whole career. To again remove autocorrelations, we also define the *h-index increase* $h_i^+(15) = h_i(15) - h_i(t_e)$.

Dependent variables are standardized like the independent ones.

Prediction models: To determine the age in which the early career of an author ends, we predict citations and the citation increase using all independent variables by varying $t_e \in [1, 15]$. The chosen value is then used

in the following models. In *dropout prediction*, we regress dropout against the independent variables using a logistic model. In *success prediction*, we regress citation increase and h-index increase against the independent variables using a linear model. We use the elastic net variant since it contains regularization techniques to ensure that the model generalizes well (to avoid overfitting). These techniques estimate weights that penalize regression coefficients. This is useful when multiple independent variables are correlated with each other [ZH03]. There are two parameters. The mixing parameter λ controls the extent to which overfitting is avoided by L1 regularization (which makes some weights zero, i.e., selects variables to remove) as opposed to L2 regularization (which makes weights small but not zero). When $\lambda = 1$ only L1 penalties are applied; when $\lambda = 1$ only L2 penalties are applied. We use the default $\lambda = 0.5$, that is, the elastic net will perform variable selection but will keep highly correlated variables in the model. The regularization parameter α is a constant that multiplies the penalty weights. When $\alpha = 0$, the model becomes an ordinary-least-squares regression (without any regularization). The optimal value for α is learned from the data.

In all prediction models, there are 292,443 observations. Regression coefficients and their weights are learned in 10-fold cross-validation. That is, the data is randomly divided into 10 folds of 29,244 observations, and in 10 iterations the model is trained on 9 folds and tested on the remaining one [Hox17]. Regression coefficient are reported as averages across the 10 folds. When means are far from zero, effects are sizable; when standard deviations are low, coefficients are robust.

For the binary prediction model (dropout prediction), we use two scores as evaluation metrics. The *F1* score is the weighted average of the precision (proportion of predicted positives that are correct) and recall (proportion of known positives that are predicted correctly). The *average precision* summarizes a precision-recall curve as the weighted mean of precisions achieved for every highest value of recall. Both range from 0 to 1 [LKA16]. For the linear models (success prediction), we use two other goodness-of-fit measures. The *mean squared error* quantifies the mean squared distance of all observations to the regression line. The *adjusted* R^2 coefficient of determination measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It corrects for the number of independent variables that the models use. It increases only if the new term improves the model more than would be expected by chance. Both measures range between 0 and 1, where higher values are better. For all four evaluation metrics, we report the average value across 10 folds.

Appendix

4.A Vertical inequality and cumulative counting

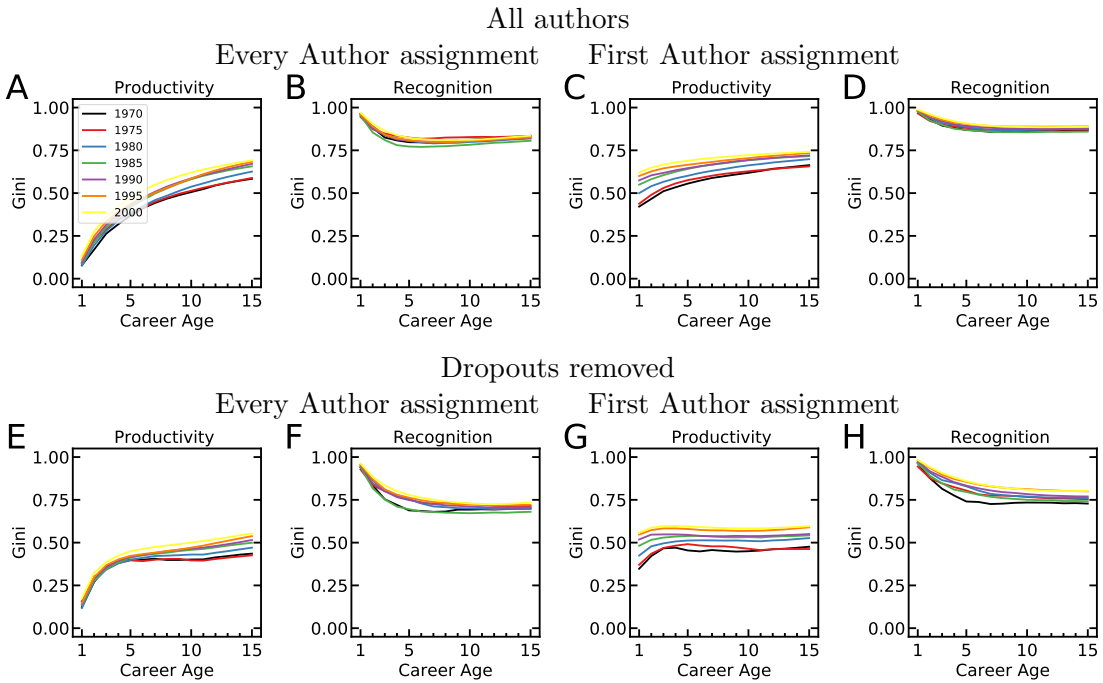


Figure 4.A.1: Inequality over career ages (cumulative counting): Vertical inequality in productivity and recognition as a function of career ages, depicted for seven cohorts between 1970 and 2000. We count publications and citations cumulatively ($P(t)$ and $C(t)$, defined in "Materials and methods: Vertical inequality"). (First two columns) Assigning publications to all authors. (Last two columns) Assigning publications only to first authors. (Second row) Authors are filtered that have not published for ten consecutive years (most likely left academia).

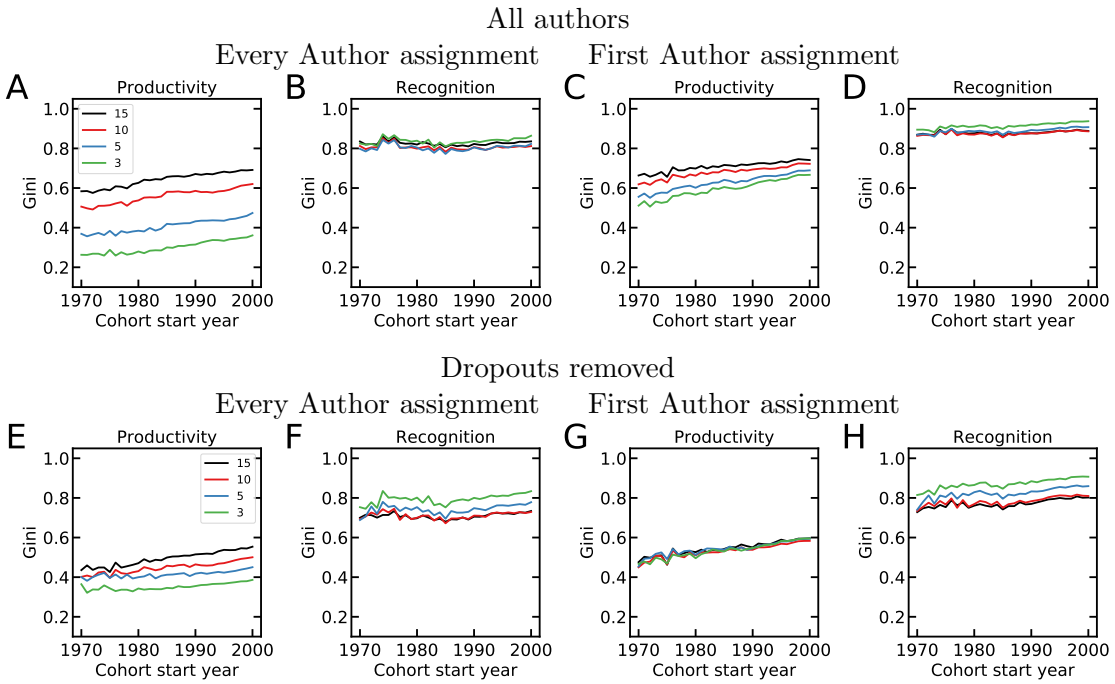


Figure 4.A.2: Inequality over cohorts (cumulative counting): Vertical inequality in productivity and recognition as a function of cohort start year, depicted for career ages 3, 5, 10, and 15. We count publications and citations cumulatively ($P(t)$ and $C(t)$, defined in "Materials and methods: Vertical inequality"). (First two columns) Assigning publications to all authors. (Last two columns) Assigning publications only to first authors. (Second row) Authors are filtered that have not published for ten consecutive years (most likely left academia).

Part II

Inequality in Music Industry

Introduction

The music world is one of many creative industries whose work has been associated with the economic boom of post-industrial cities [Flo14a, How13]. Music offers cultural and social meaning to individuals and societies, as well as pathways for self-realisation, creative expression, agency, identity building and socialisation [Bec08]. However, despite this influential role, building a career in the music world is a story of constant struggle. While few become superstars and enjoy mainstream recognition and economics success, others' careers follow a marginal path with little outcome. Success is not a mere result of musical skills, talent and creativity, but also a matter of strategies and adopting the right behavior. In this environment musicians careers follows the ethos of freelance jobs and entrepreneurship, conflated with notions of self-management, self-promotion and networking as strategies for artistic and commercial success [McR02, Blo17]. Similar to science, every musical genre and subculture may favour certain behaviours and practices based on its socio-cultural history and characteristics. However, every music genre, from production to its cultural appreciation involves a series of interactions between a number of actors such as peers, mentors, musicians, recording studios, labels, distributing companies, promoters, music venues, audiences and critics. Together they create a complex system of actors that influence how musicians' works are produced, performed, and appreciated by the public and peers. In this thesis, I focus on Electronic Dance Music (EDM) as a case study. Compared with many other music genres, EDM is a relatively young subculture that is witnessing a great cultural and social transformation in the last decade. Moving from the counter-culture movement to the center of mainstream culture, it has become one of the biggest pool of money and talents within music industry. **Chapter 5** follows a formal sociological approach based on bipartite networks to study one of the underlying mechanisms of success in this field – the hipster paradox – using digital traces of performing live and releasing music.

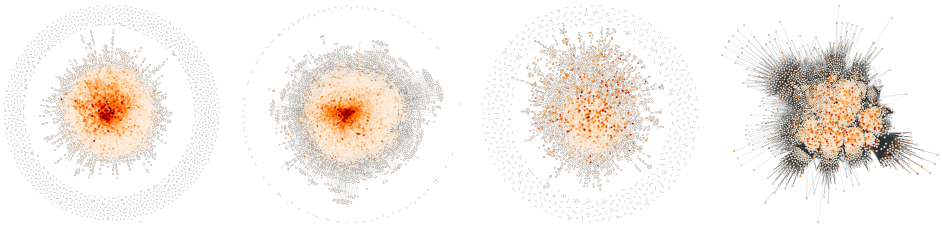
Chapter 5

The Hipster Paradox in Electronic Dance Music: How Musicians Trade Mainstream Success Off Against Alternative Status

Abstract. The hipster paradox in Electronic Dance Music is the phenomenon that commercial success is collectively considered illegitimate while serious and aspiring professional musicians strive for it. We study this behavioral dilemma using digital traces of performing live and releasing music as they are stored in the *Resident Advisor*, *Juno Download*, and *Discogs* databases from 2001-2018. We construct network snapshots following a formal sociological approach based on bipartite networks, and we use network positions to explain success in regression models of artistic careers. We find evidence for a structural trade-off among autonomy and success. Musicians in EDM embed into exclusive performance-based communities for autonomy but, in earlier career stages, seek the mainstream for commercial success. Our approach highlights how Computational Social Science can benefit from a close connection of data analysis and theory.

5.1 Introduction

Counter-cultural and anti-establishment fields legitimize themselves by distancing from the mainstream. Yet, to sustain their careers and achieve economic success, cultural producers in such fields need to strive for widespread recognition for their work. Approaching mainstream success while not becoming mainstream themselves, running the risk of alienating supporters



(a) Co-gig (b) Co-venue (c) Co-label (d) Co-style

Figure 5.1.1: Networks of musicians connected by (a) co-performing at gigs, (b) co-performing in clubs and other locations, (c) co-releasing on music labels, and (d) co-releasing in music styles. Networks are largest connected components with insignificant ties and isolated nodes removed. Node size is proportional to how close a musician is to all others (closeness centrality). Node color gives a musician's success in terms of the distance traveled between live performances (the darker the more successful). Intuitively, musicians of international renown are in demand in venues that are distant from each other. These snapshots uncover that successful musicians follow the mainstream by taking central positions in networks built on gigs and venues. In this paper, we show that this strategy is associated with success in early career stages. Snapshots are for the 2013-2015 period.

from their subculture and being labeled as a "sell out," is the paradox that subcultural producers face. We refer to it as the *hipster paradox*, borrowing the term from the phenomenon that the hipster subculture blends mainstream and alternative lifestyles [Gre10].

Scholars in sociology and the *science of success* have studied what makes for a successful career among Jazz [PD09] and Punk musicians [Cro15], painters [Giu99], and writers [dN03]. In recent years, a magnitude of studies have investigated the working condition and success of creative careers using large-scale digital behavioral data [REB10, ADJ15, JMBI20]. Yet, we know little about how producers in counter-cultures deal with the dilemma the hipster paradox poses to agents in creative industries.

Electronic Dance Music (EDM) makes for an interesting case study because its history is one of non-conformity with mainstream music culture—mainly white Rock music. It started as a collective action by those who felt alienated by the mainstream: mainly the black and gay population [McL01]. EDM's increasing popularity in the last decade has brought it from the margin and underground culture to an industry with a global value of 7.3 billion US dollars in 2019 [Wat20]. Autonomy from mainstream values is itself a central value in EDM. Unregulated, unlicensed, anti-establishment, and exclusive parties, organized by communities of enthusiast, served as a safe space for personal expression and liberty [AK07].

Large-scale behavioral data from digital platforms enable unobtrusive,

longitudinal analysis that may help to uncover behavioral patterns and mechanisms. Such studies on EDM have highlighted the importance of community embeddedness for value creation and success [ADJ15, JMBI20]. However, insights into how musicians deal with the hipster paradox are mainly derived from qualitative interviews with musicians. These diagnose a success/autonomy trade-off that consists of rooting commercial practices in exclusive and alternative performance-based communities [Rei11, LB13, Ort18, WvV19]. The ethnographic method allows for in-depth insights, but it relies on retrospective accounts of field participants that suffer from memory and desirability biases.

In this paper, we study the hipster paradox in EDM using large-scale and longitudinal digital traces of musicians. Grounding our observations in the careers of over 4,000 artists over almost two decades, we study how their relationship with mainstream appeal affects their success. Using digital trace data has the benefit that our observations are unobtrusive accounts that unfold over time. Inspired by sociological field theory [Bou93], we identify two primary practices that embed artists within the EDM subculture and are not mainstream or alternative *per se*: performing live and releasing music. Whereas mainstream labels and live venues are a conduit to widespread popularity and economic success, alternative releases and performances reinforce and legitimize the artists' belonging to the EDM subculture.

How important is it for musicians to be embedded into a community? How important is it to belong to the mainstream? Are bridging or redundancy-avoiding strategies associated with success? And how does all that change over an artist's career? To answer these questions, we construct a large data corpus by harvesting the *Resident Advisor*, *Juno Download*, and *Discogs* platforms. For the 2001-2015 period of observation, we construct four analytical networks that convey how similar musicians are in terms of practicing EDM [BC13, EM18]. We quantify positions in these networks, devise a measure of success that is based on long-distance travels, and regress success on network variables in linear mixed models. Figure 5.1.1 gives an impression of these networks and the position of successful musicians.

We find evidence of a structural trade-off between revenue and autonomy. Musicians in EDM embed into exclusive performance-based communities for autonomy but, in earlier career stages, seek the mainstream for commercial success. Our results show that successful musicians gain a sufficient support base early in their careers at the risk of "selling out," while established artists that assert their alternative status find long-term success.

5.2 Related Work

Electronic Dance Music

The hipster paradox can be rooted in the sociological theory of "fields of cultural production," a framing that is useful for understanding the conflict of art and money. According to this idea, legitimacy in fields of art (i.e., sub-field of restricted production) springs from autonomy from the economic order (i.e., from sub-field of large-scale production) [Bou93, ch. 1]. In EDM, the relationship of art and money is complex (and subject to our modeling). The history of EDM shows that a polar distinction between those that do "art for art's sake" and those that work for the "creative industry" are too simple. For example, the EDM subfield in the UK is much more centralized and commercialized than the US subfield, but it emerged from the latter's reluctance to partner up with the record industry [WvV19].

Nowadays, EDM is home to the "notion that, equipped with the right set of tools, skills, and talent, one individual can 'make it' alone" [Ort18, p. 156]. [Rei11] finds that musicians in EDM seem to embody this "Me Inc." ideology, that is, they do strive for commercial success day by day, and concludes that it calls into question the supposed autonomy of cultural producers. This situation makes the hipster paradox an existential problem for musicians.

There are two main practices in EDM that allow them to face the dilemma. The practice of *performing live* is strongly related to the notion that EDM enshrines a love of music and dancing. In *gigs* such as club nights and raves, performance and participation meld, and music acts as a gravitational force for social relations [Tur09]. Serious and aspiring professional musicians must carefully choose in which *venues* to play. On the one hand, larger venues pay higher wages, but, on the other hand, since mass production is considered "selling out," performances in big clubs are endowed with a negative label [Ort18, ch. 4].

Interviews with musicians suggest that they address the paradox via a particular kind of network sociality: "As individualistic entrepreneurs, grassroots musicians often find themselves in weak positions, having less power to negotiate conflicts, bargain for better opportunities, and navigate the social structures and groups that organise EDM musical activities. To compensate, many aspiring professional participants join networks who function as 'defensive exclusionary networks' ..., and in the process distance themselves from others." [Ort18, update: p. 156] During live performance events, strategic relationships occur in settings that correspond to musicians' natural state of being [LB13]. Musicians embed into systems of intersubjective ties that are "informational, ephemeral but intense, and ... characterized by an assimilation of work and play." [Wit01b, p. 71]. Since these networks maintain familiarity and mutual valuation, commercial success is not stig-

matized [Rei11]. Local club scenes are the vivid faces of these dynamics. For one, they form around geographical locations where cities like London and Berlin take core positions in the field [ADJ15].

Besides performing live, *releasing music* is the other main practice in EDM. Songs and records are a way to express autonomy. Other than performing, which requires at least access to a venue, musicians are, in principle, free to produce in whatever *style* or music genre they want. Musicians are free to just release their music online or to start their own label [Rei11]. This informal "do-it-yourself" culture drives the evolutionary dynamics of EDM. For example, "drum-n-bass" is a main genre that differentiated into "abstract drum-n-bass," "ambient drum-n-bass," and "intelligent drum-n-bass" [McL01, p. 60]. Like venues, styles are crystal nuclei of exclusionary practices in communities [Ort18, p. 219].

By released music, musicians demonstrate their seriousness and gain access to the inner social circles of communities which opens new pathways to making a career [McL01, Rei11]. To produce and release at a large scale, musicians have to secure deals with music *labels*. Labels function as gatekeepers of the creative industry: They sift through the pool of cultural producers and select those that are promising to meet the current taste of the community or field. This asymmetric power over the boundaries gives them influence over the tastes, opinions, and reputations of producers, performers, and participants [Ken08, Rei11]. It has been found that a small fraction of star artists help other musicians into top ranks via mentorship and recording collaboration. Which musicians these are is, in turn, influenced by their styles, i.e., changes in the social cores of communities mirror the cultural drift of styles [JMBI20].

Literature identified ways in which EDM musicians employ performance-based practices to navigate the dilemma. We build upon this literature, finding corroborating evidence of how embedding in communities of musicians facilitate this process. We further expand upon these insights and show how stages in EDM careers mediate which of these practices are successful. Though, while the relationship between success and performance practices is well established, we know less about the role of practices related to releasing music.

Although studies on the dilemma highlight the importance of network effects, they are largely qualitative studies based on interviews. In contrast, we perform an empirical analysis of the network of musicians based on the digital traces of their practices. Hence, we next discuss the related literature at the intersection of network science, art, and success.

