Socioeconomic inequality in academia

Machine learning-based data perspectives and empirical findings on junior researchers in Germany

Dissertation zur Erlangung des akademischen Grades Doktor der Wirtschafts- und Sozialwissenschaften (Dr. rer. pol.)

Vorgelegt im Fachbereich Wirtschaftswissenschaften

der Universität Kassel

von

Andreas Rehs

Datum der Disputation: 06.07.2021

Erstgutachter:	Prof. Dr. Guido Bünstorf
Zweitgutachter:	Prof. Dr. Ivo Bischoff
Drittgutachter:	Dr. Cornelia Lawson

Danksagung

Trotz des Downloads von 80.000 Online-Dissertationen und der Inspizierung dessen, was man daraus wohl scrapen und zu Paperideen weiterverarbeiten könnte, bin ich leider kein Experte für Danksagungen geworden. Wem man wann an welcher Stelle dankt und welche Phrase hierzu eher zu bevorzugen bzw. zu vermeiden wäre, „vielen Dank für die Überlassung des Themas“, ist in etwa so kompliziert wie Arbeitszeugnisdeutsch.

Ich probiere es trotzdem mal: Zuallererst möchte ich meiner lieben Frau Verena danken. Du hast mir große Geduld entgegengebracht und mich an vielen Abenden „meinem Thema überlassen“. Ohne dich hätte ich das nicht geschafft! Gleiches gilt meinen Eltern und meiner Familie, auf die ich seit Beginn meines Studiums zählen konnte.

Ebenso möchte ich meinem Betreuer und frühen Förderer Guido Bünstorf danken. Du hast mir kein Thema überlassen, sondern mir die akademische Freiheit geschenkt meine ganz eigene Agenda zu entwickeln. Ohne dein Vertrauen und vielfältige Unterstützung seit Beginn meiner Hilfskrafttätigkeit im Jahr 2013 hätte ich diese Arbeit weder begonnen noch erfolgreich beendet.

An dem Erfolg dieser Arbeit hast jedoch nicht nur du Anteil, sondern maßgeblich auch das gesamte Lehrstuhlteam. Auch euch möchte ich danken. Durch die akademische Sozialisation in der Mittagspause mit euch und den „Bischoffs“, habe ich eigentlich erst verstanden, wie VWL funktioniert und auch immer wieder Spaß dabei gehabt zwischen Erdbeerjoghurt und Pommes empirische Forschungsdesigns zu diskutieren. An dieser Stelle möchte ich auch Ivo Bischoff, der jede meiner Arbeiten von Bachelorarbeit bis Dissertation zweitbegutachtet hat, sowie Cornelia Lawson als Drittgutachterin danken. Zu guter Letzt bedanke ich mich herzlich bei Karin und dem gesamten Sekretariatsteam, wiss. Hilfskräften, VWL3-Tutoren, der R-community (ganz besonders auf stackoverflow.com), #econtwitter und wohlwollenden anonymen Referees.

August 2021

Table of contents

List of tables VIII
List of figures IX
1 Introduction 1
1.1 Overview 1
1.2 The economics and sociology of inequality in academia 2
1.2.1 The economics of science 3
1.2.2 The sociology of inequality in academia 4
1.3 The measurement of socioeconomic inequality in academia 6
1.3.1 The scientometrics of inequality 6
1.3.2 Labor economics and the measurement of inequality 7
1.4 The case of junior scientists in Germany 9
1.5 Contribution of this dissertation 10
1.5.1 The detection of thematic differences between author populations 11
1.5.2 Web-scraping and machine learning-based development of author databases 12
1.5.3 Inequalities in academic and nonacademic outcomes of young German scientist: The case of place of birth and advisor gender 14
1.6 Databases 18
1.6.1 The electronic catalog of the DNB and its linked datasets 18
1.6.2 Online Dissertations 18
1.6.3 The Web of Science publication database 20
1.6.4 Database relations 21
2 A structural topic model approach to scientific reorientation of economics and chemistry after German reunification 23
2.1 Preface 23
2.2 Introduction 23
2.3 Historical background 25
2.4 Data 27
2.5 Topic modelling in large-scale text analysis 29
2.6 Discussion and Conclusion 39
3 A supervised machine learning approach to author disambiguation in the Web of Science 42
3.1 Preface 42
3.2 Introduction 42
3.3 Characteristics of disambiguation approaches 44
3.3.1 Disambugiation methods 44
3.3.2 Features for disambiguation 45
3.4 Data 46

3.4.1 Data sources, sampling strategy and preprocessing 46
3.4.2 Feature generation 50
3.5 Machine learning approach to author disambiguation 54
3.5.1 Random forest 54
3.5.2 Logistic regression 55
3.5.3 Machine learning results 55
3.6 Graph-based author community detection 57
3.6.1 Methodology 57
3.6.2 Application to full block “Muller, M.” 60
3.7 Discussion and Conclusion 61
4 The scientific productivity of German PhD graduates: A machine learning-based author name disambiguation and record linkage approach 64
4.1 Preface 64
4.2 Introduction 64
4.3 Literature 66
4.3.1 The scientific productivity of PhD students and graduates 66
4.3.2 Policy perspective on PhD publications 66
4.3.3 Occupational perspective 67
4.3.4 Scientometric perspective 67
4.3.5 German PhD productivity 68
4.4 Methods 68
4.4.1 Author disambiguation and record linkage of German dissertation authors 68
4.4.2 Disambiguation of German authors in the WoS 70
4.5 Probabilistic record linkage of publication data to German dissertation authors 74
4.6 Results 78
4.7 Discussion and conclusion 80
5 Protégé-advisor gender-pairings in academic survival and productivity of German PhD graduates 83
5.1 Preface 83
5.2 Introduction 83
5.3 Gender pairings and outcomes of doctoral advisory 85
5.4 Data and methods 86
5.5 Results 90
5.6 Discussion and conclusion 92
6 Career paths of PhD graduates in eastern and western Germany: Same qualification, same labor market outcomes? 95
6.1 Preface 95
6.2 Introduction 95
6.3 Overeducation among PhD graduates 98

6.4 Empirical setting 100
6.4.1 Data 100
6.4.2 Main variables 102
6.4.3 Control variables 104
6.5 Results 105
6.5.1 Descriptive evidence 105
6.5.2 Econometric results 108
6.6 Robustness checks 111
6.7 Conclusions 112
7 Conclusion 114
8 Appendices 119
References 127

List of tables

Table 1. Dataset structure 35
Table 2. Cosine similarities by subgroups 37
Table 3. Chemistry OLS regression 38
Table 4. Economics and business administration OLS regression 39
Table 5. Characteristics of the test and training set 49
Table 6. Top 5 blocks by frequency in training and test sets 50
Table 7. Descriptive statistics features synthetic test and training set 53
Table 8. Pairwise prediction results synthetic test set 57
Table 9. Feature importance random forest 57
Table 10. Cluster algorithm results synthetic test set 60
Table 11. Application: Cluster results for block “Muller, M.” 61
Table 12. Data structure DNB 69
Table 13. Data structure WoS 69
Table 14. WoS and DNB homonym frequencies 70
Table 15. Summary statistics of the disambiguated dataset at block level 74
Table 16. Identifier weighting in probabilistic record linkage 75
Table 17. Prepared datasets for linkage 77
Table 18. Descriptive statistics 89
Table 19. Negative binomial regression and average marginal effects for productivity during PhD study 91
Table 20. Complementary log-log model – Yearly hazard of exit from research 92
Table 21. Selected average marginal effects for holding an expert job in t+5 109
Table 22. Average predictive margins for birthplace_east at different values of workplace_east (holding an expert job in t+5) 109
Table 23. Selected average marginal effects for exceeding the contribution assessment ceiling in t+5 110
Table 24. Average predictive margins for birthplace_east at different levels of workplace_east (exceeding the contribution assessment ceiling in t+5) 111

List of figures

Figure 1. Database relations 22
Figure 2. Two title-topic distributions of the economics and business adm. topic model .. 33
Figure 3. Topical prevalence of two economics and business administration topics 33
Figure 4. Mean topic prevalence before and after German reunification 34
Figure 5. Yearly mean cosine similarity between topic distributions in dissertation pairs . 36
Figure 6. Data processing schematic 47
Figure 7. Number of missing first name in 100K paper-author sample of WoS 49
Figure 8. Illustration of true and false predicted paper pairs 59
Figure 9. Number of disambiguated publications by year and missing first names 73
Figure 10. Coverage of dissertations in the DNB by discipline 78
Figure 11. Coverage of dissertations in the DNB by year 79
Figure 12. Average productivity 2 and 5 years after dissertation 79
Figure 13. 2- and 5-year productivity by gender 80
Figure 14. 2- and 5-year productivity by birthplace in eastern or western Germany 80
Figure 15. Database schematic 87
Figure 16. Mean number of cumulated publications before and after dissertation by gender pairing 88
Figure 17. Kaplan Meier Curve and risk table for final publication after PhD completion 91
Figure 18. Spatial distribution of the PhD graduates` birthplaces (left) and location of their PhD universities, by birthplace in eastern of western Germany (right) 106
Figure 19. Workplace according to planning regions five years after dissertation, by birthplace in eastern and western Germany 107
Figure 20. Share of PhD graduates with wages above the contribution assessment ceiling (modifications) 112

1 Introduction

1.1 Overview

The global pandemic of 2020 has shown that science can heal the world. The rapid development and deployment of vaccines has slowed the spread of the virus and saved millions of lives. Two German scientists contributed greatly to this achievement. Özlem Türeci and Uğur Şahin, children of Turkish immigrants, developed one of the leading and most potent vaccines against the Coronavirus. Being children of immigrants and, in the case of Özlem Türeci, female, their success story is an exception in German academia.

In Germany and other western knowledge societies, a large number of young scientists find their talents sacrificed to a lack of opportunities. This divide follows traditional socioeconomic boundaries like gender and parental education and income, and negatively affects long-term national economic growth, as innate talents cannot pursue their comparative advantage (Acemoglu, 1995; Hsieh, Hurst, Jones & Klenow, 2019).

A natural starting point to understand socioeconomic inequality in academia is to explore the early career stages of a scientist. Women's careers, for instance, are affected by the leaky pipeline effect (Blickenstaff, 2005). The number of women at each advanced career stage in academia decreases gradually, finally reaching its minimum in the professor stage. Only 20% of the professors at German universities are female. Similar observations have been made concerning privileged social backgrounds. German professors are more likely to come from academic households (Möller, 2015).

What are the causes and effects of socioeconomic inequality in academia? This question is an interdisciplinary research problem that involves economics, sociology and information science. As in any other applied context, the development of databases and methods is a prerequisite to this endeavor. Accordingly, this dissertation's first goal is to develop machine learning-based tools that help establish new databases on socioeconomic inequality in academia. Germany in particular lacks individual-based scientometric indicators and will therefore be the focal country of my analysis (Konsortium Bundesbericht Wissenschaftlicher Nachwuchs, 2017). I address two relevant and partly pending issues in applied questions of socioeconomic inequality in academia, and focus on junior scholars. In summary, my research question is: How can methods of machine learning and social sciences jointly help to establish new databases on and provide subsequent insights into socioeconomic inequality among junior researchers in German academia?

The remainder of this dissertation is as follows: In the next sections of the introduction, I will review the foundations of the economics and sociology of inequality in academia. Section 1.3 discusses the measurement of socioeconomic inequality in academia and introduces the research fields of scientometrics and labor economics. Subsequently, I address the system of higher education in Germany and the state of inequality for young scientists within it. Section 1.5 describes the contributions of my dissertation, how each chapter and underlying paper is embedded in the current literature and why the topics investigated are relevant to socioeconomic inequality in academia. Finally, Section 1.6 presents my datasets of online dissertations, the German National Library (DNB) catalog and Web of Science (WoS) publications, and how all databases are related to each other.

Chapters 2, 3, 4, 5 and 6 are the central part of my dissertation and come from five different papers. Two have been published in international journals, one is at major revisions and two have been presented at a leading conference on scientometrics. Chapter 2 discusses the detection of thematic differences in dissertation titles. The subject of Chapter 3 is the development of a supervised machine learning approach to author name disambiguation in the WoS publication database. Chapter 4 builds on that approach, applying the method to German author names and thereby creating a dataset of about 11 million disambiguated publications. Chapter 4 also links the disambiguated author dataset to German dissertation authors in the DNB catalog. This linkage becomes relevant in chapters 5 and 6, which refer to the DNB and other linked datasets. These chapters of my dissertation are the applied empirical contribution to inequality research. Chapter 5 thereby investigates the career outcomes of protégés by different advisor-protégé gender pairings, and Chapter 6 investigates the career paths of PhD graduates in eastern and western Germany. The last chapter of my thesis is dedicated to concluding remarks.

1.2 The economics and sociology of inequality in academia

Inequality in academia is an interdisciplinary research problem. It concerns all disciplines directly related to higher education research, such as economics, sociology and political science. Because of its multifaceted background, the concept of inequality varies with every disciplinary perspective and requires some delineation. In my dissertation, I choose the unusual terminology “socioeconomic inequality” in order to combine the closely related but separate concepts of social and economic inequality. The UN refers to economic inequality as “how economic variables are distributed — among individuals in a group, among groups in a population, or among countries.” (United Nations, 2015, p. 2). One can separate economic inequality into two dimensions: inequality of outcomes and inequality of opportunities. Inequality of outcomes concerns the distribution of variables like wealth and income, whereas inequality of opportunities concerns differences in circumstances and preconditions that affect life outcomes (Sen, 1995). Social inequality is an extension of economic inequality, and concerns the distribution of various social variables, such as health, nutrition and political freedom. However, social variables are often also economic variables, which makes social and economic inequality two very closely related concepts. In the following paragraphs, I want to briefly review the theoretical foundations of inequality in academia with respect to economics and the sociology of science.

1.2.1 The economics of science

Science is a versatile and interesting research subject for economists. It is an endogenous source of long-term economic growth (Romer, 1994; Romer, 1990) and includes particular markets, goods and incentives (Stephan, 1996). Those markets, goods and incentives build a complex of causal mechanisms and empirical patterns that influence socioeconomic inequality in academia. The academic labor market for junior scientists is one example. It is characterized by standardized career steps, like earning a PhD, and a relatively high degree of transparency in junior researchers' skills and productivity (through their published work). Since labor markets function on merit selection, the academic labor market for junior scientists is characterized by a shortage of tenured positions. This shortage leads to selective pressure after completion of a PhD and causes the majority of junior scientists to depart from the academic labor market (Cyranoski, Gilbert, Ledford, Nayar, & Yahia, 2011; Stephan, 1996). Socioeconomic inequality thereby emerges when certain groups, like women, are disproportionally affected (Fox & Stephan, 2001). Socioeconomic inequality also develops in other outcomes of the market for junior scientists. I return to this topic in the section 1.3.2, where I will elaborate on the labor economics of the academic job market.

Scientific output is special from an economic perspective. It has characteristics of a public good and is free to use by others. In the same sense, scientific work is most often indifferent in its availability and access. Scientific work, therefore, theoretically fulfills some of the basic premises for social equality. The public good properties of scientific research allow every scientist to refer to academia’s body of knowledge, leading to equal chances. This captures the nature of scientific work only partly. Scientific work includes in no small part implicit (or tacit) knowledge that cannot be codified and which is particular to individuals (Nelson & Winter, 1982; Polanyi, 2015). Dasgupta and David (1994) state that scientific work does not become a public good by its publication alone. It requires expert knowledge or extensive codification in a manner that non-specialists can process the underlying knowledge.

Scientific progress is strongly linked to public expenditure and related funding policy. However, science also has its own currency and reward systems, where monetary incentives play a minor role. First, scientists get intrapersonal rewards for solving complex problems or “puzzles” (Kuhn, 2012). Second, scientists earn reputation and recognition in the scientific community. In several disciplines, publications and the prominence of the journals in which they have been published are an accepted proxy for scientific recognition. Also, promotions, prizes, calls and university and home departmental reputation play a role. The academic reward system works by the principle of “winner takes all” (Stephan, 1996). Credit is allocated primarily to those who have priority of discovery (Merton, 1957). Scientific discovery “races” are similar to innovation in industry but do not include (temporal) regulation of the good knowledge by intellectual property rights. In contrast to innovators in industry, academic scientists have an economic incentive to disclose their knowledge to gain recognition (Stephan, 1996). Socioeconomic inequality in academia develops in the market for recognition. The chances to publish and actual publications are distributed unevenly among individual scientists (e.g., Lotka, 1926; Merton, 1968), among groups in a population, such as between men and women, (e.g., Larivière, Ni, Gingras, Cronin, & Sugimoto, 2013; Prpić, 2002) or among countries (May, 1997). I address the issue of publications and related indicators in section 1.3.1.

1.2.2 The sociology of inequality in academia

Inequality is a key topic in sociology and its theories of societies. Sociology differentiates two socioeconomic inequality perspectives: the functionalist theory and the conflict theory (Huaco, 1966). The functionalist theory is known as the Davis-Morre hypothesis and considers inequality as inevitable, desirable and functional for prosperous societies (Davis & Moore, 1945). Ability is central in the functionalist view. It requires skills and training and should be rewarded by society. The functionalist theory is oriented to a meritocratic society.

Conflict theory opposes functionalist theory and is associated with classics of economics and sociology (e.g., Marx & Engels, 1848). It presumes that the allocation of power between social groups leads to suppression of the less powerful groups. Suppressive behavior thereby maintains the status quo, and is transmitted via social and economic institutions. Conflict theory assumes inequality to be dysfunctional and harmful for the prosperity of society.

Both theories can be applied to the socioeconomic inequality in academia. Merton (1973) introduced the functionalist perspective of science by proposing *universalism* as one of the four ethical principles, or “Mertonian norms,”[1] of science. Universalism claims that “the acceptance or rejection of claims entering the lists of science is not to depend on the personal or social attributes of their protagonist; his race, nationality, religion, class, and personal qualities are as such irrelevant” (Merton, 1973, p. 270).

[1] Besides universalism, Merton also introduced communism, disinterestedness and organized scepticism. Scientific communism is related to “Open Science” and claims that scientific knowledge should be free and shared in the scientific community. The concept of scientific communism overlaps the concept of knowledge as a public good. Disinterestedness concerns scientific institutions and states that they should act for the benefit of a common scientific enterprise and not for personal interest. Organized scepticism claims that science and its methods and institutions should be organized to foster critical scrutiny.

In practice, the *universalism* principle in science is overcome by several factors. Cumulative advantages are among the most discussed (Merton, 1968; de Solla Price, 1976). Merton uses the term *Matthew effect* to describe cumulative advantages, referring to the Gospel of Matthew: “For to every one who has will more be given, and he will have abundance; but from him who has not, even what he has will be taken away.” (Matthew 13:12). Prominent scientists receive disproportionally more credit for their scientific works than lesser-known scientists. Merton argues that the Matthew effect may affect a vast number of scientific publications; scientists quickly screen the quality of publications they intend to read and use the author’s or journal’s reputation as a proxy. This procedure prefers prominent scientists and thereby establishes cumulative advantages. Empirical evidence suggests the Matthew effect is also present in funding (Bol, De Vaan, & Van De Rijt, 2018), citations before and after the award of academic prizes (Azoulay, Stuart, & Wang, 2014), paper and author-level citations (Birkmaier & Wohlrabe, 2014; Tol, 2009, 2013) and institution prestige related to author-level citations (Medoff, 2006). The Mathilda effect is a related to phenomenon. It describes that the work of women scientists is attributed to their male colleagues (Rossiter, 1993).

Bourdieu’s *Homo academicus* (1984) is another significant contribution to the sociology of science and reflects on the social structure of science. Bourdieu is a proponent of global theories of societies and is different from the previously discussed sociologist Merton, who takes a middle-range approach to sociological phenomenons. Power, habitus and hierarchies in science are the central topics in Bourdieu’s book. He assumes that science is a hierarchical social system in which the actors attain different levels of power. The actor’s social position is relational and depends on the position of others in the social space of science. Power in this social system is thereby determined by the possession of three different types of capital: economic capital, social capital and cultural capital.

Bourdieu defines economic capital as “immediately and directly convertible into money and institutionalized in the form of property rights” (Bourdieu, 1988[2], p.242). Social capital is “the aggregate of the actual or potential resources which are linked to the possession of a durable network of more or less institutionalized relationships of mutual acquaintance and recognition” (Bourdieu 1988, p. 247). Cultural capital comprises knowledge, education, behavior and skills and is defined as “familiarity with the legitimate culture within a society” (Bourdieu 1988, p. 242). There are three subclasses of cultural capital: institutionalized, objectified and embodied cultural capital. Objectified cultural capital refers to the possession of cultural goods (e.g., books), embodied cultural capital comprises language, values, taste and knowledge and institutionalized cultural capital refers to educational attainment, such as academic degrees or positions.

[2] English translation of Bourdieu, 1984.

A social class in this system consists of actors with similar levels of power and compositions of capital. The members of this social class share a certain habitus and try to reproduce their social class, which establishes and perpetuates social inequality. Bourdieu illustrates social classes in academia with the example of disciplines. He distinguishes between disciplines whose researchers predominately possess social capital (e.g., medicine, law), and those whose predominately possess cultural capital (e.g., philosophy, psychology). Within faculties there exists, again, different social classes and stereotypes of researchers. Those researchers who have relatively low social capital and high cultural capital identify themselves as pure researchers and aim to attain greater recognition in the scientific community. Their socioeconomic background tends to be unprivileged. On the other hand, researchers with high social capital and low cultural capital strive for leading positions in their field and in society. Their socioeconomic background tends to be privileged. To summarize, social inequality is a phenomenon rooted in homosocial reproduction and related possession and composition of different sorts of capital.

1.3 The measurement of socioeconomic inequality in academia

The practical measurement of inequality is, like its theoretical foundations, interdisciplinary and strongly dependent on the research question. In the following paragraphs, I will introduce the two fields of research most relevant to this dissertation and illustrate how they measure socioeconomic inequality in academia. These fields are scientometrics and labor economics.

1.3.1 The scientometrics of inequality

Scientometrics, as a research field, deals with the quantitative features and characteristics of science and scientific research. It is an interdisciplinary research field that overlaps significantly with information science, mathematics, statistics and sociology and economics of science. Modern scientometrics began with the pioneering work of Derek de Solla Price and Eugene Garfield, who were faced with the need for systematic research and databases caused by the growing amount of scientific literature in the 20th century. They established citation and publication-based analysis as one of the key concepts in scientometrics (Garfield, 1972; de Solla Price, 1963). Garfield (1972) introduced the Science Citation Index, upon which the WoS database was developed.

Scientometrics is centered primarily on the bibliometric analysis of scientific publications and citations but also includes alternative metrics, such as web-based indicators and scientific prizes. Leydesdorff and Milojevič (2015) define five major research issues in scientometrics: measurement of impact, delineation of reference sets for measuring the impact of journals or institutional units, theories of citation, mapping of science and policy and management context-related research. For socioeconomic inequality research in academia, policy-related research in conjunction with the measurement of impact is especially relevant. The measurement of scientific impact is built on publications and their citations and is closely related to scientific productivity measurement. Related indicators address the publications themselves (e.g., number of publications or average citation rate per year of a publication), scientific journals (e.g., impact factor) or the individual researcher (e.g., Hirsch's (2005) h-index).

In scientometric studies addressing inequality in academia, these indicators have been used to produce several stylized facts (see Meyer, 2011). Scientific productivity by gender is among the most important ones and addressed in this dissertation (e.g., Cole & Zuckerman, 1984; Leahey, 2006; Prpić, 2002). Female scientists write on average fewer papers than their male counterparts. The causes and reasons for this discrepancy include social selection, such as discrimination by gender; self-selection mechanisms, such as selection into active parenthood; and gender differences in career commitment. Finally, the cumulative advantages and disadvantages discussed in Section 1.2.2 also play a role in the gender productivity gap.

Lotka's law is another stylized fact related to scientometrics and inequality in academia (Lotka, 1926; Stephan, 1996). It states that the number of authors contributing papers to a particular field follows a power law: a small number of authors publish many papers, whereas a large number of authors publish individually only a few papers. Lotka's law is presumably related exclusively to differences in scientific ability and not to socioeconomic circumstances. Gupta, Kumar and Aggarwal (1999), for instance, find no statistical difference between female and male scientists in the parameters of the distribution of the number of publications, and that Lotka's law does not apply to gender. Scientometrics, its methods, and their impact on and application in science policy are subject to much debate (Frey, 2008; Weingart, 2005). The bibliometric evaluation of publications is being used more often in decisions on individual hiring, tenure and funding (Heckman & Moktan, 2020). The journal impact factor has become one of the most important criteria for choosing publication venues (Haustein & Larivière, 2015) and is used to create influential researcher rankings (Sturm & Ursprung, 2017). Bhattacharya and Packalen (2020) argue that the focus on scientometric indicators incentivizes researchers to conduct incremental science at the expense of risky research projects. Incremental research projects, however, may not lead to breakthroughs and may decelerate scientific progress.

1.3.2 Labor economics and the measurement of inequality

Labor economics provides a rich set of theoretical and methodical approaches to address inequality. In this regard, the economics of discrimination and the human capital theory are two of the major neoclassical approaches. I want to briefly review these two concepts, beginning with the economics of discrimination. Discrimination refers to a pejorative distinction or differentiation made among groups or individuals in labor markets. It is socially unacceptable and economically inefficient (Oaxaca, 2001).

There are two main theories concerning the cause of discrimination. The first focuses on the “taste” for discrimination and presumes that employers are driven by prejudice in hiring and other job-related decisions (Becker, 1957). It states that employers voluntarily sacrifice profit by discriminating against equally qualified employees by characteristics such as gender, age and race. Taste-based discrimination has been addressed empirically by several studies[3] (e.g., Agan & Starr, 2018; Levitt, 2004; Lippens, Baert, Ghekiere, Verhaeghe & Derous, 2020; Quillian, Pager, Hexel & Midtbøen, 2017).

The second neoclassical theory on discrimination is called statistical discrimination, which occurs when employers have imperfect information about potential employees. It claims that employers make their hiring decisions based on generalizations about demographic groups. If they consider certain demographic groups to be less productive, employers will prefer not to hire individuals from those groups. Statistical discrimination, therefore, acts as a heuristic for decisions under uncertainty. It becomes problematic when societal beliefs and prevailing prejudice make the process of statistical discrimination self-perpetuating (Aigner & Cain, 1977; Arrow, 1973; Phelps, 1972). Statistical discrimination has been addressed in several empirical studies of social inequality (Levitt, 2004; List, 2004; Thijssen, Coenders & Lancee, 2021). Spence (1973) proposes that statistical discrimination can be circumvented by quality or productivity signals, such as university degrees.

The human capital theory assumes that individuals invest in their skills and education in order to maximize their future returns in labor markets. Employers maximize their profits and accordingly hire those groups of workers that have the highest productivity. (Becker, 1957; Mincer, 1958). From the human capital theory perspective, the gender wage gap, for instance, is explained with the more intermittent attachment to the labor force of women (e.g., childbirth, Mincer & Polachek, 1974).

Economists measure inequality in labor markets using the four dimensions of wages, hiring, unemployment and the attainment of skill-adequate jobs. To measure inequality in wages, Blinder (1973) and Oaxaca (1973) developed the econometric method of wage decomposition, which statistically determines mean outcome differences between groups. Inequality in hiring decisions was adressed by Carlsson and Eriksson (2019) and Thijssen et al. (2021); unemployment inequality was the research subject for Gilman (1965) and Bergmann (1971). Bender and Heywood (2011) studied inequality in the attainment of skill-adequate jobs. In the last two decades, field experiments have become the

[3] For overview see Bertrand and Duflo (2016).

predominant type of empirical study adressing inequalty and discrimination in labor markets (e.g., Bertrand & Mullainathan, 2004; Leibbrandt & List, 2015).

1.4 The case of junior scientists in Germany

Socioeconomic inequality in academia can be traced back in part to childhood and early adolescence (Dasgupta & Stout, 2014), but this "early" connection to inequality is beyond the focus of my dissertation. This inequality and its underlying social and individual selection processes persist throughout school and university. For these reasons, I focus my research in this dissertation on junior scientists. I define junior scientists as persons who qualify to pursue a career in academia, which can be PhD students or PhD graduates who are active in research and do not depart academia for the non-academic labor market. It is difficult to distinguish academic from non-academic labor markets, as this requires detailed and individual job profiles of the junior scientist under investigation (Konsortium Bundesbericht Wissenschaftlicher Nachwuchs, 2017).

The German science system is divided institutionally into the university sector, with regular universities and universities of applied sciences, and the strong non-university sector, with public research institutes and industry research. German higher education is thereby embedded in the Bologna Process: students usually enter three- to four-year bachelor's degree programs that are accepted and harmonized throughout the European Bologna system; for further university education, students can pursue a master's degree, which usually qualifies them to start a doctorate.

The doctorate is the highest academic grade in the German education system. For simplicity, I use the terms "PhD" and "doctorate" interchangeably in this dissertation.[4] The time to obtain a PhD varies by discipline, ranging from an average of less than a year in medicine to 7 years in engineering. The majority of doctoral students are usually employed directly at a departmental chair, and structured PhD programs and scholarships account for only a minority of PhD students in Germany. There are also a considerable number of PhD students employed at nonuniversity research institutes (Konsortium Bundesbericht Wissenschaftlicher Nachwuchs, 2017). Most doctoral students are in the field of medicine, with about 41,000 students of the total 180,000 in 2019 (Destatis, 2019). Other notable disciplines are chemistry and biology. Between 1999 and 2019, Germany has had a stable number of PhD graduates, on average graduating 25,000 young scientists a year. Germany is one of the countries with the highest number of doctorates per capita (Hachmeister, 2019; OECD, 2019). After the completion of a doctorate, a young scientist who pursues an academic career typically applies for postdoctoral positions at departmental chairs or nonuniversity research institutes to qualify for further career steps, like professor positions. Before 2002, the qualification for professorship was

[4] When explicitly referring to German doctorates in medicine, I will explain related difficulties.

usually attained by habilitation. Since then, young scientists have also been able to apply for junior professorships. Junior professorships are, depending on the German state, combined with tenure track positions. The final career step for a young scientist is the attainment of professorship.

Socioeconomic inequality among junior scientists in German academia can be found between several demographic groups. The most striking inequality is between men and women. Although women account for 48% of graduates qualified to start a PhD, they account for only 45% of completed doctorates. In the further career stages of junior professorship and habilitation, the percentage of women drops to only 40% and 28%, respectively (Konsortium Bundesbericht Wissenschaftlicher Nachwuchs, 2017). Heinisch, Koenig and Otto (2020) find that female German PhDs work part-time more often than male German PhDs.

Socioeconomic background in terms of parental education, wealth and occupation is another determinant of inequality among junior scientists. Children from academic households enter into academic careers more often than those from non-academic households (Jaksztat, 2014). Hartmann and Kopp (2001) find a similar relationship with respect to German PhDs in top industry positions.

1.5 Contribution of this dissertation

This thesis contribution lies in the different fields. Very generally, I am concerning questions in scientometrics and in the economics of science. Thereby, economic depends on scientometrics and its methods and databases to answer questions such as those related to socioeconomic inequality in academia.

For this reason, the first three chapters of my dissertation are dedicated to scientometrics. My contribution in this respect is, therefore, of methodical and database related nature. At first, I am developing and applying machine learning-based methods to address the preceding questions problems in databases. Those preceding problems lie in detecting thematic differences, disambiguation of author names, and record linkage methodology. Similarly, I establish novel author-level databases that enhance the investigation of doctoral graduates' academic and non-academic career outcomes in Germany.

The second main contribution comes with the application of the generated databases in questions in economics of science. I focus on two currently debated topics in socioeconomic inequality in academia that partly lack empirical evidence. These topics are eastern and western German doctoral graduates' career outcomes and doctoral graduates' scientific survival and productivity by different advisor-protégé gender-pairings in German academia. In the following subsections, I will discuss the relevance and contribution of my thesis more precisely.

1.5.1 The detection of thematic differences between author populations

The exploration of thematic differences is an unconventional yet promising perspective on socioeconomic inequality in academia. It tackles the essence of scientific work, namely what is researched and how this is embedded into the body of scientific literature. Observed differences between academic outcomes, such as career opportunities and trajectories, therefore, need to be investigated in conjunction with thematic aspects. This thesis contribution lies in the development and application of machine learning-based methods that address thematic differences. Thereby, I overcome the shortcomings of traditional thematic classification approaches, such as keyword assignments, expert-based classification of subjects, and forward and backward citations to a publication (De Bellis, 2009). These traditional methods include high levels of complexity reduction and a loss of knowledge in the scientific publications' content. Practically, subtle but often decisive differences between two papers on the same topic can hardly be addressed without expert-level knowledge in the respective scientific field.

Similarly, topical overlaps between loosely related papers cannot be detected without having expert knowledge in both papers' fields. The addition of more and more papers will eventually constrain the ability of experts to detect differences and similarities between papers. Therefore, the large-scale quantification and detection of thematic differences in research topics is an open problem in scientometric research. Advances in machine learning, especially in the statistical analysis of large text collections, alleviate these issues under certain circumstances. In this way precise difference detection between scientific texts can be feasible without having deep knowledge in the respective field.

My main effort was to train and test a probabilistic text model ("structural topic model"), aggregating the outcomes and then incorporating them into a linear regression framework. This aggregation procedure allows me to calculate the level of difference between dissertation titles by regional and temporal origin of the dissertation. My approach demonstrates how to identify and track differences between scientific work on the level of individual researchers, but also larger entities of the scientific system, such as different scientific disciplines or parts of a country.

The machine learning approach was applied to the case study of dissertation titles written at eastern[5] and western[6] German universities in economics and business administration and chemistry before and after German reunification. German reunification is especially suited for investigating differences in research topics because the transition of the political

[5] Eastern Germany refers to the territory of the former German Democratic republic and today includes the German states of: Thuringia, Saxony, Saxony-Anhalt, Mecklenburg western Pomerania and Brandenburg. The city of Berlin was separated in eastern- and western-Berlin during the German division.

[6] Western Germany refers to the territory of the Federal Republic of Germany from 1949 to 1990. Western Germany includes the states: Hesse, Lower-Saxony, Bavaria, Baden-Württemberg, Saarland, Rhineland-Palatinate, North-Rhine Westphalia Bremen, Hamburg and Schleswig-Holstein.

system in eastern Germany went hand in hand with the scientific system’s transition. German reunification led to the dismantling in eastern Germany of a large number of chairs, institutes and research organizations, as well as a broad institutional restructuring in academia. Reasons included political motives and a mismatch between what had been researched under the old (socialist) system and what was considered interesting in the new one. This change affected social sciences more severely than natural sciences and therefore provided two different structures to investigate thematic differences and topical reorientation. In these two structures, motives and incentives for individual scientists in the two disciplines and parts of Germany to change research topics differed substantially and may have manifested in minor and major thematic differences.

1.5.2 Web-scraping and machine learning-based development of author databases

Author-level scientometric indicators are central to investigate questions in academic inequality among junior scholars. They are an accepted measure in evaluating scientific output (Abbott et al., 2010; Hicks, 2012) and are in their aggregated form able to uncover social inequalities by investigating social group differences in scientific output measures. As already discussed in section 1.3., author-level scientometric indicators are, for instance, used in studies that address gender inequality (e.g., Leahey, 2006; Prpić, 2002). However, for the investigation of other questions in social inequalities and economics of science, especially concerning Germany, there current databases either miss information or does not exist at large scale. The federal report on junior scientist in Germany especially points towards this substantial research gap (Konsortium Bundesbericht Wissenschaftlicher Nachwuchs, 2017). Chapter 3 and 4 of my dissertation address these issues.

The first contribution in this regard lies in enriching the DNB-dissertation base with author-level micro-data scrapped from online dissertations. In the DNB, which includes almost all dissertations handed in at German universities (see section 1.6), there is only limited information on important sociodemographic variables like the regional origin of the PhD holders and their dissertations. In the same way background information on the doctoral advisors is also most often missing in the DNB database. A wide range of questions, like social inequalities in academia that may be related to the regional origin or PhD advisor can, therefore, hardly be addressed. Other German databases that address PhDs cannot act as a substitute. They are survey based and do not include detailed regional background or advisor information.[7] To address this research gap, I am developing a new database that builds on scraping online dissertations (please see 1.6 for details).