Network Analysis of Fields of Art

Networks have been shown to be apt representations of fields. Most abstractly, a field is a space of relations among positions. Fields govern in-

dividuals' practices, and they manifest as social networks. The power of graph-theoretical approaches is that they make positions amenable to measurement and computation [dN03, BC11b]. One way to construct these analytical structures is by way of bipartite (2-mode) networks. By modeling practices as relationships of agents and symbolic facts (e.g., music venues or styles), formal frameworks allow for constructing fields as networks from practices [EM18]. This approach involves a projection of the two-mode network to a binary or weighted one-mode network [BC13]. Agents with similar patterns of choices in the initial two-mode network have similar patterns of ties and, hence, similar positions in the projected one-mode network. The structure of this network can then be analyzed and visualized using the graph-theoretical repertoire of Social Network Science [MW03b, BE06].

Music involves a series of relations between a variety of agents such as artists, mentors, recording studios, labels, distributing companies, promoters, music venues, audiences, and critics [Sma99]. A number of studies uses networks to explore music fields and musicians' careers [ADJ15, Cro20, EC18, MWH17]. For example, an analysis of the bipartite network of artists and festivals shows that Turkey's Metal music field exhibits a core-periphery structure. Bands with a stronger affiliation to the Rock style, a larger number of festivals played, and support from major labels are more likely to occupy central positions in the network [EC18]. Similar work on Punk [Cro15] and Jazz [Ved17] suggests that artists who occupy central positions in co-gig networks and form open cliques have higher chances of success. However, most studies consider only one aspect of musicians' careers, namely the affiliation to either gigs, venues, labels, or music styles. Our work contributes to this line of research by analyzing the career of EDM artists using all four networks.

5.3 Materials and Methods

5.3.1 Datasets

Our research design calls for measuring the field of EDM via the practices that constitute it. To study the hipster paradox in a large-scale quantitative way, we collect data from digital platforms. While all data come with limitations, which we discuss towards the end of the paper, these are especially suitable to our design because they capture the digital traces left by the practices of performing live and releasing music. We build a corpus of traces from three platforms, each providing partial information (*Resident Advisor* for performances, *Juno Download* and *Discogs* for releases). Once combined, this corpus offers a holistic view of the field.

Table 5.3.1: Dataset statistics. Each dataset offers partial information about practices of artists in EDM. RA consist of a larger number of artists and serves as the primary source for musicians. JD and *Discogs* together provide release information for abut half of the musicians. While JD has of more releases, *Discogs* provides richer information on music styles.

	Musicians	Gigs	Venues	Releases	Labels	Styles
Live performances						
Resident Advisor (RA)	63,543	728,850	50,410	-	-	-
Releases	39,042					
Juno Download (JD)	35,844	-	-	259,147	30,488	69
Discogs	23,663	-	-	160,130	30,281	339
Total	63,543	728,850	50,410	332,162	39,661	347

Live Performances

We use *Resident Advisor* (RA, `residentadvisor.net`) as a primary source for selecting a large sample population of EDM musicians and information about their live performances. RA is an online music magazine and platform dedicated to EDM. It serves as one of the main information hubs for EDM events and culture worldwide. Musician profiles contain information about their "gigography" including event venue, date, and lineups. Similarly, each venue has a profile page that includes information such as its address, social media links, and archived past events.

We infer the geo-coordinates and location of venues by using the combination of four geo-location APIs, namely Nominatim, HERE, Google and GeoNames APIs. We manually assigned the city to 150 venue and found 78% correct assignment from the APIs.

Music Releases

We compile a discography of musicians by combining data from two major online music discographies and stores. *Juno Download* (JD, `junodownload.com`) is considered one of the largest independent dance music download stores worldwide. It provides a large catalog of electronic music styles with over 6 million tracks. Each track is attributed with artist name, label name, release name, release date, and music genre(s). *Discogs* (`discogs.com`) is a crowdsourced discography platform, the largest and most comprehensive music database and marketplace with 10 million releases across various genres. With a share of 14.26%, electronic music is the second largest genre (after Rock with 23.68%) in the platform.[1] The platform provides information about musicians and bands, namely a short biography, social media and internet pages (e.g., Wikipedia, personal website), band members, aliases, and name variations.

[1]Matt Larner, "State of Discogs 2017," *Discogs BLOG*, February 14, 2018, `https://blog.discogs.com/en/state-of-discogs-2017/`, retrieved June 11, 2021.

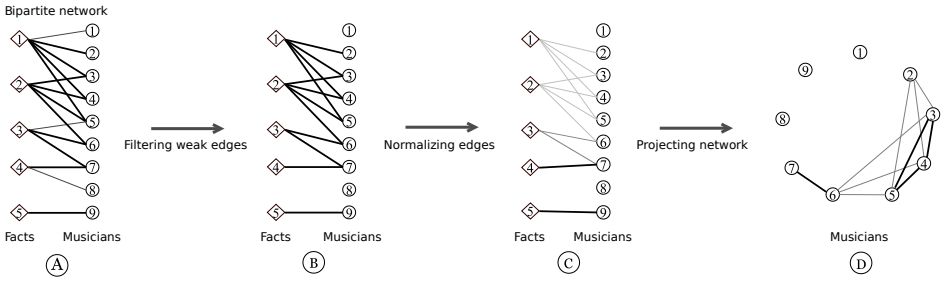


Figure 5.3.1: Construction of networks representing the field of EDM from bipartite networks representing practices in EDM. (A) Bipartite networks of facts (EDM venues, labels, or styles) and musicians with weighted edges (i.e., facts can be selected multiple times). (A→B) Weak edges are removed where a fact is selected only once. (B→C) Edges are normalized so all facts have unit weighted degree. (C→D) The network is projected to obtain the network of musicians where edge weights give their similarity in terms of selecting the same facts (performing in the same venues, releasing on the same label, or releasing in the same style). Thicker and darker edges indicate larger weights.

To identify RA artists in JD and collect their discographies, we query the website using artist names extracted from RA. To find the *Discogs* page for each RA artist, we first check if there is a link in her RA profile page. For the remaining artists, we use the *Discogs* search API to query for artist names. For musicians with multiple aliases or projects, we combine all the releases under the name with the highest number of gigs. To avoid name ambiguities and duplicate entries, we match label names from JD and *Discogs* using release and artist names.

Final Sample and Period of Observation

Up until 2000, the number of active musicians in RA is smaller than 1,000. All data was collected in 2018. Hence, we limit our period of observation to 2001-2018. Table 5.3.1 reports the numbers of musicians, gigs, venues, releases, labels, and styles derived from these practices. The full sample consists of 63,543 musicians that play 11 gigs on average. For 61% of them, we also found releases.

5.3.2 Methods

Network Construction

To analyze fields as networks, we follow a formal approach [EM18]. It consists of representing the traces of practices in bipartite networks where one

part is a *musician*, an artist who performs live and may also release music. The other part is a *fact* (gig, venue, label, or style). Bipartite networks of musicians and gigs or venues derive from the performing practice; bipartite networks of musicians and labels or styles derive from the releasing practice.

Analytical networks representative of the field of EDM are constructed from these bipartite networks [BC13]. Figure 5.3.1 schematizes their construction. There are two preprocessing steps. First, we remove all facts that musicians chose only once since those introduce noise. Second, we normalize edges in bipartite networks in such a way that the number of selections by all musicians sums to one for all facts. The analytical networks are then obtained by projecting bipartite networks in such a way that musicians become the nodes. For the creation of gigs networks, the first preprocessing step is different because the original edges are not weighted (a gig is a one-time event). Instead, we remove gigs with only one musician. A check reveals that such events are mainly data artifacts caused by missing information in the lineup listings. We also remove gigs (such as festivals) with a large number of musicians (three standard deviations over the mean). These are rare events that entail many but, due to normalization, weak edges that overshadow network analysis. This method results in musician *co-gig*, *co-venue*, *co-label*, and *co-style* networks (snapshots for the 2013-2015 window are depicted in figure 5.1.1).

As a result of normalization, two musicians can be similar either if they co-perform in many popular gigs or venues, or if they co-perform in fewer but more alternative ones (and similarly so for releasing music on labels and in styles). Correspondingly, communities can emerge in two different ways that map to mainstream and alternative practices. We shall give an example. In figure 5.3.1D, there are two communities constituted by strong ties: The first consists of nodes 3, 4, and 5; the second consists of nodes 6 and 7. The first is a *mainstream community* since it derives from all nodes selecting the popular facts 1 and 2; while the second is an *alternative community* that derives from the selective focus on the otherwise unpopular fact 3.

Cohorts and Careers

Following a longitudinal research design means that we construct the four networks described above for sliding time windows of three years. A *career* is then a sequence of positions in these networks [Bou93, p. 18]. Our period of observation is 2001-2018. Musicians are the units of observation. Musician a belongs to a *cohort* defined by the first year t_a in which they perform live. In year $t \geq t_a$, a musician has a *career age* $\tau(t) = t - t_a$. We differentiate among different *career stages*. A musician can be in the early stage ($\tau < 5$), mid stage ($5 \leq \tau < 10$), or late stage ($10 \leq \tau$).

We also attribute musicians to one of five success-based *career types*: sta-

ble successful, stable mediocre, stable unsuccessful, upward, and downward. To do so, we use travel distance (which we introduce in section "Measuring Success") to compute the percentile rank $r_a(t) \in \{1, 2, 3, 4, 5\}$ of each musician in a given year (musicians in the first 20% quantile have rank 1, in the second 20% quantile rank 2, ...). Next, we compute the average rank $\overline{r}_a$ over a career and the rank difference δ_a between the first and last years of a career. We consider careers with $\overline{r} \leq 2$ as "stable successful," $\overline{r} \geq 4$ as "stable unsuccessful," $\delta > 0$ as "upward," $\delta < 0$ "downward," and the remaining as "stable mediocre." Upward and downward careers do not include careers already assigned to one of the first two categories. As Figure 5.4.1 shows, "stable successful" is the largest category.

5.3.3 Research Design

Research Questions

The literature on EDM proposes that musicians solve the dilemma posed by the hipster paradox by embedding into alternative and exclusive communities in which they can pursue commercial activities without being stigmatized [Rei11, Ort18]. Quite generally, *communities* are cohesive social formations that have the purpose of reducing uncertainties for its constituents [Whi08]. In EDM, musicians join communities for informal reasons (i.e., the love of music and dancing) and formal reasons (e.g., to strategically forge ties to market intermediaries) [Tur09, Rey13, JMBI20]. First, we want to know if there is empirical evidence for the cohesive nature of EDM.

Research question 1: To what extent is community embeddedness associated with success?

Next, we address the aspect of belonging to the *mainstream* culture. As we have seen in the methods section, communities can have their origins in both mainstream and alternative practices. Is it true that only the alternative path leads to success, as the literature suggests? Or is the mainstream path also viable, despite its inherent risk of losing legitimacy?

Research question 2: To what extent is mainstream belonging associated with success?

An important part of the explanation how musicians solve the dilemma they face is that the communities they embed into emerge from exclusionary practices, that is, musicians distance themselves from others [Ort18, p. 156]. This implies that successful musicians take positions in communities that have rather impermeable boundaries. Embedding into multiple communities would then not be associated with success. On the other hand, positions in boundaries can be sources of creativity and success thanks to the opportunities of *bridging* structural holes that exist between communities [Bur92].

Research question 3: To what extent is bridging associated with success?

The structure of a musician's immediate network neighborhood may also have an effect. It has been shown that dense ego networks are detrimental to creativity, likely because they are correlates of rather indistinguishable and redundant node neighborhoods [US05a]. As such, they have a *constraining* effect on a node [LNP13].

Research question 4: To what extent is constraint associated with success?

Finally, the literature, being largely based on interviews and ethnographic work, has not touched upon how changes in network positions may be associated with success over musicians' careers. That means, we seek answers to the questions above by differentiating between the early, mid, and late career stages of musicians.

Measuring Success

We derive the success measure from live performances. Live performance is the main source of income in popular music [MPCG11] and particularly in EDM [Ort18, ch. 4]. Our rationale is that musicians who perform in gigs around the world cover long geographical distances. Our measure is based on the trajectory of musicians' travels among gig locations. Each venue has a dedicated page in RA. We use its address to obtain the city where it is located. Let $C_a = \{c_1, c_2, ..., c_N\}$ be the *travel trajectory* of musician a who makes N visits to cities c ordered in time. The same city can be visited multiple times. The success variable is then the summed travel distance $d_a = \sum_{i=1}^{N-1} \epsilon(c_i, c_{i+1})$ where ϵ is the Euclidean distance function.

This proxy for success finds anecdotal validation in the fact that the most-traveled musicians are indeed enormously successful acts, and top EDM musicians Tiësto and Paul van Dyk lead the ranking even before Rock icons Bob Dylan and Metallica.[2] The variable also passes a formal evaluation test: It is able to predict which musician belongs to the top 100 in two annual international ranking polls. The average predictive accuracy is 85% from 2008-2018, on average. A comparison with other travel-based success measures shows that it is important to consider the order of city visits. This is mirrored in reality where it is common practice by grassroots artists who build their music career next to a day job to arrange for multiple live performances when they travel to far-away cities. This way, they can reach larger audiences and save time and money.

[2] Jacob Shamsian, "The 10 most-traveled musicians have toured over 11 million miles around the world — here's the full list," *Insider*, February 15, 2017, `https://www.insider.com/musicians-who-travel-the-most-2017-2`, retrieved September 13, 2021.

Measuring Positions in the Field

The advantage of our network approach to field theory is that we can operationalize the four different types of positions addressed by research questions 1-4. The first two types serve to diagnose the importance of *network closure* for success. The core positions in a network represent its mainstream behavior. We operationalize the construct of mainstream belonging as the closeness centrality in a network. The closeness of a node is the inverted sum of its distances to all other nodes [OAS10]. It is close to 1 for core nodes and close to 0 for peripheral nodes. This is a global measure because a node's position is characterized with respect to to all other nodes.

Communities are cohesive network substructures with the density (i.e., the ratio of the numbers of observed and possible ties) increasing from the periphery of a community to its core [MW03b]. We operationalize the construct of community embeddedness as the maximum k-core that a node belongs to, where the k-core is a maximal subgraph whose nodes are all connected to at least k others [BZ11]. Musicians in the core (periphery) of a community will have large (small) values. Compared to the global closeness centrality measure, this is a local measure because it takes nodes at an intermediate distance of an observed node into account. While closeness centrality makes use of edge weights, the k-core algorithm assumes an unweighted graph.

The other two types of positions refer to the importance of *network openness* for success. The first is bridging which we operationalize with node betweenness centrality, the extent to which the shortest paths among all node pairs pass through a node [BP07]. Again, we contrast this global measure with a local one: The clustering coefficient [WS98] is our measure for the last construct of constraint. It is close to 1 (0) for strongly (weakly) constrained nodes. Note that this is the only network variable where an inversely proportional relationship with success is expected. Both measures, bridging and constraint, are computed using edge weights.

Linear Regression of Success

We regress the dependent success variable on 16 independent network variables (4 types of positions for 4 analytical networks) and baseline variables. The analysis is longitudinal, that is, we use independent variables aggregated in 3-year time windows to explain success in the ensuing 3 years: The independent variables are computed for rolling time windows $[t-2, t]$; The dependent success variable is computed for travel trajectories in windows $[t+1, t+3]$. Observations are collected for $t \in \{2003, 2004, ..., 2015\}$. We exclude musicians that never reach a career age of 5 years as well as musicians for which there are less than 5 observations. This way, we put a focus on serious and aspiring professional musicians [Ort18, p. 8] and pro-

Table 5.3.2: Results of mixed-effects regressions of success. Model 1 contains baseline and network-based variables, model 2 also includes interactions with career stage dummy variables. Independent variables and their interactions computed for moving 3-year time windows explain success in the ensuing 3 years. Effect sizes are log odds ratios (i.e., for a one-unit increase in an independent variable x, there is a $\exp(x) - 1$ percent increase in the likelihood of success). In model 2, variables without an interaction term represent the population average effect. For the interpretation of interaction effects, coefficients must be summed (example in the text). Intervals are reported for the 95% confidence level.

	Model1	**Model2**
Intercept	**9.007** [**8.781**; **9.233**]	**8.971** [**8.742**; **9.200**]
Number of gigs	**.454** [**.357**; **.551**]	.049 [−.104; .203]
Number of releases	**.132** [**.059**; **.205**]	**.097** [**.001**; **.194**]
Mid career	**−1.355** [**−1.423**; **−1.287**]	**−1.324** [**−1.393**; **−1.256**]
Late career	**−2.090** [**−2.259**; **−1.921**]	**−2.233** [**−2.411**; **−2.055**]
Number of gigs*mid career		**.499** [**.331**; **.667**]
Number of gigs*late career		**.608** [**.307**; **.910**]
Number of releases*mid career		.045 [−.067; .157]
Number of releases*late career		.217 [−.064; .497]
Co-gig		
Community	**1.210** [**1.124**; **1.297**]	**1.188** [**1.074**; **1.302**]
Mainstream	**.328** [**.254**; **.402**]	**.514** [**.412**; **.616**]
Bridging	−.046 [−.110; .018]	−.042 [−.153; .068]
Constraint	**−.288** [**−.335**; **−.241**]	**−.476** [**−.540**; **−.411**]
Community*mid career		.117 [−.013; .247]
Community *late career		**.560** [**.247**; **.872**]*
Mainstream*mid career		**−.321** [**−.451**; **−.191**]
Mainstream*late career		**−.627** [**−.945**; **−.310**]
Bridging*mid career		−.003 [−.130; .124]
Bridging *late career		−.024 [−.231; .183]
Constraint*mid career		**.379** [**.292**; **.466**]
Constraint*late career		.171 [−.038; .380]
Co-venue		
Community	−.005 [−.078; .067]	−.084 [−.181; .012]
Mainstream	.071 [−.005; .147]	−.020 [−.128; .089]
Bridging	.003 [−.033; .039]	.002 [−.079; .084]
Constraint	.008 [−.045; .061]	**.104** [**.026**; **.181**]
Community*mid career		**.125** [**.014**; **.236**]
Community*late career		.247 [−.014; .509]
Mainstream*mid career		**.168** [**.033**; **.304**]
Mainstream*late career		**.373** [**.054**; **.692**]
Bridging*mid career		.005 [−.084; .094]
Bridging*late career		−.066 [−.277; .145]
Constraint*mid career		**−.172** [**−.272**; **−.073**]
Constraint*late career		**−.245** [**−.482**; **−.008**]
Co-label		
Community	.023 [−.030; .076]	.037 [−.042; .116]
Mainstream	.045 [−.015; .105]	**.095** [**.004**; **.186**]
Bridging	.009 [−.031; .049]	.027 [−.038; .092]
Constraint	−.016 [−.064; .032]	−.013 [−.085; .059]
Community*mid career		−.022 [−.117; .073]
Community*late career		−.058 [−.265; .149]
Mainstream*mid career		−.080 [−.194; .034]
Mainstream*late career		−.083 [−.329; .162]
Bridging*mid career		−.040 [−.122; .041]
Bridging*late career		.067 [−.084; .218]
Constraint*mid career		.000 [−.091; .092]
Constraint*late career		−.082 [−.287; .122]
Co-style		
Community	**−.117** [**−.186**; **−.048**]**	**−.099** [**−.193**; **−.005**]**
Mainstream	**.092** [**.018**; **.165**]**	.106 [−.001; .213]
Bridging	.007 [−.029; .043]	−.002 [−.054; .051]
Constraint	.001 [−.059; .061]	−.057 [−.145; .030]
Community*mid career		−.012 [−.124; .100]
Community*late career		**−.277** [**−.541**; **−.013**]**
Mainstream*mid career		−.010 [−.139; .119]
Mainstream*late career		−.192 [−.485; .101]
Bridging*mid career		.036 [−.036; .108]
Bridging*late career		−.136 [−.308; .035]
Constraint*mid career		.100 [−.012; .211]
Constraint*late career		.088 [−.166; .341]
Marginal R^2	.242	.245
Conditional R^2	.617	.625
AIC	137,811	137,613
Variance: Musicians (Intercept)	6.753	6.875
Variance: Start years (Intercept)	.097	.100
Variance: Residual	6.989	6.891

Bold coefficients: Null hypothesis value outside the confidence interval.

* Effect not significant in corresponding model that excludes release-based variables.

** Effect not significant in corresponding model that excludes performance-based variables.

vide reasonable numbers of observations to capture trends and variations in careers. To mitigate the impact of censoring bias in our analysis, we exclude musicians with at least one gig or release before 2001. We only keep observations that contain both release and live performance activities. These filters reduce the number in the overall data set (table 5.3.1) to 4,224 musicians and 27,077 observations (musician-career age combinations). We use linear mixed models [BMBW15] with musicians and cohorts as random effects. Musicians differ in individual characteristics like skills or creativity. In addition, they are likely effected by cohort-specific conditions. For example, musicians whose start year coincides with the widespread popularity and internationalization of EDM culture are likely to have a higher average number of gigs and longer travel trajectories. Similarly, self-promotion on digital platforms is a rather new practice. By fitting musicians and start years as random effects, we account for variations within these variables. Independent variables are z-standardized, i.e., they indicate how much a score of socio-cultural capital deviates (in terms of standard deviations) from the average value of all musicians in a year t. This transformation partially accounts for the year-specific variations in the field. The dependent variable is logged. Career analysis is implemented via interaction effects. We use a musician's career stage as an interaction term with dummy coding to evaluate the association of each independent variable with success at different phases of a career.

We report the marginal (just fixed effects) and conditional (fixed and random effects) pseudo-coefficients of determination (R^2) [NS13], the Akaike Information Criterion (AIC), and several statistics related to the random effects.

5.4 Results

Table 5.3.2 reports the results from two regression models, where interaction terms for career stages are added in the second one. To ease understanding, we report the percent changes that can be obtained from the table. First, we report the results regarding non-network variables. Then, answers to the four research questions are given in dedicated subsections whose headlines sum up the answers.

The baseline model 1 shows that a one-standard-deviation increase in the number of gigs increases the likelihood of success by a factor of $\exp(0.454) = 1.57$ (a 57% increase). Releasing more music, on the other hand, is less associated with increased success (14%). The largest effects we find pertain to how success changes as musicians advance in their careers. Musicians in the mid career stage are 73% less likely and musicians in the late career stage are even 88% less likely to be successful than early-career musicians. That means, success is mostly an early-career phenomenon. One explanation is

(a) Stable successful (b) Stable mediocre (c) Stable unsuccessful (d) Upward (e) Downward

Figure 5.4.1: Artists can be grouped into five categories according to their career trajectories. Curves depict the average travel distance with 95% confidence interval. The number of artists within each group are (left to right): 1362, 394, 985, 393, and 1090.

that the dependent variable is a travel-based proxy of success. The finding then is that musicians travel less the more their career advances. However, figure 5.4.1 shows that decreases of success with career age are just the average effect. In fact, there are quite a few musicians with stable successful and even upward career trajectories. Correspondingly, when we consider interaction effects (model 2), we find that playing more gigs is associated with larger increases of success the more careers advance: Although the impact of number of gigs in early career is not clear, it is likely to increase the chance of success in mid and late career dramatically. With this in mind, we move on to answering the research questions. The first one asks about the association of success with community embeddedness.

Successful Musicians Embed Into Communities at Gigs

Embedding into communities that result from social relations at gigs is most strongly associated with success (235% increase, model 1), particularly in the late career stage (474% increase, model 2). Co-venue networks are indicative of the importance of place. Due to our bipartite network approach, musicians that perform in core venues are core musicians in co-venue networks. We find that performing in core venues becomes significantly more important in the mid-career stage, but the effect is very small (4% increase). There are also significant effects regarding the importance of music style. Interestingly, community embeddedness is negatively associated with success (11% decrease, model 1), with decreases rising from 9% in the early career to 31% in the late career stage (model 2).

The conditional R^2 states that model 2 can explain 62.5% of the variance in success. But since the marginal R^2 is at 24.5%, most of the variance is explained by individual characteristics which we do not measure. Also, the marginal R^2 of models (performed as robustness checks, not reported here) that exclude performance-based variables (number of gigs, co-gig and co-venue network variables) is a mere 2.8%. That means, the practice of releasing music is practically not relevant for success, while most explanatory power comes from live performances. Correspondingly, no effects related to music styles are robust.

To answer the first research question, we found that only communities formed at gigs are associated with success, but strongly so. We next contextualize this result by answering how success is associated with mainstream belonging. Is success all about alternative communities, as the literature suggests? Or is embedding into mainstream communities a path to success after all?

Successful Musicians Avoid Mainstream Gigs but Seek Mainstream Venues Over Time

Belonging to the co-gig mainstream is associated with an overall 38% increase in success (model 1), but there is a significant trend over an average career. The effect is strongest in the early career (67% increase, model 2) but becomes modest in the mid career (21% increase) and even turns into a 11% decrease in the late career stage. Opposing this trend, performing in mainstream venues is slightly associated with success increases after the early career stage. They amount to 16% and 42% increases in the mid and late career stages, respectively. There are very small effects ($< 10\%$ increases of success) that releasing on mainstream labels and in mainstream styles is beneficial in the early career stage. However, for lack of explanatory power and robustness we will not discuss these.

The answer to the second research question, thus, is that mainstream belonging actually is associated with success with opposing trends for gig-based and venue-based networks. The emerging picture is that, while embedding into gig-based communities is important throughout successful careers, these communities transform from mainstream to alternative communities as careers progress (or musicians move between them accordingly). Whereas the first two questions detailed the role of network closure, we next investigate network openness. The third research question asks about the association of success with bridging.

Successful Musicians are at Home in One Exclusive Community

Bridging otherwise disconnected parts is never associated with success in any of the four networks. From the perspective of the general networks literature, this is surprising because bridging positions are often found to be sources of creativity. However, from the perspective of the EDM literature, this null result is perfectly expected. It indirectly suggests that the communities that successful musicians embed into have an exclusive character. In other words, positions in multiple, or between, communities are not rewarded. Successful musicians are at home in one community that is walled off from others. This finding begs the question of whether musicians need to distinguish themselves while belonging to one, exclusive community, to find individual success.

Successful Musicians Avoid Redundant Connections at Gigs

As expected, dense co-gig ego networks have a constraining effect (25% decrease of success). Adding interaction effects does not yield a trend over career stages (decreases jump from 38% to 9% and 26%). Turning to venue-based network variables, high constraint means that musicians cluster by

playing in a redundant and, hence, indistinguishable set of venues in terms of musicians playing there. Constraint turns from making success slightly likely in the early career (11% increase) to making it slightly unlikely in the mid career (7% decrease) and late career (13% decrease) stages.

The answer to the last research question is that too dense ego networks constrain success in all performance-based networks and career stages. The exception is that it is beneficial to start careers by playing in venues that host a redundant set of musicians.

5.5 Discussion

Summary

Field theory posits that agents in markets strive for revenue while agents in artistic fields strive for autonomy. However, many artists in EDM do strive for commercial success [Rei11]. This hipster paradox creates a dilemma. On the one hand, artists strive to make a living from performing live and releasing music; On the other hand, commercial success is collectively despised due to the counter-cultural roots of EDM. It has been proposed that musicians solve this dilemma by embedding into alternative and exclusive communities in which work and play fuses [Ort18].

We find that embedding into communities that derive from social relations at live gigs is, indeed, most strongly associated with success for an average musician. This is particularly the case in the late career stage where, on average, success tends to decrease. However, in the early and mid career stages, it is mainstream communities in the core of the field, not alternative communities in the periphery, that increase the likelihood of success. It is only in the late career stage that mainstream belonging is negatively associated with success. This finding gives nuance to the explanation that embedding into alternative communities is the path to success all the way through. Yet, we do find indirect empirical evidence that distancing from others is important as positions between communities are never associated with success. Boundaries around exclusive communities, in other words, matter. In addition to all explanations proposed so far, we find that it is also important that gig communities avoid redundancy so that musicians can leverage the creative potential of varied contacts.

Our findings become even more nuanced if we contrast gigs with venues. Venues are known to be drivers of communities where musicians meet their exclusive crowds [LB13]. Here, we find weak evidence for a crossover effect. In the early career stage, successful musicians play in venues that host a redundant community of artists. As their careers progress, it becomes increasingly important to perform in the mainstream venues of the field. Finally, by releasing music artists demonstrate their seriousness [McL01].

We do not find this practice to contribute to an explanation of travel-based success.

In sum, our results constitute evidence of a structural trade-off among revenue and autonomy. Musicians in EDM embed into exclusive performance-based communities for autonomy but, in earlier career stages, seek the mainstream for commercial success.

Sociological Interpretation

Our results find sound sociological interpretation in a network approach to fields [Bou93, Whi08]. Agents in fields have strategic selection principles called *habituses* which generate concrete practices in the face of collective behavior. In this theoretical framework, the trade-off among revenue and autonomy is a result of habituses operating in the field of EDM. Musicians observe which peers perform in which venues or release on which labels, what success they achieved at which risk to legitimacy, take into account past experiences, and perform the next steps based on these meanings. The inevitable result of the operation of habituses is that fields have a core that harbors its mainstream behavior [Whi08, pp. 147]. This means that, for the field of EDM, even if musicians try to be alternative and avoid doing what most do, some kind of mainstream behavior will always emerge. This is the essence of the hipster paradox at large.

Methodologically, the formal approach we have followed [EM18] models behavior as a *duality* of practices and fields. That means, fields are emergent outcomes of collective practices, but they also set expectations for, and influence, future practices. We have operationalized this model using bipartite networks. The four projected networks we analyze are not social networks of manifest social relations but meaning structures of symbolic relations. That means, the graph-theoretical approach is true to the original idea of field theory [Moh13].

Limitations and Future Work

Our dependent success variable is derived from travel trajectories and is, hence, a proxy for success. We have done so because, in our research design, success must be measured for rolling time windows. In principle, success can also be defined in various other ways such as record sales, record label deals, prices, or online popularity such as the number of followers on social media platforms. If historic data can be leveraged, future studies could use different non-proxy metrics or combine multiple metrics in a compound measure.

This study considers the practices of performing and releasing. However, self-promotion is becoming ever more important [ADJ15, Ort18]. Social media platforms such as *SoundCloud* and *Instagram* allow artists not only

to promote themselves on a global scale, but also to connect and interact with their peers in new ways. What is more, our study is restricted to the production side of cultural objects. But their consumption also leaves digital traces, for example, in the form of likes, mentions, and purchases. Future studies could also account for the impact of self-promotion and cultural consumption.

The dataset comes with a number of limitations. For the most, the recency and self-selection in RA may bias the results of this study to certain musicians and music genres and a particular time period. For example, the number of gigs, artists and venues that register in the website show an exponential increase over time. The self-selection results in over representation of certain artists. For example Tech house, Techno, Minimal, Deephouse, and House account for more than 50% of releases and events in the datasets. Furthermore, name ambiguity, inaccurate and faulty content, and APIs errors could introduce errors in our dataset. However, our manual evaluations show these errors are marginal.

5.6 Conclusion

We have studied the hipster paradox as it yields an interesting behavioral dilemma in the field of EDM. Our results support the explanation offered by the EDM literature, namely, that musicians embed into exclusive performance-based communities to be autonomous in their quest for success. Our longitudinal study allows to refine this explanation since we find behavioral differences between musicians in different career stages. In earlier career stages, musicians seek the mainstream for commercial success. Cultural production in the field of EDM cannot be explained by a polar distinction between art and money. Instead, our results point towards a structural trade-off among revenue and autonomy.

We hope our approach highlights that large-scale digital behavioral data, together with computational methods and social theories, allow to gain new insights into social phenomena such as the hipster paradox. Besides the explanations offered in the EDM literature, we also relied on general theories like field and network theory and the formal methods they provide. We believe that our approach is quite generic and can be used to study other fields of cultural productions, especially music.

Part III

Inequality on Social Media

Introduction

Social media platforms hold a strong position in guiding public opinion. They have become essential channels, among others, for communication and information consumption. People turn to these platforms to receive and share information with their peers, communities, and the public at large. Similar to traditional media, the information generated or disseminated by social media users might be biased toward a particular ideology or beliefs. However, the dynamic of information generation and distribution in social media platforms could constitute new challenges to deal with this issue. On the one hand, contrary to traditional media, in which content is generated by a group of professional and domain knowledge experts, every social media user could participate in providing content about any topic. As a result, social media users could transfer their existing biases into the online space. On the other hand, the algorithms that run these platforms could amplify such biases by promoting them further. For example, Facebook's news feed algorithm could recommend content that conforms with existing users' beliefs in order to increase their engagement with the platforms [BYF+22].

Wikipedia's popularity, its influential role in shaping public opinion and its collaborative and community-based content production processes make it an interesting use case to study online biases. The Wikipedia editor community could enforce their own existing biases to the public opinion through the content and structure of Wikipedia articles. **Chapter 6** looks at the issue of gender bias in Wikipedia and asks if successful men and women – those who are recognized for their achievements – receive equal treatment and attention by the Wikipedia community.

Chapter 6

It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia

Abstract. Wikipedia is a community-created encyclopedia that contains information about notable people from different countries, epochs and disciplines and aims to document the world's knowledge from a neutral point of view. However, the narrow diversity of the Wikipedia editor community has the potential to introduce systemic biases such as gender biases into the content of Wikipedia. In this paper we aim to tackle a sub problem of this larger challenge by presenting and applying a computational method for *assessing gender bias on Wikipedia along multiple dimensions*. We find that while women on Wikipedia are covered and featured well in many Wikipedia language editions, *the way women are portrayed* starkly differs from the way men are portrayed. We hope our work contributes to increasing awareness about gender biases online, and in particular to raising attention to the different levels in which gender biases can manifest themselves on the web.

6.1 Introduction

Wikipedia aims to provide a platform to freely share the sum of all human knowledge. It represents an influential source of information on the web, containing encyclopedic information about notable people from different countries, epochs and disciplines that is used for learning and educational purposes worldwide. Wikipedia is also a community-created effort driven by a self-selected set of editors. The demographic characteristics of this set of editors is known: it is predominately white and male [LUD+11, CB12, HS13].

This known gender bias in the population of editors has the potential to introduce gender biases into the contents of Wikipedia as well. For example, the population bias might lead to differences in the ways women and men are

portrayed on Wikipedia. It might also mimic or even exaggerate inequalities that are already existing in the real world. At the same time, assessing the manifold and subtle ways in which gender biases can manifest themselves has been challenging, and we know little about the different dimensions of gender biases on Wikipedia. Yet, due to the influential nature of Wikipedia, it is important to reveal, assess and correct such biases, if they exist. This paper tackles a sub-part of this larger challenge.

Objectives: In particular, the overall goal of this work is to *assess potential gender inequalities in Wikipedia articles* along different dimensions.

Approach: To assess the extent to which Wikipedia suffers from potential gender bias, we analyze articles about notable people in six language editions along four different gender bias dimensions: coverage bias, structural bias, lexical bias and visibility bias. *Coverage bias* determines differences between the number of notable women and men portrayed on Wikipedia. For example, one might hypothesize that notable men are more likely to be covered by Wikipedia. *Structural bias* quantifies gender homophily/disassortativity, i.e. gender-specific tendencies to preferably link articles of notable people with the same or different gender. For example, one might hypothesise that articles about women have more links to men than vice versa. *Lexical bias* reveals inequalities in the words used to describe notable men and women on Wikipedia. For example, articles about women are potentially more likely to mention their family (husband or kids) than articles about men. *Visibility bias* reflects how many articles about men or women make it to the front page of Wikipedia. Again, one can hypothesize that articles about men might have better chances to be selected.

Contributions & Findings: We present and apply a computational method for *assessing gender bias on Wikipedia along multiple dimensions.* We find that most Wikipedia language editions exhibit a slight over-representation of women, but the proportional differences in the coverage of men and women are not significant. That means, men and women are covered equally well in all six Wikipedia language editions. Also on the visibility level, we do not find any evidence for male-bias in the selection procedure of articles that are featured on the startpage of the English Wikipedia. These are encouraging findings suggesting that the Wikipedia editor community is sensible to gender inequalities[1] and covers notable women and men equally well. However, we also find that *the way women are portrayed* on Wikipedia starkly differs from the way men are portrayed. We find evidence for both structural and lexical gender biases. On a structural level, we observe an asymmetry: Women on Wikipedia tend to be more linked to men than vice versa. On a lexical level we find that especially romantic relationships and family-related issues are much more frequently discussed on Wikipedia articles about women than men.

[1] also cf. `http://meta.wikimedia.org/wiki/Gender_gap`

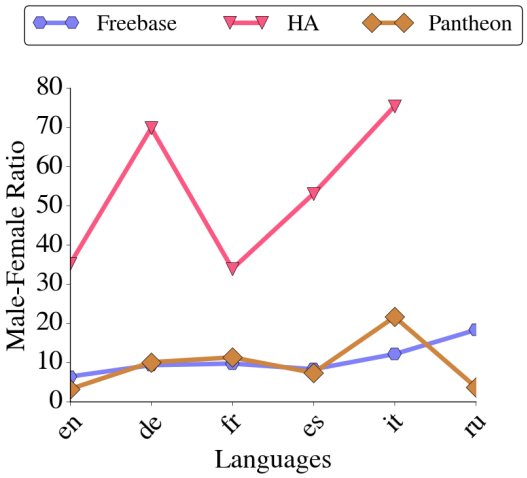


Figure 6.1.1: **Male-Female Ratio:** The ratio of men and women in our reference datasets that are born in a country where one of the six languages is predominantly spoken. Across all language editions the local heroes of a country tend to be predominantly male. For example, if we look at notable people in freebase we find between 7 and 12 times more men than women depending on which countries we consider.

6.2 Materials & Methods

In the following we discuss our data collection and our methodology that allows to systematically explore gender inequalities on Wikipedia on multiple dimensions.

6.2.1 Datasets

To estimate the bias on Wikipedia that goes beyond the bias in the offline world, ideally one would have a complete list of notable people available that is (a) not biased and (b) independent from Wikipedia. Since it is impossible

Table 6.1.1: **Statistics of the datasets:** The number of articles and median article length of all Wikipedia articles that belong to one of the notable people from our three reference datasets.

	Freebase	HA	Pantheon
Total Num Articles	109,481	4,002	11,341
Female Articles	12,685	88	1,496
Male Articles	96,796	3,914	9,845
Median Num Words Female	458	1,121	1,106
Median Num Words Male	412	820	1,017

to obtain such a list, we use the following three collections of notable people as *reference datasets*, each having different strength and weaknesses:

Freebase: We use a collection of around 120k notable people that has been used in previous research for studying the mobility of notable people [SSA+14] and was obtained from freebase. Freebase contains data harvested from sources such as Wikipedia, NNDB, FMD and MusicBrainz, as well as individually contributed data from users. We only take individuals into account for which gender and basic bibliographic information (i.e.,full birth and death date and birth and death location) is available. Freebase directly links to Wikipedia articles in different language editions, if articles about the entity are available.

Pantheon: Pantheon is a project developed by the Macro Connections group at the MIT Media Lab that is collecting, analyzing, and visualizing data on historical cultural popularity and production. The Pantheon dataset [YRH+16] contains information on 11,340 biographies that have presence in more than 25 languages in the Wikipedia (as of May 2013) and provides links to Wikipedia articles about these people.

Human Accomplishment: The third dataset which we use is compiled from a book called "Human Accomplishment" [Mur03] (short HA) and contains information on 4,002 eminent individuals from arts and sciences who made a significant contribution prior to 1950. The inventories were constructed by Charles Murray using linguistic records, such as encyclopedia entries from a number of different languages and sources. Also this dataset has biases since e.g. Murray relied mainly on materials in Roman-alphabet languages. To find Wikipedia articles about those individuals, we use the Wikipedia search API and search for the full name. To select the right search result from the list we compare the birth date, birth location, death date and death location of the candidates in the search results with the person we are looking for.