[7] I refer to the DZHW PhD Panel.

The second contribution comes with the development and application of a machine learning-based author name disambiguation approach. Disambiguation generally concerns whether references to entities belong to the same entity or different entities (Talburt, 2011). Author name disambiguation or namesake problems are prevalent issues in scientometric research (D'Angelo & van Eck, 2020; Weingart, 2005). They prevent the clear attribution of scientific output, like papers, to their authors. Accordingly, high-quality and extensive databases that are needed to research social inequality in academia and questions in the economics of science cannot be build.

In current studies of social inequality and economics of science the author name disambiguation techniques are methodically straightforward, limited to certain disciplines and specific samples and do not include a systematic evaluation of the disambiguation performance (e.g., Gaule & Piacentini, 2018; Heinisch & Buenstorf, 2018). This can lead to subsequent problems. Schulz (2016) shows that database quality can invalidate bibliometric indicators, such as the number of papers per author. Schulz shows that those indicators are strongly depend on the performance of author name disambiguation approach. Therefore, author-level indicators, such as the h-index or number of paper-based rankings can significantly change in their value and ultimately become invalid.

With the advancements in methodology and the increase in computational capacities, machine learning methods are especially suited in providing an advanced disambiguation approach that can address these problems. Machine learning methods are state of the art in scientometrics and information science (Tekles & Bornmann, 2019). They can detect complex relationships in publication data and thereby disambiguate author names and their publications. I contribute to this literature by developing a supervised machine learning approach with graph-based methods that can handle missing data and rapidly disambiguates large author sets. These two characteristics are currently not properly addressed in other literature and are impediment to disambiguate full publication platforms like the WoS in a reasonable amount of time. I also combine traditional features with the thematic feature presented in Chapter 2. A strength of my approach vis-à-vis the literature is that I provide a detailed feature assessment that identifies relevant paper attributes. This may help other disambiguation approaches in the future.

Finally, I put my disambiguation algorithm into practice and develop an author-level scientometric database for German authors and their publications in the WoS. I linked this database with the to be described DNB dissertation database and thereby tackle a considerable data gap on junior researchers in Germany. The latest federal report on junior scientists in Germany indicates a lack of comprehensive databases on German PhDs' publications. It emphasizes that, currently, no conclusion on the scientific contribution and output of PhDs students can be drawn (Konsortium Bundesbericht Wissenschaftlicher Nachwuchs, 2017, p. 35). Existing German sources only address particular and PhD populations in specialized contexts. The study of Bornmann and

Enders (2001), for instance, is outdated and only address a small, survey based sample. My database addresses this gap and provides extensive information on the productivity of PhDs in Germany. This database is also linked to birthplace, gender and advisor information of the PhDs and allows to investigate a wide range of questions in the economics of science and of social inequality in academia.

1.5.3 Inequalities in academic and nonacademic outcomes of young German scientist: The case of place of birth and advisor gender

My dissertation contributes to applied inequality research of academia by investigating two strongly debated topics where literature and empirical evidence are scarce. These topics are: The career outcomes of eastern and western German PhD graduates and the protégés academic career outcomes by different protégé-advisor gender pairings in German academia. The developed databases in the previous chapters allow to investigate of these topics on the basis of valid and extensive databases.

Career outcomes of eastern and western PhD graduates

Starting with eastern and western German PhD graduates, the literature shows that PhD graduates are highly relevant for the economic prosperity and growth of knowledge societies. They create and transfer knowledge from universities to industry. A crucial issue in this respect is whether they can fully exploit their investment in education in their next jobs or whether they are at risk of mismatch on the labor market. For the PhD graduates themselves, part of their investment in education is unproductive, which translates into lower returns on investment in the form of employment below their skill level and lower wages. Findings on the labor market performance of PhD graduates and on the obstacles they face in using their abilities are therefore highly relevant not only for the individuals themselves, when considering their subsequent career paths, but also for policy makers and governments that finance the education of this group and support their integration into the innovation system.

From the social inequality perspective, PhD graduates belong to a country’s educational and economic elite, holding top positions in academic, economic, political or cultural spheres, while representing certain values and attitudes (Dahrendorf, 1965; Dee, Dee, & Thomas, 2004; Hartmann & Kopp, 2001). For Germany, this is even more the case than in other countries, as a PhD is not only a prerequisite for a scientific career, but is also associated with a high reputation and appreciation outside academia. Moreover, in more general terms, a high level of human capital such as that acquired by PhD graduates can generate positive externalities for the general public by strengthening social cohesion and political participation in a democracy (Auer et al., 2017). Hence, any factors that diminish PhD graduates’ returns to education may lead to adverse consequences for the individuals concerned, such as inadequate jobs and wages, and ultimately social inequality.

Focusing on the regional background as an inhibiting factor, eastern Germany constitutes an especially intriguing case. Unlike in other Central and eastern European transformation economies, the incorporation of the former German Democratic Republic into the western democracy and market economy was undertaken very rapidly, with western German institutions being extended to and implemented in the new eastern part of Germany (Salheiser, 2012, p. 123). As a result, a considerable number of the old eastern German elites were replaced by western Germans, which went hand in hand with the breakdown of the old Socialist elite recruitment regime (Best, 2005; Geißler, 2014). This profound exchange of elites continues to have an effect today. Bluhm and Jacobs (2016, p. 30) note that eastern Germans occupy only 2% of Germany's top positions, although eastern Germany accounts for 17% of the whole population. In eastern German public discourse, the underrepresentation of eastern Germans in top positions and the consequences for social and political coherence have frequently been the topic of lively discussions (e.g., Lukas & Reinhard, 2016), indicating that the transformation process in eastern Germany is still in progress. In the light of the ongoing public debates, it is surprising that there is very little representative empirical evidence on the underrepresentation of eastern Germans in top positions in Germany.

Against this background, Chapter 6 investigates whether having an eastern or western German background impacts whether or not PhD graduates can fully capture the returns on their education. It is unclear whether being from eastern Germany plays an important role for the employment trajectories of highly educated individuals, since the processes of acquiring social and cultural capital changed dramatically for eastern Germans in the course of reunification (Salheiser, 2012). Therefore, Chapter 6 traces the employment trajectories of eastern and western German PhD graduates in order to analyze whether the eastern German graduates fare less well than their western German counterparts and whether their eastern German background can explain this. In order to exclude any detrimental effects that might arise from systematic differences between the doctoral education systems in the German Democratic Republic and the Federal Republic of Germany, my coauthor and me only consider individuals who completed their dissertation after 1994. We compare the two groups with respect to two main labor market outcomes, thereby contributing to related findings for PhD graduates (e.g., Auriol et al., 2013; Di Paolo & Mañé, 2016; Koenig, 2019). First, we investigate whether an eastern German background is associated with a higher probability of being overeducated for the current job, taking up the conjecture that eastern German PhD graduates might be less likely than their western peers to work in jobs that fully exploit their human capital. Second, we examine whether an eastern German background is associated with a lower probability of achieving high wages as compared to a western German background. Hereby we take into account the persisting labor market differences between eastern and western Germany that specifically concern wages (Schnabel, 2016). To differentiate between an eastern or western German background we use the place of birth as the most straightforward measure. Since birth could be overshadowed by the location of the university where the PhD was completed or the subsequent place of work, we also consider these two measures.

The results reveal no significant negative impact on labor market success either for a birthplace in eastern Germany or for a dissertation submitted to an eastern German university. In that respect, the same qualification level results in the same labor market outcomes. It is more the place of work that matters, which indicates the profound impact of the still divergent economic conditions in the two parts of Germany on PhD graduates' employment prospects. In particular, a place of work in eastern Germany substantially reduces the chances of achieving high wages. This result is confirmed when the different regional differentiations are controlled for.

Protégé-advisor gender-pairings in academic survival and productivity of German PhD graduates

Protégé career outcomes by different advisor-protégé gender pairings is the second applied topic in inequality research and is addressed in chapter 5. Doctoral advisors are chosen as a research object because they are often the most influential persons at the beginning of an academic career. They transfer knowledge, attitudes, norms, and behavior to their protégés and influence their protégés' academic socialization and success (Barnes & Austin, 2009). Several studies have addressed the various scientific and socioeconomic characteristics of the advisors and their protégés to point out what makes these relationships mutually successful. Gender pairing in advisor-protégé relationships repeatedly stands out in this regard. It has diverse effects on career attainment and publication output of protégés (Gaule & Piacentini, 2018; Hilmer & Hilmer, 2007; Pezzoni, Mairesse, Stephan, & Lane, 2016).

Especially for pairings involving women this question is of high societal and scientific interest in Germany. As observed for other countries, women are underrepresented in advanced career stages of German academia (Larivière et al., 2013). Although they account in 2017 for 51,7% of the graduates, their share of PhD holders is 45,4%. Women's share even lowers to 25,6% when considering German professors (Statistisches Bundesamt, 2020). This female exit from the academic workforce indicates social inequality and a misallocation of talent (Acemoglu, 1995). Consequences imply decelerated scientific progress with negative spillovers to industry and the economy in general. Women may also be individually affected. If they are equally qualified to start and pursue an academic career, but at some point quit, their educational investment cannot be fully exploited (McGuinness, 2006).

Gaule and Piacentini (2018) argue that this under-representation of women in academia perpetuates itself through the lower availability of same-gender advisors for female students. They argue that underrepresentation works through a productivity channel or a preference channel. In the productivity channel, students are less productive when collaborating with an advisor of the opposite gender. As productivity is generally the primary driver of academic career success, this leads to higher drop-out rates for female PhD graduates advised by men. In the preference channel, the authors argue that working with an advisor of the opposite gender is less enjoyable and leads to lower career satisfaction and a higher chance of dropping out early. Gaulle and Piacentini show that research productivity during the PhD, and the propensity to become faculty after graduating, are both related to the gender of the advisor.

I add to the findings of Gaulle and Piacentini and investigate the same realtionship for German PhD students and their advisors. In the first step, I test whether productivity during the PhD is linked to advisor-protégé gender pairings in German academia. In the second step, I focus on the disentanglement of the temporal patterns related to career outcomes and advisors' gender after the PhD. From the temporal perspective, academic careers, and careers in general, are non-dichotomous processes. They include multiple decisions and promotions that differ in their duration and in their point of time. The investigation of fixed points in time, as done in Gaule and Piacentini (2018), does not exploit the temporal dimension to its full extent. In this sense, it is an open question of how long protégés in different gender pairings remain in academia and which drop-out „risk" they take after their PhD.

These durations can be considered as survival times and allow to utilize related models such as Cox proportional hazard or complementary log-log regression. The complementary log-log regression used in this paper estimates covariates' effect upon the time a specified event takes to happen and assumes time to be discrete (Tutz & Schmid, 2016). Therefore, I can investigate how the advisor's gender and other characteristics affect the time one PhD graduate remains in academia after finishing his or her PhD. A similar methodology has been applied by Sabatier, Carrere, and Mangematin (2006) to investigate the time it takes for female and male postdocs to attain professorship.

While I find that being female has a strongly negative effect on publishing during the PhD, being advised by women does not have any effect on publication productivity during the PhD. The academic survival probability by gender and advisor gender as measured by the last year of publication is investigated with time discrete cloglog regression and represents my main finding. I find that female advisors lead to a 37% lower yearly probability to write the last publication; this effect is not different between men and women. In line with the observable female underrepresentation in academia, I find that women have a yearly 38% higher hazard to exit from research (as proxied by the author's last WoS publication).

1.6 Databases

The following subsections describe in more detail the databases used in this dissertation. I review the electronic catalog of the DNB and its linked datasets, the WoS publication database and my dataset of online dissertations. Finally, I explain the intersections of these databases and how I employed them in this dissertation.

1.6.1 The electronic catalog of the DNB and its linked datasets

The DNB is Germany’s central archival library. The DNB collects, documents and archives all printed publications and sound recordings issued in Germany together with works that were composed in the German language or that relate to Germany (Deutsche Nationalbibliothek, 2019). Since PhD graduates are required by law to supply a copy of their dissertation to the DNB, it holds an almost complete set of dissertations submitted to German universities since the 1970s. The electronic catalog of the DNB features information on dissertation authors, university name, year of publication and subject and therefore is a highly suitable data source for research on PhD graduates and young scientists in Germany (see e.g., Buenstorf & Geissler, 2014a; Buenstorf & Heinisch, 2020). I refer to a 2015 copy of the DNB catalog processed by the Chair of Economic Policy, Innovation, and Entrepreneurship.

With the exception of Chapter 3, the DNB is part of every chapter of this dissertation. The DNB was used either alone, such as when exploring dissertation thematic differences in Chapter 1, or used in conjunction with other databases. In this regard, Chapter 4 concerns the linkage of disambiguated WoS author data to the DNB and represents a unique database for the publishing activity of junior scholars in Germany. The database used in Chapters 3 and 5 is based on online dissertations that were downloaded either from the DNB directly, or from university servers and then linked to the DNB. In the chapters of my dissertation, the DNB catalog always acts as the point of departure, since it includes essential PhD graduate characteristics, such as the year and place of dissertation.

Chapter 6 uses the database from the IAB-INCHER project of earned doctorates (IIPED). The IIPED combines information on dissertations in the DNB electronic catalog with individual labor market history from the Integrated Employment Biographies (IEB) of the Institute for Employment Research (IAB) (see Heinisch et al., 2020 for more details).

1.6.2 Online Dissertations

One drawback of the DNB catalog is that the PhD graduates’ place and date of birth, advisor information and other important author-level characteristics are very rarely reported. In order to retrieve this essential information, I used online dissertations. In many faculties, PhD students must report their place and date of birth, as well as their advisor's name, on the front page of their dissertation.[8] Their basic publication and author information, including a URL link to their dissertation, are indexed in the DNB catalog in the same manner as for printed publications. I used this link to download 40,000 dissertations from the DNB server for which I could find essential information online. However, not all dissertations in the DNB database have working URL links. I therefore resorted to systematically searching individual university servers for online dissertations as a second strategy.[9] These were linked with the dissertations in the DNB catalog with approximate matching of the author's name, the university name and the year in which the dissertation was submitted.[10] These two procedures yielded a total of 79,000 dissertations for which I know the unique identifier in the DNB catalog. Because the DNB catalog version used is from 2015, online dissertations after 2015 remain unmatched.

The PhD graduate's birthplace and advisor were retrieved with text-pattern matching. Typical keywords on front pages or curriculum vitae, like "place of birth" or "supervisor", indicate the subsequent mention of information of interest. For dissertations written in English, I systematically searched for words like "born in", "birthplace" and "advisor". For dissertations in German, I repeated this procedure with corresponding German terms.[11] I searched for these keywords on the front pages or in the curriculum vitae of every dissertation from the DNB and various university servers and saved the three words following each keyword. In the next step, I manually cleaned the resulting text string of frequent errors.

To find information on birthplaces, me and my colleague Maria Theissen searched the text string in Google Maps to obtain a unique address and more general information, such as country, state and zip code. The Google Maps search engine is advantageous because it accepts diverse spellings and ambiguous German city names.[12] I was able to identify the birthplaces of 27,321 German PhD graduates with this procedure.

I proceeded differently with advisor information. The fundamental problem with advisor names is that they are subject to a disambiguation and record linkage problem. The advisor name "Müller, Matthias", for instance, appears on 50 different online

[8] Sometimes the dissertations also include a curriculum vitae.

[9] These servers include the full set of online dissertations (as of August 2017) from the universities of Kassel, Munich (TU and LMU), Braunschweig, Freiburg, Frankfurt am Main, Greifswald, Darmstadt, Düsseldorf, HU Berlin, Halle-Wittenberg, Magdeburg, Regensburg, Rostock and Ulm; all universities in Saxony and Thuringia; and the Karlsruher Institut für Technologie.

[10] I used a fuzzy-string matching procedure based on the Levenshtein distance for the author's name and allowed a time window of 2 years before and after the date of the dissertation in order to compare the year of submission to the DNB with the years stated on the university server website. This is necessary because the two dates do not necessarily coincide. To correct mismatches, in the name-matching procedure I also checked whether the matched name appears on the front page of the dissertation.

[11] The German expressions are "geb. in", "geboren", "aus", "Geburtsort" and variations of these terms.

[12] Since some German town names occur more than once in Germany, the nearby river is added to their names in order to avoid confusion. However, the attachment of the river name is not used consistently; for example Halle/Saale, Halle a. d. Saale and Halle Saale.

dissertations. In the DNB there are again 100 different dissertations with the advisor name “Mueller, Matthias”. Which advisor dissertation now relates to which protégé dissertation? Some of the 50 dissertations cannot plausibly function as dissertation advisors since the dissertation year is greater or equal to the protégés dissertation year. The remaining ones, however, can function as an advisor and require to use some matching approach. Since I don’t have any matching variable other than the name, I dismissed all ambiguous advisor names. Therefore, common advisor names like “Muller, M” were not considered in my approach. Instead, I linked only distinct advisor names, such as “Bünstorf, G”. This strategy incurred a loss of valuable advisor information but ensured high-quality advisor-protégé pairs. This resulted in 13,315 advisor-protégé pairs where the advisor has a unique name in the DNB catalog. The advisor-protégé pairs are used in Chapter 5.

1.6.3 The Web of Science publication database

The WoS is one of the leading bibliometric platforms in the world. It dates back to the creator of the science citation index, Eugene Garfield (Garfield, 1972), and is used in numerous bibliometrics studies (see e.g., Bornmann & Mutz, 2015; Fudickar, Hottenrott & Lawson, 2018). In this dissertation, I use a 2017 version of the WoS where author addresses have been disambiguated by the Komeptenzzentrum Bibliometrie (Rimmert, Schwechheimer & Winterhager, 2017). The 2017 WoS includes about 52 million publications from 1980 to 2017, related to 178 million author names. There are several peculiarities in this version. First, the WoS has low coverage of arts and humanities publications and is focused on English-language journals (Mongeon & Paul-Hus, 2016). In the social sciences, especially in journal publication-dominated disciplines, the coverage is significantly better. The WoS has the best coverage in natural sciences, and has improved its coverage across all disciplines in the last two decades. The quality of the bibliometric information in the WoS has also improved (Liu, Hu & Tang, 2018). From about 2006 onwards, the WoS has generally provided full coauthor information for the included papers. Before 2008, only the corresponding author was required to fill in his or her address and basic information, such as first name.

This missing information, in conjunction with the author name disambiguation, is one of the major drawbacks of the WoS database. I address this issue by developing a machine-learning method for author name disambiguation using the WoS *Researcher ID*. Using the Researcher ID, authors can assign papers to their user account in the WoS. The Researcher ID has previously been shown to provide true authorship information by Tekles and Bornmann (2019) and is amply available in the WoS database. There is a Researcher ID available for 21 million paper-author relationships, which makes them distinguishable from other authors. The WoS database is used in Chapters 3, 4 and 5.

1.6.4 Database relations

This subsection describes the relations between my created and external databases. Most of the datasets used come from a cascade of processing and linking steps and build an interrelated data complex. Figure 1 illustrates this complex. The chronological point of departure is the DNB and the detection of thematic differences of eastern and western German doctoral graduates addressed in Chapter 2. I also used the methodical approach from Chapter 2 to create a machine learning feature for the author disambiguation algorithm in Chapter 3. Besides that, however, Chapter 2 remains isolated from a linkage perspective.

The next step was scraping the online dissertations from the DNB and various university servers. I linked the dissertations from university servers to the DNB by approximate matching of year, university and first and last name. I checked a sample of 100 linked dissertations for matching plausibility. I then scraped advisor and birthplace information from the downloaded and linked dissertations. This information is the basis for Chapters 4, 5 and 6. In Chapter 6, I refer to the IIPED dataset from Heinisch et al. (2020). The IIPED dataset links the DNB with social security data from the IAB. It includes the birthplaces retrieved via their linkage to the DNB. The birthplaces also play a role in Chapter 4, where they are used to show differences in bibliometric outcomes of eastern and western German doctoral graduates.

Chapter 5 uses the linkage complex of advisor information scraped from the online dissertations, the DNB data and the author disambiguated dataset. Chapter 3 and 4 are strongly related. In Chapter 3, I develop and test a machine learning-based author disambiguation algorithm. In Chapter 4 I apply an older, but nearly identical version of this algorithm and disambiguate 50% of the German author name blocks. This is because the paper upon which Chapter 3 is based was subject to revisions after submission to the *Journal of Informetrics*. For reasons of improvement and currency, I included the revised version of the paper in this dissertation.

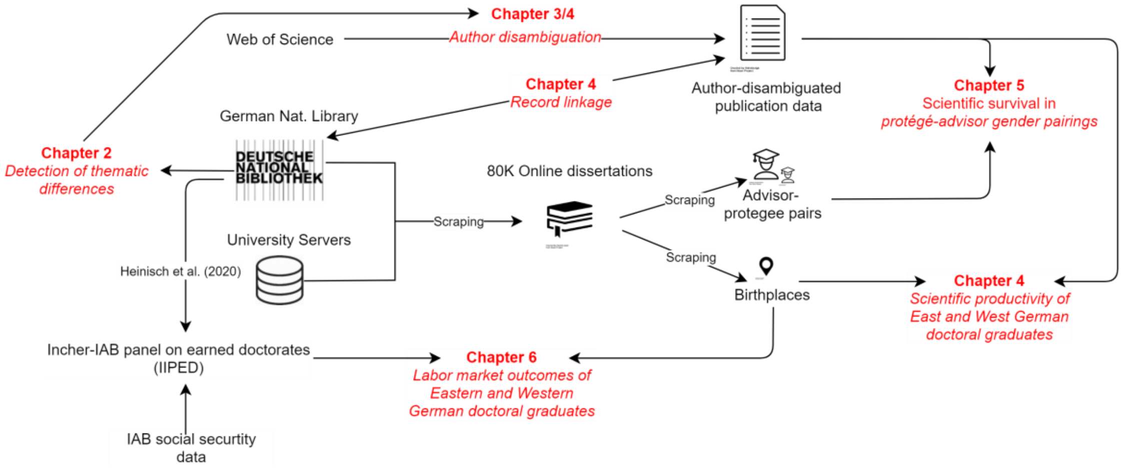


Figure 1. Database relations

Source: Own depiction. Icons by RockIcon, David Lopez and thirddesgin from NounProject.

2 A structural topic model approach to scientific reorientation of economics and chemistry after German reunification

2.1 Preface

This chapter builds on the paper: Rehs, A. (2020). A structural topic model approach to scientific reorientation of economics and chemistry after German reunification. *Scientometrics 125*(2), 1229–1251. https://doi.org/10.1007/s11192-020-03640-0. The paper is reproduced here in its published form, with only minor editorial changes to make it consistent in style with the remainder of the thesis. I already made a first modest attempt to study thematic convergence of eastern and western German research after reunification in my bachelor thesis, however based on much more limited data and completely different methods than are used in this chapter.

2.2 Introduction

Growth of science, growth of topical difference identification issues?

Classification systems of scientific literature play a central role in bibliometrics (Glänzel & Schubert, 2003) and will become more and more important with the exponentially growing amount of scientific literature. From World War II to the early 2000s, the stock of scientific literature is estimated to have doubled about every 9 years (Bornmann & Mutz, 2015) and in 2009 amounted to over 50 million publications (Jinha, 2010). These growth rates and underlying numbers raise concerns that the large current and future stock of knowledge will become more and more difficult to structure for single scientists (Landhuis, 2016) and established databases (Larsen & von Ins, 2010). Traditional classification systems rely on keyword assignments, expert-based classification of subjects, and forward and backward citations to embed a publication in the network of knowledge flows in scientific literature (De Bellis, 2009). These methods include high levels of complexity reduction and therefore a loss of knowledge in the content of the scientific publications. Practically, subtle but often decisive differences between two papers on the same topic can therefore hardly be addressed without having expert-level knowledge in the respective scientific field. In the same manner, topical overlaps between loosely related papers cannot be detected without having expert knowledge in both papers’ fields. The addition of more and more papers will eventually constrain the ability of experts to detect differences and similarities between papers. The large-scale quantification and detection of thematic differences in research topics is therefore an open problem in scientometric research. Advances in machine learning, especially in the statistical analysis of large text collections, alleviate these issues under certain circumstances. In this way precise difference detection between scientific texts can be feasible without having deep knowledge in the respective field.

The case of scientific reorientation in eastern and western Germany

In this chapter, I therefore develop and apply such a machine learning approach to difference detection based on the case study of dissertation titles written at eastern and western German universities in economics and business administration and chemistry before and after German reunification. German reunification is especially suited for investigating differences in research topics because the transition of the political system in eastern Germany went hand in hand with the transition of the scientific system. German reunification led to the dismantling in eastern Germany of a large number of chairs, institutes and research organizations, as well as a broad institutional restructuring in academia. Reasons included political motives, but in several instances also a mismatch between what had been researched under the old (socialist) system and what was considered interesting in the new one. This change affected social sciences more severely than natural sciences and therefore provided two different structures to investigate thematic differences and topical reorientation. In these two structures, motives and incentives for individual scientists in the two disciplines and parts of Germany to change research topics differed substantially and may have manifested in minor and major thematic differences. The section "Historical background" will therefore elaborate on the disciplinary and general historical circumstances before and after the reunification.

Dissertation as a data source

Journal publications and their linked indicators, such as citations, are the main subject of investigation in scientometric research and have contributed to substantial advances in the field (e.g., Garfield, 1972; Hirsch, 2005). However, under certain historical, institutional and disciplinary circumstances, such as in my case, journal articles are not the best means of inquiry[13]. Therefore, I use dissertation titles as an alternative source of information to identify and track the differences in the two disciplines in Germany before and after reunification. Dissertation titles offer several potential advantages for my approach and are, despite limited use in scientometrics (Morichika & Shibayama, 2016), amply available in Germany. This is because every doctoral student is mandated to send in a copy of his or her dissertation to the DNB (Deutsche Nationalbibliothek). The DNB

[13] At the most general historical level, journal publications have not been the dominant medium for scientific communication in disciplines where they are nowadays the standard form of publication. This especially applies to one of my subjects of investigation, namely economics and business administration in western Germany (Hicks, 1999; Leininger, 2008). I found no literature that reflects on the publication system and culture in economics and business administration in the German Democratic Republic. Regarding chemistry in western Germany, Weingart, Strate, & Winterhager (1991) indicate that journal publications were in the 80s and today still are the most popular means of publication (Hahn, 2009). In the German Democratic Republic, publications in chemistry were common, but due to isolation the German Democratic Republic underperformed in comparison to western Germany in terms of relative publications per capita.

archives the dissertation and stores some basic author and dissertation information in its electronic catalog. I have access to this catalog, which provides me an almost complete list of dissertations that were submitted in both parts of Germany, since 1970. Thus, I have a good picture of the thematic landscape during my period of investigation in Germany. My work is based on a number of presumptions: First, in Germany the doctoral advisor (often dubbed the "Doktorvater") has a strong influence on the doctoral student and their choice of research topic. Moreover, the advisor is usually required to have a chair at a university, as only they are entitled to award PhDs. Therefore, the dissertation topics most likely represent the research topics present at a chair. Second, the title of a dissertation represents its content in a very condensed form. Together, these assumptions lead to the conjecture that the research focus of a chair is reflected in the titles of dissertations submitted at an a university with which he or she is affiliated. This allows me to draw conclusions on the general thematic landscape of university research in Germany during my period of investigation.

A structural topic model approach to differences in dissertation titles

My main effort was in applying a probabilistic text model ("structural topic model") to these dissertation titles, aggregating the outcomes and then incorporating them into a linear regression framework, which allows me to calculate the level of difference between dissertation titles by regional and temporal origin of the dissertation. In this way my approach demonstrates how to identify and track differences between scientific work on the level of individual researchers, but also larger entities of the scientific system, such as different scientific disciplines or parts of a country. In my case study, I find in economics and business administration research topics considerable differences between eastern and western Germany before reunification. After reunification, I observe a strong and rapid conformation. In chemistry there are few differences between eastern and western before reunification. Afterwards, the results suggest a moderate thematic convergence.

2.3 Historical background

The scientific system and doctoral education in the German Democratic Republic

Since the birth of the two Germanies in 1949, the intra-German relationship has been characterized by a competition of political (and economic) systems. Walter Ulbricht, prominent veteran socialist politician of the German Democratic Republic (GDR) was renowned for his saying "overtaking without catching up". The early socialists strove to demonstrate the superiority of socialism over capitalism, with scientific and technological achievements playing a central role. Even the constitution of the GDR (§ 2, Abs 1) claimed that the foremost aim of a socialist society was to increase the effectiveness of scientific and technological development and labor productivity (Volkskammer der DDR, 1976). This orientation of scientific advancement on aspects of productivity dated back at least to Lenin and had consequences for the academic landscape of the GDR. Industrial application of research findings was heavily emphasized. Basic research was carried out almost exclusively by universities, but free choice of the research subjects was increasingly restricted and almost non-existent beginning in the 1960s (Gruhn & Lauterbach, 1977). PhD candidates had minimal freedom in choosing their research subjects. In the case of Humboldt University in Berlin shows that roughly two-thirds of dissertation topics followed the five-year research plans of the government (Wollgast, 2001). Furthermore, international contact was more or less limited to other socialist states and access to western world academics and their publications was difficult to gain (Mann, 1979). Limited financial resources made internationally competitive research impossible in the majority of scientific fields. However, the conditions of career advancement in academia closely resembled those in western Germany. The average student in the GDR had to complete a basic and an advanced (or specialized) part of his study to earn a degree. Afterwards, a dissertation (Promotion A) had to be written to obtain the title “Dr.” in a scientific field. In contrast to the Federal Republic of Germany, the GDR had universal requirements for the award of a PhD degree, which included a fair amount of ideology (Deutsche Demokratische Republik, 1968, § 5, Abs. 1). PhD degrees could be earned through research studies (2-3 years long, similar to a graduate school), employment at a university chair (usually four years’ contract) or distinction in industrial and societal engagement (similar to an external PhD candidate) (Belitz-Demiriz, Voigt, & Gries, 1990; Guenther, 1989).

Unlike in the GDR, the scientific system of western Germany during my period of investigation was (and still is) free of ideological constraints. The constitutional (basic law) “freedom of teaching and research” (§5, Abs. 3) guaranteed vast autonomy for university researchers. Regarding factors that could have implicitly constrained freedom of research in western Germany in the 1980s and early 1990s, Peisert and Framheim (1994) argue that, in the case of third party funding, there was no strong influence from semi-public and public institutions on research topic choices. The systems of doctoral education in eastern and western Germany closely resembled each other; both countries doctoral students were predominately employed at the chairs directly; graduate schools played a minor role. However, the level of involvement of ideology in doctoral education clearly distinguished the two.

The transition and political change in Germany in 1990 had a deep impact on academic institutions, most notably in scientific fields that were heavily affected by socialist ideology. The prime example is economics and business administration, which was almost completely dismantled and rebuilt from the ground up, often involving new personnel, structures and research agendas. Kolloch (2001) reports that by 1994 90% of the economics and business administration chairs at the biggest eastern German university (HU Berlin) were replaced with western Germans.

In chemistry the historical preconditions were quite different. In the GDR, the discipline was considered to be a crucial scientific productive force that would directly and indirectly increase economic output. Chemistry and other natural sciences were therefore oriented to the requirements of the local industry (Meske, 2004), which led to a much greater focus on applied research in eastern Germany. GDR policymakers, for example, built a technical college in the centre of the eastern German chemistry cluster Leuna-Buna-Bitterfeld. The GDR chemical industry and, in consequence, the discipline of chemistry was dependent on crude oil deliveries from the Soviet Union to produce precursors and final chemical products. The GDR, however, used the dominant share of crude oil deliveries from the Soviet Union to refine petrol, which was to a large extent exported in order to bring in much-needed hard, foreign currency. This petrol-focused production caused a shortage in the production of other products based on crude oil (e.g., rubber and plastic). eastern German chemistry therefore researched non-oil-based ways of producing such goods. Lignite was a viable alternative, since eastern Germany had large lignite resources and existing processing facilities dating back to World War II. For the scientific discipline of chemistry this lignite based “business model” of the GDR resulted in a strong emphasis on related research problems. Chemistry as a discipline was therefore politically determined, applied and focused foremost on the special demands of eastern German chemical industry. For western Germany I find no indication of any profound specialization or a general focus on applied topics in chemistry. This may be a consequence of the constitutional right of freedom in teaching and research and a conservative industrial policy.

2.4 Data

The two disciplines, economics and business administration and chemistry, and their historical background before and after German reunification are therefore suited for my analysis of identifying research topic differences. They provide two structures: for economics and business administration, a structure with substantial topical heterogeneity before and after reunification; and for chemistry, one with relative topical homogeneity. In the following section, I will describe the processing steps used to obtain the final dataset of thematic differences in dissertation titles (Rehs, 2020b).

I use the online catalog of the DNB as the basis for my analysis. The catalog lists the vast majority of PhD dissertations submitted at German universities, including the GDR. There are entries for approximately one million PhD dissertations, which are classified by subject. I use this classification to distinguish between economics and business administration and chemistry. Due to the peculiarities of German medical dissertations, I eliminate dissertations which are cross-listed in chemistry and medicine. Furthermore, I employ information on university location (cities, name of university or a combination of both) to separate eastern from western dissertations[14]. I assume that reorientation of research topics after the reunification continued until 2010. To obtain a picture of the thematic landscape before reunification, I consider the years 1980 to 1989. The years 1990 to 1994 are eliminated from my data, since the replacement of eastern German chairs took several years and the number of observations from eastern German university dissertations dropped significantly during this time period.

In the next step, I paste every dissertation title and subtitle into one string and standardize this string. My pre-processing includes standard text-mining methodology: transformation to lowercase, removal of punctuation, language detection and removal of non-German titles, stemming, n-gram detection and removal of very frequent words, rare words, stopwords and short titles. Different languages in a text collection can considerably distort the outcomes of the topic modelling algorithm to be presented due to problems with (text-mining) token recognition. Although differently spelled words can have the exact same meaning in two languages, they are considered statistically as different tokens in text machines. Solutions based on translation cause more problems than they solve. My approach is therefore to exclude all titles written in English. I am aware of the downsides of this procedure and might miss some important dissertations that are addressed to an international audience. Dissertations written in German might also differ in quality. Nevertheless, as my language identification algorithm (Ooms, 2018) shows, English titles only account for roughly 10% of the dissertations. The small number of English titles would therefore distort the statistical inference based on topic modelling. All titles identified as neither German nor English are defaulted to German.

Mentioned n-grams are applied because some words are by nature bounded, like “United” and “States”. To improve the performance of the topic model to be presented, I want the algorithm to treat these words as one character. Bigrams are two bounded words and trigrams three bounded words. In both corpora I count the most frequent bi- and trigrams. I assume that only the top bi- and trigrams add relevant context for the subsequent algorithm. For both economics and business administration and chemistry, I set the boundary for relevant n-grams at the top 1%. I proceed by searching these n-grams in every string. If they occur, I add them to the string and remove the words that composed them.

I remove very frequent and very rare words for reasons of complexity reduction and minor relevance for topic modelling. Very frequent have the same properties as stopwords, but are not included in standard stopword dictionaries since they are dataset specific. They don’t add relevant context; rather, they are commonly used terms within a dataset and identically distributed across all documents (e.g., for dissertation titles, “investigation” or

[14] I exclude observations which are labelled “Uni Berlin”, since it is uncertain whether the university is in east or west Berlin.

"method" may appear very frequently). I set the threshold for removal at the upper 0.1% limit of the most frequent words. The same holds for very rare words. Because of their low frequency, they don't add context, and are removed if they appear fewer than 3 times in total.

Finally, I delete very short titles from my data set. Since topic modelling infers the topic distributions by drawing words from each title numerous times, titles consisting of only few words can be problematic because there is less room for randomness in each title. I therefore exclude titles containing fewer than five words.

2.5 Topic modelling in large-scale text analysis

The latent Dirichlet allocation

To address my research question I use topic modelling, which is a family of probabilistic methods for analyzing large text collections. Topic modelling has found various applications in scientometrics, such as in investigating the topics that construct scientific publications (Blei & Lafferty, 2007). Any topic modelling algorithm is, in general, a generative model of word counts. In my case that means I define a data-generating process for each dissertation title and then use the data to find the most likely values for the parameters within the model.

The most common topic modelling algorithm is the latent Dirichlet allocation (short: LDA, Blei, 2012; Blei, Ng, & Edu, 2003). In the LDA algorithm my dissertation titles are represented as mixtures of topics. In these mixtures, each word within a given dissertation title belongs to exactly one topic. Single dissertation titles can therefore be considered as vectors of topic proportions, which indicate the percentage of words belonging to each topic. In the following section I will describe the statistical methodology and orient on the notation and description of (Roberts, Stewart, & Airoldi, 2016; Roberts, Stewart, & Dustin, 2019; Roberts et al., 2014).

The generative process in LDA starts by considering each dissertation title (index: Diss) as a distribution over topics (θ_{Diss}), which is drawn from a global prior distribution. In the next step, for each word in the dissertation title (indexed by n), the LDA algorithm draws a topic (z) for that word from a multinomial distribution based on its distribution over topics ($z_{Diss,n} \sim \text{Mult}(\theta_{Diss})$). Depending on the topic selected, the observed word $w_{Diss,n}$ is drawn from a distribution over the vocabulary $z_{Diss,n}$ ~Mult($\beta_{z_{Diss,n}}$), where $\beta_{k,v}$ is the probability of drawing the v-th word in the vocabulary for topic k.

A hypothetical pre-1990 eastern German title in economics might therefore be represented as a mixture over 10 topics. Topics are, again, a distribution over words that are more or less likely to be related to that topic (e.g., "Marx", "worker", "class" might each have high probability in the same topic). The LDA is completed by assuming a Dirichlet prior for the topic proportions such that $\theta_{Diss} \sim \text{Dirichlet}(\alpha)$. However, there are disadvantages that come along with the application of LDA. The resulting posterior distributions can have many local modes. That means that different initializations can produce different solutions. In order to address this issue, I use the spectral initialization procedure described in (Arora et al., 2013), which is also implemented in the R package on structural topic modelling (Roberts et al., 2019)

Structural topic modelling

Structural topic modelling is an extension of the LDA process described above which allows covariates of interest (such as the temporal origin or university of the dissertation) to be included in the prior distributions for dissertation-topic proportions and topic-word distributions. Thus, the covariates offer a method of “structuring” the prior distributions in the topic model, including additional information in the statistical inference procedure. The topic prevalence (as described in the LDA section) can therefore be influenced by some set of covariates X through a standard regression model with covariates $\theta \sim$ LogisticNormal $(X\gamma, \Sigma)$. In contrast to the described LDA algorithm, I abolish the assumption that topical prevalence (how much a topic is discussed by a covariate) is constant across all dissertation titles. This is a major improvement in comparison to LDA and allows the parameters that generated the dissertation title to be reconstructed more precisely.

I use university and year dummies of the dissertation as topical prevalence variables in my structural topic models on chemistry and economics and business administration. I argue that these variables are best suited to capture temporal and university level variation in dissertation titles and are different from the main independent variables in the regression framework to be presented. Year dummies as topic prevalence variables should capture trends and temporarily popular topics in the 25-year span of my investigation. For universities, irrespective of their eastern or western German background, I presume that there are regionally bound topics. This is because the chairs at universities might have inherent topics that are reflected in the dissertations they produce. Therefore, I include university dummies as the second set of topical prevalence variables in my topic model. In structural topic models, proportions (θ) can also be correlated (see also Blei & Lafferty, 2007); i.e., in a given dissertation title, the high proportion of a topic that is related to socialism might also increase the likelihood of high proportion of a related topic (e.g., a topic related to Leninism).

In my structural topic modelling, I stopped at the point where θ can be influenced by some set of covariates X through a standard regression model with covariates $\theta \sim$ LogisticNormal $(X\gamma, \Sigma)$. The next step in the structural topic model algorithm is described as: “For each word (w) in the response, a topic (z) is drawn from the response-specific distribution, and, depending on the topic, a word is chosen from a multinomial distribution over words parameterized by β, which is formed by deviations from the baseline word frequencies (m) in log space ($\beta_k \propto \exp(m + K_k)$)" (Roberts et al., 2014, p. 4). This distribution can include a second set of covariates that can model how word frequencies between values of that covariate can differ. Within a "socialistic" topic, this allows GDR dissertations, as indicated by a variable, to use the word "Marx" more frequently than dissertations from western Germans (they might use "Engels" more often instead). Since the used version of the R package (Roberts et al., 2019) allows the inclusion of only one variable for such "topical content" and my approach would require several other variables, I don't include any variable of such kind.

When it comes to finally fitting the structural topic model, the major problem is in the mathematically intractable posterior distribution. To solve such a problem, (Roberts et al., 2019, 2014) developed a method for approximate inference based on variational expectation-maximization algorithms (Blei, Kucukelbir, & McAuliffe, 2017; Dempster, Laird, & Rubin, 1977) that, upon convergence, give estimates of the model parameters. Convergence is achieved when the change in the approximate variational lower bound between the iterations becomes very small. I accordingly set the value for convergence to 1e-06.

In conclusion, there are two major improvements that structural topic modelling provides for my setting as compared to LDA. First, topics can be correlated, which much better reflects the "true" data-generating process behind dissertation titles and science in general. The second major improvement is that each dissertation title has its own prior distribution over topics defined by covariate X, rather than sharing a global mean.

Topic model application and cosine similarity regression framework

In the next step, I estimate two separate structural topic models – one for economics and business administration and one for chemistry. For both I consider the whole period of investigation from 1980 to 1989 and 1995 to 2010. For each topic model I use 75% of the dissertations to estimate the model parameters. For the remaining 25%, I apply the topic models. This separation of training and test datasets is a standard procedure in machine learning and aims to detect overfitting of my models. Overfitting means that my topic model learns the data generating process of the underlying titles too well. In this way I lose model flexibility, which has negative impacts on the performance of the topic model on new, unseen dissertation titles. The final training and test set sizes in economics and business administration are a randomly sample of processed dissertation titles and include 6,855 observations for the training and 1,767 for the test set. In chemistry, sizes are 10,361 and 2,580. Eastern German test titles account for 317 dissertations in chemistry and 338 dissertations in economics and business administration (training and test set). In economics and business administration this broadly reflects the population size of eastern Germany (about 18% that of western Germany). In chemistry I find no explanation for the proportionally smaller number of dissertations in eastern Germany (9%).

When finally fitting my topic models, I arrive at 76 topics in chemistry and 69 in economics and business administration. One result of the two topic model applications is that I obtain a topic distribution for every title. Figure 2 illustrates the topic distribution of two titles in my topic model for economics and business administration[15] [16]. Figure 3 now represents the top words with the highest β probability of two topics. I choose topics 11 and 40 in economics and show their yearly mean probability across all titles because they show how two, probably very antagonistic topics change in prominence over time. While topic 11, which may indicate socialism, loses importance after 1990, topic 40, as a probable proponent of capitalism, on average gains importance. A list of words associated with other topics can be found in Appendix A. Since every topic is a probability distribution over words, top words may provide some indication of the underlying subject. However, interpretation should be done very cautiously, since the most probable words only represent a small fraction of the probability distribution. Moreover, most probable words are not necessarily the most exclusive words to a topic.

Figure 4 is similar to Figure 3, but shows the distribution of the mean topic probability before and after the reunification for all topics. For economics and business, I can observe a high popularity of a small number of topics in eastern Germany before the reunification (such as topic 11). After reunification, high mean probability for single topics in one part of the country disappear. In direct comparison to economics and business, the mean probabilities for single topics in chemistry are small. However, there are still differences in some topics between eastern and western Germany before the reunification. Remarkably, topics that weren't popular before the reunification in one part of the country became popular after the reunification. The popularity of topic 71, for example, increased considerably after the reunification in eastern Germany.

[15] To improve readability, I show the original title without the cleaning steps described in the previous chapter.

[16] Translation for title 1: Category management in multi-channel-retailing marketing and market behaviour. Translation for title 2: The institutions of the Swiss real estate market. An analysis under consideration of transaction costs with suggestions to market efficiency increase.

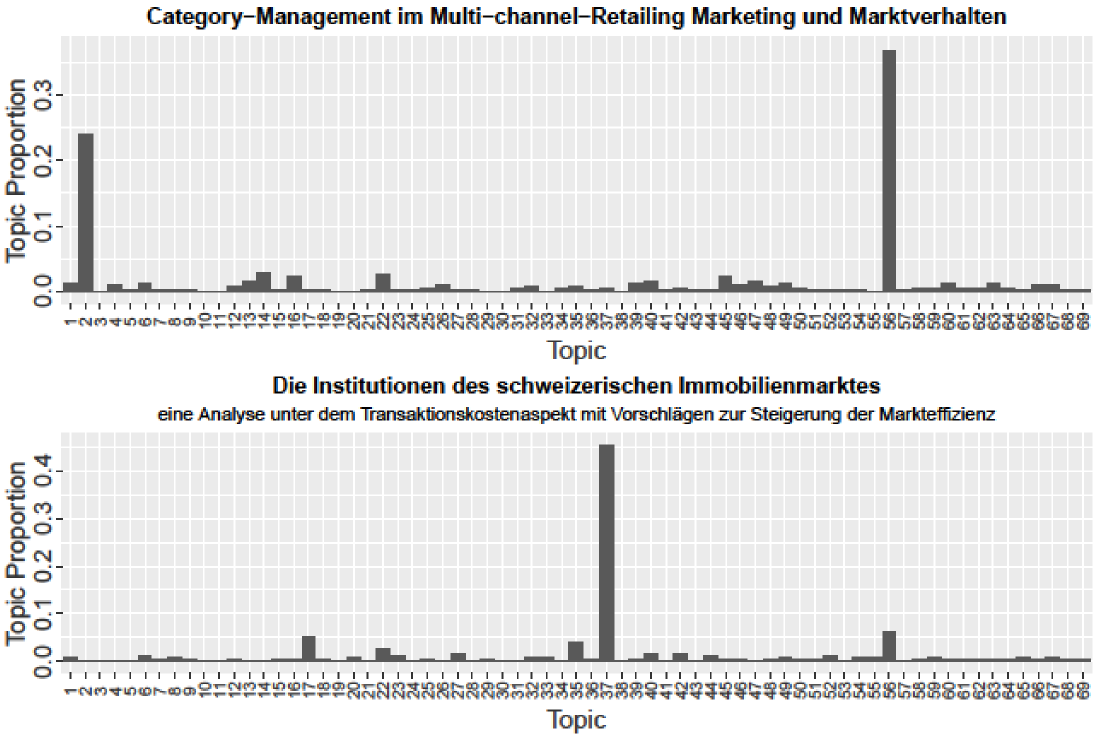


Figure 2. Two title-topic distributions of the economics and business adm. topic model

Source: Own depiction.

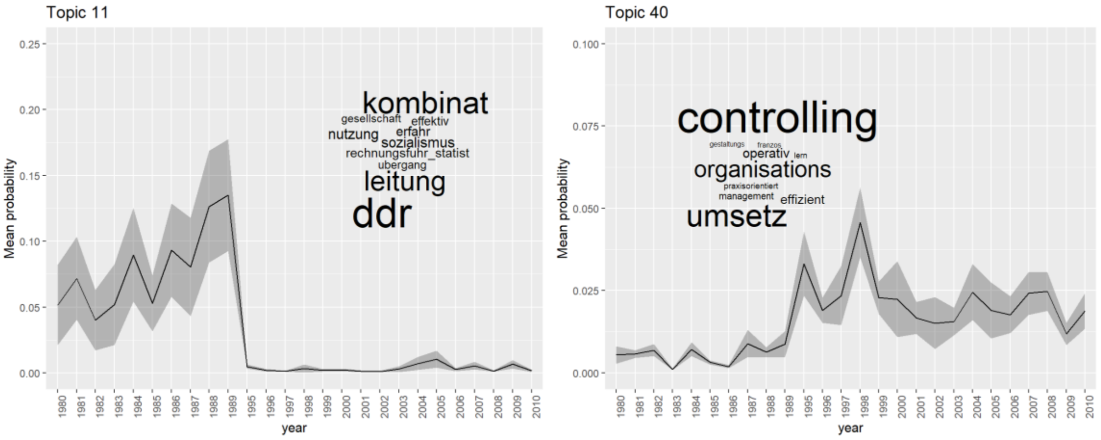


Figure 3. Topical prevalence of two economics and business administration topics

Source: Own depiction.

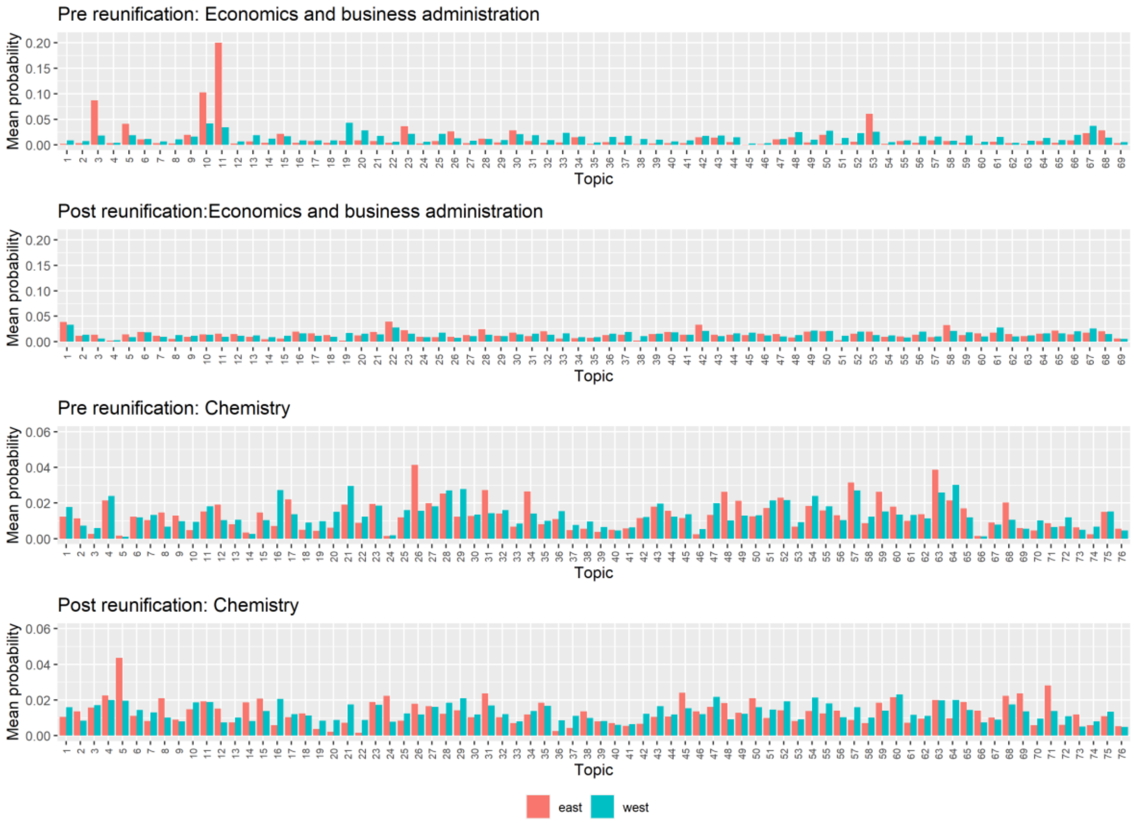


Figure 4. Mean topic prevalence before and after German reunification

Source: Own depiction.

In order to compare the retrieved topic distribution of every title, I now use the cosine similarity measure, which has various applications in the comparison of topic model outcomes (see e.g., Ramage, Dumais, & Liebling, 2010). The cosine similarity is a measure for the distance between two vectors and is defined between zero and one; values towards 1 indicate similarity. As topic proportions per dissertation title are vectors of the same length, the cosine similarity allows a comparison of the topic distribution between two documents. For the two exemplary dissertations I obtain a cosine similarity of .14.

In the next step, I calculate the cosine similarity between all topic-document distribution pairs (see dataset: Rehs, 2020b). This means the topic distribution of title 1 is compared to title 2, title 3 and so on. I drop duplicate observations (e.g., when cosine similarity between 2 and 3 is the same as between 3 and 2). Since I know for every observation of the cosine similarity where both dissertations titles were written, I can employ this information in creating variables that can be attached to these similarity pairs (see Table 1 for an illustration of my dataset). I create a dummy *diff_part* that describes whether the two underlying dissertations for every similarity score are from different parts of Germany. The dummy variable *post95* indicates whether a dissertation was written after 1995.

Finally, I add university dummies to address differences in similarity scores arising at the university level. As the similarity score is calculated between two dissertations that were most often written at different universities, I consequently add dummies for both. The dummy *sameuni* indicates whether both titles in a pair are from the same university. In order to ease the interpretation of my dataset, I require both titles to be from the same year.

Table 1. Dataset structure

Title 1	Title 2	*Cosine sim.*	University 1 = *Univ_k*	University 2 =*Univ_k*	*same uni*	*same year*	*diff_ part*	*same part*	*post 95*
Die Institutionen des…	Catergory Management …	0.14	1	0	0	0	1	0	1
…	…	…	…	…	…	…	…	…	…

Source: Own data.

In the next step, I build different subsets of the data in order to address the peculiarities of my case study. For dissertations in chemistry, I build five data subsets: dissertations written before 1990 in both eastern and western Germany, dissertations written after 1990 in both Germanies, all dissertations written in western Germany and all dissertations written in eastern Germany for the time period studied. For economics and business administration, I proceed accordingly.

These subsets allow me to apply a linear regression framework, where the similarity score for each pair of dissertations is the dependent variable, and *diff_part, post95, same_uni* and the university dummies are the independent variables. This approach aims to aggregate the cosine similarities in order to demonstrate relationships between the underlying groups of dissertation titles from eastern and western Germany and different periods. The regression formula is given by (1).

$$Cosine_{\ i,j} = \beta_0 + \ \beta_1 Diffpart_{i,j} + \beta_2 post95 + \beta_3 Diffpart * \beta_4 post95 + \beta_5 sameuni + \beta_n X_{i,k} + \beta_n X_{j,k} + \varepsilon_i \tag{1}$$

j = dissertation 1 in pair

i = dissertation 2 in pair

k = university

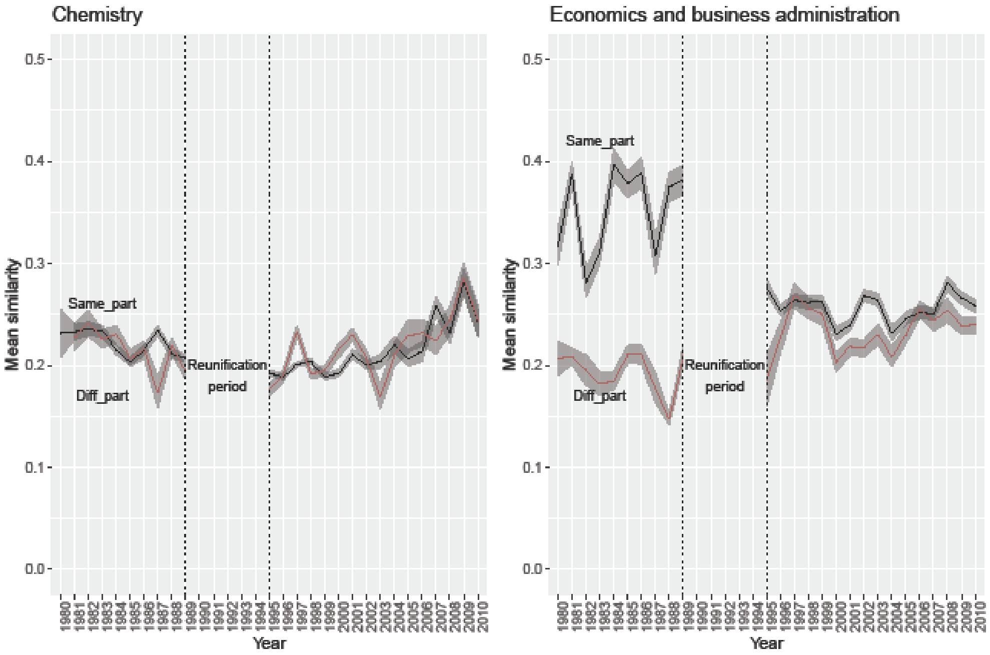


Figure 5. Yearly mean cosine similarity between topic distributions in dissertation pairs

Source: Own depiction.

Table 2 and Figure 5 show descriptive results for the cosine similarity by certain variables. In Figure 4 I depict the mean similarity (with 95% conf. interval) of *diff_par t* = 0 and *diff_part* = 1. The graph shows that the convergence in economics and business administration seems to have happened very quickly. In chemistry there was no convergence, as the average dissertation pair similarities by regional origins were never very different in my period of investigation.

Regarding the mean similarity of *diff_part* = 1 in Table 2, I observe in economics and business administration a significant increase from before to after 1990 and in chemistry a slight decrease. Chemistry topics were therefore, on average, more similar between eastern and western before the reunification than after the reunification. Nevertheless, the visual pattern of the mean by single years presented in Figure 5 does not obviously support this finding. The results for *sameuni* in Table 2 also deliver interesting insights. Within a university, topics in both disciplines were considerably more similar than topics in different universities.

Table 2. Cosine similarities by subgroups

Economics and business							
	n	min	1st	Median	mean	3rd	Max
cosine similarity all	62,586	0.0068	0.1462	0.2265	0.2578	0.3318	1
Eastern = 1	10,458	0.0141	0.1460	0.2309	0.2843	0.3682	0.9960
Western = 1	52,128	0.0068	0.1462	0.2258	0.2524	0.3266	1
diff_part = 1	18,607	0.0070	0.1235	0.1921	0.2213	0.2857	0.9951
diff_part = 0	43,979	0.0068	0.1589	0.2421	0.2732	0.3503	1
Sameuni = 1	1,308	0.0132	0.2377	0.3426	0.3844	0.4927	1
post95=1 and *diff_part* = 1	56,504	0.0068	0.1527	0.2342	0.2647	0.3398	1
post95=0 and *diff_part* = 1	6,082	0.0121	0.1052	0.1631	0.1933	0.2406	0.8582
post95 = 0	12,784	0.0121	0.1433	0.2330	0.2819	0.3719	1
post95 = 1	49,802	0.0068	0.1470	0.2252	0.2561	0.3241	0.9974
Chemistry							
	n	min	1st	median	mean	3rd	Max
cosine similarity all	148,647	0.0027	0.0970	0.1691	0.2065	0.2729	1
Eastern = 1	16,739	0.0034	0.1071	0.1769	0.2125	0.2765	0.9932
Western = 1	131,908	0.0027	0.0968	0.1680	0.2057	0.2724	1
diff_part = 1	29,922	0.0034	0.1053	0.1751	0.2100	0.2740	0.9989
diff_part = 0	118,725	0.0027	0.0961	0.1675	0.2056	0.2726	1
same uni = 1	3,744	0.0195	0.1606	0.2610	0.3023	0.3999	1
post95 = 1 and *diff_part* = 1	139,469	0.0027	0.0974	0.1684	0.2060	0.2724	1
post95 = 0 and *diff_part* = 1	9,178	0.0034	0.1070	0.1800	0.2131	0.2800	0.9867
post95 = 0	47,329	0.0028	0.1046	0.1813	0.2164	0.2893	1
post95 = 1	101,318	0.0027	0.0953	0.1639	0.2018	0.2653	0.9998

Note: Similarities of 1 are due to rounding.

Source: Own data.

Table 3 aggregates the chemistry cosine similarities in a linear regression framework. The pre models in both tables show the differences between eastern and western topics before reunification. Both pre models in Table 3 arrive at significantly negative coefficients of the variable *diff_part*. This indicates lower cosine similarity between two chemistry dissertations written in different parts of Germany. Full period model 1 in shows the differences between eastern and western Germany after reunification. The interaction of *diff_part* and *post95* in Table 3 is positive and statistically significant. This indicates increasing similarity between eastern and western German chemistry dissertations after the reunification. However, the effect diminishes after including university dummies and the variable *sameuni*, as shown in full period model 2. The last approach in chemistry concerns the thematic change within eastern or western German dissertations and is shown in models 5, 6, 7 and 8. The results suggest that there is no thematic change from before to after the reunification in eastern German chemistry dissertations. For western German chemistry dissertations, surprisingly, there is a negative change. This means that western German dissertations became more dissimilar while eastern ones didn’t.

Table 3. Chemistry OLS regression

	Dependent variable: Cosine similarity							
	Full period (1)	Full period (2)	Pre (3)	Pre (4)	eastern (5)	eastern (6)	western (7)	western (8)
diff_part	-0.004**	-0.009***	-0.004**	-0.007**				
	(0.002)	(0.002)	(0.002)	(0.004)				
post95	-0.017**	-0.020***			0.009	-0.002	-0.018***	-0.020***
	0.009	(0.001)			(0.008)	(0.010)	(0.001)	(0.001)
*diff_part*post95*	0.013***	0.002						
	(0.002)	(0.002)						
Sameuni		0.098***		0.092***		0.119***		0.096***
		(0.002)		(0.004)		(0.011)		(0.003)
Constant	0.217^{***}	0.266^{***}	0.217^{***}	0.182^{***}	0.237^{***}	0.228^{***}	0.217^{***}	0.260^{***}
	(0.001)	(0.012)	(0.001)	(0.030)	(0.007)	(0.024)	(0.001)	(0.013)
Uni dummies	NO	YES	NO	YES	NO	YES	NO	YES
Observations	148,647	148,647	47,329	47,329	2,029	2,029	116,696	116,696
R^{2-}	0.002	0.029	0.0001	0.041	0.001	0.080	0.003	0.029
Adjusted R^2	0.002	0.028	0.0001	0.039	0.0001	0.066	0.003	0.028

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

Source: Own data.

For economics and business administration, the results are presented in Table 4. Here, regression results of models 3 and 4 show a large decrease in cosine similarities for topic distributions of dissertation titles written in different parts of Germany before the reunification. Models 5 and 6 of Table 4 present the regression results of topics in economics and business administration before and after reunification in eastern Germany. In both models I reach significance and a substantial effect of -.27 and -.22, respectively. The last approach, which is presented in full model 1 and 2 of Table 4 shows the similarity between eastern and western after the reunification. The positive interaction term of *diff_part* and *post95* in full model 2 suggests that there is an increasing similarity, and the coefficient sizes of cosine similarity indicate that the effects observed in economics and business administration are of relevant magnitude. This could have been expected, as the discipline underwent a drastic reorientation after German reunification. In chemistry, the statistically significant effects are much smaller. Chemistry may serve as an example of how even minor changes can be detected by my approach.

Table 4. Economics and business administration OLS regression

	Dependent variable:							
	Cosine Similarity							
	Full period	Full period	Pre	Pre	eastern	eastern	western	western
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
diff_part	-0.169***	-0.187***	-0.169***	-0.202***				
	(0.002)	(0.003)	(0.003)	(0.003)				
post95	-0.105***	-0.069***			-0.272***	-0.222***	-0.042***	-0.041***
	(0.002)	(0.001)			(0.007)	(0.009)	(0.002)	(0.002)
*diff_part*post95*	0.147***	0.129***						
	(0.003)	(0.004)						
Sameuni		0.086***		0.086***		0.105***		0.074***
		(0.004)		(0.008)		(0.007)		(0.005)
Constant	0.3623***	0.358***	0.362***	0.383***	0.520***	0.546***	0.299***	0.323***
	(0.002)	(0.009)	(0.002)	(0.017)	(0.0004)	(0.042)	(0.002)	(0.009)
Uni dummies	NO	YES	NO	YES	NO	YES	NO	YES
Observations	62,586	62,586	12,784	12,784	3,104	3,104	40,875	40,875
R^{2-}	0.069	0.123	0.202	0.416	0.332	0.419	0.008	0.037
Adjusted R^2	0.069	0.121	0.202	0.409	0.332	0.409	0.008	0.034

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

Source: Own data.

2.6 Discussion and Conclusion

In this paper I have shown how scientists' research problem choices can be detected with a machine learning approach. For this purpose, I investigated the thematic change after an unexpected political transition. I used dissertation titles in the disciplines of economics and business administration and chemistry before and after German reunification in eastern and western Germany. I applied structural topic modelling combined with cosine similarity-based regression. I found differences between the two parts of Germany in both disciplines before the reunification. These differences decrease somewhat after the reunification. My results suggest that eastern German dissertation title topics in the field of economics and business administration are significantly more different before reunification than thereafter.

The substantial differences in economics and business administration before the reunification are likely to be related to politics, and are in accordance with the historical circumstances that I described in the chapter "Historical background". Economics and business administration as a discipline was extremely important in the ideological framework of the GDR. The research of economists and business administrators, more so than in other disciplines, had to therefore be vetted and brought in line with socialist ideology. Topics related to capitalism, which were researched in western countries like western Germany, were therefore de facto impossible to research in the GDR.

Regarding my findings after the reunification, I again refer to section "Historical background". As described, massive personnel replacement, as well institutional redirection, took place in eastern German economics and business administration after the reunification. The free chairs were predominately filled with western German economists and business administration scholars (anecdotal evidence). Consequently, the dissertation topics picked by new these scientists would have been very different from the topics of the dismissed eastern German scientists and their predecessors. However, within the long time span I investigate after the reunification (15 years), other factors could have also led to declining differences within economics and business administration.

One potential explanation for the small differences in chemistry research topics between eastern and western Germany before reunification could be the industrial relevance of the discipline, which motivated the GDR government to directly and indirectly interfere with the topic choices of eastern German scientists. The prime example of direct influence was the official yearly plans for science and technology, which forced chemistry to meet the industry demands of eastern Germany (Gruhn & Lauterbach, 1977). The economic and societal restrictions in the GDR also had an influence on topic choices and therefore on the topics and results that I can observe. Collaboration, for instance, was for eastern German scientists almost exclusively possible with researchers from other socialist countries (Weingart, Strate & Winterhager, 1991). This prevented thematic spread that could have resulted from collaboration with western German colleagues. The different characteristics of economic uncertainty of the GDR in comparison to western Germany may also have had an indirect impact on scientific topic choices. The academic field in the GDR was, for instance, fully employed at any point in time, albeit with a considerable hidden unemployment rate, as it was socialist state doctrine to employ everyone (Gutmann, 1979). Picking risky research problems was possibly not associated with risky labor market outcomes for eastern German chemists and scientists in general. Nevertheless, the choice of risky topics was contradicted by the aforementioned science and technology plans, which forced eastern German researchers to pick applied topics that met industry demands. Lastly, the small differences between eastern and western Germany in chemistry before the reunification could also be attributed to western German peculiarities.

The method presented and developed in this paper – structural topic modelling and a cosine similarity-based regression approach – are its main contributions, and aimed to detect differences in research topics of eastern and western German scientists before and after German reunification. As demonstrated, this turned out to be successful; my trained model detects reasonable differences in a set of unseen titles. The inclusion of dissertation level variables, like affiliation to single universities or dissertation year information, in training a topic model can be considered as a decisive advantage of my approach. Research problem choice is dependent on various factors, such as regional and temporal origin of the dissertation. In the topic modelling process, which tries to reconstruct the data-generating process behind the dissertation title, these factors should therefore not be considered constant across all dissertations in the training set (as done by the LDA topic model algorithm).

The incorporation of paired cosine similarities into a regression approach has, to my knowledge, never been used before and is therefore a methodical innovation of my paper. The regression framework presented in this paper provides not only an easily interpretable aggregation of the cosine similarities, but also a way to test hypothesis. In this sense, other contexts and datasets in scientometric research could be addressed by my approach, which may deliver new perspectives on thematic and, therefore, scientific change in general.

From the visual inspection of the most probable words in economics and business administration, I conclude that my model was able to discover meaningful relationships. The usage of short documents – in my case, dissertation titles – did not turn out to be a problem. In the application to the unseen documents, which were the basis for validation, my algorithm worked well. As topic modelling does not aim to label the detected topics, I can sometimes only guess what the found differences and their underlying topics most likely refer to. This is a major disadvantage of any sort of topic modelling. The foundation of this problem arises from language as a dynamic, complex and strongly context-related semantic system. Topic models can only find the relations in this system, but not understand and label them accordingly. It is therefore beyond the scope of my paper to find reasonable labels for topics I detected.

The linkage of my data to measures of scientific success and impact could provide interesting further research questions. The topical choices that are associated with academic rewards for PhD students, for example, could be investigated. Also, my method could be promising for the investigation of other types of documents; abstracts and scientific articles may contain document-level information which could shift topic proportions in the same way as the variables in my paper. Because of increased document length in these cases, the topic model algorithms would exponentially increase calculation time, but gain statistical properties and topic quality. Therefore, my method of structural topic modelling combined with a cosine similarity-based regression framework offers potential, generally, for applications in scientometrics and higher education research.

3 A supervised machine learning approach to author disambiguation in the Web of Science

3.1 Preface

This chapter builds on an early version of a paper with the same title, published in the *Journal of Informetrics* as: Rehs, A. (2021). A supervised machine learning approach to author disambiguation in the Web of Science. *Journal of Informetrics*, *15*(3), https://doi.org/10.1016/j.joi.2021.101166. This version is the revised one from December 2020.

3.2 Introduction

Author-level scientometric indicators have become an important object of research and help to understand the fundamentals of the scientific system. They are, for instance, crucial in investigating scientist productivity (Hirsch, 2005), scientific collaboration (Glänzel & Schubert, 2006) or mobility patterns in academia. Similarly, author-level scientometric indicators are being used more often to evaluate individual researchers and the scientific system in general, thereby providing a vital basis for the decision-making of university administrators and policy makers (Abbott et al., 2010; Hicks, 2012). The databases upon which these indicators are calculated should therefore be of adequate quality and represent the actual author publications as well as possible (D'Angelo & van Eck, 2020; Weingart, 2005).

One of the key challenges to the validity of indicators is author name ambiguity, also called the namesake problem (Shin, Kim, Choi & Kim, 2014). The namesake problem belongs to the universal problem of entity resolution and concerns problems of whether references to entities belong to the same entity or different entities (Talburt, 2011). One issue in author name ambiguity is called the block problem and typically occurs with common names like „Zhang, Ying“, which appeared in 6,124 publications in the 2017 version of the publication database WoS. Two papers written by „Zhang, Ying“ might be written by either the same „Zhang, Ying“ or two different people, each with the name „Zhang, Ying“. Blocks are even more problematic when using abbreviations and initials in author names. „Zhang, Ying“ becomes „Zhang, Y.“, which is not distinguishable from „Zhang, Yong“ or other names starting with the letter Y. This abbreviation practice increases the number of publications by „Zhang, Ying“ to be disambiguated in the WoS from 6,124 to 256,554. Synonyms are also a problem in name ambiguities. Synonym problems occur when the same author appears under different names, such as the German name „Müller“, which is sometimes written as „Muller“ or „Mueller“. The manual disambiguation of such block and synonym cases in small datasets is usually the best way to account for all problems. However, when it comes to large datasets, such as for „Zhang, Ying“, manual disambiguation is impossible in a reasonable amount of time and may include errors as well (Shin et al., 2014). Accordingly, adequate publication databases that represent all publications belonging to an author cannot be built, and the validity of subsequent scientometric indicators is endangered.

A substantial amount of literature therefore deals with computational methods to solve this problem. With the advancements in methodology and the increase in computational capacities, machine learning methods came into the focus of research. Machine learning methods can detect complex relationships in publication data and can thereby disambiguate author names and their publications. I contribute to this literature by developing a supervised machine learning approach with elements of graph-based methods that can handle missing data and rapidly disambiguates large author sets. In this way, I develop the methodological foundation to disambiguate full publication platforms like the WoS in a reasonable amount of time. I also combine traditional features with a new, topic model-based, thematic feature. A strength of my approach is that I provide a detailed feature assessment that identifies relevant paper attributes. This evaluation may help future disambiguation approaches in selecting features.