Data Collection Procedure: We crawled the content of articles about people in our reference datasets using Wikipedia's API in November 2014. For the English Wikipedia, the articles that have been featured at the front page in the last few years were extracted from the "Today's Featured Article" archive[2]. Table 6.1.1 provides the basic statistics for each dataset and Figure 6.1.1 shows the ratio between men and women that are born in a country where one of the six languages we studied is predominantly spoken. The overlap between the three reference datasets is very low. For example, for those people from our reference datasets which we could map to the English Wikipedia the Jaccard coefficient is 0.016 for freebase and HA, 0.035 for freebase and pantheon and 0.097 for pantheon and HA. The six language editions that we explore in this study are those which had the highest coverage of notable men and women from our largest reference dataset,

[2] `http://en.wikipedia.org/wiki/Wikipedia:Today%27s_featured_articlef`

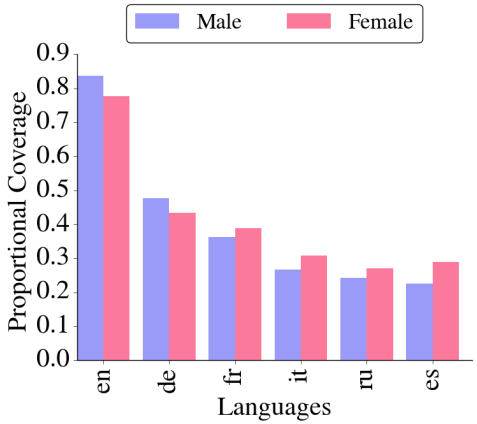


Figure 6.2.1: Freebase

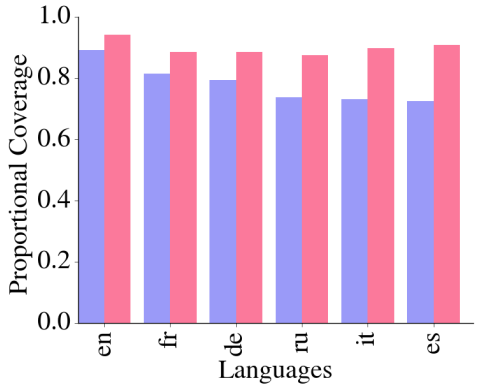


Figure 6.2.2: HA

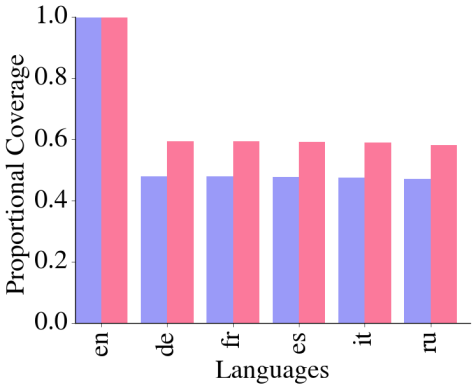


Figure 6.2.3: Pantheon

Figure 6.2.4: **Coverage Bias:** Proportional coverage of notable women and men. Surprisingly, in most language editions the proportion of notable women covered is slightly higher than the proportion of notable men.

freebase.

6.2.2 Measuring Gender Inequality

We propose to analyze gender inequality on Wikipedia on the following four dimensions: which notable men or women are presented on Wikipedia (coverage bias)? How are they presented (lexical bias)? What structure emerges from the hyperlink network of articles (structural bias)? And which articles get featured on the startpage of Wikipedia (visibility bias)?

Coverage Bias: To estimate coverage bias we compare the proportions of notable men and women of different reference datasets that are covered by Wikipedia. Ideally, a reference dataset consists of an unbiased list of people who should be presented on Wikipedia. It is important to understand that a biased reference dataset will obviously impact our results. If, for example, our reference dataset is already biased towards men (i.e., it covers only extremely famous women but also less famous men) than the proportion of women who are represented on Wikipedia would probably be higher than the proportion of men. To address this issue we analyze the coverage using several independent reference datasets (Jaccard coefficient between the three datasets ranges from 0.0 to 0.12 for different language editions), assuming that each of them will have a different bias and seeking patterns that exist across all three datasets.

Further, gender-differences in the extent to which men and women are covered on Wikipedia may exist. Therefore, we also analyse the article length distribution of men and women.

Structural Bias: We analyze the patterns of gender assortativity based on the probability that an article about a person of one gender links to an article about a person of the other gender. We compare the probability that a link ends in an article of gender g_2 given that it comes from an article of gender g_1 with the probability that a link ends in an article of gender g_2 regardless of the gender of its origin:

$$L(g_1, g_2) = log\left(\frac{P(to = g_2|from = g_1)}{P(to = g_2)}\right) \tag{6.1}$$

where $P(to = g_2|from = g_1)$ is the conditional distribution that an edge links to an article of gender g_2 given that it comes from an article of gender g_1, and $P(to = g_2)$ is the probability that any link ends in an article of gender g_2 regardless of the gender of its origin. L measures the log likelihood ratio between edge probabilities, comparing the posterior probability of finding a gender at the edge of a link given that we know the gender of its origin, and comparing it with the base rate of linking to an article of gender g_2. This way, positive values of L indicate increased connectivity from g_1 to g_2, and negative values the opposite, and define a c assortativity matrix of the four

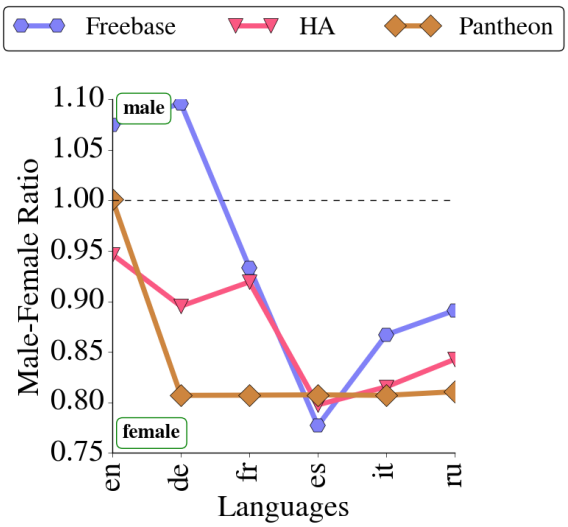


Figure 6.2.5: **Coverage Gap:** Ratio between the number of notable men and women from three different reference lists that are covered on different language editions of Wikipedia.

combinations of genders that measures the tendencies to connect within and across genders.

For the case of same gender connections we use the standard definition of assortativity [PDL18]:

$$\frac{\sum_g P(from = g, to = g) - P(from = g) * P(to = g)}{1 - \sum_g P(from = g) * P(to = g)} \tag{6.2}$$

For the case of asymmetry across genders, we compare the entries of L from one gender to the other, as $A = L(F, M) - L(M, F)$. Positive values of A will indicate a stronger tendency of articles about women to connect to articles about men than the opposite, controlling for the difference in in-degrees and sizes of both genders.

The finding of gender assortativity and asymmetry between genders requires a test that allows us to compare our empirical estimates against null models of the network. For that reason, we set up numerical simulations of three different null models: a *randomized gender* model in which we shuffle the genders of nodes; a *randomized link end* model in which we rewire links to random articles, maintaining out degrees but fully randomizing in-degree; and a *randomized link origin* model, in which we maintain link ends but rewire their origin to an article sampled at random, which maintains in-degrees but randomizes out degrees. We run each simulation 10,000 times, recording values of assortativity and asymmetry to measure the mean and 95% confidence intervals of these two statistics under each null model.

Structural biases can also manifest in the centrality measures, as suggested by the *Smurfette principle* [Mur91]. That means, women can be positioned in the periphery of a network with a core composed of men. In that case the centrality of women would be lower. We operationalize centrality on Wikipedia as a quantification of importance, measuring the in-degree and k-coreness of an article. The in-degree of article p is trivially calculated as the amount of articles that link to article p, and the in k-coreness is computed through a pruning mechanism based on in-degree [GTV13].

Lexical Bias: To explore gender-specific lexical inequalities on Wikipedia we use an open vocabulary approach, inspired by [SEK+13]. An open-vocabulary approach is not limited to predefined word lists, but linguistics are automatically determined from the text. We compute the tfidf scores of the word stems obtained from a Snowball Stemmer and use them as features to train a Naive Bayes classifier. The classifier determines which words are most effective in distinguishing the gender of the person an article is about. Log likelihood ratios $L(word, g)$ are used for comparing different feature-outcome relationships.

$$L(word, g) = log\left(\frac{P(word|g)}{P(word)}\right) \tag{6.3}$$

where $P(word|g)$ is the conditional distribution that a word shows up in an article about a person given that the person's gender is g, and $P(word)$ is the probability that a word shows up in any article regardless of the gender of the person the article is about.

The *Finkbeiner test* [Fin13] suggests that articles about women often emphasize the fact that she is a woman, mention her husband and his job, her kids and child care arrangements, how she nurtures her underlings, how she was taken aback by the competitiveness in her field and how she is such a role model for other women. Also the historian Gillian Thomas who investigated the role of women in Britannica states in her book [Kar93] that as contributors, women were relegated to matters of "social and purely feminine affairs" and as subjects, women were often little more than addenda to male biographies (e.g., Marie Curie as the wife of Pierre Curie).

We create the following three categories of words that capture some aspects that could be over-represented in articles about women according to what Thomas observed in the Britannica and what the Finkbeiner test suggest:

- *Gender* category contains words that emphasize that someone is a man or woman (i.e., man, women, mrs, mrs, lady, gentleman)
- *Relationship* category consists of words about romantic relationships (e.g., married, divorced, couple, husband, wife)
- *Family* category aggregates words about family relations (e.g., kids, children, mother, grandmother).

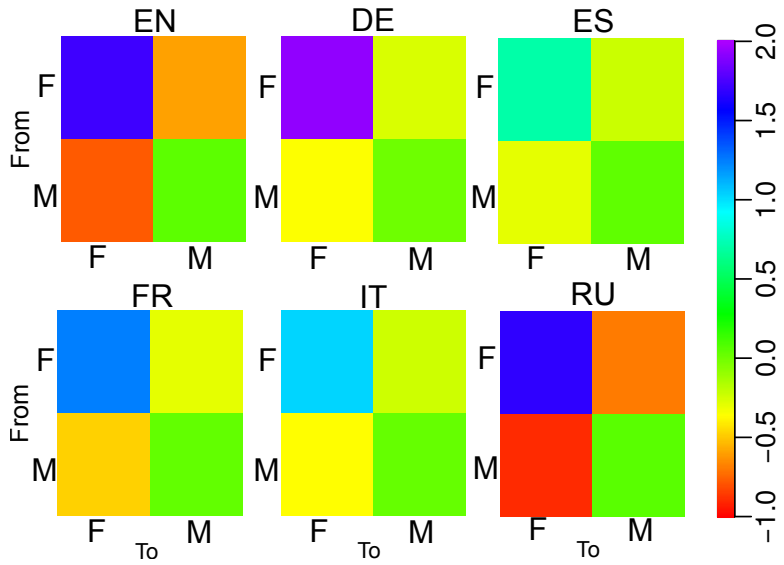


Figure 6.2.6: **Structural Assortativity and Asymmetry Bias:** Logarithmic assortativity matrices for the hyperlink networks of articles about notable men and women in six language editions of Wikipedia. Assortativity of connections within genders becomes apparent for the minority class, women. All language editions show an asymmetry of connectivity across genders. The strongest assortativity and asymmetry is visible in the English and Russian Wikipedia.

All other words that cannot be assigned to the above mentioned categories fall into the category *Others*. To gain further insights into the types of words that have the highest log likelihood ratio for articles about men or women, native speakers of each language manually code the 150 words which are most useful for differentiating articles about men and women in each language edition.

Visibility Bias: To estimate visibility bias we simply compare the proportions of notable men and women of different reference datasets that got featured on the startpage of the English Wikipedia. We test the significance of the difference in proportions between men and women that got featured using a Chi-Square test.

6.3 Results

In the following, we present our empirical results on gender inequality on Wikipedia.

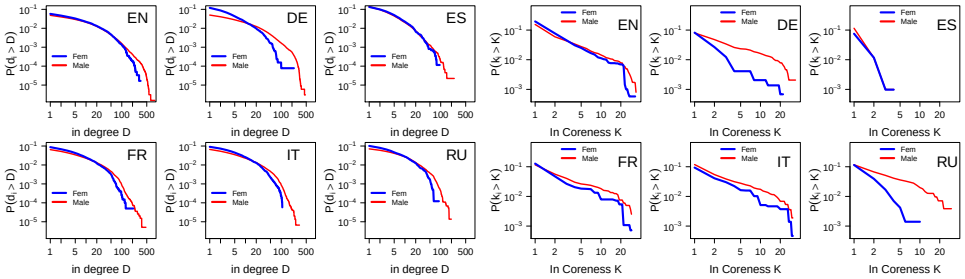


Figure 6.3.1: **Structural Centrality Bias:** Complementary cumulative density function of the in-degree distributions (left) and in k-core decompositions (right) of articles about men and women in six language editions. In some language editions like the English (EN), the Russian (RU) and the German (DE) one, men are always significantly more central than women, no matter how we measure centrality, while in others like the Spanish (ES) one, women and men are either equally central or women are more central.

6.3.1 Coverage Bias

Figure 6.2.4 shows that the best coverage across languages is achieved for people that made significant contributions to science and arts before 1950 and are therefore listed in the HA reference dataset. Across all three reference datasets we consistently observe that *women are* not - as initially hypothesized - underrepresented on Wikipedia, but are even *slightly overrepresented* (cf. Figure 6.2.5). Also when looking at article notable distributions of men and women, we see that *articles about women tend to be longer* than articles about men (cf. Table 6.1.1) in all three datasets. This could potentially be the result of the effort of Wikipedians to improve the coverage of minorities such as women or it can be a side product of a bias in our reference datasets which may only include very notable women, but may also cover less notable men. We addressed the later issue by selecting several reference datasets which we hope are not all subject to the same bias.

6.3.2 Structural Bias

Figure 6.2.6 shows the logarithmic assortativity matrices of articles about men and women in six different language editions of Wikipedia based on our largest reference dataset, Freebase. The assortativity of connections within genders becomes apparent for the minority class, women, in all cases (cf. high values of $L(F, F)$). The matrices also provide a comparison across genders: $L(F, M)$ and $L(M, F)$ are both slightly negative in all language edition, which means that women connect less to men and men less to women than we would expect. All language editions show an asymmetry of connectivity across genders, even when we correct for overall incidence in Equation 6.1.

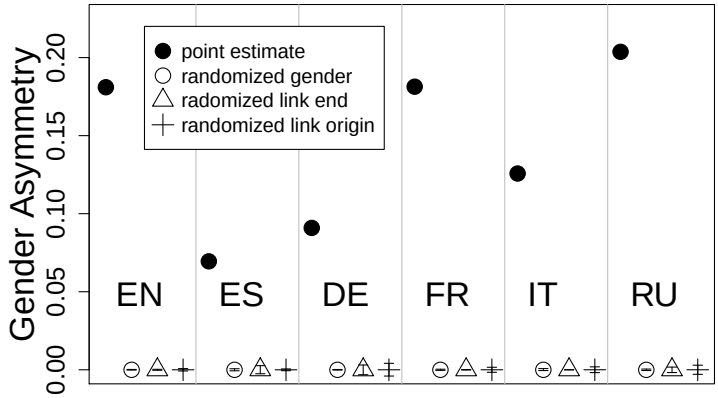


Figure 6.3.2: **Significance of Structural Asymmetry Bias:** Arithmetic mean of point estimates of gender asymmetry for men and women in six language editions and comparison with the three null reference models. Error bars (smaller than symbol size) show 95% confidence intervals over 10,000 simulations of each model. The empirical estimates are significant in comparison to the narrow confidence interval of the null models.

The value of $L(F, M)$ tends to be higher than $L(M, F)$, which means that men link even less to women than women to men.

Figures 6.3.3 and 6.3.2 show the arithmetic mean of the empirical point estimates of assortativity and asymmetry for bother gender, in comparison with the values in the three null models. It is evident that the three randomization methods destroy any kind of assortativity or asymmetry pattern, and that the empirical estimates are significant in comparison to the narrow confidence interval of the null models. Assortativity is positive in all cases, indicating that *articles about people with the same gender tend to link to each other*. For the case of asymmetry, there is a positive value of A (which we defined as $A = L(F, M) - L(M, F)$) in all six language editions, validating our observation that *articles about women tend to link more to articles about men than the opposite*.

The above results show the existence of assortativity and asymmetry across genders controlling for degree. However, structural biases can also manifest in the centrality measures, as suggested by the *Smurfette principle* [Mur91]. To test the existence of this principle, we compare in-degree and k-coreness of articles about men and women on Wikipedia. Figure 6.3.1 shows the complementary cumulative density functions $P(d_i > D)$ for in-degree and $P(k_i > K)$ for in k-coreness in the six networks. An initial observation reveals that, in general, the tail of in-degree and in k-coreness of male articles is longer than for women articles, which is specially pronounced in the case of k-coreness of German and Russian. We validate the above observations by measuring the distance between the two distributions and

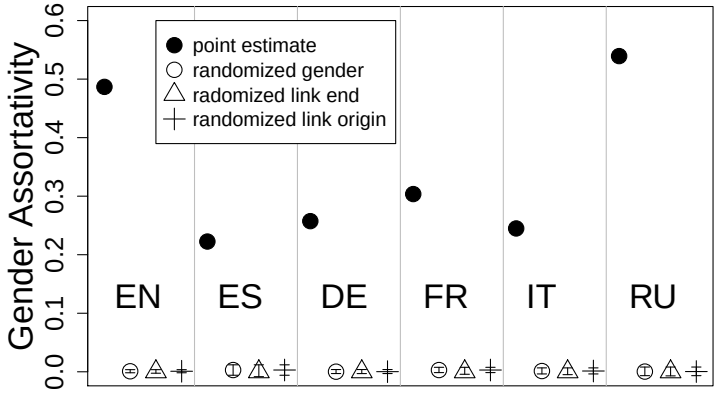


Figure 6.3.3: **Significance of Structural Assortativity Bias:** Point estimates of gender assortativity in six language editions and comparison with the three null reference models. Error bars (smaller than symbol size) show 95% confidence intervals over 10,000 simulations of each model. The empirical estimates are significant in comparison to the narrow confidence interval of the null models.

test the significance of the distance through a two-tailed Wilcoxon tests and Kolmogorov-Smirnov test (cf. Table 6.3.1). Our results highlight that, according to their in-degree distribution, men are indeed significantly more central in all language editions with $p < 0.05$ except in the Spanish one where men and women are equally central. The k-coreness distributions suggest that in all language editions except the Spanish, the Italian and the French one, men are more central then women. This indicates, in some language editions like the English, the Russian and the German one, men are always significantly more central than women, no matter how we measure centrality.

6.3.3 Lexical Bias

Our lexical analysis reveals that articles about women tend to emphasize the fact that they are about a women (i.e., they contain words like “woman”, “female” or “lady”), while articles about men don’t contain words like “man”, “masculine” or “gentleman”. The lower salience of male-related words in articles about men can be related to the concept of male as the *null gender* [Har08], which suggests that there is a social bias to assume male as the standard gender in certain social situations. This would imply that male-defining words are not necessary because the context already defines the gender of the person the article talks about. This seems to be a plausible assumption due to the imbalance between the number of articles about men and women (cf. Table 6.1.1).

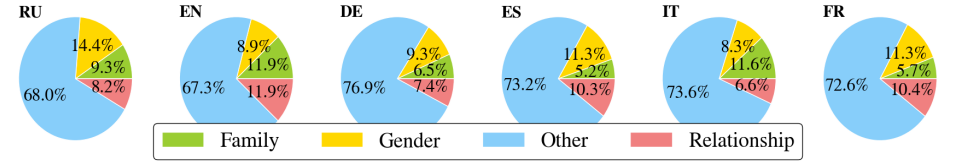
We also noticed that the relationship status and family related issues seem to be more extensively discussed in articles about woman since words like "married", "divorced", "children" or "family" are much more frequently used in articles about women. This confirms that *men and women are indeed presented differently* on Wikipedia and that those differences go beyond what we would expect due to the history of gender inequalities - i.e., the fact that it was more difficult for women to become famous in the past, amongst others because of unequal access to resources and the fact that the history was mainly documented through the eyes of men. We leave the question of investigating if the lexical bias on Wikipedia reflects the lexical bias from the general media or if the Wikipedia editor community introduces an additional bias because of their narrow demographics for future work.

We use log likelihood ratios for comparing different word-gender relationships. Not surprisingly, the most indicative words for men are often related to certain domains or fields (e.g., certain sports or professions). For example, the most discriminative word stems for men in the English Wikipedia are "basebal", "footbal" and "infantri" and an article that contains a word with the stem "basebal" is 11.5 times more likely to be about a man than a woman.