The remainder of the paper is as follows: I start by investigating the characteristics of previously applied disambiguation approaches and identify machine learning-based approaches as the most promising technique. In my subsequent data analysis, I prepare over 1.2 million paper pairs with authors that share the same last name-first initial combination, but who are distinguishable by their Researcher ID (Enserink, 2009). Using the Researcher ID authors can assign papers to their user account in the WoS. The Researcher ID has previously been shown to be useful in providing true authorship information by Tekles and Bornmann (2019) and is amply available in the WoS database. I proceed by comparing the retrieved paper pairs by their attributes and reviewing a vast set of common and novel author and paper characteristics. I emphasize the further development of name-based features that were introduced by Torvik, Weeber, Swanson, and Smalheiser (2005) and include overall first and last name frequencies, first name frequencies in a block and several other measures. I also find that the complete WoS includes up to 95% missing first names before 2006 and up to 25% missing first names after 2006. Therefore, I randomly insert missing first names and second initials to make my algorithm robust to data changes of the first name as the most decisive feature.

In the next section, I describe my methodological approach using the machine learning algorithms of random forest (Louppe, Al-Natsheh, Susik, & Maguire, 2016) and logistic regression (J. Kim & Kim, 2018), which have been successfully used in author disambiguation. The results on the more than 53,000 pairwise paper comparisons of the test set follow, which yield an F1 score of .82 in the random forest and .75 in the logistic regression. To aggregate the pairwise predictions into author clusters — in other words, all papers belonging to a single author — I apply the infomap algorithm (Rosvall & Bergstrom, 2007) which is described in the section „Graph-based author community detection“. The clustering results suggest that different authors rarely appear in the same cluster and that same authors are rarely split into different clusters. The subsection 3.6.2 addresses the external validation of my approach. I use the large block „Muller, M.“ with 11,665 papers and analyze the clustering results for the subset of papers containing Researcher IDs. The clustering results are reasonable and suggest the large-scale application potential of my approach. Finally, I discuss the results and conclude my approach.

3.3 Characteristics of disambiguation approaches

3.3.1 Disambugiation methods

The numerous disambiguation approaches in the literature differ in methods, data, scope and features used. Hussain and Asghar (2017) provide a rich survey of these approaches and distinguish on the method-level non-machine learning-based approaches and machine learning-based approaches. Machine learning techniques break down into supervised techniques, unsupervised techniques, and semi-supervised techniques. Non-machine learning-based techniques are split into graph-based and heuristic-based methods. In addition to Hussain and Asghar's categories, I include the third category probabilistic approaches. In the following paragraphs, I try to categorize the literature according to these techniques, and I also discuss the features used.

Graph-based methods use papers and their attributes as node and edge representations to detect author communities in a graph. As a result, papers with the same author block, but different real authors can be separated into connected graph-communities. Graph-based methods have been applied extensively — for example, by Fan et al. (2011), On, Lee, and Lee (2012) and Shin et al. (2014) — are visually interpretable and have been shown to disambiguate authors accurately (see Fan et al., 2011).

Heuristic approaches (e.g., De Carvalho, Ferreira, Laender & Gonçalves, 2011) use paper attributes to construct simple rules with which authors and their papers can be distinguished. For example, all papers from „Zhang, Y.“ that share at least one attribute, such as common coauthors or institution names, are assigned to be from the same „Zhang, Y.“ in heuristic approaches.

Probabilistic approaches (e.g., Tang, Fong, Wang & Zhang, 2011; Torvik & Smalheiser, 2009; Torvik, Weeber, Swanson & Smalheiser, 2005) try to set up a linkage function that gives the probability that two articles belong to the same author. Torvik et al. (2005), for example, used a reference set of true and false matches of author articles with the same name and calculated the matching probability between two articles based on several paper characteristics.

Machine learning approaches are characterized by their different requirement levels for training data. Supervised and semi-supervised methods (K. Kim, Rohatgi & Giles, 2019;

Louppe et al., 2016) require external labeling of training data regarding whether the papers in question are from the same author. With this information, the methods learn how paper attributes refer to the same or different authors. Unsupervised methods (D'Angelo & van Eck, 2020; Wu et al., 2014; Caron & van Eck, 2014) don't require labeled training data; instead, they try to find patterns in the data by themselves but are therefore more computationally expensive.

Machine learning approaches can detect complex relationships in paper attributes and adequately handle missing data. Because of their good predictive performance, they generally achieve high precision and recall rates and are therefore superior to graph-based and heuristic-based methods. In comparison to probabilistic approaches, machine learning approaches are more comfortable to implement and allow for low-cost comparison and tuning of different algorithms. For predictive power, machine learning algorithms can implicitly discover the same statistical characteristics of papers and their attributes as probabilistic approaches do.

3.3.2 Features for disambiguation

Bibliographic characteristics

All the methods or databases described above require some paper or author characteristics with which the disambiguation is performed. Bibliographic information is most frequently used, as it presents the most important characteristics of a paper and is almost always available in publication platforms, such as the WoS and SCOPUS. This information typically includes journal title and issue, coauthors, keywords, abstract, publication title, subject classifications and year of publication. Bibliographic characteristics are used to either directly perform author disambiguation, or generate other disambiguation measures Treeratpituk and Giles (2009), for example, determined journal languages in order to check if two articles of the same author block were published in the same language.

Citation characteristics

The second most frequently used type of characteristic is citation data (e.g., Louppe et al., 2016; Onodera et al., 2011; Torvik et al., 2005). Onodera et al. (2011) evaluated citation data, especially self-citation data, as highly effective for author disambiguation. However, citation-based features require adequate time to gather citations, and that the paper under examination be recognized by the scientific community. Citation-based features do not scale very well for large publication platforms since the citation databases upon which the platforms are based go well beyond 9 digits, such as the more than 600 million citation relationships in the 2017 version of the WoS.

Country and institution-based characteristics

Country, institution and department name(s) are used as characteristics in numerous studies and are similar to bibliographic information in that they are often available for at least one author of a paper. Torvik et al. (2005) reported that institutional names are extraordinarily good predictors for disambiguating author names, but require tedious preprocessing to check for stopwords and abbreviations (Rimmert, Schwechheimer & Winterhager, 2017). International or interinstitutional researcher mobility is a key challenge for country and institution-based measures, as the same author will change addresses multiple times in a scientific career. Therefore, country, institution, and department name(s) may be used only in conjunction with invariable or inert features, such as the first name or thematic characteristics.

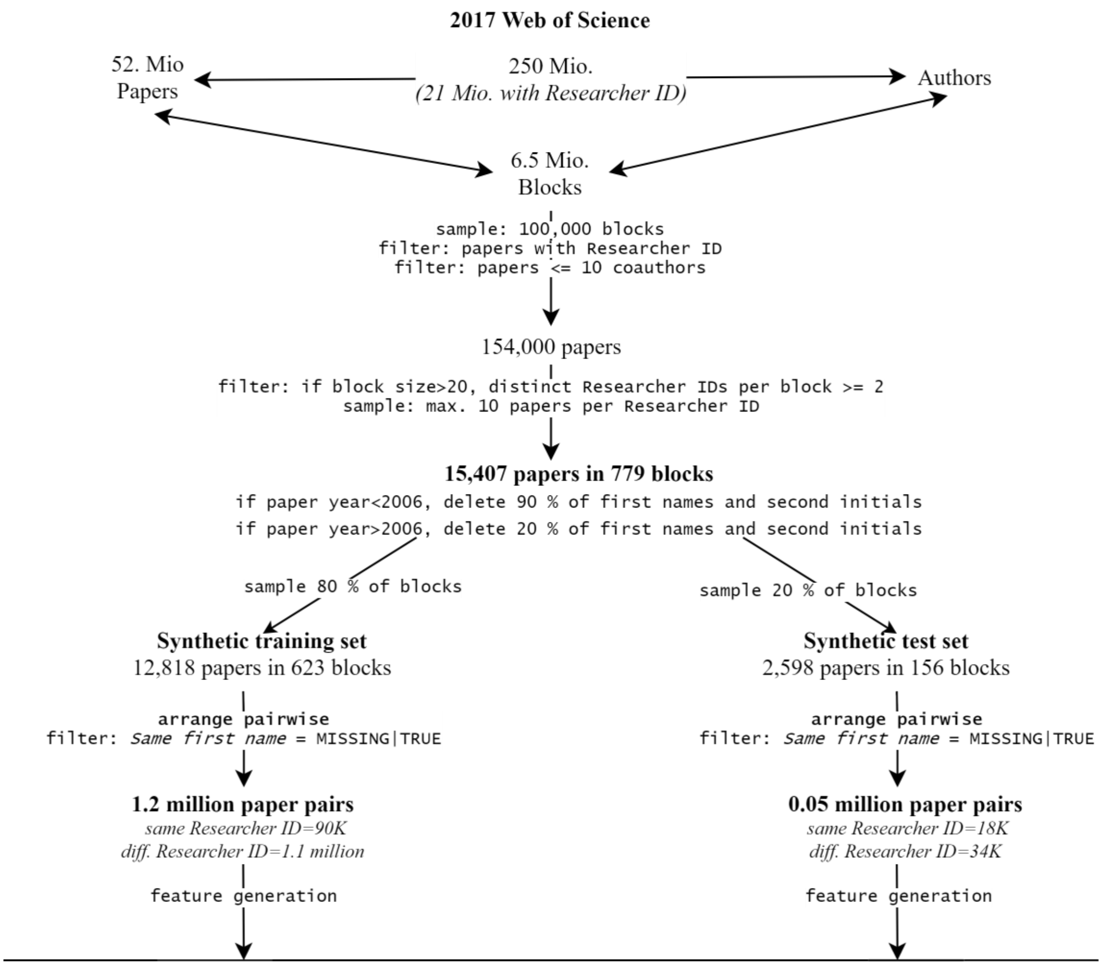
Name-based characteristics

Finally, features can be generated based on the author's name, frequency, and related statistical properties. Torvik et al. (2005) present a probabilistic model based on PubMed data. They define a similarity profile between a pair of articles in the same block, based on name attributes (middle initial, and suffix) and other paper characteristics. The similarity profile distribution is then computed from gold-standard reference sets. This reference set consists of pairs of articles that almost exclusively contain author matches versus nonmatches. In analyzing the reference set, one would, for instance, conclude that a rare similarity profile in a given block (e.g., {*non-frequent block, same suffix, same journal, same keyword}) represents a high likelihood of matching between two articles. In turn, {*frequent-block, same keyword, same journal, different suffix*} may result in a frequent profile and low matching probability. Louppe et al. (2016) and Strotmann and Zhao (2012) used another related statistical property of names in this regard to show how the determination of ethnicity and related statistical properties of a name can improve name disambiguation.

3.4 Data

3.4.1 Data sources, sampling strategy and preprocessing

In the following, I want to present my data sources and methodology. My disambiguation approach is based on pairs of papers in the same surname-initial block. For these blocks I know which paper-pairs belong to the same and different authors. The to be presented machine learning methodology then learns the relation between paper pairs from the same and different authors upon paper characteristics. Figure 6 shows the schema of my data processing.


2017 Web of Science
52. Mio Papers
250 Mio.
(21 Mio. with Researcher ID)
Authors
6.5 Mio. Blocks
sample: 100,000 blocks
filter: papers with Researcher ID
filter: papers <= 10 coauthors
154,000 papers
filter: if block size>20, distinct Researcher IDs per block >= 2
sample: max. 10 papers per Researcher ID
15,407 papers in 779 blocks
if paper year<2006, delete 90 % of first names and second initials
if paper year>2006, delete 20 % of first names and second initials
sample 80 % of blocks
sample 20 % of blocks
Synthetic training set
12,818 papers in 623 blocks
Synthetic test set
2,598 papers in 156 blocks
arrange pairwise
filter: Same first name = MISSING|TRUE
arrange pairwise
filter: Same first name = MISSING|TRUE
1.2 million paper pairs
same Researcher ID=90K
diff. Researcher ID=1.1 million
0.05 million paper pairs
same Researcher ID=18K
diff. Researcher ID=34K
feature generation
feature generation

Block	Paper ID 1	Paper ID 2	*Same Researcher ID*	*Same first name*	...	Training
Wang, Z	123	456	1	1	...	1
Wang, Z	123	789	1	missing	...	0
...	...	...	...	...	...	...

Figure 6. Data processing schematic

Source: Own depiction.

In the first step, I use a 2017 copy of the WoS processed by the Kompetenzzentrum Biblometrie (Rimmert et al., 2017). The WoS has previously been used for disambiguation purposes by D'Angelo and van Eck (2020) and Tekles and Bornmann (2019) and is considered one of the leading publication platforms. The 2017 WoS includes about 52 million publications from 1980-2017 that relate to 178 million author names. Those 178 author names again relate to 6.5 million different blocks. There is a Researcher ID available for 21 million paper-author relationships, which makes them distinguishable from other authors. The Researcher ID builds the basis for generating paper pairs of the same and different authors in a block.

To retrieve blocks and subsequently paper pairs that contain the Researcher ID, I randomly select 100,000 blocks and search for papers with Researcher IDs. In this step, I only consider papers with ten or fewer coauthors because of the computational power required and the presumably different author-characteristics of papers with more than ten coauthors. I also add restrictions on the number of distinct Researcher IDs in a block to address the unknown number of true authors in a block set. My thought is that very infrequent blocks that have a Researcher ID available represent only one real author. For example, if a block set contains 500 papers, but only 10 of them have a Researcher ID that refers to one real author, it is not reasonable to say that block sizes of 500 papers generally relate to one author. Machine learning with that data would get the data generating process wrong and may underestimate the number of real authors.

I address this problem by requiring block sets associated with more than 20 papers to have at least two distinct Researcher IDs included. For block sets associated with 20 or fewer papers, one distinct Researcher ID in a block set is sufficient. In this way, I try to mimic the true number of authors in a block set, presuming that a block set size of 21 is a reasonable threshold to indicate block sets that may represent only one real author.

In total, my dataset now includes 154,092 papers and allows me to generate paper pairs. However, my data set is imbalanced considerably towards sets of very frequent blocks and sets of single scientists who have published many papers. In the machine learning approach, this could result in a performance bias in favor of productive authors and high computational costs. I accordingly restrict the number of papers that are written by a single author to a random set of 10 papers, resulting in 15,407 papers.

When splitting paper pairs into the test and training sets, I randomly sample by block, assigning 80% of the blocks to the training set and 20% to the test set. Table 5 depicts the descriptive statistics for the preliminary test and training set in comparison to the full WoS dataset. A remarkable characteristic is the number of missing first names. While my datasets have a negligible number of missing first names, the WoS includes up to 95% missing of first names before 2006 (see Figure 7). As my goal is to provide a robust model that can handle a variety of cases, especially those with missing data, I therefore randomly delete 90% of the first names prior to 2006 in my dataset to mimic the WoS structure as well as possible. After 2006 I include 20% missing first names. I refer to my training and test sets as synthetic test and training sets because of this manipulation.

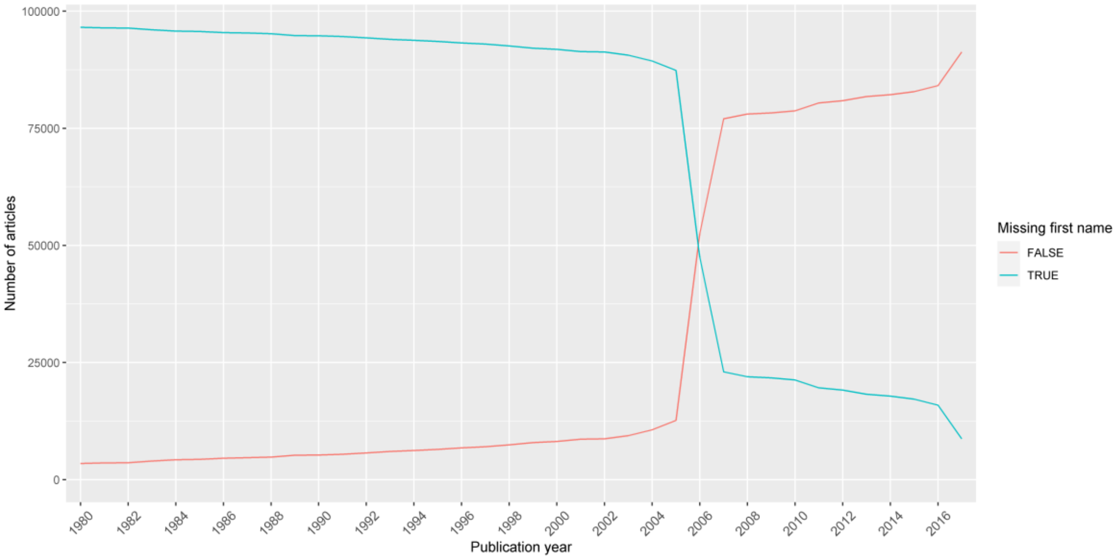


Figure 7. Number of missing first name in 100K paper-author sample of WoS

Source: Own depiction.

Table 5. Characteristics of the test and training set

	Train set	Test set	WoS 2017
Paper characteristics			
Number of distinct authors (Researcher IDs)	1,904	381	> 322,686
Number of papers	12,818	2,589	52,055,209
Mean papers per (Researcher ID)	6.73	6.79	59.50
Mean paper year	2008	2008	2004
Mean number of authors per paper	4.96	5.06	4.74
First name characteristics			
Number of papers < 2006	2,991	707	58%
Number of papers > 2006	8,468	1,585	42%
Number of missing first names	360	31	~50%
Number of missing first names <2 006	99	12	~90%
Number of missing first names < 2006 after generating missing first names	2,846	675	-
Number of missing first names > 2006	213	16	~20%
Number of missing first names > 2006 after generating missing first names	1870	318	-
Block characteristics			
Number of blocks	623	156	6,450,190
Mean number of papers per block	20.57	16.60	35.92
Mean number of Researcher IDs in block	3.05	2.44	Unknown
Paper pair characteristics			
No of paper pairs where same block and same first name = MISSING|TRUE	1,252,075	53,501	-
No of paper pairs where same block and same Researcher ID and same first name = MISSING|TRUE	91,949	18,660	-
No of paper pairs where same block and different Researcher ID and same first name = MISSING|TRUE	1,160,126	34,841	-

Source: Own data and depiction.

In the following, I focus on paper pairs as the basis of my analysis. I only use paper-pairs where the first name is either missing or the same. This constraint reduces the number of comparisons since presumably irrelevant comparisons of papers with different first names are left out. There is a considerable difference between the test and training set in the number of paper pairs that can be generated; the training set with its 1,2 million paper pairs is more than 23 times as large as the test set with its 53,501 pairs. Table 6 depicts the most frequent names in the training and test set. In the random sampling of blocks, I assigned by chance into the training set the very frequent block „Wang, Z.“, which accounts for the dominant number of papers and accordingly paper pairs.

Table 6. Top 5 blocks by frequency in training and test sets

Train set					Test set				
Rank	Block	Distinct *Researcher IDs* in set	No of papers in set	Number of papers in WoS	Rank	Block	Distinct *Researcher IDs* in set	No of papers in set	Number of papers in WoS
1	Wang, z	218	950	157,519	1	Xie, j	26	100	16,936
2	Chen, g	68	309	54,578	2	Shi, h	17	80	13,270
3	Chen, t	43	189	39,443	3	Choi, d	16	70	11,114
…	…				…				
623	Yakushevich, n	1	1	5	156	Voytenko, v	1	1	10

Source: Own data and depiction.

3.4.2 Feature generation

Name-based features

The next major step is the creation of features that provide input to the machine learning procedure. For name-based features, I use the same approach as Torvik et al. (2005), who calculated first name priors that input their probabilistic approach to author disambiguation, and provide as information to the machine learning algorithms the overall block paper counts (*Block set size*), the overall count of the first name stated on paper 1 and paper 2 in a pair (*First name count 1* and *First name count 2),* the second first name initials (*Second initial count set 1* and *Second initial count set 2),*the overall last name count *(Last name count*), the first name count within the block paper set (*First name count group 1* and *First name count group 2*) and several ratios[17] of these features.

First name count 1 and *First name count 2* complement the *Block set size* and provide additional, valuable information by indicating whether a given first name is expected to belong to multiple authors within a block. The first name „John“, for instance, can be found in the WoS 67,041 times, while the first name „Soesoe“ can be found only five times. The probability that two papers are from the same author should be much higher for the infrequent „Soesoe“ than for the frequent „John“. *First name count group 1* and *First name count group 2* allow to analyze this the first name frequency on the block level. If, for instance, in a block set „Doe, J.“ with 50 observations, the first name „John“ accounts for all 50 observations, I would have either a very productive John Doe or, more

[17] The ratios are: *Block set size* / *Last name count = Ratio Block Last, First name 1 count set* / *Last name count = Ratio First Last Group 1, First name 2 count set* / *Last name count = Ratio First Last Group 2.*

likely, two or more different John Does. If „John“ accounts for only one observation, and there are 49 other distinct names of block „Doe, J.“, I can be certain that there is only one John Doe. *First name count group 1* and *First name count group 2* account for this and provide the first name's frequency in the block.

Country and institution-based characteristics

My second major group of features comes from regional and institutional information retrieved from papers. The WoS processes the author's address string and extracts the country. I use this country information in the feature *Same country*. This feature might be especially helpful with „international“ blocks, such as „Muller, M.“, which frequently occurs in the US and Germany. I expect that the dominant share of block paper pairs can be found within a single country and face two challenges. The first problem is that the address string is missing in 28% of my papers because the block author is not the corresponding author of the paper, as only those were mandated to fill in their address string on the paper before 2008 (Liu, Hu & Tang, 2018). The second challenge consists of little deviations in the address strings and institutional mobility of researchers. This peculiarity is addressed by using the Jaccard string similarity metric between two address strings (*Institution name similarity)*.

Thematic features: Topic modeling

One major methodological innovation of my author disambiguation approach is the exploitation of latent thematic information in paper titles and abstracts. I use topic modeling, which is a group of probabilistic methods used to discover the latent semantic structures (topics) in text collections. In topic modeling, documents, such as titles or abstracts, are considered as mixtures over *K* latent topics, where each topic is again considered to be a distribution over all the words that exist in the collection of documents. I use a correlated topic model as basis (Blei & Lafferty, 2007) and 50,000 random abstracts from WoS to train a model with 89 topics. The next step is the feature generation with this correlated topic model. I apply the topic model to both abstracts and titles in a block and retrieve two topic distributions for each abstract or title. I apply cosine similarity to compare all distribution pairs. The cosine similarity measures the angle between two vectors (the topic distributions of the two abstracts or titles) projected in a multi-dimensional space. Values towards 1 indicate similarity, which allows me to see whether two abstracts or titles are semantically similar by a single number. The cosine similarity for paper pair title distributions is shown by the feature *Thematic similarity title* and for abstracts by *Thematic similarity abstract*. To complement these measures, I additionally calculate the Jaccard distance of the titles in a paper pair (*Jaccard distance title*).

Thematic features: Classifications and keywords

I also exploit thematic information retrieved from subject classifications and self-assigned keywords in the WoS. The WoS usually makes journal-level classifications, which include five broad categories and 252 disciplines. I use the 252 disciplines to create the feature *Same classification*, which checks if two block papers have at least one classification in common. To generate a feature from keywords, I use only author-assigned keywords, as I assume they are of higher quality than automatically generated keywords. I deleted very frequent keywords (count > 100.000 in the WoS) from the list of possible keywords because I assume, they don't add to the performance of the feature. Finally, the feature *Same keyword* checks if two papers have at least one common keyword.

Bibliographic features

My last set of features concerns bibliographic characteristics of the block paper pairs. The feature *Diff. in number of coauthors* shows differences in coauthorship counts and addresses individual publication behavior or disciplinary differences in coauthorship counts. *Diff. in publication year* should help to disambiguate block authors who are published in different periods of time. *Pubyear 1* and *Pubyear 2* account for yearly effects, such as the generally lower number of papers and subsequent matching probability in past years. Finally, *Same first name* compares whether two papers share the same first name or whether the information is missing in one of the papers.

Table 7 presents statistics for the synthetic test and training set and is differentiated by authors who have the same block and are either the same or different authors according to the Researcher ID. Variation between these two groups is necessary for the random forest and the logistic regression to find meaningful rules with which the same and different authors can be distinguished. For better arrangement, I separated numeric and factorial features. For all features, I find variation between the two groups of same authors and different authors. My features are, therefore, generally relevant and should be used in the machine learning procedure.

Table 7. Descriptive statistics features synthetic test and training set

	Synthetic Train Set (1,252,075 paper pairs)						Synthetic Test Set (53,501 paper pairs)					
	Same Researcher ID pairs (91,949 pairs)			Different Researcher ID pairs (1,160,126 pairs)			Same Researcher ID pairs (18,660 pairs)			Different Researcher ID pairs (34,841 pairs)		
	Miss.	Mean	Std. dev.	Miss.	Mean	Std. dev.	Miss.	Mean	Std. dev.	Miss.	Mean	Std. dev.
<u>*Numeric features*</u>												
Jaccard distance title		0.76	0.31		0.94	0.03		0.76	0.32		0.94	0.03
Thematic similarity title	171	0.30	0.36		0.12	0.17	156	0.31	0.37	47	0.14	0.21
Thematic similarity abstract	17,780	0.28	0.36	192,519	0.12	0.18	3,419	0.31	0.38	6,327	0.13	0.21
Institution name similarity		0.39	0.42		0.07	0.08		0.37	0.41		0.05	0.09
Diff. in publication year		3.24	4.34		5.61	4.80		3.18	4.26		7.10	5.77
Publication Year 1		2007	6.45		2008	5.74		2007	6.48		2005	6.77
Diff. in author position		1.33	1.56		1.76	1.66		1.30	1.54		1.79	1.67
Diff. in number of coauthors		5.02	2.26		5.01	2.20		4.98	2.26		4.91	2.19
Diff in no of citations		10.22	43.23		9.39	22.13		9.28	30.11		8.79	29.95
First name count 1		20,814	42,852		11,388	12,317		20,889	45,460		13,379	25,063
First name 1 count set		2,227	14,641		6,889	29,383		441	1873		955	2,926
Second initial 1 count set		8,977	28,755		36,511	57,593		1,792	3,406		3,481	56,306
Last name count		472,497	808,817		1,866,321	727,629		59,306	79,688		127,093	93,408
Block set size		28,276	53,081		123,568	57,499		3,995	5,007		8,489	6,183
Ratio Block Last		0.12	0.15		0.06	0.03		0.16	0.19		0.10	0.12
Ratio First Last group 1		43.99	957.15		1.89	4.97		5.74	49.99		0.12	0.32
...												
	Different	Missing	Same	Different	Missing	Same	Different	Missing	Same	Different	Missing	Same
<u>*Factorial features*</u>												
Same first name		42,581	49,368		1,119,544	40,582		9,115	9,545		32,913	1,928
Second initial	261	67,854	23,834	201,707	941,260	17,159		14,783	3,877	2,442	31,866	533
Same country	5,140	37,632	49,177	410,418	4,93,555	256,353	1,014	8,208	9,438	8,746	21,039	5,056
Same keyword		84,745	7,204		1,160,124	2		172,38	1,422		34,841	0
Same classification		63,771	28,178		1,147,807	12,319		12,969	5,691		34,152	689
Same coauthors		77,881	14,068		1,160,100	26		15,832	2,828		34,840	1
...												

Source: Own data and depiction.

3.5 Machine learning approach to author disambiguation

3.5.1 Random forest

In my research, I used random forest and logistic regression machine learning algorithms to disambiguate author names. In a pre-test I also tested a 2-layer feed-forward neural network, which delivered comparable results, but does not allow feature assessment.

Breiman (2001) proposed the use of random forests as a method for classification and regression tasks. A random forest consists of an ensemble of decision trees, each of which uses a random sample of the data and features to find a given tree's best split. By introducing random sampling of data and features, random forests generally avoid overfitting the training data. Random forests can be applied in both supervised and unsupervised approaches. One of the major advantages of random forests is their ability to evaluate feature performance.

I use a supervised implementation of a random forest model with 200 trees and tune the hyperparameter *mtry*, which is the number of random features tried at each split of a tree. I tune *mtry* by incremental increases and choose the value with the best out-of-bag error. The out-of-bag error exploits that not all trees use every observation in training. In this way, the out-of-bag error measures the mean prediction error for all samples not included in the predictions. The out-of-bag error, therefore, allows validation of the performance without using an external test set. I obtain the best result with a *mtry* value of 10. The „winning“ class for an observation in a random forest is the one with the maximum ratio of the proportion of votes to cutoff. To find the best trade-off between precision and recall, I also tuned this cutoff value and find 0.8 proportion of trees that vote for two papers in pair of being from the same author as the best cutoff value.

Finally, random forest performance can suffer from large class imbalances in the training data (Kim & Kim, 2018). Here, inference of the majority class is easier since there are more examples of how best to split a decision tree. In my case, the negative class (same block and different person) is prevalent and accordingly results in a biased model. To address this issue, I sample in each tree as many positive training examples as negative examples. Because I could build only trees that take at maximum 107,002 observations (two times 53,501 Same Researcher IDs in the synthetic training set) into their inference procedure, which is less than 10% of the total dataset, I may undersample some small but important sub-feature classes. To make my model robust, I would need in each tree observations that refer to small, medium, and large blocks. Large blocks with more than 10,000 papers, however, account for 95% of the observations; medium sets from 500 to 9,999 papers, for 3%; and small sets of fewer than 500 papers, for the remaining 2%. Accordingly, only a very small number of observations in each sample can be drawn from medium and small paper sets. Therefore, the tree would most likely not learn meaningful rules related to those sets. Because this repeats in every tree, my random forest model would generally underperform on small and medium sets. I address this issue by implementing a stratified sampling strategy for each decision tree. I establish five block-size groups in this strategy and sample 6,000 positive and 6,000 negative paper-pairs from each group.[18]

3.5.2 Logistic regression

In addition to the random forest model[19], I also train a logistic regression classifier with most of the same features. Logistic regression is based on different concepts and therefore might arrive at different conclusions on feature relevance, and also different predictions. Unlike random forest, logistic regression estimates the explicit probability that two papers are from the same author. It uses a logistic function to find the model parameters via maximum likelihood estimation. For multicollinearity issues in this estimation, I cannot use highly correlated variables, such as the *Last name count* and *Block set size* in the same model. I identify these problematic correlations by running the model and then calculating each variable's variance inflation factor. A variance inflation factor greater than five generally indicates collinearity. I remove these variables and estimate the model again. The full model can be found in Appendix B along with the estimation of average marginal effects.

3.5.3 Machine learning results

Table 8 presents the results of the random forest and the logistic regression. The random forest algorithm delivers precision and recall rates of >.75 values, beating the logistic regression in precision but not in recall. Table 9 depicts the importance of the features, as shown by the mean decrease in accuracy for the predicted class *Same author* = T and *Same author* = F. In other words, how the number of correct classifications decreases for the given class when the feature is excluded. The column Mean Decrease Accuracy summarizes this measure for both classes. The final column, Mean Decrease in Gini, indicates the average of a given variable's total decrease in node impurity, weighted by the proportion of sampled observations reaching that node in each decision tree. This measures how important a variable is for estimating whether two papers are from the same author for all of the trees that make up the forest. A higher Mean Decrease in Gini indicates higher variable importance.

Same first name is the most powerful feature for the Mean Decrease in Gini. For *Mean Decrease Accuracy where Same author* = F and *Same author* = T, the *Jaccard distance*

[18] For the smallest group of block sets with less than 100 papers, I can sample only 500 positive and negative observations.

[19] Other machine learning models, such as neural nets, Naïve Bayes classifiers and SVMs were also considered, but they do not fulfil at least two criteria of: predictive power, possibility of feature assessment and training time and scalability (prediction time).

title and *Same classification* are the most important features. The performance of the *Jaccard distance title* can be explained by its universal applicability, as every paper must have a title. Although the *Jaccard distance title* was intended to complement the cosine similarity of the title and abstract, it outperforms both measures substantially. However, the insignificance of the cosine similarity of title and abstract and other features should not be overemphasized, as there might often be a high correlation between the randomly drawn features in each tree. Random forests can capture this correlation and therefore use only the most powerful features when splitting a node. Highly correlated variables appear irrelevant although they might be only slightly worse predictors.

It is noteworthy that *Same coauthors* and *Same keyword* do not play a role, and the same keyword even impacts average accuracy negatively, perhaps because of my stratified sampling strategy. Since there are only 7,206 observations with the *Same keyword* and 14,094 with the *Same co-author*, it is unlikely that enough of those observations are sampled in each tree, and subsequently the trees cannot build meaningful splitting rules.

The logistic regression significantly outperforms the random forest in identifying true negative cases and false-positive cases. Assessing the marginal effects in Appendix B, I again can observe that the first name plays a major role in determining whether two papers are from the same author. As indicated by the average marginal effects (sample) column of Appendix B, the average effect of the *Same first name* on matching probability is about 8 pp. Other relevant features are *Jaccard distance title* (-37 pp), *Same second initial* = T (10 pp) and *Institution name similarity* (13 pp). Unlike with the random forest, the feature *Same coauthors* plays a role in the logistic regression model, and its effect is about 12 pp. This is likely because the logistic regression utilizes all observations where *Same coauthors* = T when estimating the model. As explained above, random forests may suffer from sampling not sufficiently enough observations where *Same coauthors* = T in each tree.

Table 8. Pairwise prediction results synthetic test set

	Random Forest	Logistic regression
Paper pairs	53,501	53,501
True negative	33,531	34,233
False negative	4,664	7,047
False positive	1,310	806
True positive	13,996	11,613
Pairwise precision	0.91	0.95
Pairwise recall	0.75	0.62
Pairwise F1	0.82	0.75
Classification threshold	0.80	0.80

Note: Random forest parameters: *mtry* = 11, stratified sampling by block set classes, 200 trees.

Source: Own data and depiction.

Table 9. Feature importance random forest

	Mean decrease accuracy *Same author=T*	Mean decrease accuracy *Same author=F*	Mean Decrease Accuracy	Mean decrease Gini
Same first name	26.16	37.29	38.16	8130.88
Institution string distance	70.59	62.38	73.30	7266.78
Jaccard distance title	192.54	121.09	155.86	7100.02
Ratio First Last group 1	55.18	28.77	58.80	3281.55
Same classification	138.81	102.91	129.61	2494.7
First name count group 1	25.02	18.92	20.94	2342.46
Ratio First Last group 2	44.91	32.41	48.27	2130.46
Same second initial	89.79	26.24	38.00	1958.98
First name count group 2	69.88	39.7	45.05	1478.85
Diff. in no of citations	93.75	33.83	38.47	1333.61
Diff. in publication year	71.52	35.12	39.38	1327.15
Publication Year 2	63.92	28.26	33.49	1224.83
Second initial count set 1	42.04	33.46	40.12	1202.55
Ratio Block Last	39.17	27.96	29.64	1193.52
Block set size	23.10	18.86	21.16	1129.82
Last name count	23.04	20.88	22.59	1096.66
Second initial count set 2	37.73	18.74	22.29	1021.48
Publication Year 1	50.11	21.86	26.06	983.34
First name count 2	16.04	12.94	14.72	965.83
Thematic similarity abstract	63.43	25.07	31.16	953.04
Thematic similarity title	62.51	24.50	30.81	933.60
Same country	28.21	16.92	23.56	857.13
First name count 1	15.17	11.92	13.82	821.16
Diff. in no of coauthors	81.70	27.25	41.36	696.53
Diff. in author position	72.50	28.03	34.83	556.80
Same coauthors	7.55	14.05	14.18	11.25
Same keyword	-3.81	6.75	-0.35	0.76

Note: Ordered by the mean decrease in Gini.

Source: Own data and depiction.

3.6 Graph-based author community detection

3.6.1 Methodology

At this point, I cannot draw author-paper sets from the predicted paper pairs; they must first be aggregated in order to show all papers belonging to an author. This can be done by graph-based community detection, where papers represent nodes and edges, the predicted matches of my machine learning approach. Edges can be weighted by their attributes, such as *Same first name* or *Same second initial*. I determine edge weights by a score that adds one point to an edge weight when both papers have the same first name, country, classification and second initial; makes no change if the first name/country/… information is missing; and subtracts one when the two papers have different first names/countries/… . I also sum up the given values for the *Jaccard distance title*, *Thematic similarity title*, *Thematic similarity abstracts* and *Institutional distance* and add them to the score. Finally, I consider the time difference between two papers in the edge weighting by adding the time difference's reciprocal value.

Figure 8 shows an illustration of the exemplary graph communities. Red lines indicate false predictions and line thickness, the edge weight. Although many cohesive communities represent all true papers from a single author, the small number of erroneously assigned paper pairs connects different communities and therefore misclassifies all papers in both communities. Algorithms that can detect incorrect linkages are referred to as community detection methods and are well described in graph theory (Newman, 2018).

In the following section, I test the infomap algorithm (Rosvall & Bergstrom, 2007). The basic premise of the infomap algorithm is that random walks from a given node are more likely to stay within the same community rather than leave the community. This, however, only holds true when the community is cohesive, when there are many connections within the community and wrong predictions to other communities are uncommon. The included edge weights support the infomap clustering because they indicate which edges are most likely to go for the random walk. In order to provide a reference to the infomap algorithm, I assume that all connected paper pairs belong to the same author. Results for this reference clustering are depicted in Table 10.

To evaluate the clusters, I use the average cluster purity (ACP), average author purity (AAP) and the K-Metric. ACP, AAP and the K-Metric are frequently used cluster metrics in author disambiguation (Kim, 2019). ACP is defined by Equation (1) and evaluates the purity of author clusters on the block level. An ACP of 1 indicates whether the calculated clusters each only contain papers of the same Researcher ID. The APP evaluates the fragmentation of the calculated clusters and analyses whether the papers of a given Researcher ID can be found in different clusters (Ferreira, Veloso, Gonçalves, & Laender, 2014). AAP is defined by Equation (2). An AAP value of 1 indicates no fragmentation of Researcher IDs into different clusters. K-metric is defined as the ACP and the AAP's geometric mean and is defined by Equation (3).

$$ACP = \frac{1}{N} \sum_{i=1}^{e} \sum_{j=1}^{t} \frac{n_{ij}^2}{n_j} \quad (1)$$

$$APP = \frac{1}{N} \sum_{i=1}^{t} \sum_{j=1}^{e} \frac{n_{ij}^2}{n_i} \quad (2)$$

$$K = \sqrt{ACP * AAP} \quad (3)$$

In (1) and (2), N is the number of papers in a block, t is the number of distinct Researcher IDs in a block, e is the number of detected clusters in a block and n_i is the number of papers in detected cluster i. Finally, $n_{ij}$is the total number of papers in detected cluster i, which also belong to the Researcher ID j.

The clustering results shown in Table 10 indicate that after the application of the infomap algorithm, both approaches suffer from an increase in fragmentation. Clusters are more often divided into different sub-clusters. At the same time, the new clusters contain, on average, fewer different Researcher IDs as compared to the reference clustering approach. The K-Metric, which combines both measures, suggests two different patterns. Only the random forest predictions profit from the application of the infomap algorithm. The logistic regression clusters are, on average, of the same quality when using the reference clustering of connected components.

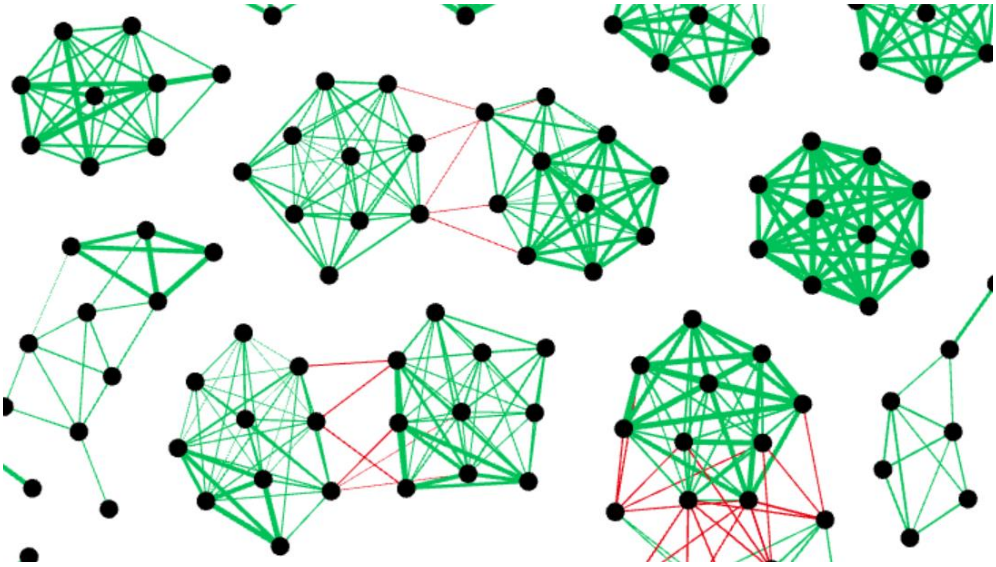


Figure 8. Illustration of true and false predicted paper pairs

Source: Own dataset, graph visualization with Gephi.

Table 10. Cluster algorithm results synthetic test set

	Random forest		Logistic regression	
	Connected components (reference)	Infomap	Connected components (reference)	Infomap
Number of papers	2,589	2,589	2,589	2,589
Number of clusters	385	484	594	636
Number of papers per cluster	6.72	5.34	4.39	4.07
ACP	0.6487 (0.26)	0.9379 (0.09)	0.8141 (0.12)	0.7763 (0.14)
APP	0.9208 (0.08)	0.8546 (0.11)	0.9108 (0.14)	0.9580 (0.10)
K-Metric	0.7496 (0.16)	0.8915 (0.07)	0.8529 (0.09)	0.8561 (0.08)

Note: Standard deviation in parentheses.

Source: Own data.

3.6.2 Application to full block “Muller, M.”

The test of my algorithm on real-world data is my final step. I aim to demonstrate the applicability and predictive performance using the typically German block „Muller, M.“. The block „Muller, M“ takes the 572nd rank of blocks in the WoS and accounts for 11,655 papers that have less than 10 coauthors. I process these papers and generate 87 million paper pairs where one or both first names are missing in a pair or where the first names are the same. I apply the random forest model and the logistic regression to the paper pairs in the next step. The processing takes 2,700 seconds in the random forest and 80 seconds in the logistic regression. Both calculations have been done in R and on one core of a Intel Xeon Bronze 3204.

In the next step, I build the graph and apply the infomap community detection. The infomap takes about 30 seconds to process the predictions from the random forest and the logistic regression and delivers 1,136 clusters in the random forest and 723 clusters in the logistic regression. I measure the quality of these clustery by evaluating the ACP, APP and K-Metric of the 851 papers that have a Researcher ID. Table 11 shows the results. For both approaches, I find perfect author cluster purity indicating that the clusters only contain papers of the same Researcher ID. The AAP, which describes the fragmentation of a Researcher ID into different clusters, is 0.61 in the random forest and 0.66 in the logistic regression. Compared to the test set results, both algorithms perform worse in the AAP measure and create too many clusters. The K-Metric, which reports the geometric mean of AAP and ACP, decreases significantly as compared to the test set results.

Table 11. Application: Cluster results for block "Muller, M."

	Random forest	Logistic regression
Number of papers	11,655	11,655
Number of paper pairs	88,164,593	88,164,593
Number of clusters	1,136	723
Mean number of papers per clusters	10.25 (32.87)	16.12 (44.13)
Papers with Researcher ID		
Number of papers with *Researcher ID*	851	851
Number of different Researcher IDs	27	27
ACP	1	1
AAP	0.61	0.66
K-Metric	0.78	0.81

Note: Standard deviation in parentheses.

Source: Own data and depiction.

3.7 Discussion and Conclusion

In this paper, I developed a supervised machine learning approach to author disambiguation in large publication databases. I used 12,818 publications from 1,904 scientists with Researcher ID available in the WoS and compared pairwise all authors and papers that have the same last name-first initial block. In this comparison, I exploited bibliographic, country, institution, thematic and author name-based characteristics of papers to train a random forest and a logistic regression classifier. I applied the model to an unseen test set of 2,589 publications and obtained pairwise F1 scores of over .75. In my post-estimation, I used graph-based clustering to combine predictions to find all papers belonging to a single author. To cancel incorrect predictions from the logistic regression and the random forest, I applied infomap community detection, thereby significantly improving my results. Finally, I also applied the model on the full block „Muller, M." with 11,655 papers in the WoS. I evaluated this application by analyzing the clustering for papers that have a Researcher ID and find perfect ACP and acceptable APP cluster metrics.

My choice of test and training data allowed me to collect a rich dataset of true authorship information, which I manipulated to mimic the WoS as well as possible. This manipulation may not have been completely accurate, as I made some important assumptions about the data-generating process of the WoS. I had no information about the true number of authors in a block paper set and assumed that the number of true authors is reflected in the block set size. While this assumption may hold true for medium and large block sets, it may not for small block sets, requiring further refinement of my approach. For example, if an author with a very atypical name published 100 papers, this author would not have been included in my sample because I required block sets with more than 20 papers to have at least two different Researcher IDs, and therefore two different real authors. Therefore, I missed an important aspect of the true data-generating process.

I also had to make restrictions on the number of sampled papers per Researcher ID in order to keep the computational effort manageable. This is problematic in that a block set consisting of 100 papers and two authors, where author 1 has 90 papers and author 2 has 10 papers, would result in exactly 10 papers sampled for each author. Therefore, the resulting prior matching probability for a pair of papers of this set would be lower and again not reflect the true data-generating process.

One solution to these problems is manually disambiguating full block sets and using this information as training data. As described in the introduction, this is a major effort, likely includes human-coder error and is not feasible for large block sets of thousands of papers. However, these large sets are needed to design training data that resembles the true data-generating processes of the WoS and other publication platforms. Therefore, a promising direction for further research might be pre-structuring the manual disambiguation process to save resources and time. In this pre-structuring, my algorithm could be applied to publication data that does not contain identifier information. As the logistic regression predictions can be interpreted as probabilities, predictions that are close to my cutoff values can be filtered out and disambiguated manually, as they are likely more prone to errors.

Finally, my approach's major drawback is the difficulty in appropriately handling synonyms and author name changes. Synonym issues are especially common for German names that include umlauts, which are either not included in English character sets or were inconsistently transformed by the authors. For instance, the German „Müller“ is sometimes written „Muller“ or „Mueller“. Currently, both names are treated as different blocks in my approach. In the future, this and similar problems in other languages could be solved by considering the string distance of names when creating block sets. To address issues in handling author name change, I propose searching the co-author network of a given author for persons with the same first name and the same characteristics, such as institution name.

I decided not to use citation-based features since they require high computational effort and therefore don't allow me to apply the algorithms to vast sets of publication data in reasonable amounts of time. Many of my generated features are sparse, such as *Same coauthors*, or suffer from missing data, such as *Thematic similarity abstract*. This was not the case with first names. I therefore intentionally deleted 90% of first names before 2006 and 20% after 2006 to get a quota of missing data similar to that in the full WoS.

Despite their artificially created, missing observations, first names are one of the most decisive features in the logistic regression and the random forest approaches and confirm the previously discovered relevance of this feature (Torvik & Smalheiser, 2009). Some of my newly designed features, like the thematic similarity of abstracts and titles, did not lead to significant improvement and in fact underperformed compared to similar and simpler features, like the *Jaccard distance title*. Finally, the name-based features, like *First name count group 1* and *First name count group 2*, performed well and were the most important group of features.

The random forest performs better than logistic regression in the F1 measure. This advantage, however, becomes smaller when community detection methods are applied. This may be because the logistic regression is more restrictive on positive classifications. This is beneficial in the later infomap community detection because it incurs fewer incorrect linkages between different author clusters, which must be canceled, but enough correct linkages between papers of the same author. The results of the community detection are currently not considered when tuning the random forest and the logistic regression. It is therefore uncertain if random forest disambiguation performance can be improved in more highly tuned machine learning approaches. Tuning the sampling strategy of the random forest may cause other improvements in performance.

The infomap algorithm, in combination with edge weighting, was a decisive post-estimation technique in author name disambiguation and considerably improved performance. Tuning of infomap parameters and the edge weighting need systematic evaluation. The application of other community detection algorithms may further improve the results. However, both issues are beyond the scope of this paper and should be addressed in future studies.

My application to a full block („Muller, M.“) in the WoS was successful. By evaluating the subset of papers with Researcher ID, I find perfect ACP and acceptable APP for both algorithms' clustered predictions. The short processing durations of the 87 million comparisons in the block „Muller, M.“ on a small-scale server system suggest my approach's scalability. Therefore, my approach can be applied to large amounts of publication data such as the full WoS database in reasonable amounts of time. The results of such large-scale application may then be evaluated with the numerous remaining Researcher IDs that were not used in this paper. Other datasets, such as SCOPUS, DBLP or PubMed, may be similarly appropriate for testing since they include reliable identifiers other than the Researcher ID, like Orcid and PubMed IDs. Finally, I want to emphasize the need for a high-quality disambiguated publication database. My approach can help overcome this challenge and potentially improve bibliometric insights and subsequent science and technology policy recommendations.

4 The scientific productivity of German PhD graduates: A machine learning-based author name disambiguation and record linkage approach

4.1 Preface

This chapter builds on an early version of the paper in Chapter 3. Because of the revision process, I changed the methodology in the paper of Chapter 3. I could not update the results presented in this paper. The 30-day processing durations of the author name disambiguation algorithm collided with the submission of this thesis. However, since the methodology was only subject to minor changes, I don't expect any substantially different result. A shortened version of this chapter also appeared in the conference proceedings of the ISSI 2021 conference in Leuven, Belgium as: Rehs, A. (2021). The scientific productivity of German PhD graduates: A machine learning-based author name disambiguation and record linkage approach. *Proceedings of the 18th conference of Scientometrics & Informetrics,* 1531-1533.

4.2 Introduction

Although PhD students and graduates[20] play an important role in scientific knowledge production, technology transfer, and economic growth (Stephan, Sumell, Black, & Adams, 2004), we have little information on their publications and related bibliometric indicators. This lack of data makes it challenging to answer current research questions related to the increasing number of PhD graduates in western countries (Cyranoski et al., 2011), the substantial scientific impact of PhD graduates (Larivière, 2012) and the continuous underrepresentation of female and minority PhD graduates in academia (Larivière et al., 2013).

Therefore, there is a need for new data sources to identify PhD students' publications and a need to link them to other databases of interest (Morichika & Shibayama, 2016). The problem of missing publication information is especially prevalent in Germany. The latest federal report on young scientists indicates a lack of comprehensive databases on German PhDs' publications. It emphasizes that, currently, no conclusion on the scientific contribution of PhDs students can be drawn (Konsortium Bundesbericht Wissenschaftlicher Nachwuchs, 2017, p. 35). Other German sources only address particular and survey based PhD populations in specialized contexts (e.g., Bornmann & Enders, 2001). This lack motivates this paper's primary goal: developing an extensive

[20] In the following I will refer to PhD graduates as those who have just finished their PhD.

database with publication information on German PhD graduates, upon which further research can be built.

This database must include many PhD graduates and link them precisely to their publications, solving an author name disambiguation and a record linkage problem. In the author name disambiguation problem, author names on publications can belong to different persons and need to be resolved through other paper attributes, such as disciplinary information or name frequency, into different author entities (Torvik et al., 2005). Typically, this problem is prevalent with common names, such as "Thomas Schmidt", which account for thousands of publications in the WoS publication database. Several studies have addressed this disambiguation problem with various machine learning methods in recent years, yielding highly accurate results. However, machine learning has so far not been used to develop an extensive, interdisciplinary publication database that builds the foundation for applied scientometric research on PhD graduates and other groups of interest.

Therefore, this paper's first step is to apply a machine learning model to disambiguate author names in the WoS. In the next step, these disambiguated authors and their publications are linked to dissertations in the DNB. A lack of useful unique identifiers makes linking the two databases challenging. I address this problem by using a probabilistic approach for frequent names and deterministic record linkage approach for infrequent names. Probabilistic record linkage attempts to link databases using multiple, possibly nonunique keys, such as first names or institutional affiliations (Fellegi & Sunter, 1969; Sayers, Ben-Shlomo, Blom, & Steele, 2016). In probabilistic record linkage, each identifier is weighted by its ability to determine matches and non-matches between two databases. In this way, the probability that two records refer to the same entity can be estimated, which allows the researcher to filter potential matches according to different thresholds of uncertainty. Deterministic record linkage works in a much simpler way. It compares one or more identifiers across databases and links the entities if the identifiers agree (e.g., do last names between entities match). For datasets that are likely to include only one to one matches, such as linking a person with an infrequent name across two databases, this is a fast and highly performant alternative to probabilistic linkage.

My chapter is structured as follows. In the next section, I review the literature, namely the studies that address the scientific productivity of PhD graduates, and try to sort the literature into the different research types of systemic research, career and occupational research, and scientometric research. I end the literature discussion with a particular focus on studies of German PhD graduates and their publications. In the section *Author disambiguation and record linkage of German dissertation authors*, I describe my databases, the author disambiguation, and the record linkage methodology and procedure. I display the disambiguation results in a separate section directly thereafter. I processed a total of 184,703 author name homonyms into 1.9 million authors and 10.6 million publications. From the 960,000 relevant dissertations in the DNB catalog, I was able to link 61,640 to a corresponding author in my disambiguated publication database. I find that, over time, more PhD graduates are publishing, but this increase may be attributable to better data quality and subsequent linkage performance. I achieve the best linkage of PhD graduates and authors in natural sciences, and can link up to ten percent of the dissertation authors in a discipline. Finally, I investigate the productivity patterns related to the PhD and early career stage. I observe in my 35-year period of investigation a generally stable rate of productivity by graduates two and five years after completing their PhD. To demonstrate my dataset's possible applications, I also investigate productivity by gender and by birthplace in eastern or western Germany. Women underperform compared to men both two and five years after PhD completion. For eastern and western Germans, my results are inconclusive. The paper ends with my discussion and directions for further research.

4.3 Literature

4.3.1 The scientific productivity of PhD students and graduates

In this section, I provide an overview of the different research types of PhD graduates affected by publication data and related bibliometric indicators. Despite their often interdisciplinary nature, I broadly categorize the research types into occupational- and career-related research, systems of higher education research, and scientometric research. This categorization aims to illustrate how publication data from PhD graduates can answer relevant research questions in various fields, and emphasizes the need for corresponding databases.

4.3.2 Policy perspective on PhD publications

PhD graduates foster scientific progress and economic growth, and, therefore, their research output and its underlying mechanisms are relevant from a general policy perspective. On a macro level, PhD publication data has been used to investigate the growth of scientific research (Andersen & Hammarfelt, 2011; de Solla Price, 1963) and to measure how much PhDs contribute to scientific knowledge production. For example, Larivière (2012) found that PhD graduates accounted for about 30% of the scientific publications in the Canadian region of Quebec. PhD publication data can also help researchers to investigate changes in science, such as demonstrated by Rehs (2020a), in which I examined German dissertations before and after the "shock" of German reunification. Here, the topics that construct dissertations in economics and business changed significantly in eastern Germany after the reunification.

The use of PhD publication data in scientific policy research is due to its availability, either restricted to highly aggregated groups of interest, such as certain countries or disciplines, or done at a micro level. Accordingly, I can hardly draw general implications on issues that potentially manifest in the PhD study stage. Here, a current research topic that could benefit from PhD student publication data is the underrepresentation of women and minorities in academia (Fisher et al., 2019; Larivière et al., 2013).

4.3.3 Occupational perspective

From the occupational perspective, completing a PhD marks the start of the first, but also one of the most decisive stages of an academic career. In these early years, many personal and external factors interact and shape the quality, productivity, impact, and topics of the PhD candidate's scientific work well beyond their doctoral graduation. Doctoral advisor(s) and their characteristics have repeatedly been identified as influential for the research topics of PhD graduates (Buenstorf & Geissler, 2014; Waldinger, 2010), and beneficial to their productivity and related career outcomes (Baruffaldi, Visentin, & Conti, 2016). In the same manner, graduate training programs (Buchmueller, Dominitz, & Lee Hansen, 1999), funding (Horta, Cattaneo, & Meoli, 2018; Larivière, 2013) and the disciplinary context and its reward structure (Levin & Stephan, 1991; Millar, 2013) shape the productivity and career outcomes of PhD graduates. I did not find any studies addressing faculty prestige related to PhD graduate productivity. Likewise, personal factors, such as cognitive ability and resilience, have been examined only in general and not concerning PhD graduate publication activities. The occupational perspective on PhD graduates publications also includes nonacademic outcomes, where PhD publications play an important role in knowledge transfer from universities to industry (Buenstorf & Heinisch, 2020; Stephan et al., 2004),

Another type of career and occupational research concerns the difficulties and effects related to publishing before PhD completion. Merga, Mason, and Morris (2020) find that publishing before PhD completion is challenged by thesis cohesion, time pressures, and the publication process. Merga and her coauthors identify positive effects concerning accessibility and dissemination of findings, external feedback through peer review, research skills, career and reputation, emotion and motivation, and ease of examination. Horta and Santos (2016) identify a positive relationship between publishing during PhD study and future productivity, impact, and career outcomes. Tregellas, Smucny, Rojas, and Legget (2018) support this finding and further show that publishing during PhD study is related empirically to employment status, gender, and the time since completing the doctoral degree.

4.3.4 Scientometric perspective

The use of PhD publication data in scientometric research dates back to the founding of modern scientometrics itself when Derek J. de Solla Price used dissertation counts to measure the growth and saturation of science (de Solla Price, 1963). In the following decades, as the use of traditional scientometrics indicators related to citations and journal publications increased, the interest of the scientometric community in PhD data declined.

But recently, it has come into focus again. In addition to the traditional count and related growth of dissertations (Andersen & Hammarfelt, 2011), the scientometric interest in PhD data is manifold. Research focuses on the measurement of impact from dissertations (Kousha & Thelwall, 2019; 2020), the journal publications arising from dissertations (Hagen, 2010; Igami, Bressiani, & Mugnaini, 2014), dissertations formats and their online publication (Bangani, 2018), the dissertation topics (Rehs, 2020a), and the general productivity during the PhD stage. Related to these scientometric research questions are often the development of methods and databases (Kim, Hansen, & Helps, 2018; Morichika & Shibayama, 2016).

4.3.5 German PhD productivity

There are only a few studies that investigate in particular the scientific productivity of German PhD graduates. These studies are survey-based and try to represent German PhD graduates' general population as best as possible. Bornmann and Enders (2001) investigate a sample of 1,259 PhD graduates from biology, electrical engineering, German philology, math, social sciences, and economics and business. They asked their participants to give their discipline and the count of publications during their PhD study. 56% of the students in economics and business published at least one paper. The highest rate of students who publish at least one paper is achieved in electrical engineering (86%). On average, PhD students in German philology produce the lowest average number of publications (4.4) and those in the social sciences the highest (7.3). Jaksztat (2017) investigates the scientific productivity, gender, and parenthood status of German PhD students. He finds that women publish fewer papers than men during their PhD study. Parenthood does not affect productivity for either gender. In contrast to Bornmann and Enders, Jaksztat's study differentiates between peer-reviewed and non-peer-reviewed publications. Men publish, on average, 2.4 articles and women 1.6 peer-reviewed articles during their PhD study[21].

4.4 Methods

4.4.1 Author disambiguation and record linkage of German dissertation authors

Data sources

The dataset I used to trace the publication record of German PhD graduates is built on two sources: the DNB's electronic catalog and a 2017 version of the WoS (Rimmert et al., 2017). The DNB is mandated to collect all German dissertations, and therefore its catalog lists the vast majority of PhD theses submitted at German universities[22]. For every

[21] Statistically different at the 1% level.

[22] For various reasons a small number of theses are not archived by the DNB, e.g., dissertation of German politician Karl-Theodor zu Guttenberg.

dissertation, the catalog stores some basic information, like the dissertation's year, the granting university, title of the dissertation, the author's name. Table 12 depicts selected information on an actual dissertation entry in the catalog.

Table 12. Data structure DNB

Attribute	Value
Title	Über "w-Exklamativsätze" im Deutschen
Author	Avis, Franz Josef d'
Publisher	Tübingen: Niemeyer
Year	2001
Type	Dissertation
Classification:	German Linguistics
University	University of Tübingen

Source: DNB. Own depiction.

Regarding my second dataset, I use a 2017 version of the WoS as the basis to retrieve publication data. In the WoS, all indexed publications have a set of low-level information, which includes the title of the publication, author surnames plus initials, publication year, journal title, and issue. Other information differs in its coverage for a number of reasons, including whether the underlying author is the corresponding author and whether the publication is from before 2008[23]. The WoS's basic structure is illustrated by the example of one exemplary publication in Table 13.

Table 13. Data structure WoS

Attribute	Value
Publication title	Podolski has Contract until 2007, regardless of whether I play in the First or Second Division" A Question on the Acceptibility [sic] of a new Construction with Nouns without articles
Year:	2013
Journal:	Zeitschrift für Germanistische Linguistik *41*(2):212-239
Authors:	D'Avis, F.; Finkbeiner, R.
Author 1:	Franz D'Avis
Institution	Johannes Gutenberg Univ Mainz, Deutsch Inst, FB 05, Jakob Welder Weg 18, D-55099 Mainz, Germany
Institution 2	Johannes Gutenberg Univ Mainz, Deutsch Inst, D-55099 Mainz, Germany
Email	davisf@uni-mainz.de
Researcher ID	-
Author 2:	Finkenbeiner , R
Institution	...
Abstract	In German, singular count nouns usually are accompanied by…
Keywords	…
Classification	Lingusitics

Source: WoS. Own depiction.

To link the German dissertation authors to the WoS publications, I now have to perform an author name disambiguation and a record linkage task. The author name disambiguation task concerns the WoS database and aims to resolve the described homonym and synonym problems. The resulting author-publication sets then need to be

[23] The WoS changed the data collection process and standards around 2008 (Clarivate Analytics, 2021).

linked to the dissertations in the DNB catalog. This task refers to a record linkage problem. Here, again, homonym and synonym problems occur since multiple dissertations may match one author publication set by name, dissertation year, or other characteristics. In the following section, I describe the methodology I applied to approach these two tasks.

4.4.2 Disambiguation of German authors in the WoS

I consider all surname-initial combinations (in the following: homonyms[24]) that appear in the DNB dissertation database as worthwhile to disambiguate. There are 533,197 distinct homonyms in the DNB; 153,213 appear more than once, which means there is more than one dissertation related to that homonym. The left panel of Table 14 shows the most common homonyms and their corresponding numbers of dissertations. In the right-most column, the same homonyms are ranked by their number of papers in the WoS.

Table 14. WoS and DNB homonym frequencies

Surname - initial block	Frequency DNB	Frequency percentile DNB	Rank DNB	Frequency WoS	Frequency percentile WoS	Rank WoS
Müller, M	877	99.99	1	12,184	99.99	579
Müller, H	668	99.98	2	8,539	99.98	950
Müller, A	609	99.8	3	7,553	99.98	1,144
Müller, S	555	99.8	4	6,471	99.97	1,401
Müller, J	543	99.8	5	9,271	99.98	850
…						
Wang, Y	127	96.59	169	222,984	99.99	1
Zhang, Y	125	96.50	178	184,703	99.99	2
	N=1,068,533		N=533,197	178,540,513		N=6,450,191

Note: The German Müller was transformed to Muller and Mueller in the WoS.

Source: DNB and WoS; own calculation.

As Table 14 indicates, there is a divergence in ranks between the WoS and the DNB. This divergence is plausible since traditional German names are more prevalent in Germany than elsewhere. Moreover, highly ranked homonyms in the WoS might also reflect that countries with much larger populations might naturally produce more research papers. This divergence requires us to put a large amount of computational effort into the disambiguation of large and non-typical German homonym sets, such as “Zhang, Y.,” which apply only to a relatively small number of German PhD graduates. However, I must ensure that “Zhang, Y”’s 125 dissertations are correctly assigned to the 184,703 publications and their actual authors.

Machine learning-based disambiguation

I return to the disambiguation of the WoS publication database. Here, I need to build features that allow us to identify similarities between papers with which I can

[24] Upon a disputable referee request, I referred to homonyms as “blocks” in Chapter 3. In both Chapters and throughout my dissertation both terms are interchangeably.

disambiguate them into author-publication sets. In the following, I use my machine learning approach developed in the previous chapter. I use the Researcher ID that many authors assign to their WoS account in order to create a training set of paper pairs from authors that share the same name, but are different persons, according to the Researcher ID. In this machine learning approach, I then apply logistic regression and random forest to paper and name attributes to learn a discrimination function. The paper and name attributes used include author characteristics (such as first name and second initial, stated on the two papers in a pair), thematic characteristics (such as subject classification, abstract similarity), and institutional and regional characteristics (country, name of the university, name of the department).

The predictions now represent how papers of the same homonym relate to each other. To create author-publication sets of all papers that belong to a single author, I build a graph where papers are represented by nodes and positive predictions by edges. In this graph, connected components or communities are likely to represent author-publication sets. However, false predictions may connect different communities and therefore misclassify all papers in both communities. Algorithms that can detect incorrect linkages are referred to as community detection methods (Newman, 2018). They implicitly require densely connected communities to detect the erroneous linkages. In my approach, I use the infomap algorithm (Rosvall & Bergstrom, 2007). The basic premise of the infomap algorithm is that random walks from a given node are more likely to stay within the same community rather than leave the community.

I modify my approach from the previous chapter in several aspects to increase performance and to minimize computational efforts. For performance, I overwrite any false predictions made by the logistic regression if both papers have the same Researcher ID. I also cancel positive predictions if the authors of both papers have different first names, different second initials, or if the papers are more than 25 years apart. I interfere in the graph construction and delete all edges from nodes that connect to more than 200 other nodes. Those seldom-occurring highly connective nodes (or papers) are likely to include a high number of wrong predictions because they imply that a researcher has taken part in more than 200 publications. I find this unlikely, and tested the 200-paper threshold on some homonyms. This threshold greatly reduced the number of implausibly large author-publication profiles.

In the next step, I optimize processing time by leaving out several variables that turned out to be irrelevant or computationally expensive in Chapter 3. Those are keywords and citation counts. Finally, I reestimate the model with the remaining variables and the original training data. I do not find any severely different estimators as compared to the original model from Chapter 3.

In applying the algorithm from Chapter 3 I find that disambiguation of very frequent homonyms is beyond my computational capacities. For “Wang, Y.” I would have to disambiguate 222,984 papers. That disambiguation would process over 49 billion paper-to-paper comparisons. I deal with this problem partially by considering only papers that have fewer than 11 coauthors. For those papers, an author’s contribution might be negligible, and the disambiguation of those papers, therefore, does not add to my overall goal of providing an individual scientific productivity database.

In applying my algorithm, I also try to exploit my computational capacities as much as possible. Therefore, I parallelize parts of my algorithm and store the database in a distributed SQL server setup. I optimize the SQL configuration through indexes and buffer parameters and run the disambiguation in three instances of R[25] . In one instance of R, I disambiguate predominantly small homonyms. Two other disambiguate predominately medium and large homonyms. After 30 days of processing, I stop the algorithm.

Disambiguation of publications in the WoS: Results

My algorithm disambiguated 184,783 of the 533,197 homonyms in the DNB and ended up with 10.6 million publications processed into about 1.9 million different authors. These processed homonyms cover about 51% of the 960,000 relevant dissertations[26] in the DNB. Concerning the complete WoS, my approach processed 2.8% of the 6.5 million homonyms. Those homonyms account for 4.2% of the 178 million homonym-publication relationships in the WoS. Figure 9 shows the total number of papers by year and the number of missing first names in the WoS. I can observe a significant increase in first names after 2006/2007/2008.

[25] Run on a microserver with 256GB RAM and 10 cores.

[26] I consider dissertations between 1975 and 2015 and those outside the disciplines: history, philosophy, theology, and arts and music as relevant.

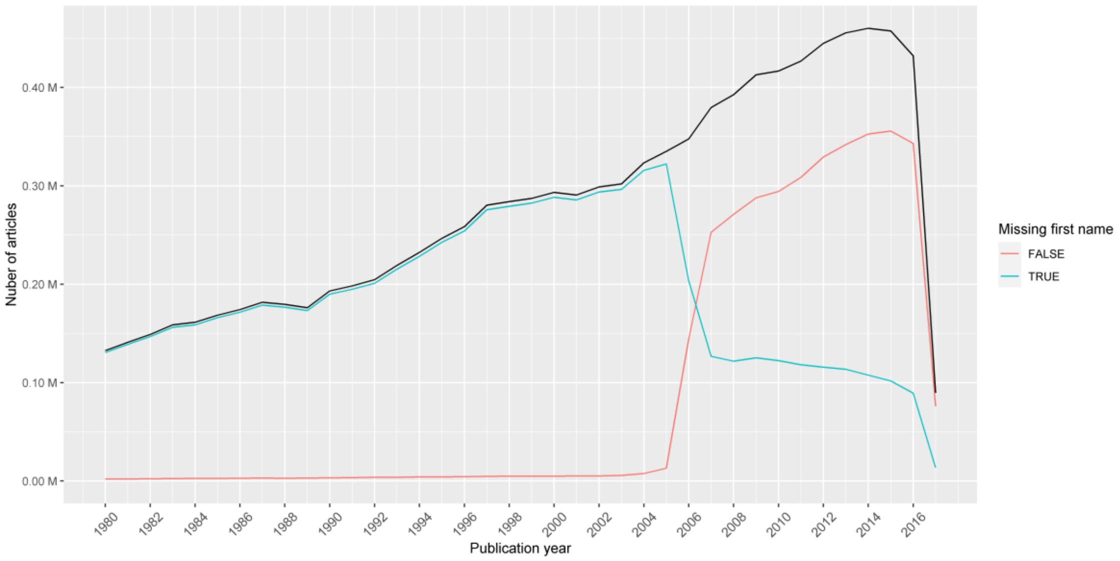


Figure 9. Number of disambiguated publications by year and missing first names

Source: WoS and own calculation. Own depiction.

Table 15 shows summary statistics for the disambiguated dataset at the homonym level. When comparing the most common names here with those in Table 14, I can see that my algorithm did not include the DNB's most common homonyms. This is because my previously described strategy of disambiguating small, medium, and large homonym sets in separate R instances was not completed. I presume that the left-out names and scientific characteristics are unrelated and do not result in any bias of my sample. The same holds for the processed names. The choice of processed names was based on frequency and some random order keys in the initial database (Rimmert et al., 2017). My disambiguation returns authors that have, on average, 4.63 publications, with a mean publication year of 2000.

Table 15. Summary statistics of the disambiguated dataset at block level

Rank	Homonym	No. of papers	No. of authors	Mean no. papers per author	Number of distinct first names	Mean publication year
1	*schmidt, h*	7,617	1,485	5.13	101	1999
2	*schwartz, m*	6,975	1,788	3.90	90	2000
…	...					
500	*bauer, d*	1,937	426	4.55	31	2001
..	..					
5,000	*horst, j*	407	52	7.83	21	2000
…	...					
50,000	*denzinger, a*	25	4	6.25	2	2006
…	...					
184,738	*zywzok, w*	1	1	1	-	1980
N = 184,738		N = 10,694,018	N = 1,987,400	Mean 4.63 (7.71)	Mean 8.98 (15.84)	Mean 2000 (9.95)

Note: Std. dev. in parentheses.

Source: Own data.