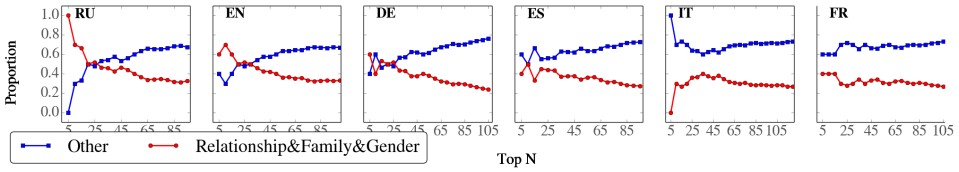
For women the picture is different since among the most discriminative words for women, words like "husband", "female" and "woman" can be found. To gain more insights into those difference, we use the previously introduced categories of words and manually code the words with the highest likelihood ratio for men or women. Our results clearly show that across all language editions almost all words that fall into the category Family, Relationship or Gender, reveal a high likelihood ratio for women. Figure 6.3.4a shows that between 32% and 23% of the 150 most indicative words

Table 6.3.1: **Significance of Structural Centrality Bias:** Differences between the in-degree distributions (W_i) and k-coreness distributions (W_k) of men and women. A positive difference (+) indicates that women are more central, while a negative difference (−) indicates that men are more central. The significance of the difference as suggested by the Wilcoxon test ($p_i <$) and by the Kolmogorov-Smirnov test ($ks_i <$). In some language editions like the English (EN), the Russian (RU) and the German (DE) one, men are indeed significantly more central than women according to both centrality measures.

	W_i	$p_i <$	$ks_i <$	W_k	$p_k <$	$ks_k <$
EN	−	10^{-15}	10^{-15}	−	0.03	10^{-4}
ES	+	0.17	0.02	+	10^{-4}	10^{-4}
DE	−	10^{-15}	10^{-15}	−	10^{-12}	10^{-8}
FR	−	10^{-9}	10^{-5}	−	0.07	0.09
IT	−	10^{-6}	10^{-3}	+	0.95	10^{-4}
RU	−	10^{-4}	10^{-7}	−	0.55	0.003



(a) Proportion of the 150 most discriminative words for women per category



(b) Proportion of the N most discriminative words for women per category

Figure 6.3.4: **Lexical Bias:** The proportion of the 150 most discriminative words of articles about women that belong to different categories. In all language editions between 32% and 23% of the 150 most indicative words for women belong to one of the three categories, while only between 0% and 4% of the most discriminative words for men belong to one of these categories. In some language edition, like the Russian (RU), the English (EN) and the German (DE) one, the proportion of the most discriminative words that belong to one of these three categories is especially high among the top words.

for women belong to one of the three categories. Note that for men only 0% and 4% of the most discriminative words belong to one of these categories. That means, words that fall into one of those categories indeed indicate that an article is about a woman which suggests that lexical gender inequalities are present on Wikipedia. Especially, in the Russian and English Wikipedia, we can see that the majority of the 25 most discriminative words of females fall into one of those three categories (cf. Figure 6.3.4b).

What are these words that fall into the categories Family, Relationship or Gender and discriminate men and women? Table 6.3.2 and 6.3.3 show the word stems with the highest gender-specific log-likelihood ratio that belong to one of the three categories. Almost all of them are indicative for women which means that words which are indicative for men tend not to fall into these categories. One can further see that, for instance, in the English Wikipedia an article about a notable person that mentions that the *person is divorced is* 4.4 *times more likely to be about a woman* rather than a man. We observe similar results in all six language editions. For example, in the German Wikipedia an article that mentions that a person is divorced is 4.7 times more likely about a women, in the Russian Wikipedia its 4.8 time more likely about a woman and in the Spanish, Italian and French Wikipedia it is 4.2 times more likely about a women.

Table 6.3.2: **English Gender-specific Likelihood Ratios:** Word stems with the highest gender-specific likelihood ratio in the English Wikipedia that belong to one of the three categories (Family, Relationship and Gender).

Category	Term	Female	Male
Relationship	husband	9.2	1.0
Gender	female	8.2	1.0
Relationship	aunt	6.5	1.0
Gender	women	6.4	1.0
Gender	madam	6.1	1.0
Gender	woman	5.6	1.0
Family	grandmoth	5.5	1.0
Gender	girl	5.3	1.0
Gender	mrs	4.9	1.0
Relationship	divorc	4.4	1.0
Gender	ladi	4.4	1.0
Relationship	wed	4.3	1.0
Relationship	marriag	3.8	1.0
Relationship	lover	3.8	1.0
Family	babi	3.7	1.0
Family	sister	3.5	1.0
Family	child	3.0	1.0
Family	mother	3.0	1.0

This example shows that a lexical bias is indeed present on Wikipedia and can be observed consistently across different language editions. This result is in line with [BS14] who also observed that in the English Wikipedia biographies of women disproportionately focus on marriage and divorce compared to those of men.

6.3.4 Visibility Bias

Figure 6.3.5 shows the proportion of notable men and women that showed up at the front page of the English Wikipedia in the past few years. One can see that proportions of men and women that got selected are very small and therefore also the differences are marginal. Though we observe across all years that the proportion of men that were selected and featured at the startpage was slightly higher, the Chi-Square test suggests that the difference in proportions is not significant. Therefore, we conclude that the *selection procedure of featured articles of the Wikipedia community does not suffer from gender bias.*

6.4 Discussion

While Wikipedia's massive reach in coverage ensures that notable women have high likelihood of being represented on Wikipedia, evidence of gender bias surfaces from a deeper analysis of the content of those articles. Our re-

Table 6.3.3: **Spanish Gender-specific Likelihood Ratios:** Word stems with the highest gender-specific likelihood ratio in the Spanish Wikipedia that belong to one of the three categories (Family, Relationship and Gender)

Category	Term	Female	Male
Family	embaraz	9.6	1.0
Gender	mrs	6.1	1.0
Gender	femenin	5.3	1.0
Gender	madam	4.4	1.0
Gender	dam	4.4	1.0
Family	tia	4.4	1.0
Relationship	divorci	4.2	1.0
Relationship	bod	4.0	1.0
Gender	mujer	3.9	1.0
Gender	girl	3.9	1.0
Gender	lady	3.7	1.0
Relationship	parej	3.2	1.0
Relationship	enamor	3.0	1.0
Relationship	matrimoni	2.9	1.0
Relationship	marido	2.7	1.0
Relationship	viud	2.7	1.0
Relationship	amant	2.6	1.0
Relationship	hereder	2.5	1.0
Relationship	sexual	2.4	1.0
Family	niet	2.3	1.0

sults clearly show that subtle lexical and structural gender biases are present on Wikipedia.

Potential explanations for these biases are the following: it is possible that biases are a consequence of (i) the predominantly male editor community and the software design in general that might encourage male contributors and/or (ii) historic and present inequalities between men and women that manifest e.g. in unequal access to resources, unequal media presentation and historic documentation and implicit gender stereotyping (which has been shown to give men an unfair advantage in fame judgements [BG95]). It seems to be plausible that certain biases such as the coverage or structural bias can be explained by historic inequalities and implicit cognitive biases due to gender stereotypes that may lead to the fact that notable men seem to be more present in our minds than notable women. Other biases such as the lexical bias (e.g. the fact that articles about women disproportionately focus on marriage and divorce compared to articles about men) can more likely be explained by the narrow demographics of the Wikipedia editor community and the media portrayal of men and women. We leave the question of exploring the extent to which different factors explain different biases for future research.

Implications: The low coverage and visibility bias suggest that the Wikipedia community covers notable women and men equally. However,

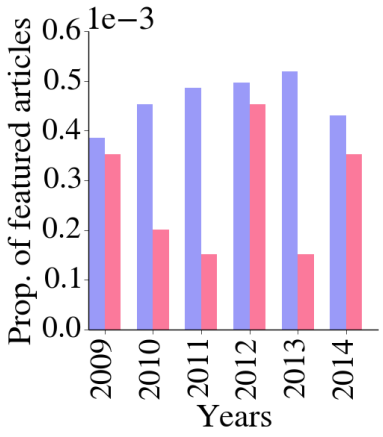


Figure 6.3.5: **Visibility Bias:** The proportion of notable men and women that were featured on the front page of the English Wikipedia in the past few years. One can see that the proportion of men is consistently higher, but the difference is marginal.

our results highlight that editors need to pay attention to the ways women are portrayed on Wikipedia. In particular, the community needs to evaluate the gender balance of links included in articles (e.g., if an article about a woman links to the article about her husband, the husband should also link back), and to adopt a more gender-balanced vocabulary when writing articles about notable people. These existing biases might put women at a practical disadvantage: For example, because modern search and recommendation algorithms exploit both structural and textual information, women might suffer from lower visibility when it comes to ranking articles about notable people or in terms of their general visibility on Wikipedia (at least if we only take links between articles about people into account; see Figure 6 in [EAL+15] for preliminary comparison of ranking algorithms).

Cross-lingual Analysis: We observe the strongest structural bias for the English and Russian Wikipedia. Also on the lexical dimension the strongest bias becomes visible in the English and Russian Wikipedia. Surprisingly the Spanish Wikipedia reveals the lowest structural bias. Comparing our results with the Gender Inequality Index of the World Economic Form (WMF) [The13] shows that a positive correlation exists between the bias in the offline world and the bias on Wikipedia. However, one needs to note that it is difficult to compare our Wikipedia based gender bias rankings of languages with the ranking of countries according to the gender inequality index since countries where the same language is predominantly spoken often reveal very different positions in the WMF ranking. We use the weighted average of the WMF rank positions of countries where the same language is

spoken[3] and weight countries by the size of the internet population[4]. The Spearman rank correlation between the ranking of the 6 languages according to the WMF index shows a correlation of 0.89 with the coverage bias based ranking, 0.37 with the structural bias ranking and 0.09 with the lexical bias ranking. This indicates that to a certain extent gender inequalities of the real world manifest on Wikipedia. However, since the Wikipedia editor community is not representative for the larger population in a country, it is also not surprising that certain biases like the lexical bias do only reveal a very limited relation with the WMF ranking. Although Wikipedia may only reflect certain aspects of gender inequalities of the real world, gender biases that are introduced by the editor community of Wikipedia may effect the larger population and therefore it is important to investigate them.

Reference datasets: Our findings with regard to coverage bias are effected by the (unknown) biases inherent in the reference datasets used. Due to this, we can not make any absolute statements about coverage inequality on Wikipedia. However, regardless of this problem, we can assert that *Wikipedia covers women and men from our reference datasets better equally well.* Using external reference datasets that represent collections of notable people to prune down the number of biographies in Wikipedia rather than studying all of them further helps to uncouple lexical bias and structural bias from coverage bias and ensures that only people that are notable from a global perspective become the subject of study. An alternative would be to select all people from Wikipedia using category pages such as "Births by Year"[5] or "Deaths by Year"[6] as starting point. However, these category pages do not exist in all language editions and therefore the selection would be based on the categories of the English Wikipedia only, which introduces a bias since every language editions tends to focus on their "local heros" [CH11, HG10].

6.5 Related Work

Gender Inequalities in Traditional Media: Feminist often claim that news is not simply mostly about men, but overwhelmingly seen through the eyes of men. In [RC11] the authors analyze longitudinal data from the GMMP (Global Media Monitoring Project) which spans over 15 years. The authors conclude that the role of women as a producer and subject of news has seen a steady improvement, but the relative visibility of women compared to men has stuck at 1:3 which means that the world's new agencies still consider the life of men three time more worth to write about it as those of

[3] http://www.infoplease.com/ipa/A0855611.html

[4] http://en.wikipedia.org/wiki/List_of_countries_by_number_of_Internet_users

[5] http://en.wikipedia.org/wiki/Category:Births_by_year

[6] http://en.wikipedia.org/wiki/Category:Deaths_by_year

women. Gender inequalities also manifest in films that are used for education purposes, as revealed by the application of the Bechdel test to teaching content [SFI12]. In [LNG+13] the authors present a cross-disciplinary, global, bibliometric analysis of the relation between gender and scientific output (i.e., number of papers, citations per paper and internationality of collaborations) using data from more than 5 million scientific publications. They find that the research output in most countries is dominated by males and that the few countries that are dominated by females have lower research output which indicates that barriers are present.

Gender Inequalities on Wikipedia: Our work is not the first work which recognises the importance of understanding gender biases on Wikipedia [RR11, EAL+15, CH11, ALKV12]. In [RR11] thousands of biographical subjects from six reference sources (e.g., The Atlantic's 100 most influential figures in American history, TIME Magazine's list of 2008's most influential people) are compared against the English-language Wikipedia and the online Encyclopedia Britannica with respect to coverage and article length. The authors do not find gender-specific differences in the coverage and article length on Wikipedia, but Wikipedia's missing articles are disproportionately female relative to those of Britannica. Our findings on the coverage dimension confirm their findings and further we also analyze the content of articles on Wikipedia which they left for future work.

In [BS14] the authors present a method to learn biographical structures from text and observe that in the English Wikipedia biographies of women disproportionately focus on marriage and divorce compared to those of men, which is in line with our findings on the lexical dimension. Recent research showed that most important historical figures across Wikipedia language editions are born in Western countries after the 17th century, and are male [EAL+15]. On average only 5.2 female historic figures are observed among the top 100 persons. The authors use different link-based ranking algorithms and focus on the top 100 figures in each language edition. Their results clearly show that very few women are among the top 100 figures in all language editions, but since the authors do not use any external reference lists it remains unclear how many women we would expect to see among the top 100 figures.

Previous research has also explored gender inequalities in the editor community of Wikipedia and potential reasons for it (cf. [LUD+11, CB12, HS13]). Also among Wikipedians, the importance of this issue has been acknowledge for example through the initiation of the "Countering Systemic Bias" WikiProject[7] in 2004.

Gender inequalities in Social Media: In [ST13] the author study a communication network in a MMOG and find a similar effect as [SL00]. Fe-

[7] `http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Countering_systemic_bias`

male players send about 25% more messages (0.74 per day) than males (0.60 per day). Consequently, females show a significantly higher average degree in their communication networks, however, the communication partners of females have a significantly lower average degree than those of males, i.e. females have more communication partners, while males tend to have better connected ones. Recent research [MW14] suggests that in Twitter and Google+ online inequality is strongly correlated to offline inequality, but the directionality can be counter-intuitive. In particular, they consistently observe women to have a higher online status, as defined by a variety of measures, compared to men in countries such as Pakistan or Egypt, which have one of the highest measured gender inequalities. In [GWG14] the authors show that subconscious biases which contribute to the creation of inequality are not only present in movie scripts but also in Twitter conversations. Also the viewing and sharing patterns of youtube videos reveal differences in which content is consumed and discussed by different genders [AGGW14]. This kind of differences also manifest in wall discussions in MySpace, where emotional expression patterns differ across genders [TWU10].

6.6 Conclusions

Wikipedia seems to have successfully established processes that ensure that notable women have a high likelihood of being portrayed on Wikipedia. At the same time, our work surfaces evidence of more subtle forms of gender inequality. In particular, women on Wikipedia tend to be more linked to men than vice versa, which can put women at a disadvantage in terms of - for example - visibility or reachability on Wikipedia. In addition, we find that womens' romantic relationships and family-related issues are much more frequently discussed in their Wikipedia articles than in mens' articles. This suggests that there are gender differences w.r.t. how the Wikipedia community conceptualizes notable men/women. Because modern search and recommendation algorithms exploit both, structure and content, women may suffer from lower visibility in social networks (or article networks) where men (or articles about men) are more central and include more links to other men than to other women. To reduce such effects, the editor community needs to evaluate the gender balance of links included in articles (e.g., if an article about a woman links to the article about her husband, the husband should also link back), and to adopt a more gender-balanced vocabulary when writing articles about notable people. Further, engineers and researchers need to develop a deeper understanding of how different types of search and recommendation algorithms impact the visibility of minorities.

In summary, the contributions of this work are twofold: (i) we present a computational method for assessing gender bias on Wikipedia *along multiple dimensions* and (ii) we apply this method to several language editions

of Wikipedia and share empirical insights on observed gender inequalities. We translate our findings into some potential actions for the Wikipedia editor community to reduce gender biases in the future. We hope our work contributes to increasing awareness about gender biases online, and in particular to raising attention to the different levels in which these biases can manifest themselves. The methods presented in this work can be used to assess, monitor and evaluate these issues on Wikipedia on an ongoing basis.

Chapter 7

Conclusion

The promise of significant economic and employment growth within the meritocratic and barrier-free world of work has attracted the attention of policy makers to the creative industry as the catalyst of the socio-economic development of countries, cities, and villages. Consequently, the creative sector has become one of the fastest-growing parts of the economy worldwide. Yet, inequality remains a major issue among creative workers. The working culture and conditions, production practices, and historical biases contribute to this issue. For instance, in a system with a hostile culture towards minorities the prospect of "full opportunity and unfettered social mobility for all" vanishes [Flo14b]. At the same time, in a system where the main resources are trapped within the boundaries of a closed group of elites, newcomers and less advantage groups and individuals would have a lower chance to succeed in their career. Among others, social relations are the vital resources and defining factor of success within creative careers. Their significance is inherited in the lack of formal structure of, and standard promotion and demotion processes in the creative industry compared to traditional organizations. Instead, careers are shaped as a series of formal and informal interactions and affiliations, resulting in some degree of reputation and recognition. Here, understanding inequality requires investigation into the range of individuals' activities and their production practices over the course of their careers.

This thesis aims to advance our understanding of the dimensions and mechanisms of inequality within creative careers. To reach this objective, this thesis brings together two strands of research. First, it engages with rich theoretical and empirical insights from social sciences to lay the theoretical foundation of this thesis. Second, it taps into the scale and granularity of digital behavioral data to address some of the limitations of previous studies. In particular, it investigates inequality 1) over a large population and demographics to account for the diverse experiences and actions of actors involved in a specific social system; 2) over longer time spans to study the

evolution of inequalities and identify its underlying patterns and mechanisms; 3) along multiple dimensions using multiple constructs and finally 4) over both online and offline spaces. Particular focus is put on identifying patterns of successful careers with respect to social, cultural, and human capitals. Success is primarily defined and measured in the forms of reputation and recognition. Social network analysis and complex network theories are used to operationalize and measure the concepts of capital, career, and social structure. Here, temporal relational data on collaboration and co-affiliation practices provides the essential signals to capture the similarity of individuals and their practices over time. Hence, allowing us to measure the positions of individuals within a specific socio-cultural structure at different points in time.

Obtaining demographics information such as gender are essential to study horizontal inequality. Therefore, this thesis start by offering a novel approach to infer the gender of large heterogeneous population of scientists from their names and images. Chapter 3 investigates horizontal and vertical inequality in the entire computer science discipline using both relational and distributive analyses. It highlights multiple dimensions that the gender gap manifests itself in academic careers and shows how they evolve over time. Chapter 4 follows the distributive approach to further examine the evolution of horizontal and vertical inequalities in computer science. It highlights the role of early-career performance on the chance of future success and shows how people who start their careers at a different point in time might be subjected to varying levels of inequality. Chapter 5 takes a purely relational approach to assess the vertical inequality in Electronic Dance Music (EDM) club scene. It introduces the "hipster paradox" as one of the underlying mechanisms of success in EDM, and discusses how community belonging help aspiring musicians deal with the trade-off between autonomy and mainstream success. The last chapter turns toward the online space as a new domain in which social inequality can arise. It examines Wikipedia articles about notable people and shows that existing gender bias manifests online. To conclude, I provide an overview of the results and contributions of the present work, address its limitations, and briefly elaborate on promising directions for future research.

7.1 Results and Contributions

The main contributions of this work are three folds. First, this thesis collects and offers publicly available datasets to support future research on creative careers. These datasets offer rich historical and relational data on the working practices of groups and individuals within the two sectors of creative the industry: science (i.e. computer science) and music industry (i.e. EDM). Second, the empirical findings in this thesis provide new evidence of the

dimensions and mechanisms of inequalities among creative careers. These findings shed new light on the historical manifestation of inequalities, their social and cultural dimensions, and their structural aspects and properties. Third, it offers a computational social science research framework that aims to fill a methodological gap in inequality studies (see 1.5). That is, to develop interdisciplinary approaches that take various perspectives and methods from different disciplines into account. The proposed framework brings together theories and methods from social and computational sciences that guide different stages of my research to answer the research questions posed in Section 1.3. The following section summarizes the contributions made by this thesis by addressing each question.

RQ1: What are the dimensions of inequality in creative careers?

This thesis identifies multiple dimensions of inequality in creative careers. *Vertical inequality* exists along multiple dimensions, namely productivity, success measured as popularity and recognition, social capital and dropout. Productivity and recognition in each range of practices that makes up the careers in that field. In science, rapid and continues publication of high impact research papers is the main imperative of academics careers. The distribution of productivity (i.e. number of publication) and recognition (i.e. h-index or number of citations) defines the hierarchy of academic fields in which authors take positions. Chapter 4 reveals a high degree of inequality in recognition and productivity in computer science. In music, productivity can be measured as the number of live performances (i.e., gigs) and music releases. They offer musicians visibility, recognition, and financial rewards. In chapter 5 success is measured as the sum of distance that artists traveled for their live performances. The results hint toward a high degree of inequality in terms of productivity and success in EDM careers.