I evaluate my disambiguation results by using the inter- and intra-author-publication set occurrence patterns of ORCID (Open Researcher and Contributor ID) iDs. An ORCID iD is a unique identifier for authors and similar to a Researcher ID, which authors can assign to their papers via an online platform. I now check how much a given ORCID iD spreads across different author-publication sets and how many different ORCID iDs can gather in one author-publication set. For all papers that have an ORCID iD, I find that, for one author, there are an average 1.004 (SD = 0.06) ORCID iDs; in turn, one ORCID iD on average relates to 3.11 different author-publication sets (SD = 18.43). Only the latter inter-author-publication set is high, but this does not mean my disambiguation is low quality. This is because I removed edges from nodes that connected to more than 200 papers. For the 1,000 papers with ORCID iD "0000-0003-4978-4670," and similar cases, I, therefore, falsely removed all correct predictions. Therefore, I produced a high number of isolated author profiles that refer to the same person. The edge removement procedure may also lead to several biases in my resulting database. Younger researcher that have not published more than 200 papers may be favored. In the same sense very eminent and researchers in disciplines with high-frequency publication cultures are underrepresented.

4.5 Probabilistic record linkage of publication data to German dissertation authors

In this section, I want to connect entries in the DNB with those in my disambiguated WoS database. However, I lack unique identifiers in the two databases that could train a supervised machine learning algorithm, as done for author disambiguation in the WoS. Therefore, I use a record linkage techniques.

Record linkage is either deterministic or probabilistic. In deterministic linkage, records are matched if linkage fields agree, or unmatched if they disagree. Heinisch and Buenstorf (2018), for example, use first names and university affiliation to match DNB and WoS data. However, this approach, and deterministic approaches in general, are not robust to measurement error (e.g., related to misspellings or divergent character sets) and missing data. Additionally, uncertainty in the merging procedure cannot be quantified. Instead, arbitrary thresholds determine the similarity sufficient for matches.

The Fellegi & Sunter model of probabilistic record linkage

Probabilistic record linkage approaches can account for this uncertainty (Fellegi & Sunter, 1969). They estimate the probability that two given records refer to the same or different entities. In this process, each identifier, such as first names, is weighted by its performance in determining matches and non-matches between two databases.

Two probabilities are of interest for the determination of weights: unmatched probability (u-probability) and matched probability (m-probability). The u-probability gives the probability that a variable between two datasets agrees by chance. For example, I use the address string to compare records between the DNB and the disambiguated WoS dataset that share the same block “Muller, M”.

If I assume ten unique and uniformly distributed addresses in the two databases, the u-probability would be 1/10, or 0.1. The m-probability describes the probability that a variable in matching pairs will agree. Again using the example of addresses and abstracting from peculiarities of the two datasets, which I will address later, the matched probability between two records may be exactly 1. That means the address of a dissertation always agrees with the address stated on the corresponding publication.

However, I don’t know much about this probability and the related true matches. I can only estimate the m-probability using iterative methods, such as the expectation maximization algorithm (Dempster et al., 1977). Given that I would estimate 0.95 for the m-probability, the address variable weights would be calculated as described in Table 16. The described calculation can be repeated for all other possible variables between the two datasets. Finally, the total weight for matching and non-matching are determined by adding all variable weights. The posterior probability for matching is then derived from the total weight (Blakely & Salmond, 2002).

Table 16. Identifier weighting in probabilistic record linkage

Variable	Outcome	Proportion of links	Proportion of non-links	Frequency ratio	Weight
Address	Match	m = 0.95	u ≈ 0.1	m/u ≈ 9.5	ln(m/u)/ln(2) ≈ 3.25
Address	Non-match	1−m = 0.05	1-u ≈ 0.9	(1-m)/(1-u) ≈ 0.05	ln((1-m)/(1-u))/ln(2) ≈ -4.17
First name	…				

Source: Depiction based on Blakely and Salmond (2002).

I use an extension of this framework (Enamorado & Fifield, 2019; Enamorado, Fifield, & Imai, 2020). This “fastlink” framework allows the incorporation of partial agreements and missing data, which relaxes the conditional independence condition of the Fellegi-Sunter model. The conditional independence condition requires the variables to be independent of each other, which is seldom the case with real-world data. In my case, a publication’s address field and the author’s first name correlate in their submissions. This is because on journal publications of the 80s and 90s, often only the corresponding author filed in its first name and address. The adapted setup allows me also to incorporate the prior probability of a match across all pairwise comparisons and a weighting of this match vis-à-vis the likelihood derived from the data.

In Bayesian statistics, prior probability refers to the matching-probability of two records before any data is observed. I find 0.1 a reasonable prior matching probability between two records since only a small to medium fraction of dissertation authors might have ever published a paper. Therefore, they don’t appear in the disambiguated WoS database and cannot be matched. I assign this prior probability a weight of 0.5 and the likelihood derived from the data a weight of 0.5. The fastlink framework, however, also allows modeling a second prior probability. The second one determines the prior probability that two records can have identical values for some variables, even though they do not represent a match. Here, I find 0.02 a reasonable prior probability, because the year variable or the first name variable may sometimes coincide but not necessarily represent a match between two records. I weigh this prior probability with 0.5. Finally, I set a threshold for matching two records based on the posterior distribution and discard all matches that have a posterior probability below 0.8.

Record linkage preparation

The next step is the linkage preparation of the disambiguated WoS and the DNB. Both databases have peculiarities. While the DNB, as a dissertation database, usually captures a single event, my disambiguated WoS database covers a researcher’s lifetime record of publications, sometimes including dozens of publications. Therefore, I have to aggregate all WoS publications belonging to a disambiguated author and build a synthetic author profile consisting of the first papers year, the mode value of the address, and the mode value of the first name. This aggregation provides me with a single record for each disambiguated author in the WoS.

Some other peculiarities of the two databases include their coverage of disciplines and regions. The WoS contains, mostly, relevant journal publications in sciences, but also social sciences publications. The WoS yields low coverage rates for arts and humanities, as journal articles are not the dominant academic medium in this discipline. In turn, the DNB includes almost all German dissertations and, therefore, more accurately represents the disciplinary spectrum of (German) academia than my disambiguated WoS database.

I deal with this issue by excluding DNB publications listed as history, philosophy, theology, and arts and music dissertations. This strategy aims to reduce the number of non-matches. Regarding regional coverage, I may have, in no small extent, disambiguated non-German authors. Therefore, I restrict the synthetic WoS author profiles to those with an address from Germany or surrounding countries, or to be missing. In this way, I sort out nonrelevant author profiles and reduce the number of non-matches.

Another peculiarity of the DNB is the level of information needed to link the two datasets. In the DNB, only the dissertation author's first name and surname, granting institution of doctorate, dissertation title, and dissertation year are useful for linkage. These pieces of information are almost always available. In the disambiguated WoS, the corresponding information bears substantial missing data. In my author-publication sets, 60% of the profiles have no first name and 47% no address information. Moreover, the address string and the granting institution of the doctorate naturally never agree. Therefore, I process the DNB's institution string and extract the city name (e.g., Universität Kassel becomes Kassel). This extraction becomes the variable *city* in the DNB dataset. In the next step, I create a list of all possible city names and test if one of the cities occurs in the address string of the WoS synthetic author profiles. If this is the case, I assign the found city name to the variable *city* in the synthetic author profiles. In the next step, I block the data by homonyms to include synthetic author profiles where the first name is missing. My dataset now looks as depicted in Table 17.

Table 17. Prepared datasets for linkage

	DNB data			WoS synthetic author profiles		
Homonym	First name	City	Year	First name	City	Year
Müller, M	Michael	Kassel	2010	M.	-	2005
Müller, M	Mira	Kassel	2012	M.	Bremen	2001
Müller, M	Manuel	Bremen	1998	M.	-	2015
…	…					

Source: Fictitious data. Own depiction.

Finally, I consider only homonyms with sufficient frequency. A minimum frequency is computationally needed to conduct reasonable probabilistic inference (Enamorado et al., 2020). I do so by setting the threshold to homonyms with at least ten different dissertations and at least ten different author profiles.

I apply deterministic linkage for homonyms containing fewer observations than these thresholds. Here, I merge authors that match in all the variables *city*, *year*, and *first name*. If there are fewer than three entries in a both databases, I relax the matching to *year* and one other variable. Generally, I allow a margin of +/− 2 for the variable *year*. For *city* and *first name* I allow fuzzy matching but allow variation of only one character in the Levinstein distance used.

4.6 Results

In my probabilistic approach, I was able to link 30,840 dissertation authors to corresponding author profiles. The deterministic strategy yields 30,804 dissertations. My dataset of 61,640 linked authors covers 5.4% of the relevant DNB dissertations and 0.03% of the disambiguated WoS author profiles. Compared with previous findings on the share of German PhD graduates who publish (Bornmann & Enders, 2001) my coverage rate is small. However, the little information that was useful for linkage in the DNB, and the subsequent uncertainty in the linkage procedure, required me to set high thresholds for probability sufficient for matching. Therefore, and because of many other methodical and data problems, I find this a reasonable percentage.

In this section, I analyze the linked dataset by describing temporal and disciplinary coverage and productivity patterns. I also investigate how many dissertation authors in a cohort publish and whether this share changes over time. Finally, I utilize data that has been linked to the DNB and analyze the shares in the dataset of women and of those born in eastern Germany (Fuchs & Rehs, 2020).

Figure 10 describes the percentage of dissertations in the DNB by discipline. I obtain the highest coverage in disciplines where journal publications are traditionally popular, such as sciences and mathematics.

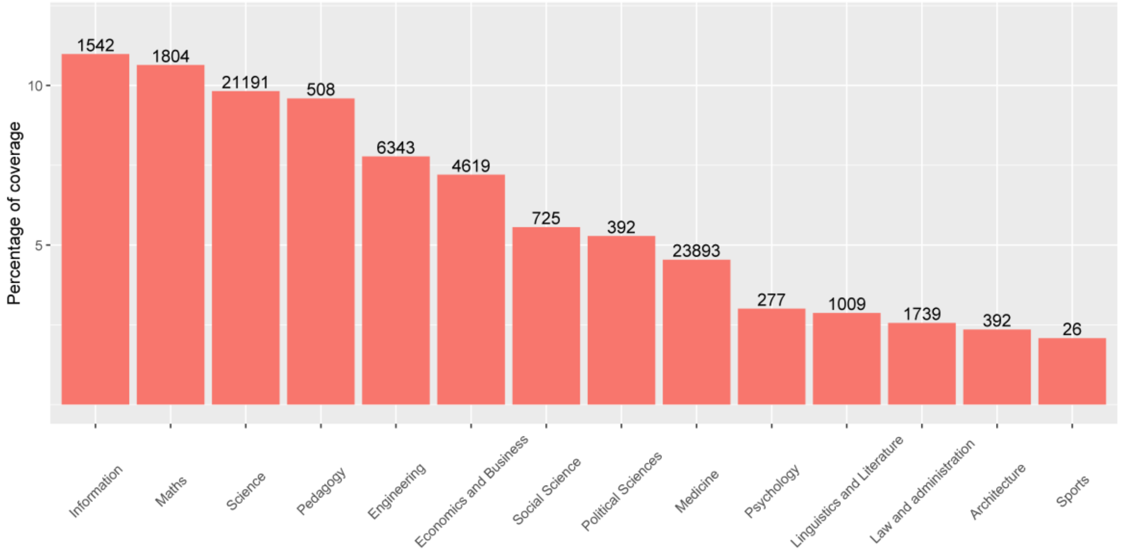


Figure 10. Coverage of dissertations in the DNB by discipline

Source: Own data and calculations. Own depiction.

Figure 11 shows the coverage of dissertations by year. The results indicate a continuously growing number of the identified dissertation. This is possibly related to the increasing number of papers published (see Figure 9) and the related availability of first names for linkage. Figure 12 shows the linked authors' productivity within a 2-year and 5-year time frame after the dissertation. The 2-year time frame should cover all publications related to the PhD stage and account for different publication lags across disciplines. The 5-year time frame aims to capture the decisive early career productivity. For the 2-year time frame, I observe stable productivity from PhD graduates over the 35-year time frame of my investigation. My linked PhD graduates publish, on average, three WoS-indexed publications during their PhD. The 5-year pattern is more heterogeneous but also includes more variability, as the standard deviation bars indicate. Again, I neither observe a significant increase nor decrease in PhD productivity over time. The declining 5-year productivity towards the end of my time frame can be attributed to my database's censoring in 2017.

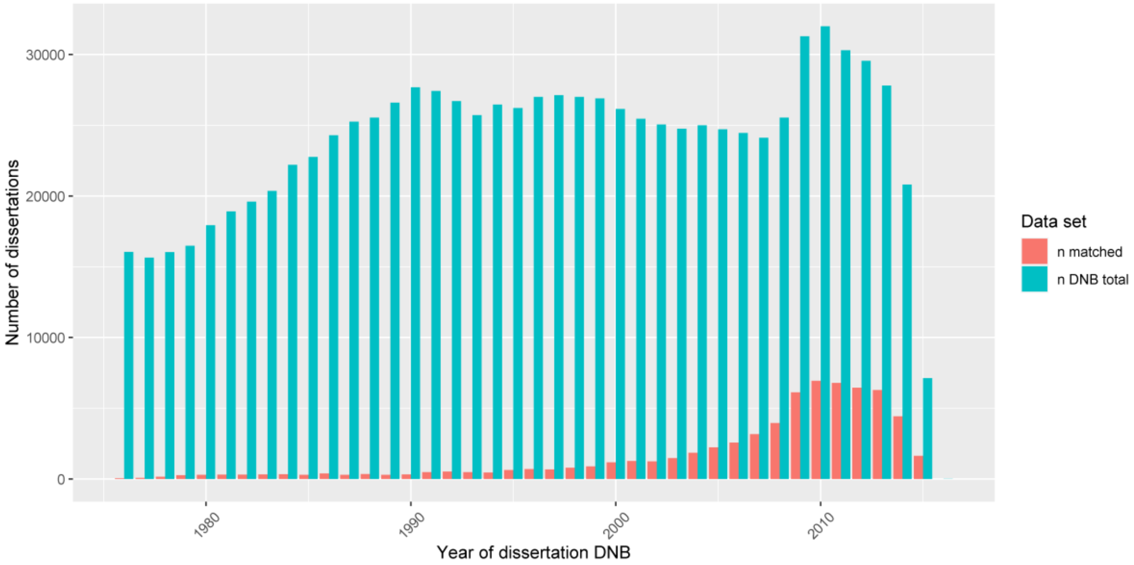


Figure 11. Coverage of dissertations in the DNB by year

Source: DNB, own data and calculations. Own depiction.

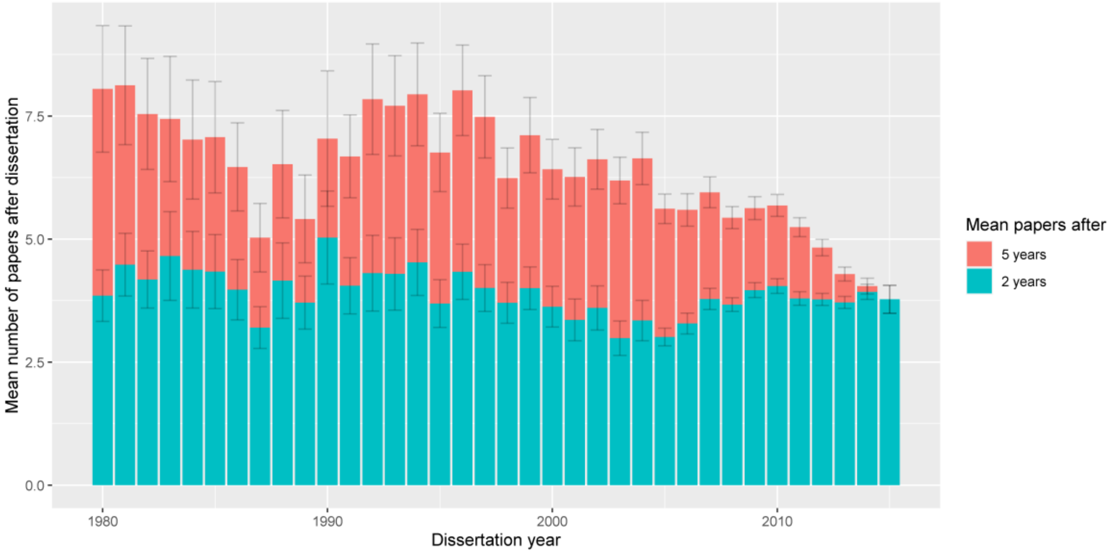


Figure 12. Average productivity 2 and 5 years after dissertation

Note: Error bar indicates a standard deviation above the mean.

Source: Own data and calculations.

Figure 13 shows the productivity of PhD graduates by gender. For both time frames after the PhD, male PhD graduates regularly outperform female PhD graduates. This pattern confirms previous findings from Jaksztat (2017) and should be investigated more precisely in further studies. Finally, Figure 14 depicts the 2- and 5-year productivity by eastern or western German birthplace of the PhD graduates.

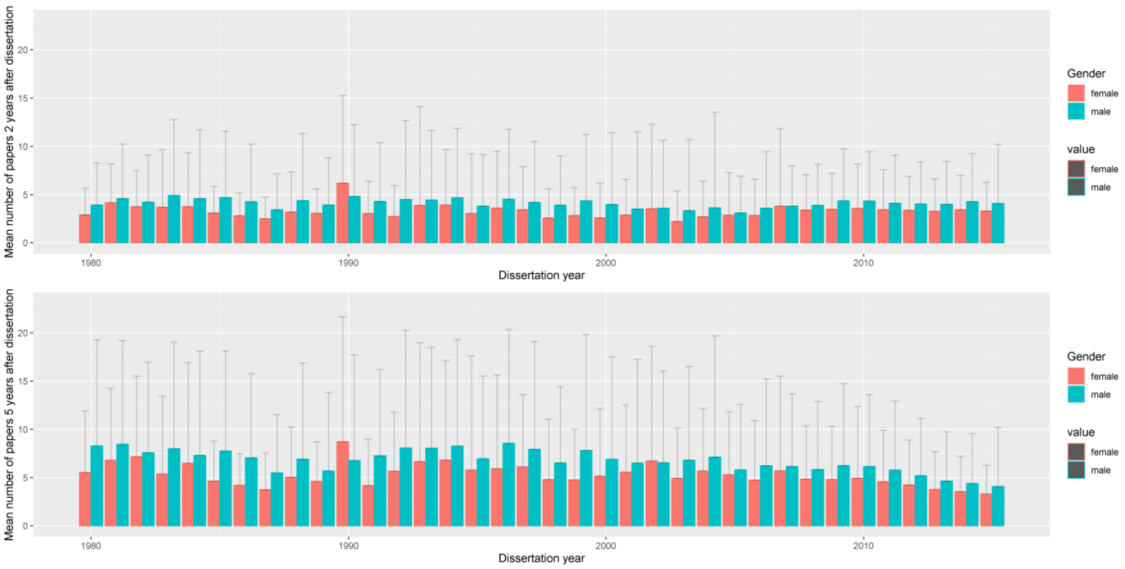


Figure 13. 2- and 5-year productivity by gender

Note: Error bar indicates a standard deviation above the mean.

Source: Own data and calculations.

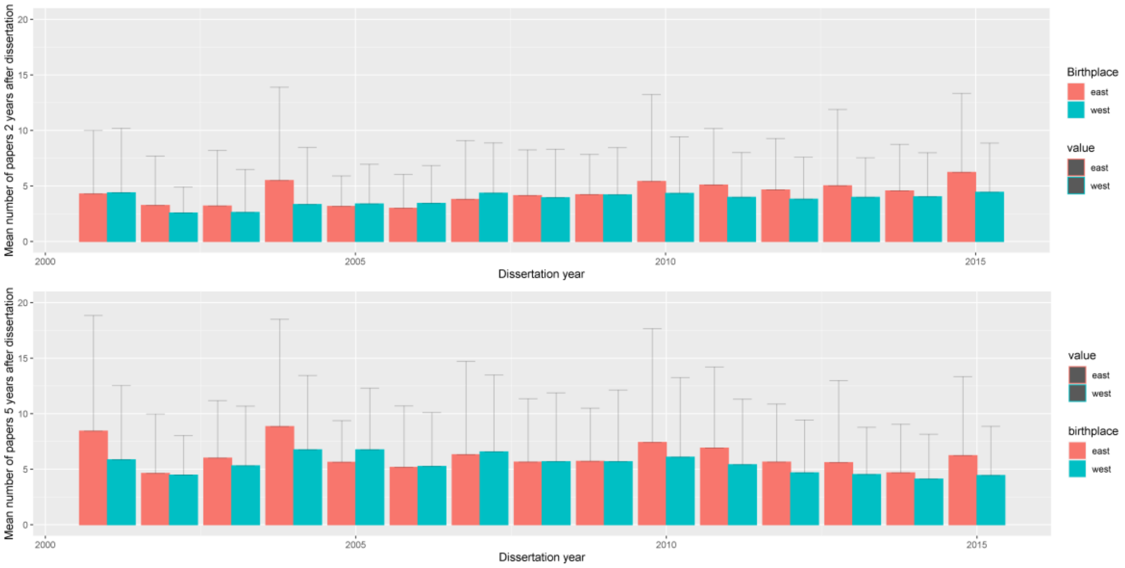


Figure 14. 2- and 5-year productivity by birthplace in eastern or western Germany

Source: Own data and calculations.

4.7 Discussion and conclusion

In this paper, I have developed a database for the scientific productivity of German PhD graduates. Using the machine learning-based approach of the previous chapter, I disambiguated about 184,000 author names and 10.6 million publications in the WoS into 1.9 million author profiles. Subsequently, I linked the authors to 61,640 dissertations in the DNB using a probabilistic and deterministic record linkage approach. This paper's primary research goal – developing a database on German PhD productivity – was therefore accomplished.

My approach has several limitations. I focus only on dissertation authors who publish in WoS-indexed journal articles and use them as a measure of productivity. This focus does not account for the wide range of publication cultures in German academia. It favors identifying publications in natural science fields over those in the social sciences and arts and humanities. Methodically, my approach is subject to limitations, as well. I adapted the author disambiguation algorithm from Chapter 3, which builds on WoS training data that contains the Researcher ID author identifier. This dataset does not fully capture the true data generating process of publication data of the WoS and may have caused some problems that I had to address. In preliminary testing setups, the algorithm produced some implausible and huge author entities. I manipulated my algorithm from Chapter 3 and set 200 as a threshold for a paper, regarding its maximum number of positive predictions to other papers. If this threshold was exceeded, I isolated the concerned papers, which produced many authors with only one paper. My evaluation of the results via the ORCID iD show that this seems to have prevented the clustering of papers from very productive authors – who take part in more than 200 publications – into an author-publication set containing only 1 paper. These problems need to be addressed from a methodical and data-oriented perspective, but are beyond this paper's scope. The adjustment of the training data from the algorithm of Chapter 3 to author names that are linked to many papers can potentially settle the issue.

My time was restricted and I had to stop the disambiguation after 30 days. Therefore, I could not disambiguate all names that appear in the DNB catalog and especially miss the DNB's most common homonyms. This is because my strategy of disambiguating small, medium, and large homonym sets in separate R instances was not completed. I presume that the left-out names and scientific characteristics are unrelated and do not result in any bias of my sample. The same holds for the processed names. The choice of processed names was based on frequency and some random ordered keys in the initial database (Rimmert et al., 2017).

My processing was also restricted by my computational capacities and, in the future, could be run on more powerful machines. I also see space for improvement by using blocking strategies. As discussed, large homonyms sometimes require comparison of over 10,000 publications at a time. For the largest homonym in my dataset, "Schmidt, H.," with 7,617 publications, this resulted in a processing duration of 2 hours. Here, blocking publications by WoS discipline, region, or publication period could reduce the number of comparisons and related processing duration. However, this strategy needs a thorough assessment of the underlying data quality. I leave this open for further study.

My record linkage linked a comparatively small number of authors from the disambiguated WoS database to German dissertations. However, I set reasonably low thresholds for uncertainty in my probabilistic setup. The inclusion of uncertainty in the record linkage helped me overcome problems of deterministic approaches that rely on arbitrary matching thresholds of underlying identifiers. This method equips me with a dataset of PhD publication data that is useful for further research purposes. In this regard, I demonstrated how publication counts could be used to investigate current topics like gender and minority underrepresentation in academia. My results show that male graduates, on average, outperform females in terms of 2- and 5-year publication productivity after PhD completion. For differences between PhD graduates born in either eastern or western Germany, my results do not show any conclusive pattern.

Finally, I conclude that the investigation of PhD productivity is a topic of interdisciplinary interest and one that requires interdisciplinary methods. Computer science and its advancements in machine learning can contribute to this by unlocking data sources, such as the WoS. Statistics provides the tool kit to link these with existing databases, and scientometrics and social sciences can give valuable indicators and applications.

5 Protégé-advisor gender-pairings in academic survival and productivity of German PhD graduates

5.1 Preface

This paper builds on the Chapters 3 and 4. Chapter 3 developed the machine learning algorithm that was used to build parts of the data base presented in Chapter 4. The data base of Chapter 4 was then used in this Chapter. A shortened version of this Chapter was published as: Rehs, A. (2021). Protégé-advisor gender-pairings in academic survival and productivity of German PhD graduates. *Proceedings of the 18th Conference on Scientometrics and Informetrics,* 955-967.

5.2 Introduction

Doctoral advisors are often the most influential persons at the beginning of an academic career. They transfer knowledge, attitudes, norms, and behavior to their protégés and influence their academic socialization and success (Barnes & Austin, 2009). Several studies have addressed the various scientific and socioeconomic characteristics of the advisors and their protégés to point out what makes these relationships mutually successful. Gender-pairing in protégé-advisor relationships repeatedly stands out in this regard. It has diverse effects on career attainment and publication output of protégés (Gaule & Piacentini, 2018; Hilmer & Hilmer, 2007; Pezzoni et al., 2016).

In this paper, I want to investigate the role of protégé and advisor gender in German PhD graduates' academic outcomes. Especially for pairings involving women, this issue is of high societal and scientific interest in Germany. As observed in other countries, women are underrepresented in advanced career stages in German academia (Larivière et al., 2013). Although they accounted in 2017 for 51.7% of graduate students, their share of PhD holders was 45.4%. Women account for only 25.6% of university professors in Germany (Statistisches Bundesamt, 2020). This departure of women from the academic workforce indicates a misallocation of talent (Acemoglu, 1995). The consequences of this departure imply decelerated scientific progress with negative spillovers to industry and the economy in general. Women may also be personally affected. If they are equally qualified with men to start and pursue an academic career, but at some point quit, their educational investment cannot be fully utilized (McGuinness, 2006). To my knowledge, there is as of 2020 no comparable study of this phenomenon for German PhD graduates.

Gaule and Piacentini (2018) argue that this underrepresentation of women in academia perpetuates itself through the lower availability of female advisors for female students. They argue that underrepresentation works through a productivity channel or a preference channel. In the productivity channel, students are less productive when collaborating with an advisor of the opposite gender. As productivity is generally the primary driver of academic career success, this leads to higher rate of dropout from academia for female PhD graduates who were advised by men. In the preference channel, the authors argue that working with an advisor of the opposite gender is less enjoyable and leads to lower career satisfaction and a higher chance of dropping out early. Gaule and Piacentini show that a PhD student's research productivity, and propensity to become faculty after graduating, are both related to the gender of the advisor.

In this paper, I build on Gaule and Piacentini's findings. In the first step, I test whether productivity during a PhD in German academia is also linked to protégé-advisor gender-pairings. In the second step, I focus on the disentanglement of the temporal patterns related to career outcomes and advisor gender after completion of a PhD. From the temporal perspective, academic careers, and careers in general, are non-dichotomous processes. They include multiple decisions and promotions that differ in their duration and in their point of time. The investigation of fixed points in time, as done in Gaule and Piacentini (2018), does not exploit the temporal dimension to its full extent. In this sense, it is an open question of how long protégés in different gender pairings remain in academia and which exit "risk" they assume after their PhD.

These durations can be considered as survival times and allow to utilize related models such as Cox proportional hazard or complementary log-log regression. The complementary log-log regression used in this paper estimates covariates' effect upon the time a specified event takes to happen and assumes time to be discrete (Tutz & Schmid, 2016). Therefore, I can investigate how an advisor's gender and other characteristics affect the time a PhD graduate remains in academia after finishing his or her PhD. In the following, I will refer to this as "academic survival". A similar methodology was applied by Sabatier, Carrere, and Mangematin (2006) to investigate the time it takes for female and male postdocs to attain professorship.

In the subsequent section, I discuss previous findings on gender pairings in academic protégé-advisor relationships. In the Data and methods section, I present my three data sources: doctoral advisor information scraped from German online dissertations, the DNB's catalog, and publication data from the WoS that I have previously disambiguated. In the Results section, I describe the specification of my econometric approach and use negative binomial regression to estimate the effect of advisor gender on PhD student productivity. While I found that women were less likely to publish during the PhD, being advised by women did not have any effect on publication productivity. The probability of academic survival by gender and advisor gender, as measured by the final year of publication after completion of a PhD, is investigated with a complementary log-log regression and represents my main finding. I find that PhD graduates who had female advisors are about 38% more likely per year to continue publishing after PhD completion; this effect is not different between male and female graduates. In line with the observable female underrepresentation in academia, I also find that women are about 37% more likely per year than men to stop publishing after completing a PhD. I end this paper by discussing the results, showing limitations, and, finally, concluding.

5.3 Gender pairings and outcomes of doctoral advisory

The topic of gender in doctoral advisory belongs to the greater literature on protégé-advisor relationships. This literature is divided into business research (Feeney & Bozeman, 2008; Noe, 1988), undergraduate (Bettinger, Long, Ehrenberg, Jacob, & Murnane, 2005), and postgraduate protégé-advisor relationships. Central to all these literature strains is some success outcome, like establishing business networks (Feeney & Bozeman, 2008) or influencing women to major in scientific fields of female academic role models (Bettinger, Long, Ehrenberg, Jacob, & Murnane, 2005; Canaan & Mouganie, 2019). In summarizing the literature across all those subdomains, I find that there is no clear support for the hypothesis that female advisors positively affect the outcomes of their female protégés.

The relation of advisor and protégé gender in postgraduate outcomes has been addressed in several studies and is central to the debate of female underrepresentation in academia (Pezzoni et al., 2016). When discussing those studies, one must account for the various disciplinary, institutional, and regional backgrounds in which the studies were conducted. The German context – which is subject of this study – is special in many regards (Kehm, 2006). German universities produce one of the highest proportions worldwide of doctorates in relation to the population (OECD, 2019). About 27% of those doctorates are awarded in the field of medicine, where dissertations are very different from other disciplines (Konsortium Bundesbericht Wissenschaftlicher Nachwuchs, 2017).

The sorting process of advisors and their protégés is the natural starting point to investigate the effects of advisor gender and gender pairings. It has repeatedly been found that same-gender pairings are clearly overrepresented (AlShebli, Makovi, & Rahwan, 2020; Gaule & Piacentini, 2018; Pezzoni et al., 2016). The causal mechanisms for this overrepresentation, however, remain unclear. The qualitative study from Gray and Goregaokar (2010) on executive coaching suggests that women prefer women because they act as a role model for business success. Azoulay, Liu, and Stuart (2017) point towards self-selection processes in the initial matching between protégés advisors. This “partially deliberate” social matching occurs on a small number of actor attributes. Azoulay, Liu, and Stuart (2017) indicate that geography and scientific focus are the main drivers of matching between advisors and protégés and propose a methodology that addresses self-selection problems.

The international literature comes to different conclusions on the effects of postgraduate advisor gender and protégé-advisor pairings. Starting with productivity outcomes, Pezzoni et al. (2016) find for protégé-advisor pairs of the prestigious California Institute of Technology that students working with female advisors publish 7.7% more articles per year while earning their PhD than those working with a male advisor. Using male students with male advisors as the reference group, Pezzoni et al. show that gender pairing matters in this regard. They find that male students working with female advisors publish 10% more articles per year than the reference group, and female students working with male advisors publish 8.5% less. They find no difference between women advised by women and the reference group of men advised by men. The results are robust for using the journal impact factor as a proxy for the quality of articles. Gaule and Piacentini (2018) study the productivity of PhD students in US chemistry programs. In opposition to Pezzoni et al. (2016), they find students with an advisor of the same gender tend to be more productive during a PhD program. Women profit more strongly from same-gender advisors than men do.

Hilmer and Hilmer (2007) and Neumark and Gardecki (1998) investigate protégé-advisor gender pairings in economics and consider activity-based success measures. When examining the first jobs of new PhD graduates, Hilmer and Hilmer (2007) find that female graduates who had male advisors are significantly more likely to accept research-oriented first jobs than male graduates who had male advisors. Neumark and Gardecki (1998) focus on time spent in and completion of graduate school. They find limited empirical evidence for the positive impact of female advisors on the probability that female students finish graduate school. However, female advisors are associated with female students spending less time in graduate school. Gaule and Piacentini (2018) address the likelihood of PhD students becoming university faculty based on different gender pairings. They find female students working with female advisors are considerably more likely to become faculty; for male-male pairings they do not find an effect.

In summary, the empirical evidence on the effects of gender pairings and advisor gender is ambiguous, scattered, and may depend strongly on the data, context, and operationalization of outcomes. I therefore abstain from forming a hypothesis about the protégé-advisor gender pairing effects in German academia.

5.4 Data and methods

My data rest upon two pillars: disambiguated WoS publication data, and PhD advisor info scraped from online dissertations. A schematic of my databases and their relations is depicted in Figure 15. I use the DNB’s 2015 electronic catalog and university library servers to build my online dissertations database. The DNB has the legal mission to collect and archive all printed publications issued in Germany and works written in German or relating to Germany. German PhD graduates are therefore required to supply a copy of their dissertation to the DNB. The DNB’s electronic catalog features information on their authors, the university name, the year of publication, subject, and, if available, a link to an online dissertation.

I use the provided online dissertation link and download the underlying PDF document. The download from the DNB was successful in 40,000 cases. To further increase my online dissertation dataset, I repeat the same exercise for dissertations stored at 20 different university library servers, collecting 80,000 online dissertations. All scraping has been done in 2017. I match the 40,000 dissertations from university library servers back to the DNB's catalog by the author name, year, and university name. I search dissertation front pages and acknowledgments for text patterns like "doctoral advisor" and its German variants in the next step. These patterns indicate the subsequent occurrence of advisor info. A similar approach was used by Fuchs and Rehs (2020) to scrape birthplaces from the same dataset of online dissertations. I find 13,315 protégé-advisor pairs where the found advisor has a unique name in the DNB catalog. The restriction to unique advisor names ensures correct protégé-advisor pairs.

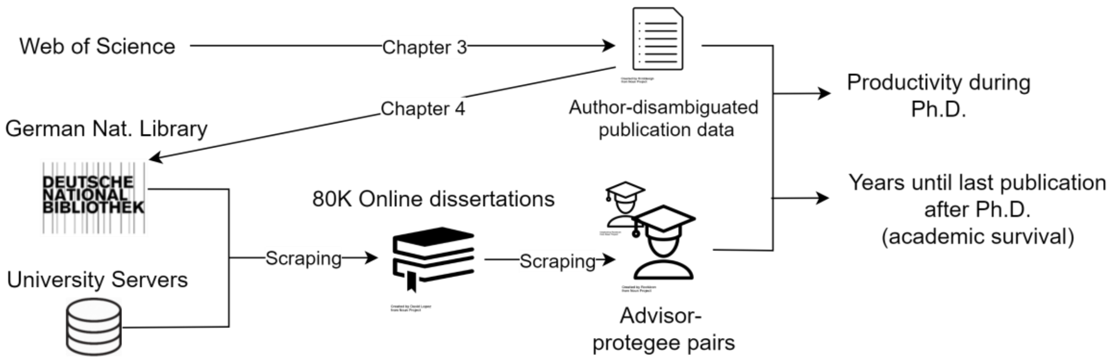

Web of Science
Chapter 3
Chapter 4
German Nat. Library
DEUTSCHE NATIONAL BIBLIOTHEK
University Servers
Scraping
80K Online dissertations
Scraping
Author-disambiguated publication data
Advisor-protegee pairs
Productivity during Ph.D.
Years until last publication after Ph.D. (academic survival)

Figure 15. Database schematic

Source: Own depiction. Icons by RockIcon, thirddesgin and David Lopez from Noun Project.