Similar to success, the chance of survival varies among the creative careers. Chapter 4 shows that most authors persist for only one year before they become inactive or drop out of computer science. Similarly, Chapter 4 shows that an average careers in EDM follows a downward trend. Meaning, for most of artists the chance of success drops as they go through their career.

This thesis also examines the structural inequality in creative careers. The configuration of direct and indirect social ties determines the positions of individuals within the structure of their field or discipline, and consequently access to social resources. The findings of this thesis suggest that creative careers are primarily characterized by a core-periphery structure - the arrangement of a network into a dense core and sparse periphery. We observe such a structure in both science (chapter 3) and music careers (chapter 5.

Comparing the career of men and women, this thesis also identifies different dimensions of *horizontal inequalities*. Chapter 3 and chapter 4 highlight

five dimensions of gender inequality in science: participation, collaboration behaviour, productivity, drop-out and success. Women tend to have lower a degree of participation, productivity and success, and higher drop-out rates. The collaboration behavior of men and women researchers also embed into different ego networks. Female networks are significantly smaller, much more clustered, contain fewer brokerage opportunities and are more short-lived than those of men. The gender issue manifests itself in the online space as well. Chapter 6 shows that the content and information structure of online platforms such as Wikipedia can mirror or amplify the existing biases in the offline world. For example, the centrality of pages about notable men compared to women in the information structure of Wikipedia, limit the visibility of women in the online space. This could pose a threat to creative workers such as artists and scientists who rely heavily on online attention to reach their audience.

RQ2: How do inequalities evolve over time? Like any other sociocultural phenomena, the deeper understating of inequalities entails taking into account their historical manifestation. We can gain insight into the social and cultural change in a system and arrive at postulates or facts about the evolution of a system. In this thesis, I examine different dimensions of vertical and horizontal inequalities and show how they change over *time* and *career stages*. First, I look at *horizontal inequalities* concerning the gender gap in academic careers. In Chapter 3 I show that in spite of considerable improvement in the career of women in the last decade, the gender gap still persists in the career of computer science researchers. While the proportion of women have increased, and they have become better integrated within the community, their career still characterized by a higher dropout and lower publication rate. The structural gender inequalities have been shrinking as women started to extend their collaboration, and join more central communities. When we look at the changes over career stages, we observe early-career stages exhibit a higher degree of inequality. For example, the difference between the dropout rate and the publication rate of male and female scientists is the highest in the first five years of academic careers. Chapter 4 shows that people who start their career at different points in time can experience different degree of inequalities. For example, the productivity gap – the higher publication rate of men compare to women – is stronger among people who started their career between 1990 to 2000. Nevertheless, horizontal inequality in productivity always favored male scientists among the most cohorts.

Similarly, *vertical inequality* can exhibits temporal characteristics. First, we may observe if a disciples as whole becomes more or less inclusive and coherent over time. For example temporal analysis of collaboration patterns indicates that the computer science community have become more collaborative and coherence over time (see chapter 3). Second, the degree of

productivity and recognition inequalities may change over time and career stages. Chapter 4 shows that inequality in productivity is slightly increasing over career ages while recognition inequality remains stable. Yet, inequality in recognition is generally larger than inequality in productivity. Similar to horizontal inequality, the degree of vertical inequality varies among some cohorts. Third, careers may exhibit different patterns of success over time. While some could carry out a successful career, the majority may have a little chance to succeed. Chapter 5 identifies five career trajectories and shows an average career in EDM follows a downward career trajectory. That is, success decreases with career age.

It is important to note that the inequality in the online space is also impermanent. Changes in the system design (e.g. algorithms) and content provides (e.g. human editors) could effect the information bias in online platforms. For example in chapter 6 I show the visibility bias in Wikipedia can change over time.

RQ3: What are the underlying mechanisms and process of inequalities in creative careers?

In addition to quantifying trends and evolution of certain phenomena, historical data enables us to discover processes and mechanisms behind inequalities. A change in one aspect of inequality can have an effect on the behaviour or strength of other aspects. By analyzing the interactions and intersections between different forms of inequalities over time we can identify patterns of co-evolution in a system. This thesis mainly investigates the processes and mechanisms behind the inequality in career success. First, it identifies in which ways inequality in productivity and social capital correlate with inequality in success. Productivity shows one of the strongest correlation with success. In science, higher number of publications is associated with highest increase in number of citations and h-index (see chapter 4). In EDM, higher number of live performances is associated with a higher degree of success (see chapter 5).

This thesis indicates that the configuration of direct and indirect social ties is highly associated with inequality of success in creative careers. On a personal level, the intensity, duration, variety, and topology of social relations influence the extent of social support, trust, information novelty, and recognition that individuals receive. On a collective level, those who manage to occupy the central positions and join bigger communities are likely to have a higher chance of success in their career. Hence, the choice of collaboration partners, and the type of collaboration (e.g. one-time or repeated collaboration) are central aspects of career development. Chapter 3 shows that network closure and network brokerage are co-determinants of scientific success. Successful scientists embed in large networks and build trustful relationships through repeating collaborations throughout their careers. But, at the same time, successful scientists also bridge otherwise disconnected

parts of the community (i.e. structural holes) to exploit various knowledge resources and stay innovative. Chapter 4 reveals that senior support, social support, and team size influence the likelihood of success and dropout. For example while publishing with large teams exhibits a negative effect on success, having many unique and senior co-authors increase the likelihood of success. The empirical findings in chapter 5 illustrates that joining exclusive communities that drive from social relations of live performances is associated with success in the EDM subculture.

Chapter 5 also shows despite direct social relations that result from collaborations and interactions, indirect social relations can also make up as important resources. Similarity in taste or affiliation to certain events or entities could signal the belonging of individuals to certain communities. For example, those who belong to larger and more central communities could benefit from higher degree of visibility and recognition.

Moreover, following certain production practices materialize more or better resources than others. In the case of the EDM, live performances are more beneficial that music releasing. Communities that derive from social relations at live gigs are most strongly associated with success. Therefore, knowing and cultivating a right practices of a specific working culture can be an essential part of building a successful career.

Interestingly, we observe there are no gender-specific differences in how collaboration patterns impact success. Despite having different collaboration behaviour, those women who become successful computer scientists exhibit the same collaborative behavior as their successful male colleagues.

This thesis also identifies two general mechanisms that drives inequality. Chapter 4 argues that cumulative advantages or *Mathew Effect* is one of the driving forces behind horizontal and vertical inequalities in computer science. Early-citation advantage, the increasing necessity for persistence in publishing, and the fact that the first three career ages are highly predictive of total-career success support this theory. Chapter 5 introduces *hipster-paradox* – the tension between mainstream success and alternative status– as the governing force behind the career of EDM artists. On the one hand, artists strive to make a living from performing live and releasing music; On the other hand, commercial success is collectively despised due to the counter-cultural roots of EDM. Musicians solve this dilemma by embedding into alternative and exclusive communities in which they can exercise a high degree of autonomy while enjoying recognition from their peers. However, in the early and mid career stages, it is mainstream communities in the core of the field, not alternative communities in the periphery, that increase the likelihood of success.

Finally, chapter 3 propose *leaky pipeline* as an answer to the long-standing issue of productivity puzzle in academia. The lower productivity rate of women stems from their higher dropout rate at different stages of academic careers. Faster dropout translates to shorter career length, lower chance of

occupying senior academic positions, and ultimately lower publication rate.

7.2 Implications and Applications

In general, the methods and findings from this thesis could be beneficial for decision-makers, practitioners, and participants in the creative industry to facilitate equal opportunities in career development. The conceptual and empirical contributions mentioned above could provide some guidelines toward three goals. First, investigating issues such as discrimination, inclusion, and diversity. Second, developing strategies to minimize the existing inequalities. And finally, building and maintaining successful careers. This section gives a short discussion of the potential implications and applications of the insights presented in this work.

Digital behavioural data for social goods. The mainstream debate about the power of social media and digital behavioral data is often discussed in the framing of privacy and digital surveillance, and draws a dystopian picture of democracy and society. The importance of such debates is to keep the stackholders, decision makers and those in power accountable for their decisions and actions. The story that gets less attention is that these systems and the data they generate could provide us with tremendous opportunities to tackle some of the long-lasting issues of our society. This thesis picks up one such issue: inequality within the fast-growing creative industry. Access to temporal data that spans over decades, and granular relational data of millions of interactions, could help us uncover some of the aspects and mechanisms of inequalities that would otherwise stay unnoticed. We can use a similar approach to study other phenomena like sustainable production and consumption, health, and well-being.

Uncovering the socio-cultural structure of creative careers: One of the unique characteristics of creative careers is the lack of a formal organizational structure. In the case of traditional organizations, the hierarchy and career path are clear. People join a company by taking on an entry-level position. Over time they climb the ladders by occupying senior positions within the same organization or other organization with similar roles and structures. Here, it is possible to hold specific people or groups accountable for their decision-making power. However, the complex, hidden, and dynamic nature of structures in creative careers requires a deeper assessment of individuals' and groups' actions and possible consequences. What leads to marginalization and exclusion of specific individuals might not be the result of direct decisions of few stakeholders but emerge from the dynamics of collective actions. Relational data of social interactions within a field provides us with rich signals to assess the content and the structure of relations in that field. Production practices such as co-authorship in science or co-acting in a movie form the basis of social relations in creative disciplines. By ana-

lyzing these web of interactions using social network analysis methods and complex network theories, we can uncover the global (e.g., core-periphery structure) and local structural properties of a discipline (e.g., nodes' degree).

Successful collaboration strategies in science. Co-authorship is arguably one of the central social processes of scientific careers. With whom and how to collaborate can determine the success of scientists to a great extent. The empirical results obtained in chapter 3 point toward some strategies to build a successful scientific career. Scientists with higher h-index or number of citations keep maintain a large network of core collaborators while simultaneously adding diverse collaborators from different of social circles. At the same time, long-lasting research partnerships can lead to collaborations that increase success through increased productivity. Collaborating with successful and senior scientists are also beneficial for a researcher.

Gender disparity in science. The issue of gender disparity in science, specifically in STEM fields has been a central topic in political and scientific discourse. The results in chapter 3 show despite some improvements in the last decades, women are still less likely to survive or succeed in their career. Women still show lower participation and productivity, and higher levels of dropout in every stage of their careers. To address this issue, efforts should be directed toward maintaining an open and barrier-free structure that not only encourages women to follow an academic career, but also ensures a fair and supportive structure that enables women to advance through their career.

Importance of community belonging for EDM musicians. Electronic dance music has been growing rapidly in the last years. More and more artists seek to build a career within this field. Often, the path of financial security is at odd with what artistic freedom - what they love to do or value. This tension creates an overwhelming condition that could put their career at risks. The results of this thesis show that one way to deal with this tension is to join communities of artists with similar production practices; performing at similar parties or releasing on similar style of music. Communities allow artists to gain recognition, support, and legitimacy, while following their interests.

Designing fair socio-technical platforms. The findings in chapter 6 suggests that socio-technical platforms could maintain and amplify the existing biases in society. As content providers, users can influence how people access, interpret, and process information. For example, the demographic bias in content providers can enforce certain views and ideas that misrepresent or suppress the ones from other groups. At the same time, platforms are not only governed by users and platform owners, but also by algorithms. Biases in the content can be picked up by the algorithms and reproduced on a larger scale and a higher speed. To avoid such issues, we first need to identify the ways social and cultural biases manifests themselves in socio-technical

platforms. This thesis proposes an approach to reveal some of these biases. Having identified the issues, we can implement procedures, or modify the design of a system. For instance, the implementation of community rules and guidelines that keep the users in check is an important step toward this goal. Additionally, there should be an effort to design algorithms that are sensitive to social and cultural issues. That is, the aim should not be only optimizing profits, but also minimizing the negative social and cultural impacts of a system.

7.3 Limitations and Future Work

- **Generalization of results.** This thesis discusses various ways in which inequality exists or could arise in certain creative careers. However, while the approach and research design can be applied to other disciplines, the empirical findings in this thesis are not readily generalizable to other disciplines, demographics, or spaces of creative careers. Every discipline has its unique history and socio-cultural characteristics that determine how individuals and groups navigate their careers and access resources. For example, the way people socialize and form ties can differ based on the norms and models of production in their field of work. While musicians meet and interact with their peers through live performances or music releases, co-authorship is the primary type of interaction in scientific careers. Such differences lead to processes that are unique to specific domains or disciplines. Comparative studies can shed light on some of the common characteristics of various creative disciplines.

- **Causality.** This work does not answer causal questions, such as if certain production practices lead to success or if the observed patterns are a consequence of success. Despite the large scale and longitudinal design of the analysis, this thesis could not account for all important variables that might mediate the relationship between dependent and independent variables. One way to achieve this is to combine survey and digital behavioral data. We could measure variables such as personality traits or talents using survey data. Additionally, a more sophisticated statistical modeling could control the duality of social interactions and field structure. While social interactions are the building block of social structure, they are also informed and shaped by it.

- **Gender binary.** The gender analysis in this thesis lacks non-binary labels and does not account for those who do not fall into the male/female gender binary. As one of the most underrepresented demographic groups in society, non-binary individuals have been subjected to vari-

ous forms of discrimination. Future research should focus on acquiring more inclusive and complete data that can be used to assess the working conditions of non-binary and gender-queer individuals.

- **Subjective success.** To study success, this thesis operationalize popularity and recognition as instances of *objective* success. The objective criteria of success are directly observable and, hence easy to measure and verify. In other words, it gives us the advantage to study the career of large populations with less effort. However, individuals' careers can be driven by less tangible factors beyond objective criteria. These factors vary among individuals based on their unique goals, motivations, incentives. Here the goal of a career is not only, or necessarily, to gain profit or higher status, but to fulfill intrinsic rewards such as flexibility or work-life balance. For example, artists who make 'art for art's sake' for personal fulfillment are not necessarily driven by economic and cultural success.

- **Interaction of online and offline space.** This thesis takes a small step toward understanding the ways in which inequality could manifest itself online. An important extension to the study here is to quantify the interaction of online and offline spaces. That is, how certain behaviors or information in the online space could harm or hinder the careers of specific individuals and groups. An example is the impact of online attention (e.g. social media exposure) on the career of male, female, and non-binary creatives. In contrast, inequalities in the offline space could lead to unfair treatment of people in the online space. For instance, the existing prejudice against people-of-color and no-male groups can lower their chance to secure projects in online freelance marketplaces where creative individuals offer services

List of Figures

List of Tables

Bibliography

[AAK12] Daniel E. Acuna, Stefano Allesina, and Konrad P. Kording. Future impact: Predicting scientific success. *Nature*, 489(7415):201–202, 2012.

[AB02] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, January 2002.

[ACORC11] Pedro Albarrán, Juan A. Crespo, Ignacio Ortuño, and Javier Ruiz-Castillo. The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, 88(2):385–397, 2011.

[ADJ15] Daniel Allington, Byron Dueck, and Anna Jordanous. Networks of value in electronic music: Soundcloud, london, and the importance of place. *Cultural Trends*, 24(3):211–222, 2015.

[AGGW14] Adiya Abisheva, Venkata Rama Kiran Garimella, David Garcia, and Ingmar Weber. Who watches (and shares) what on youtube? and when? using twitter to understand youtube viewership. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, page 593–602, New York, NY, USA, 2014. Association for Computing Machinery.

[AK07] Tammy L. Anderson and Philip R. Kavanaugh. A 'Rave' Review: Conceptual Interests and Analytical Shifts in Research on Rave Culture. *Sociology Compass*, 1(2):499–519, November 2007.

[AL90] Paul D. Allison and J. Scott Long. Departmental effects on scientific productivity. *American Sociological Review*, 55(4):469, 1990.

[AL20] Pierre Azoulay and Freda Lynn. Self-citation, cumulative advantage, and gender inequality in science. *Sociological Science*, 7:152–186, 2020.

[ALK82] Paul D. Allison, J. Scott Long, and Tad K. Krauze. Cumulative advantage and inequality in science. *American Sociological Review*, 47(5):615, 1982.

[ALKV12] Pablo Aragon, David Laniado, Andreas Kaltenbrunner, and Yana Volkovich. Biographical social networks on wikipedia: A cross-cultural study of links that made history. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, New York, NY, USA, 2012. Association for Computing Machinery.

[All78] Paul D. Allison. Measures of inequality. *American Sociological Review*, 43(6):865, December 1978.

[AMi17] AMiner. *DBLP-Citation-network V10*, 2017. `https://lfs.aminer.cn/lab-datasets/citation/dblp.v10.zip`).

[ARPS11a] Dag W. Aksnes, Kristoffer Rorstad, Fredrik Piro, and Gunnar Sivertsen. Are female researchers less cited? A large-scale study of Norwegian scientists. *Journal of the American Society for Information Science and Technology*, 62(4):628–636, 2011.

[ARPS11b] Dag W. Aksnes, Kristoffer Rorstad, Fredrik Piro, and Gunnar Sivertsen. Are female researchers less cited? A large-scale study of Norwegian scientists. *Journal of the American Society for Information Science and Technology*, 62(4):628–636, 2011.

[BA03] Jerome T. Bentley and Rebecca Adamson. Gender differences in the careers of academic scientists and engineers: A literature review. Special report, National Science Foundation, 2003.

[BC11a] Wendy Bottero and Nick Crossley. Worlds, fields and networks: Becker, bourdieu and the structures of social relations. *Cultural Sociology*, 5(1):99–119, 2011.

[BC11b] Wendy Bottero and Nick Crossley. Worlds, fields and networks: Becker, bourdieu and the structures of social relations. *Cultural sociology*, 5(1):99–119, 2011.

[BC13] Vladimir Batagelj and Monika Cerinšek. On bibliographic networks. *Scientometrics*, 96(3):845–864, 2013.

[BCR03] Helen Blair, Nigel Culkin, and Keith Randle. From london to los angeles: a comparison of local labour market processes in the us and uk film industries. *The International Journal of Human Resource Management*, 14(4):619–633, 2003.

[BD77] Alan E. Bayer and Jeffrey E. Dutton. Career age and research-professional activities of academic scientists: Tests of alternative nonlinear models and some implications for higher education faculty policies. *The Journal of Higher Education*, 48(3):259–282, 1977.

[BE06] Stephen P. Borgatti and Martin G. Everett. A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466–484, 2006.

[Bec08] Howard S Becker. *Art worlds*. University of California Press, 2008.

[BG95] Mahzarin R. Banaji and Anthony G Greenwald. Implicit gender stereotyping in judgments of fame. *Journal of personality and social psychology*, 68 2:181–98, 1995.

[BJS11] Stefan Bornholdt, Mogens Høgh Jensen, and Kim Sneppen. Emergence and decline of scientific paradigms. *Physical Review Letters*, 106(5):058701, 2011.

[BLM07] Abby Butler, Vicki L. Lind, and Constance L. McKoy. Equity and access in music education: conceptualizing culture as barriers to and supports for music learning. *Music Education Research*, 9(2):241–253, 2007.

[Blo17] Gerry Bloustien. Up the down staircase: Grassroots entrepreneurship in young people's music practices. In *Sonic Synergies: Music, Technology, Community, Identity*, pages 195–209. Routledge, 2017.

[BMBW15] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, 67(1):1–48, 2015.

[Bou83] Pierre Bourdieu. The field of cultural production, or: The economic world reversed. *Poetics*, 12(4-5):311–356, 1983.

[Bou88] Pierre Bourdieu. *Homo Academicus*. Stanford University Press, Stanford, CA, 1988.

[Bou93] Pierre Bourdieu. *The Field of Cultural Production: Essays on Art and Literature*. Columbia University Press, 1993.

[Bou09] François Bourguignon. Crime as a social cost of poverty and inequality: A review focusing on developing countries. *Revista Desarrollo y Sociedad*, 2009.

[BP07] Ulrik Brandes and Christian Pich. Centrality estimation in large networks. *International Journal of Bifurcation and Chaos*, 17(07):2303–2318, 2007.

[Bra34] Samuel C. Bradford. Sources of information on specific subjects. *Journal of Information Science*, 10(4):176–180, 1985 [1934].

[B.S83] Stephen B.Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269–287, 1983.

[BS14] David Bamman and Noah A. Smith. Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics*, 2:363–376, 2014.

[BT79] Gary S. Becker and Nigel Tomes. An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy*, 87(6):1153–1189, 1979.

[BT21] Philipp Brandt and Stefan Timmermans. Abductive Logic of Inquiry for Quantitative Research in the Digital Age. *Sociological Science*, 8:191–210, 2021.

[Bur92] Ronald S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1992.

[Bur95] Ronald S Burt. *Structural holes: The social structure of competition*. Harvard University Press, 1995.

[Bur04a] Val Burris. The academic caste system: Prestige hierarchies in PhD exchange networks. *American Sociological Review*, 69(2):239–264, 2004.

[Bur04b] Ronald S. Burt. Structural holes and good ideas. *American Journal of Sociology*, 110(2):349–399, 2004.

[Bur05] Ronald S Burt. *Brokerage & Closure: An Introduction to Social Capital*. Oxford University Press, 2005.

[BVR18] Thijs Bol, Mathijs de Vaan, and Arnout van de Rijt. The Matthew effect in science funding. *Proceedings of the National Academy of Sciences*, 115(19):4887–4890, 2018.

[BW92] Pierre Bourdieu and Loic J. D. Wacquant. *An invitation to reflexive sociology*. University of Chicago Press, 1992.

[BYF+22] Saumya Bhadani, Shun Yamaya, Alessandro Flammini, Filippo Menczer, Giovanni Ciampaglia, and Brendan Nyhan. Political audience diversity and news reliability in algorithmic ranking. *Nature Human Behaviour*, pages 1–11, 02 2022.

[BZ11] Vladimir Batagelj and Matjaž Zaveršnik. Fast algorithms for determining (generalized) core groups in social networks. *Advances in Data Analysis and Classification*, 5(2):129–145, 2011.

[CAL15] Aaron Clauset, Samuel Arbesman, and Daniel B. Larremore. Systematic inequality and hierarchy in faculty hiring networks. *Science Advances*, 1(1):e1400005, 2015.

[CB12] Benjamin Collier and Julia Bear. Conflict, criticism, or confidence: An empirical examination of the gender gap in wikipedia contributions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, page 383–392, New York, NY, USA, 2012. Association for Computing Machinery.

[CBvLvR09] Rodrigo Costas, Maria Bordons, Thed N. van Leeuwen, and Anthony F.J. van Raan. Scaling rules in the science system: Influence of field-specific citation characteristics on the impact of individual researchers. *Journal of the American Society for Information Science and Technology*, 60(4):740–753, April 2009.

[CC73] R. Jonathan Cole and Stephen Cole. *Social Stratification in Science*. University of Chicago Press, Chicago, IL, 1973.

[CH04] Paul Collier and Anke Hoeffler. Greed and grievance in civil war. *Oxford Economic Papers*, 56(4):563–595, 2004.

[CH11] Ewa S. Callahan and Susan C. Herring. Cultural bias in wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, 62(10):1899–1915, 2011.

[Cha80] Ivan D. Chase. Social process and hierarchy formation in small groups: A comparative perspective. *American Sociological Review*, 45(6):905, December 1980.

[CHOV01] David A. Cotter, Joan M. Hermsen, Seth Ovadia, and Reeve Vanneman. The Glass Ceiling Effect*. *Social Forces*, 80(2):655–681, 12 2001.

[Cit00] Marcia J Citron. *Gender and the musical canon*. University of Illinois Press, 2000.

[Cli93] Norman Cliff. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494, 1993.

[CLS17] Aaron Clauset, Daniel B. Larremore, and Roberta Sinatra. Data-driven predictions in the science of science. *Science*, 355(6324):477–480, 2017.

[Col79] Jonathan R. Cole. *Fair Science: Women in the Scientific Community*. Free Press, New York, NY, 1979.

[Cre01] Creative industries mapping document 1998,, 2001.

[Cro15] Nick Crossley. *Networks of sound, style and subversion: The punk and post–punk worlds of Manchester, London, Liverpool and Sheffield, 1975–80*. Manchester University Press, 2015.

[Cro20] Nick Crossley. *Connecting sounds: The social life of music*. Manchester University Press, 2020.

[CS91] Jonathan R. Cole and Burton Singer. A theory of limited differences: Explaining the productivity puzzle in science. In Harriet Zuckerman, Jonathan R. Cole, and John T. Bruer, editors, *The Outer Circle: Women in the Scientific Community*, pages 277–323. Norton, New York, NY, 1991.

[CSR13] Morgane Ciot, Morgan Sonderegger, and Derek Ruths. Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[CW11] S. J. Ceci and W. M. Williams. Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, pages 3157–3162, 2011.

[CZ84a] Jonathan R. Cole and Harriet Zuckerman. The productivity puzzle: Persistence and changes in patterns of publication of men and women scientists. *Advances in Motivation and Achievements*, 2:17—-256, 1984.

[CZ84b] Jonathan R. Cole and Harriet Zuckerman. The productivity puzzle: Persistence and changes in patterns of publication of men and women scientists. *Advances in Motivation and Achievements*, 2:17–256, 1984.

[CZW15] Rodrigo Costas, Zohreh Zahedi, and Paul Wouters. Do "altmetrics" correlate with citations? extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10):2003–2019, 2015.

[Dan87] Dale Dannefer. Aging as intracohort differentiation: Accentuation, the Matthew effect, and the life course. *Sociological Forum*, 2(2):211–236, 1987.

[DE06] Thomas A. DiPrete and Gregory M. Eirich. Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual Review of Sociology*, 32(1):271–297, 2006.

[DG11] Paul DiMaggio and Filiz Garip. How network externalities can exacerbate intergroup inequality 1. *American Journal of Sociology*, 116(6):1887–1933, 2011.

[DJC15] Yuxiao Dong, Reid A. Johnson, and Nitesh V. Chawla. Will this paper increase your h-index?: Scientific impact prediction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 149–158, New York, NY, 2015. ACM.

[DMS06] Waverly W. Ding, Fiona Murray, and Toby E. Stuart. Gender differences in patenting in the academic life sciences. *Science*, 313(5787):665–667, 2006.

[dN03] Wouter de Nooy. Fields and networks: correspondence analysis and social network analysis in the framework of field theory. *Poetics*, 31(5):305 – 327, 2003.

[DRB10] Jane E Dutton, Laura Morgan Roberts, and Jeffrey Bednar. Pathways for positive identity construction at work: Four types of positive identity and the building of social resources. *Academy of management review*, 35(2):265–293, 2010.

[DS04] Ed Diener and Martin EP Seligman. Beyond money: Toward an economy of well-being. *Psychological science in the public interest*, 5(1):1–31, 2004.

[DZSP+12] 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):1–11, 12 2012.

[EAL+15] Young-Ho Eom, Pablo Aragón, David Laniado, Andreas Kaltenbrunner, Sebastiano Vigna, and Dima L. Shepelyansky. Interactions of cultures and top people of wikipedia from ranking of 24 language editions. *PLOS ONE*, 10(3):1–27, 03 2015.

[EC18] Rachel Emms and Nick Crossley. Translocality, network structure, and music worlds: Underground metal in the united kingdom. *Canadian Review of Sociology*, 55(1):111–135, 2018.

[EM18] Achim Edelmann and John W. Mohr. Formal studies of culture: Issues, challenges, and current trends. *Poetics*, 68:1–9, 2018.

[Eri13] Emily Erikson. Formalist and relationalist theory in social network analysis. *Sociological Theory*, 31(3):219–242, 2013.

[EW13] Doris Ruth Eikhof and Chris Warhurst. The promised land? why social inequalities are systemic in the creative industries. *Employee relations*, 2013.

[Fin13] A. Finkbeiner. What i'm not going to do: Do media have to talk about family matters?, 2013.

[Fla17] Jessica C. Flack. Coarse-graining as a downward causation mechanism. *Philosophical Transactions of the Royal Society A*, 375(2109):20160338, 2017.

[Flo14a] Richard Florida. The creative class and economic development. *Economic Development Quarterly*, 28(3):196–205, 2014.

[Flo14b] Richard Florida. *The rise of the creative class*. Basic Books (AZ), 2014.

[FM12] Michela Ferron and Paolo Massa. Psychological processes underlying wikipedia representations of natural and manmade disasters. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, New York, NY, USA, 2012. Association for Computing Machinery.

[FS09] K. F. Ferraro and T. P. Shippee. Aging and cumulative inequality: How does inequality get under the skin? *The Gerontologist*, 49(3):333–343, 2009.

[Fuc01] Stephan Fuchs. *Against Essentialism: A Theory of Culture and Society*. Harvard University Press, 2001.

[Gar96] Eugene Garfield. What is the primordial reference for the phrase "publish or perish"? *The Scientist*, 10(12):11–12, 1996.

[GGMTY17] Ruth García-Gavilanes, Anders Mollgaard, Milena Tsvetkova, and Taha Yasseri. The memory remains: Understanding collective memory in the digital age. *Science Advances*, 3(4), 2017.

[Gib94] Michael Gibbons. *The new production of knowledge: the dynamics of science and research in contemporary societies*. SAGE Publications, Thousand Oaks, CA, 1994.

[Giu99] Katherine Giuffre. Sandpiles of Opportunity: Success in the Art World. *Social Forces*, 77(3):815–832, 1999.

[GJT09] Clifton Green, Narasimhan Jegadeesh, and Yue Tang. Gender and job performance: Evidence from wall street. *Financial Analysts Journal*, 65(6):65–78, 2009.

[GL10] Robert M Groves and Lars Lyberg. Total survey error: Past, present, and future. *Public opinion quarterly*, 74(5):849–879, 2010.

[GM11] Scott A. Golder and Michael W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[Goo20] Google. *Google Scholar Metrics*, 2020. `https://scholar.google.com/intl/en/scholar/metrics.html`.

[GP02] Carol Graham and Stefano Pettinato. *Happiness and Hardship: Opportunity and Insecurity in New Market Economies*. Brookings Institution Press, 2002.

[Gre10] Mark Greif. Positions. In Mark Greif, Kathleen Ross, and Dayna Tortorici, editors, *What Was the Hipster? A Sociological Investigation*, pages 4–13. n+1 Foundation, New York, NY, 2010.

[GT21] Annie Goh and Marie Thompson. Sonic cyberfeminisms: Introduction. *Feminist Review*, 127(1):1–12, 2021.

[GTV13] Christos Giatsidis, Dimitrios M. Thilikos, and Michalis Vazirgiannis. D-cores: measuring collaboration of directed graphs based on degeneracy. *Knowl. Inf. Syst.*, 35(2):311–343, 2013.

[GUSA05a] Roger Guimerà, Brian Uzzi, Jarrett Spiro, and Luís A. Nunes Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308:697–702, 2005.

[GUSA05b] 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.

[GWG14] David García, Ingmar Weber, and Venkata Rama Kiran Garimella. Gender asymmetries in reality and fiction: The bechdel test of social media. *CoRR*, abs/1404.0163, 2014.

[Hal02] Douglas T Hall. *Careers in and out of organizations*. Sage, 2002.

[Har08] Wendy Harcourt. Book review. *Signs*, 34(1):204–208, 2008.

[HBG04] Lars Hufnagel, Dirk Brockmann, and Theo Geisel. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences*, 101(42):15124–15129, 2004.

[HG10] Brent Hecht and Darren Gergle. The tower of babel meets web 2.0: User-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 291–300, New York, NY, USA, 2010. Association for Computing Machinery.

[HGSB20] Junming Huang, Alexander J. Gates, Roberta Sinatra, and Albert-László Barabási. Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, 117(9):4609–4616, March 2020.

[Hir05a] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005.

[Hir05b] Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569, 2005.

[HJCL20] Aniko Hannak, Kenneth Joseph, Andrei Cimpian, and Daniel B. Larremore. Explaining gender differences in academics' career trajectories. *arXiv*, page 2009.10830, 2020.

[Hol01] Constance Holden. General contentment masks gender gap in first AAAS salary and job survey. *Science*, 294(5541):396–411, 2001.

[How13] John Howkins. *The Creative Economy: How People Make Money from Ideas*. ALLEN LANE, 2013.

[Hox17] Joop J. Hox. Computational Social Science Methodology, Anyone? *Methodology*, 13(Supplement):3–12, 2017.

[HS13] Benjamin Mako Hill and Aaron Shaw. The wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PLOS ONE*, 8(6):1–5, 06 2013.

[HSFH18] Luke Holman, Devi Stuart-Fox, and Cindy E. Hauser. The gender gap in science: How long until women are equally represented? *PLOS Biology*, 16(4):e2004956, 2018.

[Hub02] John C. Huber. A new model that generates Lotka's law. *Journal of the American Society for Information Science and Technology*, 53(3):209–219, 2002.

[HWA+21] Jake M. Hofman, Duncan J. Watts, Susan Athey, Filiz Garip, Thomas L. Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J. Salganik, Simine Vazire, Alessandro Vespignani, and Tal Yarkoni. Integrating explanation and prediction in computational social science. *Nature*, 595(7866):181–188, 2021.

[Jad17] Mohsen Jadidi. Collaborations of computer scientists between 1970 and 2016, 2017.

[JKLW18a] M. Jadidi, F. Karimi, H. Lietz, and C. Wagner. Gender disparities in science? Dropout, productivity, collaborations and success of male and female computer scientists. *Advances in Complex Systems*, 21(03–04):1750011, 2018.

[JKLW18b] MOHSEN JADIDI, FARIBA KARIMI, HAIKO LIETZ, and CLAUDIA WAGNER. Gender disparities in science? dropout, productivity, collaborations and success of male and female computer scientists. *Advances in Complex Systems*, 21(03n04):1750011, 2018.

[JKTWng] Haiko Lietz Mohsen Jadidi, Daniel Kostic, Milena Tsvetkova, and Claudia Wagner. The Matthew Effect in computer science: A career study of cohorts from 1970 to 2000. forthcoming.

[JLMW21] Mohsen Jadidi, HAIKO LIETZ, Samory Mattia, and Claudia Wagner. The hipster paradox in electronic dance music: How musicians trade mainstream success off against alternative status. *Proceedings of the International AAAI Conference on Web and Social Media*, 9, 2021.

[JMBI20] Milán Janosov, Federico Musciotto, Federico Battiston, and Gerardo Iñiguez. Elites, communities and the limited benefits of mentorship in electronic music. *Scientific reports*, 10(1):1–8, 2020.

[JNB03] H Jeong, Z Néda, and A. L Barabási. Measuring preferential attachment in evolving networks. *Europhysics Letters*, 61(4):567–572, 2003.

[Kar93] Thomas A. Karel. A position to command respect: Women and the eleventh britannica (book review). *College & Research Libraries*, 54(3):282–284, 1993.

[Ken08] Mark Thomas Kennedy. Getting counted: Markets, media, and reality. *American Sociological Review*, 73(2):270–295, 2008.

[KGC14] Anna Kaatz, Belinda Gutierrez, and Molly Carnes. Threats to objectivity in peer review: the case of gender. *Trends in Pharmacological Sciences*, 35(8):371–373, 2014.

[KGW+17] F. Karimi, M. Génois, C. Wagner, P. Singer, and M. Strohmaier. Visibility of minorities in social networks. *ArXiv e-prints*, 1702.00150, February 2017.

[KT96] Svein Kyvik and Mari Teigen. Child care, research collaboration, and gender differences in scientific productivity. *Science, Technology & Human Values*, 21(1):54–71, 1996.

[KW09] Gueorgi Kossinets and Duncan J. Watts. Origins of homophily in an evolving social network. *American Journal of Sociology*, 115:405–450, 2009.

[KWL+16a] Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 53–54, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[KWL+16b] Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 53–54. International World Wide Web Conferences Steering Committee, 2016.

[LB13] Bastian Lange and Hans-Joachim Bürkner. Value creation in scene-based music production: The case of Electronic Club Music in Germany. *Economic Geography*, 89(2):149–169, 2013.

[LCMF15] Sarah-Jane Leslie, Andrei Cimpian, Meredith Meyer, and Edward Freeland. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219):262–265, 2015.

[Leh54] Harvey C. Lehman. Men's creative production rate at different ages and in different countries. *The Scientific Monthly*, 78(5):321–326, 1954.

[Ley09] Michael Ley. DBLP - some lessons learned. *Proc. VLDB Endowment*, 2(2):1493–1500, 2009.

[LF95] J. Scott Long and Mary Frank Fox. Scientific Careers: Universalism and Particularism. *Annual Review of Sociology*, 21(1):45–71, August 1995.

[LGA09] Vincent Lariviere, Yves Gingras, and Eric Archambault. The decline in the concentration of citations, 1900–2007. *Journal of the Association for Information Science and Technology*, 60:858–862, 2009.

[LH08] Timothy J. Ley and Barton H. Hamilton. The gender gap in NIH grant applications. *Science*, 322(5907):1472–1474, 2008.

[LKA16] J. Lever, M. Krzywinski, and N. Altman. Classification evaluation. *Nature Methods*, 13:603–604, 2016.

[LKW+19] Eun Lee, Fariba Karimi, Claudia Wagner, Hang-Hyun Jo, Markus Strohmaier, and Mirta Galesic. Homophily and minority-group size explain perception biases in social networks. *Nature Human Behaviour*, 3(10):1078–1087, October 2019.

[LNG+13] Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R. Sugimoto. Bibliometrics: Global gender disparities in science. *Nature*, 504(7479):211–213, December 2013.

[LNP13] V. Latora, V. Nicosia, and P. Panzarasa. Social cohesion, structural holes, and a tale of two measures. *Journal of Statistical Physics*, 151(3):745–764, 2013.

[Lon92] J. Scott Long. Measures of sex differences in scientific productivity. *Social Forces*, 71(1):159–178, 1992.

[Lot26] Alfred J. Lotka. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12):317–323, 1926.

[LPA+09] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.

[LPKL12] Anne E. Lincoln, Stephanie Pincus, Janet Bandows Koster, and Phoebe S. Leboy. The matilda effect in science: Awards and prizes in the US, 1990s and 2000s. *Social Studies of Science*, 42(2):307–320, 2012.

[LPT08] Arnaud Lefranc, Nicolas Pistolesi, and Alain Trannoy. Inequality of opportunities vs. inequality of outcomes: Are western societies all alike? *Review of income and wealth*, 54(4):513–546, 2008.

[LS16] Mark Lutter and Martin Schröder. Who becomes a tenured professor, and why? Panel data evidence from German sociology, 1980–2013. *Research Policy*, 45(5):999–1013, June 2016.

[LUD+11] Shyong (Tony) K. Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R. Musicant, Loren Terveen, and John Riedl. Wp:clubhouse? an exploration of wikipedia's gender imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, page 1–10, New York, NY, USA, 2011. Association for Computing Machinery.

[Lut05] Erzo F. P. Luttmer. Neighbors as Negatives: Relative Earnings and Well-Being*. *The Quarterly Journal of Economics*, 120(3):963–1002, 2005.

[Lut15] Mark Lutter. Do women suffer from network closure? the moderating effect of social capital on gender inequality in a project-based labor market, 1929 to 2010. *American Sociological Review*, 80(2):329–358, 2015.

[LZ86] Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, April 1986.

[Maz12] Amin Mazloumian. Predicting scholars' scientific impact. *PLoS ONE*, 7(11):e49246, 2012.

[MBD12] Rüdiger Mutz, Lutz Bornmann, and Hans-Dieter Daniel. Does gender matter in grant peer review? *Zeitschrift für Psychologie*, 220(2):121–129, 2012. PMID: 23480982.

[McL01] Kembrew McLeod. Genres, subgenres, sub-subgenres and more: Musical and social differentiation within electronic/dance music communities. *Journal of Popular Music Studies*, 13(1):59–75, 2001.

[McR02] Angela McRobbie. Clubs to companies: Notes on the decline of political culture in speeded up creative worlds. *Cultural studies*, 16(4):516–531, 2002.

[MEH+11] Amin Mazloumian, Young-Ho Eom, Dirk Helbing, Sergi Lozano, and Santo Fortunato. How citation boosts promote scientific paradigm shifts and Nobel prizes. *PLoS ONE*, 6(5):e18975, 2011.

[Mer68] R. K. Merton. The Matthew Effect in science. *Science*, 159(3810):56–63, 1968.