To retrieve publication data of the protégés and their advisors, I use the database from Chapter 4. Chapter 4 builds on Chapter 3, which develops a machine learning approach to disambiguate author names in the WoS . In Chapter 4, I use this algorithm to establish a data base of about 11 million author-name disambiguated publications and link them to 50,000 dissertation authors stored in the DNB's catalog. Using the publication profiles of German dissertation authors, I investigate how long they continue to publish in WoS journals after completing their PhD and how this duration is related to their gender and the gender of their advisor.

Productivity during the PhD and other related indicators are also calculated from this dataset. For a random set of 100 persons, I check whether the year of the final publication after completing a PhD corresponds to the actual year of dropout from academia. This is because the year of the final publication is inherently imprecise due to, e.g., publication lags. I retrieve the dropout year from online research on Xing, LinkedIn, university homepages, and other websites. In my 100-person sample, I find 63 authors left academia within a 3-year interval around the final publication year; within a 5-year interval 90 persons leave academia. The year of final publication can, therefore, used as a rough exit proxy.

Table 18 reports summary statistics for my final datasets. In line with Gaule and Piacentini (2018), I find that women disproportionally often advise women. Advisors also differ substantially in their numbers and their characteristics by gender, although advisors have, on average, the same academic age (see advisor characteristics: *Dissertation year*). No advisor occurs twice in my dataset. My earliest dissertation from protégés is from 2001; the mean dissertation year is 2005. The availability of online dissertations is the driver of this bias towards younger doctoral cohorts in my dataset. The high popularity of online dissertations may explain the sharp increase in the coverage rate of advisor info in my dataset. The difference between the PhD graduate's dissertation year and the advisor's dissertation year (see advisor characteristics: *Difference to diss. year of protégé*) is a measure for the advisor's academic age. Female and male protégés have no differences in this regard.

Figure 16 shows the temporal distribution of the mean number of (accumulated) publications of a protégé by gender pairing. The descriptive patterns suggest that productivity differences are established early, before completion of a PhD. Male protégés advised by men have the highest mean productivity and women advised by women, the lowest. After t = 0 the publication number averages may include survivor bias and can therefore not be interpreted in a meaningful way.

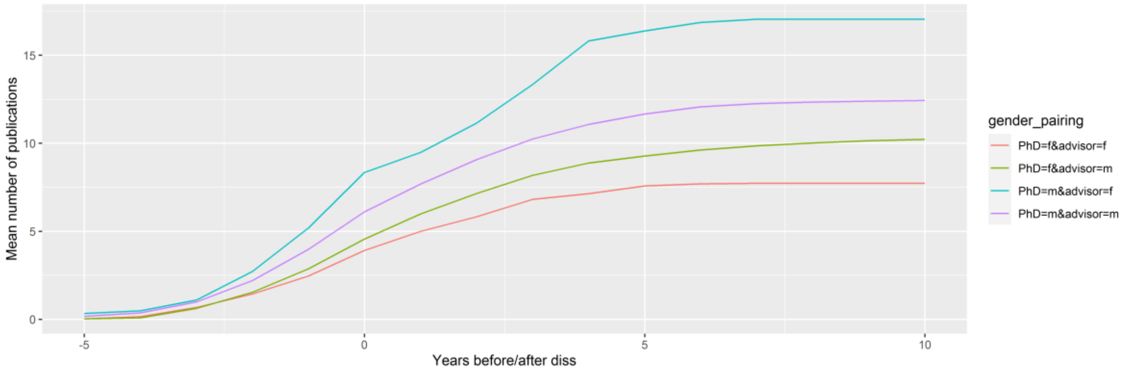


Figure 16. Mean number of cumulated publications before and after dissertation by gender pairing

Source: Own data and depiction.

Years till last pub. after diss. is my main outcome with respect to academic survival. I can observe that PhD graduates advised by men survive on average half a year longer than those by female advisors. In further differentiating the effect of female advisors, I find that their female students produce their final publication half a year earlier than male students. I observe no difference between male and female PhD graduates with male advisors in the mean number of years until the final publication.

Table 18. Descriptive statistics

Protégé characteristics

	Full sample				Protégé = male				Protégé = female			
	N	min	mean	max	N	min	Mean	max	N	min	mean	max
Gender advisor = m.	873				499				374			
Gender advisor = f.	89				21				68			
Years in academia after dissertation	962	0	2,5	15	520	0	2,5	15	442	0	2,4	15
Years till last pub. after diss. & advisor = m.	873	0	2,5	15	499	0	2,5	15	374	0	2,4	15
Years till last pub. after diss. & advisor = f.	89	0	2,1	7	21	0	2,5	7	68	0	1,9	7
Dissertation year		2001	2011	2015		2001	2011	2015		2001	2011	2015
Number of papers till 2017		1	11.5	118		1	12,6	72		1	10,2	118
Sum of papers at dissertation year		0	5,4	66		0	6,2	66		0	4,6	30
Sum of papers at diss. year & advisor = m.	873	0	5,4	66	499	0	6,1	66	374	0	4,6	30
Sum of papers at diss. year & advisor = f.	89	0	5,0	39	21	0	8,33	39	68	0	3,9	14
Number of citations		0	150	555		0	154	555		0	108	1,233
Sum of citations at dissertation year		0	6,3	78		0	6,5	78		0	5,2	63

Advisor characteristics

	Full sample				Advisor = male				Advisor = female			
	n	min	mean	max	N	min	Mean	max	N	min	mean	max
Dissertation year	962	1913	1990	2015		1913	1990	2015		1966	1991	2007
Difference to diss. year of protégé	962	0	20,5	41		-4	21,4	99		6	20,4	41
Number of papers	59	2	59,1	208		2	63,4	189		6	52,9	208

Source: Own data and depiction.

In my econometric setup, I will first investigate a student's productivity during the PhD program to address whether early productivity differences by gender and gender of the advisor exist. The number of papers written until the year after the PhD is my dependent variable. The main variables of interest are the gender of the protégé and the gender of the advisor. Since the outcome variable is a count, I use a negative binomial regression to estimate equation (1). In (1), X is a vector of control variables and includes discipline and year dummies. In addition to (1), I also estimate reduced models and models that include interaction effects of advisor and protégé gender.

$$Number\ of\ papers\ at\ year\ of\ PhD_i = \beta_0 + \beta_1 gender\ advisor_i + \beta_2 gender_i + \beta_n X_i + \epsilon_i \quad (1)$$

Modelling time discrete survival

The next step is modeling academic survival after completing a PhD. I use the year of the final publication after PhD completion as a proxy for academic survival. This outcome's operationalization is restricted in my dataset by the period of graduation cohorts from 1995 to 2015. Older cohorts can therefore be active longer than younger cohorts. My solution is to censor persons who were still active in 2017. A PhD graduate from 2014, still publishing at the cutoff of my database in 2017, is treated as right-censored after three years. I use a time-discrete survival model to estimate the risk of having the last publication at year t after PhD. In the following description of my economic approach, I will orientate on the methodology of Tutz and Schmid (2016) and van de Schoot (2020).

In my time discrete survival models the risk of event builds on the hazard h_{it}. The hazard is the conditional probability that a researcher i will exit from academia in the time period t, given the researcher did not exit earlier. The hazard function can be stated as follows:

$$h_{it} = P(T_i = t \mid T_i \geq t) \quad (2)$$

In (2), T is a discrete random variable. The equation represents the probability that the exit of a given researcher will occur in the current time period t under the constraint that it will occur now or sometime in the future. Now, the hazard in time period t can be estimated as follows:

$$\hat{h}_t = \frac{Number\ of\ exits\ from\ research_t}{Number\ of\ researchers\ at\ risk\ to\ exit_t} \tag{3}$$

In order to build a regression framework, the hazard h_{it} now needs to be linked to a linear predictor η. The relationship of the hazard function and a linear predictor can be represented as:

$$\eta = g(h_{it}) = \gamma_{0t} + x_{it}\gamma \tag{4}$$

Here, g is a link function that links the linear predictor to the hazard. In my case, the linear predictor includes a set of covariates for researcher i in period t. These covariates are protégé gender, advisor gender, and year and discipline controls. To disentangle the effects of advisor gender, I also estimate interactions of advisor gender and protégé gender. In (4), γ_{0t} represents the time-variant intercept and shows the baseline hazard. In the next step, I use a complementary log-log (cloglog) function to link the hazard to the linear predictor. Therefore, my model refers to a time discrete cloglog regression. The hazard function changes and becomes:

$$h_{it} = 1 - \exp(-\exp(\eta)) \tag{5}$$

5.5 Results

Table 19 shows the regression results and average marginal effects for productivity as measured by the number of papers published during PhD study. I do not find an effect of advisor gender in any of my full sample models. However, the coefficients and the marginal effects of protégé gender arrive at statistical significance in the baseline, in the full model, and in the full model with advisor productivity. According to the full model's marginal effects, female PhD students write 1.4 papers less than male PhD students. When including advisor productivity, as done in the full model 5, this discrepancy disappears. The advisor productivity has a statistically significant impact on the number of papers at the dissertation year. An additional publication from the advisor leads to an increase of about 0.04 publications by their protégés. Since this result is based on 59 observations, including only nine female advisors, the robustness is questionable.

Table 19. Negative binomial regression and average marginal effects for productivity during PhD study

	Dependent variable: Number of papers during PhD									
	(1)		(2)		(3)		(4)		(5)	
	Coeff.	AME	Coeff.	AME	Coeff.	AME	Coeff.	AME	Coeff.	AME
PhD student characteristics										
Gender PhD student = female	-0.327*** (0.075)	-1.767*** (0.419)	-0.291*** (0.078)	-1.786*** (0.421)	-0.262*** (0.076)	-1.417*** (0.419)	-0.2288** (0.079)	-1.436*** (0.421)	-0.4743 (0.327)	-3.476 (2.492)
Advisors gender = female	0.009 (0.129)	-0.049 (0.700)	0.311 (0.245)	0.728 (0.885)	-0.036 (0.127)	-0.198 (0.686)	0.268 (0.239)	0.526 (0.854)	-0.988* (0.418)	-7.247* (3.341)
Advisors gender = female * PhD student = female			-0.4646 (0.289)	-			-0.4399 (0.2831)	-		
Difference to dissertation year advisor					-0.006 (0.0037)	-0.032 (0.020)	0.0060 (0.003)	-0.033 (0.020)	-0.0439* (0.019)	-0.0322* (0.157)
PhD advisor characteristics										
Advisor productivity									0.0053* (0.002)	0.0396* (0.018)
Discipline dummies	NO		NO		YES		YES		YES	
Year dummies	NO		NO		YES		YES		YES	
Observations	961		961		961		961		59	

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

Source: Own data and depiction.

Figure 17 shows the Kaplan Meier survival curve for the four different protégé-advisor constellations and a table that shows the number of graduates at risk of leaving academia each year after PhD completion ($t = 0$). I observe a substantial decline in the number of persons at risk in the first two years. The dominant share of scientists, therefore, do not stay in academia after PhD. P = 0.21 indicates the log-rank test result and indicates that the time to the final publication is statistically not different between the four groups.

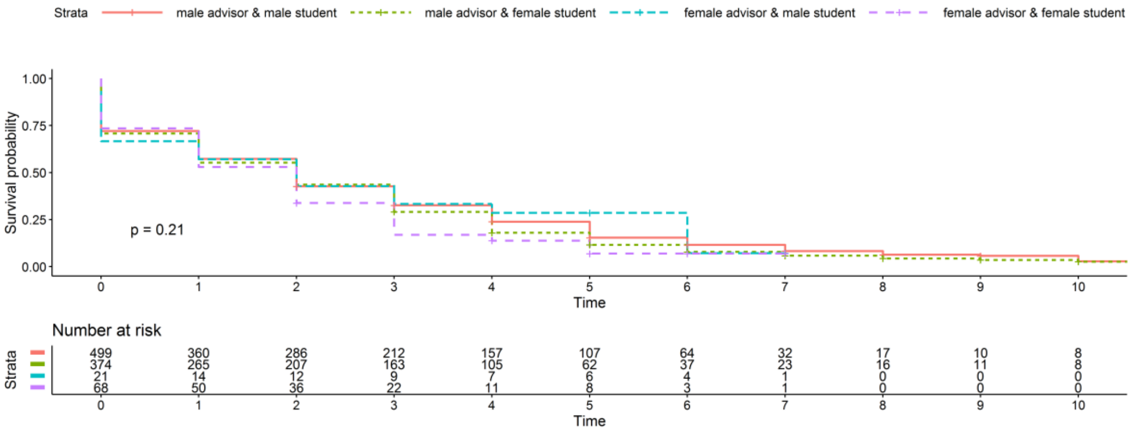


Figure 17. Kaplan Meier Curve and risk table for final publication after PhD completion

Source: Own data and depiction.

Table 20 shows the Complementary log-log regression results. The hazard for women to exit from research is, according to Model 1, 37% higher than for men. A female advisor is generally beneficial and leads to a 38% lower hazard of exit from research. Model 3 includes year and discipline dummies; it shows that women's higher hazard remains robust when including these dummies. Models 2 and 4 display the interaction of advisor gender and protégé gender. I find no statistically significant effect for the interaction.

Table 20. Complementary log-log model – Yearly hazard of exit from research

	Dependent variable:							
	Yearly hazard of writing last paper after PhD							
	(1)		(2)		(3)		(4)	
	Coef.	Exp (Coef.)	Coeff	Exp (Coef.)	Coef.	Exp (Coef.)	Coef.	Exp (Coef.)
PhD student characteristics								
Gender PhD student=female	0.315*** (0.051)	1.37	0.335*** (0.054)	1.40	0.321*** (0.054)	1.38	0.343*** (0.056)	1.41
Difference to dissertation year advisor					-0.007** (0.002)	0.99	-0.007** (0.002)	0.99
Number of papers during PhD					-0.008** (0.003)	0.99	-0.008* (0.003)	0.99
PhD advisor characteristics								
Gender of advisor	-0.474*** (0.087)	0.62	-0.293* (0.156)	0.75	-0.481*** (0.090)	0.62	-0.264 (0.163)	0.77
Gender Interaction								
*Gender PhD student *advisor gender=female:*			-0.2935 (0.156)	0.77			-0.3023 (0.196)	0.74
Intercept	1.680*** (0.063)	5.37	1.678*** (0.063)	5.36	2.006*** (0.104)	7.44	2.004*** (0.104)	7.42
Baseline risk	-0.196*** (0.009)	0.82	-0.197*** (0.009)	0.82	-0.203*** (0.010)	0.82	-0.204*** (0.010)	0.81
Discipline dummies	NO		NO		YES		YES	
Year dummies	NO		NO		YES		YES	
Observations	3324		3324		3321		3321	
Chi2	547.08		548.84		589.85		592.05	
Pseudo R2 Mc Fadden	0.19		0.19		0.21		0.21	
AIC	2334.14		2334.37		2310.34		2310.15	

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

Source: Own data and calculations.

5.6 Discussion and conclusion

In this paper, I have investigated the relation of advisor gender on protégé productivity during PhD study and on academic survival after PhD completion. In my investigation of productivity, I aimed to find if early productivity differentials are related to advisor gender. Using negative-binomial regression, I do not find any such relation. However, I find that female PhD students publish about 1.4 papers less than male students during their PhD study. My data cannot explain if this pattern is caused by other advisor and PhD student characteristics. For instance, it is still unknown if advisor productivity, quality, and protégé-advisor collaboration play a role.

My second outcome, the time between PhD completion and final WoS publication, addresses academic survival after earning a PhD. The time until the exit from research is one of the main contributions of my approach. Unlike in previous literature, which concentrates on examining outcomes at fixed points, I fully utilize the temporal dimension after PhD completion. My application of time-discrete complementary log-log regression delivers two main findings on advisor and protégé gender.

First, I find that female advisors positively affect academic survival of PhD graduates, leading to a 38% lower exit hazard. The causal mechanisms underlying this result remain unclear and are beyond the focus of my study. The female advisor effect is statistically not different between male and female PhD students. My results contrast with previous findings, such as those of Gaule and Piacentini (2018), who find that same-gender protégé-advisor pairings increase the likelihood of PhD graduates becoming university faculty. The reason for my lack of result may be a problem of low statistical power. PhD graduates advised by women make up only 89 observations in my dataset, and the effect just fails to reach the 10% statistical significance level. To obtain more observations from female advisors, one can improve the advisor scraping and linking strategy in the future. An explanation of the low number of female PhD graduates and female advisors could also be attributed to name changes after marriage. My approach does not account for women who marry and change their family name after their doctorate. Since they are then no longer observable, they are considered to have had their final publication. This problem could be solved by bibliographic coupling. If I no longer observe publications from the focal female scientist, and if a previous coauthor of hers publishes together with a person of the same first name, but different last name, this may indicate that the focal scientist has changed her name. The positive female advisor effect may be conditional to unaddressed advisor characteristics. I did not control for team characteristics, informal advisors, graduate school characteristics, funding characteristics, socioeconomic background, and many others.

The second main result I find is that women are 37% more likely per year than men to exit from research after completing a PhD. In light of the female underrepresentation, this finding is unsurprising. Nevertheless, it adds to the literature by quantifying the female dropout risk for the first time. Concerning the name changes mentioned previously, this result needs further robustness analysis. There are other factors unaddressed in my study that lead to female exit from research. In particular, the effect of private events and factors, such as childbirth and motherhood, lead to an omitted variable bias in my study. It is also unclear whether there are differences in the preference for careers in science or industry between men and women after completing their PhD, which could explain the female departure from academia.

My results potentially suffer from self-selection bias. Therefore, the initial matching process between protegees and advisors may not be random. Descriptively, the disproportionately high number of women advised by women may indicate such bias. In further study this bias can be addressed by using more advanced econometric setups (e.g., Azoulay et al., 2017).

Finally, my study is limited to young German scientists in disciplines where publication in WoS index journals is traditionally dominant. Therefore, I miss arts, humanities, and parts of social science where journal publications are not (or have not been) popular. In these disciplines, women also make up higher shares of PhD graduates, potentially leading to different productivity and survival patterns.

I conclude that female underrepresentation and its relation to advisor gender in academia is a complex empirical phenomenon of high scientific and societal relevance. My study contributed to the literature by disentangling the temporal and productivity dimensions before and after completion of a PhD. The survival operationalization offered a new perspective on female underrepresentation and is promising for application in various other questions in scientometrics and higher education research.

6 Career paths of PhD graduates in eastern and western Germany: Same qualification, same labor market outcomes?

6.1 Preface

This chapter builds on two papers. The major paper is currently under major revision at the journal *Education Economics* and has been published as working paper under: Fuchs, M., & Rehs, A. (2020). Career paths of PhD holders in eastern and western Germany: Same qualification, same labor market outcomes? *IAB-Discussion Paper, 2020*(1). The second paper concerns descriptive findings presented in this chapter and is addressed to a non-professional audience. It is published as: Fuchs, M., & Rehs, A. (2019). Erwerbsbiographien ost-und westdeutscher Promovierter nach der Wiedervereinigung: Gleiche Qualifikation, gleiche Karriereverläufe? *ifo Dresden berichtet*, *26*(06), 17-22. For reasons of consistency, I write this chapter in singular form.

6.2 Introduction

Today’s knowledge economy strongly depends on capacities for innovation, creating knowledge and solving complex problems. These capacities are associated with PhD graduates, who play a prominent role in fostering economic development and growth (Auriol et al., 2013; Stephan et al., 2004). A crucial issue in this respect is whether they are able to fully exploit their investment in education in their subsequent jobs, or whether they are at risk of mismatch on the labor market. Overeducation in the form of a level of education that exceeds the requirements for the current job has costly consequences for individuals, firms and the economy as a whole (Carroll & Tani, 2013; McGuinness, 2006). For the PhD graduates themselves, part of their investment in education is unproductive, which translates into lower returns on investment in the form of employment below their skill level and lower wages. There are diverse reasons for PhD graduates not fully reaping the returns to their education and they have not yet been exhaustively investigated (Di Paolo & Mañé, 2016; Engelage & Schubert, 2009; Steeg, Wiel, & Wouterse, 2014). Findings on the labor market performance of PhD graduates and on the obstacles they face in using their abilities are therefore highly relevant not only for the individuals themselves, when considering their subsequent career paths, but also for policy makers and governments that finance the education of this group and support their integration into the innovation system (Auer, Fichtl, Hener, Piopiunik, & Rainer, 2017; Auriol et al., 2013).

From a sociological perspective, PhD graduates belong to a country’s educational and economic elite, holding top positions in academic, economic, political or cultural spheres, while representing certain values and attitudes (Dahrendorf, 1965; Dee et al., 2004;

Hartmann & Kopp, 2001). For Germany, this is even more the case than in other countries, as a PhD is not only a prerequisite for a scientific career, but is also associated with a high reputation and appreciation outside academia. Moreover, in more general terms, a high level of human capital such as that acquired by PhD graduates can generate positive externalities for the general public by strengthening social cohesion and political participation in a democracy (Auer et al., 2017). Hence, any factors that diminish PhD graduates' returns to education may lead not only to adverse consequences for the individuals concerned, such as inadequate jobs and wages, but also to significant societal repercussions.

Focusing on regional background as an inhibiting factor, eastern Germany constitutes an especially intriguing case. Unlike in other Central and eastern European transformation economies, the incorporation of the former German Democratic Republic into the western democracy and market economy was undertaken very rapidly, with western German institutions being extended to and implemented in the new eastern part of Germany (Salheiser, 2012, 123). As a result, a considerable number of the old eastern German elites were replaced by western Germans, which went hand in hand with the breakdown of the old Socialist elite recruitment regime (Best, 2005; Geißler, 2014). This profound exchange of elites continues to have an effect today. Bluhm and Jacobs (2016, p. 30) note that eastern Germans occupy only 2% of the top positions in Germany, although eastern Germany accounts for 17% of the whole population. In eastern German public discourse, the underrepresentation of eastern Germans in top positions and the consequences for social and political coherence have frequently been the topic of lively discussions (e.g., Lukas & Reinhard, 2016) indicating that the transformation process in eastern Germany is still in progress. In the light of the ongoing public debates, it is surprising that there is very little representative empirical evidence on the underrepresentation of eastern Germans in top positions in Germany.

Against this background, this paper investigates whether having an eastern or western German background has an impact on whether or not PhD graduates are able to fully capture the returns on their education. It is unclear whether being from eastern Germany plays an important role for the employment trajectories of highly educated individuals, since the processes of acquiring social and cultural capital changed dramatically for eastern Germans in the course of reunification (Salheiser, 2012). I trace the employment trajectories of eastern and western German PhD graduates in order to analyze whether the eastern German graduates fare less well than their western German counterparts and whether this can be explained by their eastern German background. In order to exclude any detrimental effects that might arise from systematic differences between the doctoral education systems in the German Democratic Republic and the Federal Republic of Germany, I only consider individuals who completed their dissertation after 1994. I compare the two groups with respect to two main labor market outcomes, thereby contributing to related findings for PhD graduates (e.g., Auriol et al., 2013; Heinisch et al., 2020; Paolo & Mañé, 2016). First, I investigate whether an eastern German background is associated with a higher probability of being overeducated for the current job, taking up the conjecture that eastern German PhD graduates might be less likely than their western peers to work in jobs that fully exploit their human capital. Second, I examine whether an eastern German background is associated with a lower probability of achieving high wages as compared to a western German background. Hereby I take into account the persisting labor market differences between eastern and western Germany that specifically concern wages (Schnabel, 2016). To differentiate between an eastern or western German background I use the place of birth as the most straightforward measure. Since the place of birth could be overshadowed by the location of the university where the PhD was completed or the subsequent place of work, I additionally consider these two measures.

My analysis is based on a novel data set developed by (Heinisch et al., 2020) in order to follow the labor market biographies of German PhD graduates. It combines data on PhD graduates collected in the catalog of the DNB with information on their labor market biographies from the Integrated Employment Biographies of the Institute for Employment Research. This data set is then supplemented by information on the PhD graduates' places of birth, as recorded in their dissertations. My data set comprises individuals who completed their dissertations between 1995 and 2010 and their labor market outcomes for the subsequent five years. I apply logit models to assess whether an eastern German background significantly lowers the PhD graduates' probability of finding employment and earning wages that are in line with their skill level.

The results reveal no significant negative impact on labor market success either for a birthplace in eastern Germany or for a dissertation submitted to an eastern German university. In that respect, the same qualification level results in the same labor market outcomes. It is more the place of work that matters, which indicates the profound impact of the still divergent economic conditions in the two parts of Germany on PhD graduates' employment prospects. In particular, a place of work in eastern Germany substantially reduces the chances of achieving high wages. This result is confirmed when the different regional differentiations are controlled for.

The remainder of the paper is structured as follows. In section 6.3, the background on overeducation among PhD graduates and related empirical findings is discussed. Section 6.4 introduces the data used for my analysis, along with measurement issues. Descriptive evidence together with the regression results are the focus of section 6.5. The last section draws conclusions.

6.3 Overeducation among PhD graduates

Labor market mismatch and its consequences for career mobility and wages have been investigated extensively in education and labor market research (Leuven & Oosterbek, 2011; McGuinness, 2006). Due to growing numbers of higher education graduates in many countries, increasing attention has been paid to the educational attainment of PhD graduates as a special subgroup of graduates in recent years (Auriol et al., 2013). While a large body of literature deals with overeducation among graduates and highly qualified labor market participants (e.g., Carroll & Tani, 2013; Dolton & Silles, 2008; Dolton & Vignoles, 2000; Rossen, Boll, & Wolf, 2019) , empirical evidence on the labor market performance of PhD graduates has been expanding in recent years, but still leaves many research questions unanswered.

Several studies investigate the existence and consequences of a labor market mismatch for PhD graduates in specific countries, all reaching similar conclusions. Bender and Heywood (2011) examine the degree of mismatch between education and the current job among a panel of US PhD graduates. Their results show that mismatch is more likely late in careers, which is consistent with mismatch resulting from a certain evolution of the professional employment trajectory. In their study on Swiss PhD graduates, Engelage and Schubert (2009) further emphasize the role of the field of study for obtaining an adequate job. Focusing on overeducation and overskilling among Italian PhD graduates, Gaeta (2015) confirms the importance of the field of study and of job-related characteristics as conditioning factors of both forms of mismatch. Likewise, for a cohort of Spanish PhD students Di Paolo and Mañé (2016) find that many of them face involuntary mismatch accompanied by significant penalties in terms of job satisfaction and earnings. The negative impact of labor market mismatch on wages is corroborated by (Bender & Heywood, 2009) for PhD graduates in the US. Relatedly, Canal Domínguez and Rodríguez Gutiérrez (2013) study wage differences among Spanish PhD graduates and confirm that working in a job that requires higher education levels is associated with higher earnings. Steeg et al. (2014) investigate the private returns to obtaining a PhD in the Netherlands. They compare wages earned by PhD graduates to those earned by master's graduates over the first 20 years of their careers and find an average annual return to a PhD education of 6% over the entire career.

For Germany, empirical findings concerning PhD graduates' wages are provided by Heineck and Matthes (2012). They compare PhD graduates to other university graduates with respect to wages and skill mismatch and find that the monetary returns to holding a PhD are significantly higher than those to just obtaining a university degree. Furthermore, monetary returns are higher in the private sector than in the public sector. Graduates holding a PhD regard their employment as more adequately suited to their skill level than university graduates. Similarly, Falk and Küpper (2013) find that PhD graduates' wages are about 7% higher than those of university graduates. However, the wage advantages strongly depend on the field of study, with engineers having the strongest advantages. Mertens and Röbken (2013) confirm the higher monetary returns for PhD graduates compared to master’s graduates especially in the fields of economics and law. To investigate the non-academic career prospects of postdocs in German academia, Koenig (2019) uses the same data set as I do, albeit without information on the individuals’ place of birth. His results indicate that a significant number of PhD graduates remain in academia after graduation. However, there is no general wage premium in the non-academic sector for employment as a postdoc.

To my knowledge, no studies on PhD graduates in Germany have so far addressed the origin of the PhD graduates with respect to eastern or western Germany. However, I can embed my analysis in research focusing on university graduates in more general terms. Rukwid (2012) compares the extent of overqualification among university graduates working in eastern and western Germany. As a general picture, in 2010 the risk of being overqualified was higher in eastern Germany, where 23%of the graduates were in jobs for which they were overqualified as compared to 18% in western Germany. The presentation of the extent of overqualification from 1990 onwards impressively illustrates the problems faced by eastern German graduates when trying to find employment suited to their skill level in the first years after reunification. The corresponding share of overskilled eastern German graduates rose to almost 32% in 2004. At the same time, the share of western German graduates also increased, but only to relatively moderate 20% in 2004 (Rukwid, 2012, p.36). The author puts these large differences down to the severe economic aftermath of German reunification, which led to structural unemployment in eastern Germany. Large numbers of graduates lost their jobs in liquidated stated-owned enterprises and were looking for new employment in the 1990s. In addition, university degrees obtained in the German Democratic Republic were not always accepted as equivalent to degrees obtained from western German universities.

The necessity to examine eastern and western Germany separately with regard to educational mismatch also becomes evident in (Boll, Leppin, & Schömann, 2016). The authors identify the reasons for overeducation according to different measurements and for different subgroups of graduates between 1992 and 2011. For eastern German graduates, the effect of previous unemployment is more pronounced than for their western German counterparts, and they are also more likely to have been exposed to involuntary job changes. This can be put down to the poor labor market prospects in the eastern part of the country during that period. A further central finding is that overeducation exhibits a pronounced path dependency: having been overeducated in the previous year significantly increases the risk of being overeducated at present. Whereas according to individual self-assessment the probability of being currently overeducated increases more for eastern Germans than for western Germans if they exhibited this status in the previous period, the differences between eastern and western German men are quite small when measured in statistical terms. Interestingly, however, state dependency among western German women is found to be more than twice as high as for their eastern German counterparts.

In their paper on the monetary returns to a PhD, Mertens and Röbken (2013) also consider the specific economic situation in eastern Germany by including a dummy variable in the wage regressions for a place of work in western Germany. It is positive and highly significant in most of the fields of study examined, which emphasizes the higher wages earned by both regular university graduates and doctorate graduates in the western part of the country.

Summing up, the empirical evidence on overeducation specifically for eastern and western Germany reveals a higher risk of overqualification and lower wages when working in the eastern part of the country. In the following, I aim to find out whether having an eastern German background in a broader sense than just the workplace leads to potential lower returns to education in the case of the PhD graduates.

6.4 Empirical setting

6.4.1 Data

In order to obtain information on PhD graduates and their employment biographies, I make use of several data sources. My basic data set comes from the IAB-INCHER project of earned doctorates (short: IIPED, see Heinisch et al., (2020) for more details). It combines information on dissertations that are contained in the electronic catalog of the DNB (Deutsche Nationalbibliothek or DNB) with the individual labor market history from the Integrated Employment Biographies (IEB) of the Institute for Employment Research (IAB). I further enrich this information by including the PhD graduates' birthplaces, which I obtained from the online dissertations.

As Germany's central archival library, the DNB collects, documents and archives all printed publications and sound recordings issued in Germany together with works that were compiled in the German language or relate to Germany (Deutsche Nationalbibliothek, 2019). Since PhD graduates are required by law to supply a copy of their dissertation, the DNB holds an almost complete set of dissertations submitted to German universities since the 1970s. The electronic catalog of the DNB features information on the authors, the university name, the year of publication and the subject and therefore constitutes a highly suitable data source for research on PhD graduates in Germany (e.g., Buenstorf & Geissler, 2014; Heinisch & Buenstorf, 2018).

One drawback of the DNB catalog, however, is that the PhD graduates' place and date of birth are very rarely reported. In order to retrieve this essential information, I made use of URL links to online dissertations listed in the DNB catalog. In many faculties, PhD students are required to report their place and date of birth as well as the date of the examination on the front page of their dissertation.[27] However, not all universities have (working) URL links to downloadable dissertations in the DNB database. I therefore resorted to the individual university servers as a second strategy and systematically searched them for online dissertations.[28] These were matched with the dissertations in the DNB catalog via the author's name, the university name and the year in which the dissertation was submitted.[29] This yielded a total of 79,000 dissertations from the two data sources for which I know the unique identifier in the DNB catalog.

My variable of interest, a PhD graduate's birthplace, was retrieved by means of a text pattern matching approach. Typical keywords on front pages or curriculum vitae, like "place of birth", indicate the subsequent mention of a birthplace or other information of interest. In English dissertations I systematically searched for the words "born in:", "birthplace" and others. For dissertations in German I repeated this procedure with corresponding German terms.[30] I automatically searched for these keywords on the front pages or in the curriculum vitae of every dissertation from my two first data sources and saved the three subsequent words. In the next step, me and my colleague cleansed the resulting string manually of frequent errors and entered it into the Google Maps search engine in order to obtain a unique address and more general information such as country, state and zipcode for each birthplace. The Google search engine has the advantage that it takes into account diverse spellings and ambiguous German city names.[31] I was able to identify the birthplaces of 27,321 German PhD graduates with this procedure.

In the IIPED project, the data on the PhD graduates were merged with information on the graduates' labor market performance from the Integrated Employment Biographies (IEB) of the IAB.[32] The IEB contain information on employment spells, benefit receipt, participation in measures of active labor market policy, and job-search status for every person on a daily basis. Because they are not covered by the social security system, civil

[27] Sometimes the dissertations also include a curriculum vitae.

[28] These servers cover the full set of online dissertations (as of August 2017) from the universities of Kassel, Munich (TU and LMU), Braunschweig, Freiburg, Frankfurt/Main, Greifswald, Darmstadt, Düsseldorf, HU Berlin, Halle-Wittenberg, Magdeburg, Regensburg, Rostock, Ulm, all universities in Saxony and Thuringia, and the Karlsruher Institut für Technologie.

[29] I used a fuzzy-string matching procedure based on the Levinstein distance for the author's name and allowed a time window of 2 years before and after the date of the dissertation to compare the year of submission to the DNB with the years stated on the university server website. This is necessary because the two dates do not necessarily coincide. To correct mismatches, in the name matching procedure I additionally checked whether the matched name appears on the front page of the dissertation.

[30] The German expressions are "geb. in", "geboren", "aus" and "Geburtsort" and further variations of these terms.

[31] Since some German town names occur more than once in Germany, the nearby river is added to their names in order to avoid confusion. However, the attachment of the river is not used consistently, for example Halle/Saale, Halle a. d. Saale, Halle Saale and so on.

[32] For more detailed information on the IEB see Antoni, Ganzer, and Vom Berge, (2016), who provide a description of the Sample of the Integrated Labour Market Biographies based on a 2% random sample of the IEB.

servants, self-employed persons, family workers and PhD students financed solely by scholarships are not contained in the IEB. In total, the IEB covers about 80 % of the German workforce. The data are available from 1975 onwards for western Germany and from 1993 onwards for eastern Germany. For each individual, the IEB contains a range of sociodemographic characteristics (e.g., sex, date of birth, nationality, qualification level, place of residence) and job features (type of employment, occupation, industry affiliation, place of work). Although the qualification level covers vocational training or bachelor's and master's degrees, there is no information on PhDs. Consequently, it is necessary to match this with the DNB data, which includes that information, in order to trace the labor market biographies of German PhD graduates.