[Mer73] Robert K. Merton. The normative structure of science. In *The Sociology of Science: Theoretical and Empirical Investigations*, pages 267–278. University of Chicago Press, Chicago, IL, 1973.

[Mer88] Robert K. Merton. The Matthew Effect in science, II: Cumulative advantage and the symbolism of intellectual property. *Isis*, 79(4):606, 1988.

[MJA+11] Alan Mislove, Sune Lehmann Jørgensen, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. Understanding the demographics of twitter users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 554–557. AAAI Press, 2011. 5th International AAAI Conference on Weblogs and Social Media (ICWSM 2011) ; Conference date: 17-07-2011 Through 21-07-2011.

[MLB18] Marina Micheli, Christoph Lutz, and Moritz Büchi. Digital footprints: an emerging dimension of digital inequality. *J. Inf. Commun. Ethics Soc.*, 16:242–251, 2018.

[MLSS16] Benoit Macaluso, Vincent Larivière, Thomas Sugimoto, and Cassidy Sugimoto. Is science built on the shoulders of women? A study of gender differences in contributorship. *Academic Medicine*, 91(8):1136–1142, 2016.

[Moh13] John W. Mohr. *Bourdieu's Relational Method in Theory and in Practice: From Fields and Capitals to Networks and Institutions (and Back Again)*, pages 101–135. Palgrave Macmillan, 2013.

[Mol19] Christoph Molnar. *Interpretable machine learning: A guide for making Black Box Models interpretable*. Lulu, Morisville, NC, 2019.

[Moo01] James Moody. Race, school integration, and friendship segregation in America. *American Journal of Sociology*, 107(3):679–716, 2001.

[Moo04] James Moody. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological review*, 69(2):213–238, 2004.

[Mor19] Matthew D. Morrison. Race, Blacksound, and the (Re)Making of Musicological Discourse. *Journal of the American Musicological Society*, 72(3):781–823, 12 2019.

[MPCG11] Juan D Montoro-Pons and Manuel Cuadrado-García. Live and prerecorded popular music consumption. *Journal of Cultural Economics*, 35(1):19–48, 2011.

[MRDB+12] 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. *Proceedings of the National Academy of Sciences of the United States of America*, 109(41):16474–16479, October 2012.

[Mur91] Charles Murray. Hers; the smurfette principle, 04 1991.

[Mur03] Charles Murray. Human accomplishment: The pursuit of excellence in the arts and sciences. 01 2003.

[MW47] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[MW03a] James Moody and Douglas R. White. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, 68(1):103–127, 2003.

[MW03b] James Moody and Douglas R. White. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, 68(1):103–127, 2003.

[MW14] Gabriel Magno and Ingmar Weber. *International Gender Differences and Gaps in Online Social Networks*, pages 121–138. Springer International Publishing, Cham, 2014.

[MWH17] Peter Millward, Paul Widdop, and Michael Halpin. A 'different class'? homophily and heterophily in the social class networks of britpop. *Cultural Sociology*, 11(3):318–336, 2017.

[NESF05] Thomas WH Ng, Lillian T Eby, Kelly L Sorensen, and Daniel C Feldman. Predictors of objective and subjective career success: A meta-analysis. *Personnel psychology*, 58(2):367–408, 2005.

[New03] Mark EJ Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

[New09] Mark E. J. Newman. The first-mover advantage in scientific publication. *Europhysics Letters*, 86(6):68001, 2009.

[Nor97] Nigel Norris. Error, bias and validity in qualitative research. *Educational action research*, 5(1):172–176, 1997.

[NS13] Shinichi Nakagawa and Holger Schielzeth. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142, 2013.

[NSL+12] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A tale of many cities: Universal patterns in human urban mobility. *PLOS ONE*, 7(5):1–10, 05 2012.

[OAS10] Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, 2010.

[O'R96] A. M. O'Rand. The precious and the precocious: Understanding cumulative disadvantage and cumulative advantage over the life course. *The Gerontologist*, 36(2):230–238, 1996.

[Ort18] Andreas Aurelio Rauh Ortega. *?Under-the-Radar? Electronic Dance Musicians: Opportunities and Challenges with Digital Communication Technologies*. University of Leeds, February 2018.

[Pag15] Scott E. Page. What sociologists should know about complexity. *Annual Review of Sociology*, 41(1):21–41, 2015.

[Par40] Talcott Parsons. An analytical approach to the theory of social stratification. *American Journal of Sociology*, 45(6):841–862, 1940.

[PBV07] Gergely Palla, Albert-László Barabási, and Tamás Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[PD09] Diogo L. Pinheiro and Timothy J. Dowd. All that jazz: The success of jazz musicians in three metropolitan areas. *Poetics*, 37(5):490–506, 2009.

[PDL18] Leto Peel, Jean-Charles Delvenne, and Renaud Lambiotte. Multiscale mixing patterns in networks. *Proceedings of the National Academy of Sciences*, 115(16):4057–4062, 2018.

[Pel96] Alice N Pell. Fixing the leaky pipeline: Women scientists in academia. *Journal of animal science*, 74(11):2843–2848, 1996.

[Per14] M. Perc. The Matthew effect in empirical data. *Journal of The Royal Society Interface*, 11(98):20140378, 2014.

[Pet15] Alexander M. Petersen. Quantifying the impact of weak, strong, and super ties in scientific careers. *Proceedings of the National Academy of Sciences of the United States of America*, 112(34):E4671–E4680, September 2015.

[PFP+14] Alexander Michael 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. *Proceedings of the National Academy of Sciences of the United States of America*, 111(43):15316–15321, 2014.

[Phe94] Thomas J. Phelan. Striking the mother lode in science: The importance of age, place, and time. *The Journal of Higher Education*, 65(5):627–629, 1994.

[Phi04] Anne Phillips. Defending equality of outcome. *Journal of political philosophy*, 12(1):1–19, 2004.

[PJYS11] Alexander M. Petersen, Woo-Sung Jung, Jae-Suk Yang, and H. Eugene Stanley. Quantitative and empirical demonstration of the matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences*, 108(1):18–23, 2011.

[Poh02] Matti Pohjola. The new economy in growth and development. *Oxford Review of Economic Policy*, 18(3):380–396, 2002.

[PP12] John F. Padgett and Walter W. Powell. *The Emergence of Organizations and Markets*. Princeton University Press, 2012.

[PP14] Alexander M. Petersen and Orion Penner. Inequality and cumulative advantage in science careers: A case study of high-impact journals. *EPJ Data Science*, 3(1):24, Oct 2014.

[PPP+13] Orion Penner, Raj K. Pan, Alexander M. Petersen, Kimmo Kaski, and Santo Fortunato. On the predictability of future impact in science. *Scientific Reports*, 3(1):3052, December 2013.

[PPPF18] Raj K. Pan, Alexander M. Petersen, Fabio Pammolli, and Santo Fortunato. The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics*, 12(3):656–678, 2018.

[Pri63] Derek J. de Solla Price. *Little Science, Big Science... and Beyond*. Columbia University Press, New York, NY, 1986 [1963].

[Pro08] Heidi Prozesky. A career-history analysis of gender differences in publication productivity among South African academics. *Science & Technology Studies*, 21(2):47–67, 2008.

[PRSP12] Alexander M Petersen, Massimo Riccaboni, H Eugene Stanley, and Fabio Pammolli. Persistence and uncertainty in the academic career. *Proceedings of the National Academy of Sciences of the United States of America*, 109(14):5213–5218, 2012.

[RC11] Karen Ross and Cynthia Carter. Women and news: A long and winding road. *Media, Culture & Society*, 33(8):1148–1165, 2011.

[RCC14] Javier Ruiz-Castillo and Rodrigo Costas. The skewness of scientific productivity. *Journal of Informetrics*, 8(4):917–934, 2014.

[REB10] Gabriel Rossman, Nicole Esparza, and Phillip Bonacich. I'd like to thank the academy, team spillovers, and network centrality. *American Sociological Review*, 75(1):31–51, 2010.

[Rei11] Rosa Reitsamer. The diy careers of techno and drum 'n'bass djs in vienna. *Dancecult: Journal of Electronic Dance Music Culture*, 3(1):28–43, 2011.

[Rey13] Simon Reynolds. *Energy Flash: A Journey Through Rave Music and Dance Culture*. Faber Faber, 2013.

[RH79] Barbara F. Reskin and Lowell L. Hargens. Scientific advancement of male and female chemists. In Rodolfo Alvarez and Kenneth G. Lutterman, editors, *Discrimination in Organizations*, pages 100–122. Jossey-Bass, San Francisco, CA, 1979.

[RH10] F. Reitz and O. Hoffmann. Learning from the past: An analysis of person name corrections in DBLP collection and social network properties of affected entities. In *2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 9–16, Aug 2010.

[Ros09] Andrew Ross. *Nice Work If You Can Get It: Life and Labor in Precarious Times*. NYU Press, 2009.

[RPP18] Guillermo Armando Ronda-Pupo and Thong Pham. The evolutions of the rich get richer and the fit get richer phenomena in scholarly networks: The case of the strategic management journal. *Scientometrics*, 116(1):363–383, 2018.

[RR11] Joseph Reagle and Lauren Rhue. Gender bias in wikipedia and britannica. *International Journal of Communication*, 5:1138–1158, 01 2011.

[RSZ14] Ernesto Reuben, Paola Sapienza, and Luigi Zingales. How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111(12):4403–4408, 2014.

[RT15] John E Roemer and Alain Trannoy. Equality of opportunity. In *Handbook of income distribution*, volume 2, pages 217–300. Elsevier, 2015.

[RWL+06] Patrick Reuther, Bernd Walter, Michael Ley, Alexander Weber, and Stefan Klink. Managing the quality of person names in DBLP. In *Research and Advanced Technology for Digital Libraries, 10th European Conference, ECDL*, pages 508–511, 2006.

[RYSG10] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, SMUC '10, page 37–44, New York, NY, USA, 2010. Association for Computing Machinery.

[SB07] Mike Savage and Roger Burrows. The coming crisis of empirical sociology. *Sociology*, 41(5):885–899, 2007.

[SCP17] Stacy L Smith, Marc Choueiti, and Katherine Pieper. Inequality in 900 popular films: Examining portrayals of gender, race/ethnicity, lgbt, and disability from 2007–2016. *Media, Diversity, and Social Change Initiative*, 2017.

[SEK+13] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, 8(9):1–16, 09 2013.

[Sen01] Amartya Sen. *Development as freedom*. Oxford Paperbacks, 2001.

[SFI12] Cicely Scheiner-Fisher and William B. Russell III. Using historical films to promote gender equity in the history curriculum. *The Social Studies*, 103(6):221–225, 2012.

[SKS05] Elizabeth S. Spelke, Katherine Kinzler, and Anna Shusterman. Sex differences in intrinsic aptitude for mathematics and science? A critical review. *American Psychologist*, 60:950–958, 2005.

[SL00] Zbigniew Smoreda and Christian Licoppe. Gender-specific use of the domestic telephone. *Social Psychology Quarterly*, 63(3):238–252, 2000.

[SLZ+18] Anna Samoilenko, Florian Lemmerich, Maria Zens, Mohsen Jadidi, Mathieu Génois, and Markus Strohmaier. (don't) mention the war: A comparison of wikipedia and britannica articles on national histories. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 843–852, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[Sma99] Christopher Small. Musicking — the meanings of performing and listening. a lecture. *Music Education Research*, 1(1):9–22, 1999.

[SPS+14a] Emre Sarigöl, René Pfitzner, Ingo Scholtes, Antonios Garas, and Frank Schweitzer. Predicting scientific success based on coauthorship networks. *EPJ Data Science*, 2014.

[SPS+14b] Emre Sarigöl, René Pfitzner, Ingo Scholtes, Antonios Garas, and Frank Schweitzer. Predicting scientific success based on coauthorship networks. *EPJ Data Science*, 3(1):9, 2014.

[SRNM+15] Sandra Servia-Rodríguez, Anastasios Noulas, Cecilia Mascolo, Ana Fernández-Vilas, and Rebeca P Díaz-Redondo. The evolution of your success lies at the centre of your co-authorship network. *PloS ONE*, 10(3):e0114302, 2015.

[SSA+14] Maximilian Schich, Chaoming Song, Yong-Yeol Ahn, Alexander Mirsky, Mauro Martino, Albert-László Barabási, and Dirk Helbing. A network framework of cultural history. *Science*, 345(6196):558–562, 2014.

[SSB17] Christina Starmans, Mark Sheskin, and Paul Bloom. Why people prefer unequal societies. *Nature Human Behaviour*, 1(4):0082, 2017.

[ST11] Cosma Rohilla Shalizi and Andrew C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239, 2011.

[ST13] Michael Szell and Stefan Thurner. How women organize social networks different from men. *Scientific reports*, 3:1214, 02 2013.

[Sta04] Steven Stack. Gender, children and research productivity. *Research in Higher Education*, 45(8):891–920, 2004.

[Ste05] Frances Stewart. Horizontal inequalities: A neglected dimension of development. In *Wider perspectives on global development*, pages 101–135. Springer, 2005.

[STFMC11] Lindsay Shaw Taylor, Andrew T Fiore, GA Mendelsohn, and Coye Cheshire. "out of my league": A real-world test of the matching hypothesis. *Personality and Social Psychology Bulletin*, 37(7):942–954, 2011.

[SWD+16a] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. Quantifying the evolution of individual scientific impact. *Science*, 354(6312), 2016.

[SWD+16b] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. Quantifying the evolution of individual scientific impact. *Science*, 354(6312):aaf5239, 2016.

[Sø86] Aage Bøttger Sørensen. Social Structure and Mechanisms of Life Course Processes. In Aage Bøttger Sørensen, Franz E. Weinert, and Lonnie R. Sherrod, editors, *Human Development and the Life Course: Multidisciplinary Perspectives*, pages 177–197. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[The13] The global gender gap report 2013, 2013.

[The17] The DBLP team. *DBLP computer science bibliography. Monthly snapshot release of June 2017*, 2017. `https://dblp.org/xml/release/dblp-2017-06-01.xml.gz)`.

[Tur09] Thomas Turino. Four fields of music making and sustainable living. *The World of Music*, 51(1):95–117, 2009.

[TWU10] Mike Thelwall, David Wilkinson, and Sukhvinder Uppal. Data mining emotion in social network communication: Gender differences in myspace. *Journal of the American Society for Information Science and Technology*, 61(1):190–199, 2010.

[TZY+08] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 990–998. ACM, 2008.

[UBMK12] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences of the United States of America*, 109(16):5962–5966, 2012.

[US05a] Brian Uzzi and Jarrett Spiro. Collaboration and creativity: The small world problem. *American Journal of Sociology*, 111(2):447–504, 2005.

[US05b] Brian Uzzi and Jerrett Spiro. Collaboration and creativity: The Small World Problem. *American Journal of Sociology*, 111(2):447–504, 2005.

[vdLE15] Romy van der Lee and Naomi Ellemers. Gender contributes to personal research funding success in the netherlands. *Proceedings of the National Academy of Sciences*, 112(40):12349–12353, 2015.

[Ved17] Balazs Vedres. Forbidden triads and creative success in jazz: The Miles Davis factor. *Applied Network Science*, 2(1):1–25, 2017.

[VV09] Eleftheria Vasileiadou and Rens Vliegenthart. Research productivity in the era of the internet revisited. *Research Policy*, 38(8):1260–1268, 2009.

[Wat17] Duncan J. Watts. Should social science be more solution-oriented? *Nature Human Behaviour*, 1:0015, 2017.

[Wat20] Kevin Watson. Ims business report 2020, 2020.

[WC06] Martha S. West and John W. Curtis. AAUP faculty gender equity indicators. Technical Report, 2006.

[WGJS15] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1), Apr. 2015.

[Whi70] Harrison C White. *Chains of Opportunity: System Models of Mobility in Organizations*. Harvard University Press, Cambridge, MA, 1970.

[Whi08] Harrison C. White. *Identity and Control: How Social Formations Emerge*. Princeton University Press, 2008.

[Wic97] Potter Wickware. Along the leaky pipeline. *Nature*, 390(6656):202–20, 1997.

[Wit01a] Andreas Wittel. Toward a network sociality. *Theory, Culture & Society*, 18(6):51–76, 2001.

[Wit01b] Andreas Wittel. Toward a network sociality. *Theory, Culture & Society*, 18(6):51–76, 2001.

[WJK+13] 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, 07 2013.

[WJU07a] Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

[WJU07b] Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

[WLC16] Samuel F. Way, Daniel B. Larremore, and Aaron Clauset. Gender, productivity, and prestige in computer science faculty hiring networks. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1169–1179, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[WLL19] Oliver E. Williams, Lucas Lacasa, and Vito Latora. Quantifying and predicting success in show business. *Nature Communications*, 10:2256, June 2019.

[WMCL16] Samuel F. Way, Allison C. Morgan, Aaron Clauset, and Daniel B. Larremore. The misleading narrative of the canonical faculty productivity trajectory. *ArXiv e-prints*, 1612.08228, 2016.

[WMCL17] Samuel F. Way, Allison C. Morgan, Aaron Clauset, and Daniel B. Larremore. The misleading narrative of the canonical faculty productivity trajectory. *Proceedings of the National Academy of Sciences*, 114(44):E9216–E9223, October 2017.

[WMLC19] Samuel F. Way, Allison C. Morgan, Daniel B. Larremore, and Aaron Clauset. Productivity, prominence, and the effects of academic environment. *Proceedings of the National Academy of Sciences*, 116(22):10729–10733, May 2019.

[WOSMP04] Douglas R. White, Jason Owen-Smith, James Moody, and Walter W. Powell. Networks, fields and organizations: Micro-dynamics, scale and cohesive embeddings. *Computational & Mathematical Organization Theory*, 10(1):95–117, 2004.

[Wra11] K. Brad Wray. *Kuhn's Evolutionary Social Epistemology*. Cambridge University Press, Cambridge, 2011.

[WS98] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998.

[WSB13] D. Wang, C. Song, and Albert-László Barabási. Quantifying Long-Term Scientific Impact. *Science*, 342(6154):127–132, 2013.

[WvV19] Rens Wilderom and Alex van Venrooij. Intersecting fields: The influence of proximate field dynamics on the development of electronic/dance music in the US and UK. *Poetics*, 77:101389, 2019.

[WW97] Christine Wenneras and Agnes Wold. Nepotism and sexism in peer-review. *Nature*, 387:341–343, 1997.

[WZZT19] Huaiyu Wan, Yutao Zhang, Jing Zhang, and Jie Tang. Aminer: Search and mining of academic social networks. *Data Intelligence*, 1(1):58–76, 2019.

[Xie14] Y. Xie. "Undemocracy": Inequalities in science. *Science*, 344(6186):809–810, 2014.

[XS98] Yu Xie and Kimberlee A. Shauman. Sex differences in research productivity: New evidence about an old puzzle. *American Sociological Review*, 63(6):847–870, 1998.

[YRH+16] Amy Yu, Shahar Ronen, Kevin Hu, Tiffany Lu, and Cesar Hidalgo. Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific data*, 3:150075, 01 2016.

[ZCY15] Erjin Zhou, Zhimin Cao, and Qi Yin. Naive-deep face recognition: Touching the limit of LFW benchmark or not? *CoRR*, abs/1501.04690, 2015.

[ZDSP+16] Xiao Han T Zeng, Jordi Duch, Marta Sales-Pardo, João AG 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 Biology*, 14(11):e1002573, 2016.

[ZH03] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2003.

[Zhe00] Beiyao Zheng. Summarizing the goodness of fit of generalized linear models for longitudinal data. *Statistics in Medicine*, 19(10):1265–1275, 2000.

[ZM72] Harriet Zuckerman and Robert K. Merton. Age, aging, and age structure in science. In Matilda White Riley, Marilyn Johnson, and Anne Foner, editors, *A Sociology of Age Stratification*, volume 3 of *Aging and Society*. Russel Sage Foundation, New York, NY, 1972.

Appendix A

Further Publications

During the course of this dissertation, I co-authored the following publication that are not part of this manuscript.

- Article 5 [SLZ+18]: Anna Samoilenko, Florian Lemmerich, Maria Zens, Mohsen Jadidi, Mathieu Génois, and Markus Strohmaier. 2018. (Don't) Mention the War: A Comparison of Wikipedia and Britannica Articles on National Histories. In *Proceedings of the 27th International Conference Companion on World Wide Web*. 2018