From the matched data set, I select only PhD graduates who were born in Germany and whose dissertation was completed between 1995 and 2010. I set the beginning of my observation period at 1995 because good coverage of online dissertations and thus birthplaces only exist from the middle to the late 1990s onwards. In addition, for most disciplines the starting date of 1995 is justified as it represents the first cohort of PhD graduates who began their dissertation in reunified Germany. Considering earlier cohorts would inevitably also include PhD graduates who began their dissertation in the German Democratic Republic, which is not the focus of my study. I then trace their labor market performance for five years after they earned their PhD. A five-year period has been established as a good predictor of future wages in the literature. Karahan, Guvenen, Ozkan, and Song (2015) find that for US employees the bulk of earnings growth happens between the age of 25 and 35. This is especially the case for lifetime incomes in the upper percentiles of the distribution, where I expect to find doctorate graduates. Since the graduates' mean age at the time when their labor market outcome is observed is roughly 37 years (see Appendix D), I should accordingly have a good approximation of the lifetime labor wages in t+5. An additional investigation of other points in time, like t+10 and t+15, would also reduce the number of available cohorts in my data set. An additional argument pertains to the pervasiveness of fixed-term contracts in the early career stage and the postdoc phase that lasts about five years (Auriol et al., 2013). Afterwards, PhD graduates should be employed in jobs that are related to their doctoral degree. Since the DNB-IEB matching process is cut after 2015 due to the challenges involved in processing and matching the data as described in (Heinisch et al., 2020), 2010 is the last available cohort of PhD graduates. Thus, my final sample only comprises PhD graduates who gained their PhD between 1995 and 2010 and for whom I have labor market information for five years after they obtained their PhD. It includes 2,902 persons in total, 670 of whom were born in eastern Germany, 2,088 in western Germany and 144 in Berlin.

6.4.2 Main variables

I measure the labor market performance of the eastern and western German PhD graduates on the basis of two outcomes that capture the returns to education. First, I measure the potential formal overeducation due to being eastern German based on the skill level required for the occupation. This indicator is contained in the German Classification of Occupations (KldB 2010) and depicts the various degrees of complexity within those occupations which have a high similarity of occupational expertise (Paulus & Matthes, 2013, p. 9).[33] The complexity of an occupation is captured by four requirement levels that range from unskilled (low skill), specialist (medium skills) and complex specialist activities (specialist skills) to highly complex activities (expert skills). It is assumed that a certain standard of skills, abilities and knowledge must exist for practicing a certain occupation. In the case of highly complex activities, the required formal qualification encompasses university studies lasting at least four years or relevant professional experience. Corresponding jobs are typically found in research and development, teaching or on the executive boards of medium-sized and large companies. PhD graduates can therefore be regarded as being employed in line with their skill level when they work in jobs involving highly complex activities, i.e. when they are employed as experts. This indicator has regularly been used to measure formal overeducation on the basis of German administrative data (Kracke, Reichelt, & Vicari, 2018; Stüber, 2016). I encode the outcome as a dichotomous variable that is equal to one if the individual works in a job that involves highly complex activities five years after earning a PhD, and is equal to zero otherwise.

The second outcome relates to a potential wage penalty among the PhD graduates for being eastern German. To measure this, I use the nominal daily wages reported in the IEB. A general restriction here, however, is that in the IEB wages are only recorded up to the social security contribution assessment ceiling in Germany.[34] Since PhD graduates can be expected to earn wages in excess of this assessment ceiling, I construct a dichotomous variable that is equal to one if the PhD holder earns wages exceeding the inflation-adjusted social security contributions assessment ceiling in year five after earning their PhD.[35] Throughout the analysis, I only consider persons in full-time employment, because the German social security data do not contain information on the exact number of hours worked, which would be necessary to compute hourly wages.

My central explanatory variable of interest concerns the PhD graduates' regional origin, i.e. eastern or western Germany. The most straightforward differentiation is based on the place of birth. I use a dichotomous variable *birthplace_east*, which takes on the value of one if the individual was born in eastern Germany and zero in the case of western

[33] The KldB 2010 is a five-digit classification that contains two dimensions: occupational expertise is encoded in the first four digits, and the requirement level in the fifth digit.

[34] For example, in 2009 this was 157.81 euros/day in eastern Germany and 180.82 euros/day in western Germany.

[35] In 2003, there was an extraordinary sharp increase in the contribution assessment ceiling, which is taken into account in my subsequent procedure.

Germany. Since the place of birth may have a different impact on labor market outcomes in the two parts of the country and may be contorted by the individual working in eastern or western Germany, I include the place of work as a second regional distinction. The labor markets in the two parts of the country still differ in many respects due to the ongoing transformation process in eastern Germany, which is characterized by a generally higher extent of overqualification and lower wages (Fuchs, Rauscher, & Weyh, 2014; Reichelt & Vicari, 2014; Schnabel, 2016). The dichotomous variable *workplace_east* is equal to 1 when the place of work is in eastern Germany. Because Berlin constitutes a separate regional unit in the dichotomy of eastern/western Germany, PhD graduates born in Berlin are regarded as neither eastern nor western German, but are investigated separately throughout. However, I include a workplace in Berlin (*workplace_berlin*) as a separate regional distinction in order to identify the labor market effects of what is eastern Germany's largest city as well as the capital city of Germany. It is again encoded as a dichotomous variable.

A third dimension of regional origin pertains to the location of the university where the PhD was earned. I include the dichotomous variable *university_east*, that is one if the respective university is located in eastern Germany and zero in the case of western Germany in order to capture potential self-selection mechanisms in the choice of university. Since eastern German universities lag slightly behind their western German counterparts with regard to scientific productivity and recognition (Schmoch & Schulze, 2010), promising PhD candidates from both parts of the country may be more likely to take up doctoral studies in western Germany. Furthermore, the different funding structures, especially from industry (Pasternack, 2007), as well as differing research field focuses (Schmoch & Schulze, 2010) could account for selection effects. However, at the same time, research funding levels and personnel capacities in eastern German universities are similar or even higher than those in their western German counterparts (Pasternack, 2007). This would be a reason for selecting eastern German universities.

6.4.3 Control variables

In order to control for further determinants of adequate employment and wages, I consider additional groups of variables. The first group comprises individual characteristics of the PhD graduates. Age effects are covered by age in years and age squared to take any nonlinearities into account. Gender is included as a dichotomous variable that is equal to one for a female PhD graduate. Since prior work experience also impacts on subsequent labor market success, I construct a continuous variable that cumulates all employment episodes up to one year before the dissertation was published. Another important factor when conducting analysis at the small-scale regional level relates to the individuals' spatial mobility after graduation. If they look for work in regional labor markets rather than global ones, their access to suitable employment might be severely restricted (Büchel & van Ham, 2003). This is especially the case in small and rural labor markets, of which there is a disproportionately large number in eastern Germany (Granato, Haas, Hamann, & Niebuhr, 2010). Hence, mobile PhD graduates have better chances of avoiding skill mismatch if they seek employment elsewhere. I take mobility after graduation into account with a dichotomous variable. A PhD graduate is considered mobile if the location of the university where he or she completed the dissertation is in a different planning region[36] to the place of work five years later.

The second group of control variables concerns job characteristics. Since wages vary significantly between sectors,[37] I control for sectoral affiliation by considering nine economic sectors ranging from agriculture, forestry and horticulture to humanities, culture, arts and media. For analyzing the wage level only, I also include the four skill requirement levels for the job (unskilled, specialist, complex and highly complex activities). Since the regional area can also have an impact on remuneration, I further differentiate between the three broad region types of urban agglomerations, urbanized and rural regions.[38]

The third group of control variables refers to the scientific discipline in which the PhD graduate wrote his or her dissertation. The field of study strongly determines the future wages of university graduates (Grave & Goerlitz, 2012). In order to uncover any east-west specific selection effects, I encoded 17 different field dummies from the subject classification for each dissertation contained in the DNB. They include natural sciences, literature and linguistics, and economics and business. I excluded dissertations in the field of medicine, because they would account for the majority of observations and have a very particular labor market situation in Germany. Last, I take into account year dummies in order to control for a general time trend. Appendix C contains detailed definitions of all variables, and Appendix D provides descriptive statistics for the dependent and explanatory variables.

6.5 Results

6.5.1 Descriptive evidence

My final sample comprises 2,758 PhD graduates that are traced for five years after earning their PhD. Concerning eastern German backgrounds, I observe 670 PhD graduates born in eastern Germany, 637 graduates working in eastern Germany, and 918 persons who gained their PhD from an eastern German university. My coverage of PhD cohorts and

[36] See https://www.bbsr.bund.de/BBSR/DE/Raumbeobachtung/Raumabgrenzungen/deutschland/regionen/Raumordnungsregionen/raumordnungsregionen-node.html for further details (accessed 30.11.2019).

[37] Wages also vary significantly between occupations. Since many occupations are concentrated in just a few sectors, I only consider sectors in order to avoid multicollinearity.

[38] See https://www.bbsr.bund.de/BBSR/DE/Raumbeobachtung/Raumabgrenzungen/deutschland/kreise/Kreistypen4/kreistypen_node.html (accessed 03.08.2019).

their labor market outcomes in t+5 improves in the late 2000s. This is due to the improved availability of dissertations online. Regarding my outcomes, 2,016 persons have an expert job status five years after earning their PhD, and 1,051 have an income above the social security contribution assessment ceiling.

Figure 18 depicts the spatial distribution of the birthplaces and university locations of the PhD graduates in my sample. As can be expected from the spatial distribution of the population, many of the birthplaces are located in typical agglomerations, such as the Rhine/Main region in western Germany and Berlin in eastern Germany. When the location of the university is differentiated according to the place of birth, my data suggest that both eastern and western Germans tend to opt for universities in the part of Germany where they were born. Native eastern German graduates predominantly attended universities in the federal states of Saxony and Thuringia and in Berlin. Some native western German graduates also enrolled in these universities.

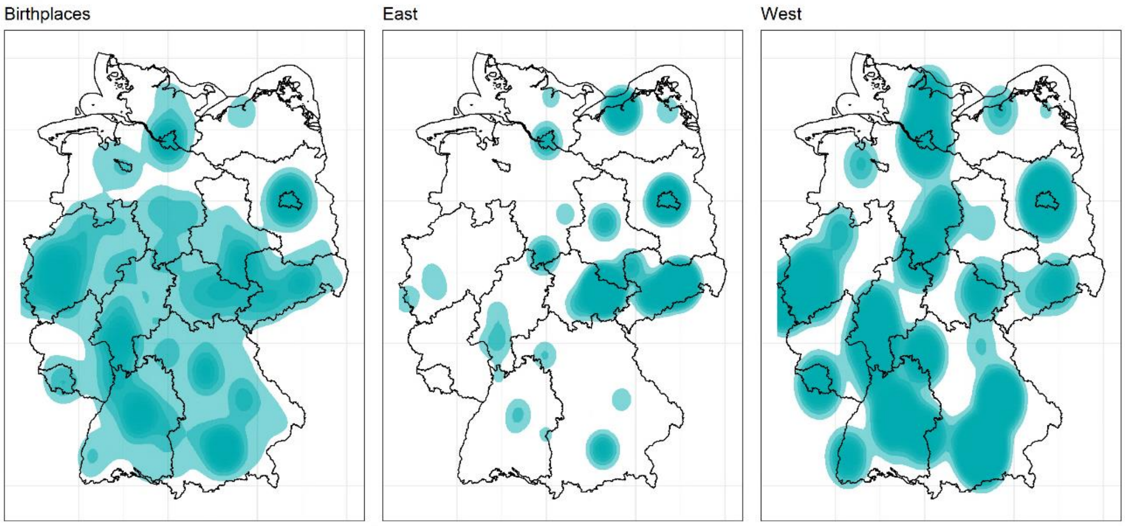
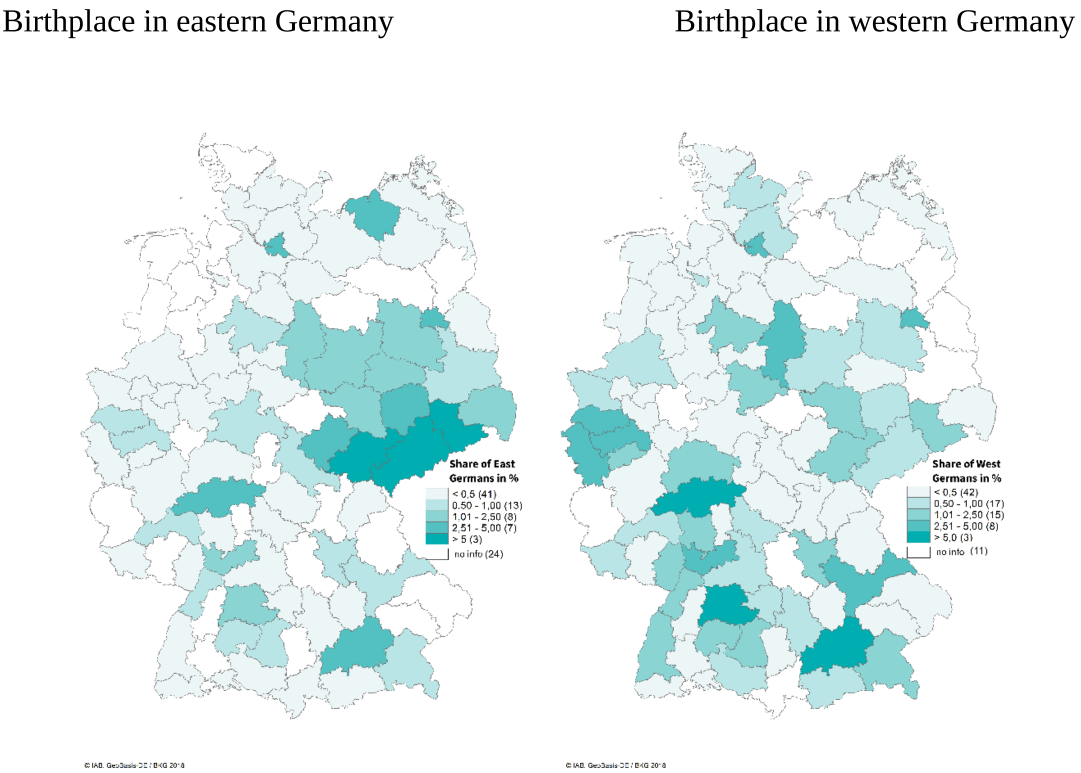


Figure 18. Spatial distribution of the PhD graduates` birthplaces (left) and location of their PhD universities, by birthplace in eastern of western Germany (right)

Source: IIPED data set, own birthplace data from online dissertations (geo-referenced by Google maps); own compilation.

Similarly, this also holds for the places of work five years after gaining a PhD, as depicted in Figure 19 – eastern Germans largely remain in eastern German regions and western Germans largely remain in western German regions. This is consistent with empirical evidence on the internal migration of graduates, which finds that the longer the graduates stay in the region of their university, the less likely they are to leave afterwards (Busch & Weigert, 2010; see also Teichert, Niebuhr, Otto, & Rossen, 2018). However, some features are noteworthy. The PhD graduates born in western Germany tend to work in the large agglomerations of the Rhine/Main area around Frankfurt, the greater Stuttgart area and the greater Munich area. In slight contrast, the workplaces of the PhD graduates born in eastern Germany tend to be concentrated in the southern parts of Saxony and Thuringia rather than in Berlin, which is eastern Germany’s largest agglomeration.



Birthplace in eastern Germany

Birthplace in western Germany

Figure 19. Workplace according to planning regions five years after dissertation, by birthplace in eastern and western Germany

Note: Shares of eastern/western German PhD graduates in relation to all eastern/western German PhD graduates in the sample.

Source: IIPED data set, own birthplace data from online dissertations (geo-referenced by Google maps); own compilation.

Regarding my two main labor market outcomes, obtaining an expert job and earning a wage above the social security contribution assessment ceiling, descriptive evidence shows considerable differences between graduates with eastern and western German backgrounds, especially with regard to the second variable. 40.8% of the PhD graduates born in western Germany, but only 30.0% of those born in eastern Germany earn wages above the contribution assessment ceiling five years after completing their PhD. However, this may be mainly associated with the current workplace and not so much with the birthplace. The PhD graduates in my sample that work in eastern Germany exceed the contribution assessment ceiling in only 23.5% of the cases, while in western Germany this is the case for 43%. This difference can be explained by the profound wage disparities that still exist between the two parts of Germany (Fuchs et al., 2014; Schnabel, 2016). Since eastern German PhD graduates generally remain in their own part of Germany rather than moving to western Germany (see Figure 19), they cannot benefit from the higher western German wages to the same extent as their western German counterparts.

The group differences in the first labor market outcome, relating to an expert job status, are not so pronounced. The shares of eastern and western German PhD graduates holding such a job are almost identical (72.5% and 73.8% respectively). Differentiating by a place of work in eastern or western Germany does not change the picture (72.6% and 73.8% respectively). Appendix D provides further information on the distribution of the PhD graduates across age groups, work experience, the sector of the economy, and the discipline in which the PhD was earned.

6.5.2 Econometric results

I now turn to econometric techniques in order to test my conjectures regarding the labor market outcomes of eastern German PhD graduates in a multivariate setting. Using a logit model, I estimate whether having an eastern German background has a statistically significant negative impact on the probability of (1) obtaining an expert job and (2) achieving wages above the social security contribution assessment ceiling five years after gaining the PhD. The general specification of the logit model is given by:

$$\pi_i = P(Y_i = 1\,|x_{i1}, \dots, x_{ik}) = F(\eta_i) = \frac{\exp(\eta_i)}{1-\exp(\eta_i)} \tag{1}$$

$$\eta_i = \alpha + \beta_1 birthplace\_east_i + \beta_2 workplace\_east_i + \beta_3 university\_east_i + \beta_4 control_i \tag{2}$$

In this specification, *birthplace_east*$_i$ denotes the place of birth, *workplace_east*$_i$ denotes the place of work, and *university_east*$_i$ denotes the location of the university where individual *i* gained his or her PhD. All three variables denoting the regional origin are constructed as dichotomous variables with the value of one for eastern Germany. Control variables are contained in *control*$_i$ and include individual, job-related and scientific characteristics as well as a time trend as described in section 6.4.3.

Depending on the model, π_i denotes either the probability of currently having a job with the highest skill requirement level (expert) or the likelihood of earning wages that are above the social security contribution assessment ceiling. Robust standard errors are estimated throughout. As the sign, magnitude and significance level of regression coefficients in non-linear models can often be misleading and thus lead to false conclusions, especially concerning interaction terms (Ai & Norton, 2003), I calculate average marginal effects and predicted margins for the covariates of interest.

Table 21 shows the results for the regional background variables when the PhD graduate has an expert job in t+5 (full results can be found in Appendix E). In the model the marginal effects of an eastern German background show a statistically insignificant impact on the likelihood of achieving an expert job status. This holds for all three delineations of the regional background as well as for the separate consideration of a place of work in Berlin.

Predictive margins for *birthplace_east* at different levels of *workplace_east* are shown in Table 22. When estimating the average predictive margins, I compute the change in the probability of having an expert job in t+5 when *workplace_east* remains fixed at 0/1 and *birthplace_east* changes for each observation to 0/1. Holding all other variables constant, the results in Table 22 show probabilities for the combinations that are of similar magnitudes to those in Table 21. The probability of native western Germans holding an expert job in t+5 when working in western Germany is 72.8%, while the corresponding value for native eastern Germans working in western Germany is 73.5%. For western Germans working in the eastern part of the country, the probability is 75.0% and for eastern Germans working in eastern Germany is it 75%. The overlapping confidence intervals indicate that there are no differences between the respective margins at the levels of *workplace_east*. Therefore, I conclude that an eastern German background in terms of birthplace and place of work has no impact on whether or not the PhD graduate achieves an expert job status in t+5.

Table 21. Selected average marginal effects for holding an expert job in t+5

Variable	dy/dx	Std. err.	z-score	p- value	95% conf. interval
			Basic model		
birthplace_east	0.069	0.023	0.30	0.76	-0.039-0.052
workplace_east	0.023	0.025	0.92	0.36	-0.026-0.073
workplace_berlin	-0.007	0.042	-0.16	0.87	-0.090-0.076
university_east	-0.033	0.043	-1.48	0.14	-0.079-0.011

Note: Delta method, Model VCE: robust, dy/dx for factor levels is the discrete change from the base level. */**/*** indicates statistical significance at the 10/5/1% level, respectively.

Sources: IIPED data set, own research; own calculation.

Table 22. Average predictive margins for birthplace_east at different values of workplace_east (holding an expert job in t+5)

	Margin	Std. err.	z-score	p-value	95% conf. interval
	Basic model				
birthplace_east = 0 at workplace_east = 0	0.728***	0.010	72.21	0.00	0.708-0.747
birthplace_east = 1 at workplace_east = 0	0.735***	0.021	35-16	0.00	0.694-0.776
birthplace_east = 0 at workplace_east = 1	0.750***	0.022	34.07	0.00	0.707-0.794
birthplace_east = 1 at workplace_east = 1	0.757***	0.021	34.49	0.00	0.714-0.800

Note: Delta method, Model VCE: robust. */**/*** indicates statistical significance at the 10/5/1% level, respectively.

Sources: IIPED data set, own research; own calculation.

Table 23 shows selected average marginal effects for achieving wages that exceed the social security contribution assessment ceiling in t+5 (full results can be found in the Appendix F). The coefficient for an eastern German birthplace is insignificant, which does not suggest any influence of an eastern German origin. However, an eastern German place of work seems to be decisive. It leads to a probability of achieving wages above the contribution assessment ceiling that is 20 percentage points lower than is the case for a place of work in western Germany.[39] This result is in line with the findings obtained by Mertens and Röbken (2013), who find that a western German place of work has a positive and significant impact on wages.

Just like for the first labor market outcome, Table 24 contains the average predictive margins for *birthplace_east* at different levels of *workplace_east*. In the basic model, the probability of achieving a wage above the contribution assessment ceiling is 42.8% for native western Germans working in western Germany and 42.6% for native eastern Germans working there. When the place of work is in eastern Germany, the probabilities of earning high wages are much lower. Native eastern and western Germans have the same probability (22%) of earning wages above the social security contribution assessment ceiling in t+5. Again, overlapping confidence intervals suggest no statistical differences between the predictive margins at the different levels of *workplace_east*.

Table 23. Selected average marginal effects for exceeding the contribution assessment ceiling in t+5

Variable	dy/dx	Std. err.	z-score	p- value	95% conf. interval
			Basic model		
birthplace_east	-0.001	0.027	-0.04	0.96	-0.055-0.052
workplace_east	-0.203***	0.026	-7.65	0.00	-0.255-0.015
workplace_berlin	-0.069	0.044	-1.58	0.11	0.015-0.017
university_east	-0.021	0.026	-0.82	0.41	-0.073-0.030

Note: Delta method, Model VCE: robust, dy/dx for factor levels is the discrete change from the base level. */**/*** indicates statistical significance at the 10/5/1% level, respectively.

Sources: IIPED data set, own research; own calculation.

[39] Note that the substantially lower contribution assessment ceiling in eastern Germany is already taken into account (see section 4.1).

Table 24. Average predictive margins for birthplace_east at different levels of workplace_east (exceeding the contribution assessment ceiling in t+5)

	Margin	Std. err.	z-score	p-value	95% conf. interval
	Basic model				
workplace_east = 0 at *birthplace_east* = 0	0.428***	0.011	35.97	0.00	0.404-0.451
workplace_east = 0 at *birthplace_east* = 1	0.426***	0.026	15.98	0.00	0.374-0.479
workplace_east = 1 at *birthplace_east* = 0	0.223***	0.022	9.78	0.00	0.179-0.269
workplace_east = 1 at *birthplace_east* = 1	0.223***	0.022	9.78	0.00	0.178-0.268

Note: Delta method, Model VCE: robust. */**/*** indicates statistical significance at the 10/5/1% level, respectively.

Sources: IIPED data set, own research; own calculation.

6.6 Robustness checks

Although 1,051 of the 2,758 persons in my sample earn wages above the social security contribution assessment ceiling (see Appendix D), this threshold might generally be set too high for the majority of PhD graduates. As a consequence, considerable variations between eastern and western could exist below the threshold, which is not addressed with my approach. I therefore check whether the percentage of individuals within my two groups changes substantially when the contribution assessment ceiling is modified. Figure 20 depicts the results of a reduction by 5% and 10% respectively. The graph shows an increase in the number of observations occurring in all regional delineations. However, I find no noticeable differences between the ratios of this increase between the single groups. This implies that a reduction in the social security contribution assessment ceiling affects the two groups in the same manner, regardless of the regional origin. I therefore conclude that the contribution assessment ceiling is a valid measure.

Further robustness checks address the different sectoral composition in eastern and western Germany. As eastern Germany has a more pronounced service sector, I repeated my regression procedure for the manufacturing sector only. Again, the birthplace does not play a role, but the place of work does. It leads to a lower likelihood of obtaining an expert job status and achieving a wage above the contribution assessment ceiling.

Finally, my separate estimate for PhD graduates born in Berlin does not deliver any robust results, since the number of observations is too small. All details on the robustness checks are available from the authors upon request.

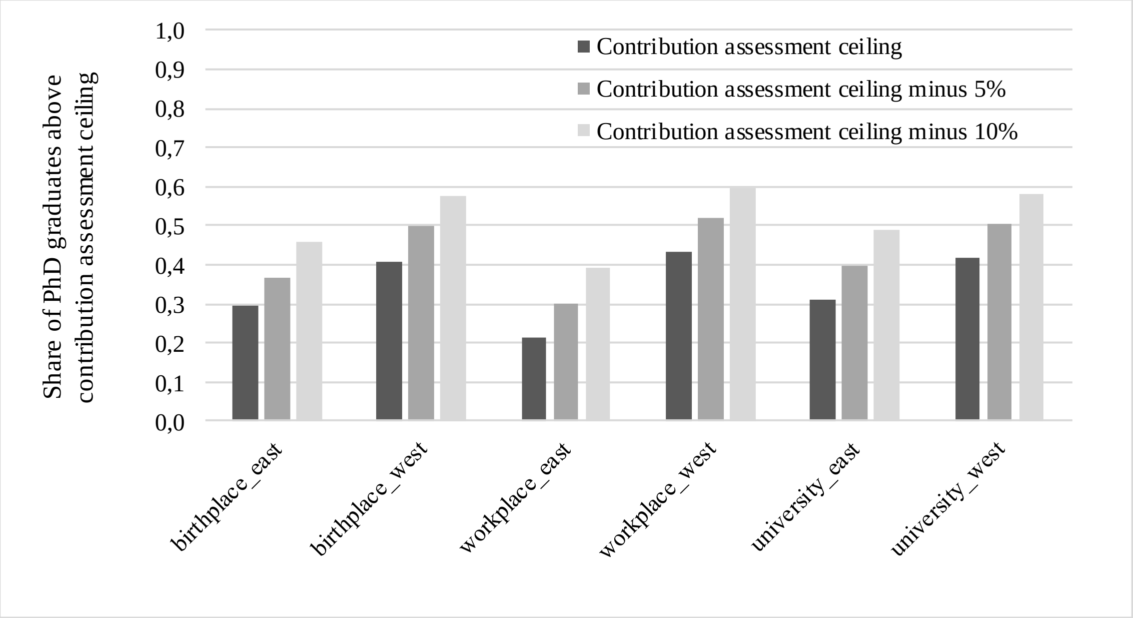


Figure 20. Share of PhD graduates with wages above the contribution assessment ceiling (modifications)

Source: IIPED data set, own birthplace data from online dissertations; own compilation.

6.7 Conclusions

Are eastern German PhD graduates prevented from fully exploiting their investment in education and thus from getting top positions nationwide because of their regional background? 30 years after the fall of the Berlin wall, the question is discussed at length in the societal reappraisal of German reunification. This paper provides novel findings on this topic by examining the labor market outcomes of PhD graduates with eastern or western German backgrounds. I differentiate between the place of birth, the location of the university at which the PhD was earned, and the subsequent place of work. The analysis uses a novel data set on the employment biographies of PhD graduates, enriched with geo-referenced information about their place of birth.

My results yield no statistical evidence suggesting that eastern German PhD graduates have poorer labor market outcomes than their western German counterparts as a result of their birthplace when it comes to obtaining a job suited to their qualification level or achieving high wages. Nor does the location of the university in eastern or western Germany have any explanatory power. Hence, the results confirm that equal qualification levels lead to equal labor market outcomes. It is more the place of work that makes a difference. In particular, a place of work in eastern Germany substantially reduces PhD graduates’ chances of earning high wages, regardless of which part of the country they were born in. This result suggests that the still divergent economic conditions in the two parts of Germany impact on PhD graduates’ labor market prospects.

The results of this paper leave ample scope for further research. One issue is the spatial mobility patterns of eastern and western German PhD graduates. In the regressions, I included an indicator for spatial mobility after gaining a PhD, which is highly significant in the case of PhD graduates with a job that is suited to their qualification level. Obviously, the degree of mobility especially from eastern to western Germany seems to matter for achieving equal labor market opportunities. A deeper investigation of this issue is open to future study. Likewise, I have refrained from considering the profound gender/region disparities among the PhD graduates that arise especially between eastern and western German women. For example, there are fundamental differences concerning labor market attachment among female graduates (Boll et al., 2016) that might also be of relevance for PhD graduates. Finally, an investigation of earlier cohorts might be of interest. Graduate and/or doctoral education that took place in the German Democratic Republic may have led to a substantial skill mismatch in some disciplines and consequently to poorer labor market outcomes for eastern German doctoral cohorts before 1995.

7 Conclusion

The research question of this dissertation asked how methods of machine learning and social sciences can jointly help to create new databases and provide subsequent insights into social inequality in German academia. To answer this question, I investigated five related research problems.

In Chapter 2, I showed how a machine learning method based on topic modeling could detect and understand thematic differences between author populations. I analyzed the dissertation titles of PhD graduates from eastern and western German universities in chemistry and economics and business administration. For dissertations in economics and business administration, my results suggest wide variety in research topics before German reunification and rapid conformation thereafter. For dissertations in chemistry, there is no apparent difference in research topics between the periods before and after reunification. My approach also has limitations. As topic modelling as the underlying methods does not aim to label the detected topics, I can sometimes only guess what the found differences and their underlying topics most likely refer to. This is a major disadvantage of any sort of topic modelling. The foundation of this problem arises from language as a dynamic, complex and strongly context-related semantic system. Topic models can only find the relations in this system, but not understand and label them accordingly. Nevertheless, the machine learning method provides a reliable methodological approach for a range of applications in social inequality research. My open-access dataset of pairwise thematic similarities between dissertation authors (Rehs, 2020b) may be used, for example, to explore thematic differences between men and women in academia. Retrospectively, Chapter 2 and the paper on which it is based provide a relevant contribution to methods and databases in social inequality research of scientific systems.

Chapter 3 again dealt with machine learning methods. In it, I tackled the author-name disambiguation problem in the WoS publication database with a supervised approach using the Researcher ID author identifier. I used this identifier to generate paper pairs of different authors who have the same name. I then used this dataset to train and test a random forest and logistic regression classifier. I clustered the resulting pairwise predictions with infomap graph-community detection. The retrieved author clusters suggest good performance of the supervised approach. My approach adds to the already extensive literature on author disambiguation by providing detailed feature assessment, handling missing data and demonstrating applicability.

However, there are also several limitations. My main difficulty in appropriately handling synonyms and author name changes (German “Mueller vs. “Muller”). Other problems consist in not fully capturing the data-generating process of the WoS. I made some important assumptions about the same and designed my sampling strategy accordingly.

This may have caused problems in in the disambiguation of authors with only few papers. With respect to my research question, Chapter 3 suggests a method upon which informative author-level databases can be generated. These may then be used in applied research questions on socioeconomic inequality in academia.

I generated such author-level databases in Chapter 4. The approach I present in this chapter is based on an older version of the algorithm from Chapter 3 and establishes a database that includes more than 10 million disambiguated publications. These publications are linked to dissertation authors in the DNB by using a probabilistic record linkage procedure. The linked dataset contains about 61,00 German PhD graduates and their publications. Finally, I analyze the scientific productivity between eastern and western German PhD graduates and between men and women. My descriptive patterns suggest that female PhD graduates write significantly fewer papers than their male counterparts after earning their PhD. Whether this is due to survivor bias or not remains unclear and is beyond the focus of Chapter 4. Between eastern and western Germans, I don't observe any difference in productivity. Chapter 4 adds to the research on socioeconomic inequality in academia by providing an extensive bibliometric database on young German scholars. The illustrative application on these two groups – eastern and western Germans and men and women – opens a direction for further research. My database and approach presented in Chapter 4 has several limitations. I focus only on dissertation authors who publish in WoS-indexed journal articles and use them as a measure of productivity. This focus does not account for the wide range of publication cultures in German academia. It favors identifying publications in natural science fields over those in the social sciences and arts and humanities. Methodically, my approach is subject to limitations, as well. I adapted the author disambiguation algorithm from Chapter 3, which builds on WoS training data that contains the Researcher ID author identifier. This dataset does not fully capture the true data generating process of publication data of the WoS and has caused some problems that I had to address.

Chapter 5 was the first chapter on applied research in social inequality. I investigated the scientific survival and productivity of German PhD graduates by different advisor-protégé gender-pairings. For this purpose, I used the databases created in Chapter 4 and advisor information scraped from online dissertations. My analysis was based on time-discrete survival regression and explored the time until an individual's final publication after earning a PhD. I showed that protégés of female advisors are 38% less likely than protégés of male advisors to drop out of academia after earning a PhD, regardless of protégé gender. However, women have a generally higher yearly dropout rate after earning a PhD. Moreover, I found that women are less productive while earning their PhD. Therefore, Chapter 5 has added to the literature on socioeconomic inequality by narrowing the research gap related to advisor-protégé gender-pairings. In this regard, the temporal patterns after earning a PhD and the context within Germany have not been previously investigated.

Chapter 5 also has several limitations. The most important limitations are the unaddressed factors that influence female exit from research. In particular, the effect of private events and factors, such as childbirth and motherhood, lead to an omitted variable bias. It is also unclear whether there are differences in the preference for careers in science or industry between men and women after completing their PhD, which could explain the female departure from academia. My results in Chapter 5 potentially also suffer from self-selection bias. Therefore, the initial matching process between protegees and advisors may not be random.

Finally, Chapter 6 concerned the labor market outcomes of eastern and western German PhD graduates. The chapter investigated the extent to which the returns on earning a PhD depend on region of birth and region where the degree was earned (eastern or western Germany), and the place of work. Me and my co-author examined the career paths of eastern and western German PhD graduates who completed their dissertations between 1995 and 2010 and estimate the returns to obtaining a job suited to their skill level and with high wages. Our dataset combines information on PhD graduates and their place of birth collected from data on PhD dissertations in Germany with data from administrative social security records. The findings show that labor market success is affected neither by being born in eastern Germany nor by earning a PhD at an eastern German university, but rather by the place of work in eastern Germany. With respect to inequality research in German academia, this chapter and its findings represent important contributions. They show that place of birth is not associated with worsened labor market outcomes for highly skilled individuals. However, there remain several open questions, such as individual mobility patterns and self-selection processes related to PhD student skills and university or departmental reputation.

In summary, I was able in my chapters and findings to contribute to my research question regarding socioeconomic inequality in academia (“How can methods of machine learning and social sciences jointly help to establish new databases on and provide subsequent insights into socioeconomic inequality among junior researchers in German academia?”). I developed two machine learning-based approaches that established new scientometric databases on junior researchers in Germany. Moreover, I carried out applied research with these databases on two currently discussed issues in socioeconomic inequality. There remain, however, a number of open questions and tasks that could not be addressed within the time frame of this dissertation. First, my databases are not fully interconnected and require further record linkage procedures. Especially the connection of scientific productivity with labor market outcomes and thematic characteristics would provide new insights on the career trajectories of junior scholars. Another problem is the external data quality and availability and related recall in my existing record linkage procedures. It is unknown if further machine learning approaches can alleviate this issue. I see great potential in the information retrieved from online dissertations. But because my dissertation is based on a 2015 version of the DNB catalog, online dissertations after 2015 are still unmatched, and I was unable to employ all existing birthplace and advisor information.

Socioeconomic inequality remains one of the most demanding and urgent problems in academia. Science and higher education policy must address this problem in order to foster scientific, economic and societal progress. The combination of research in economics and scientometrics on socioeconomic inequality can, as shown in this dissertation, identify current issues and provide advisory support.

8 Appendices

Appendix A. Top 4 words highest β probability

Topic	Economics and business administration	Chemistry
1	‘and,‘the’,‘portfolio’,‘development’,‘model’	‘radikal’,‘selektiv’,‘alk’,‘alkohol’,‘additions’
2	‘integration’,‘kost’,‘internationalisier’,‘beruf’,‘option’	‘kohlenwasserstoff’,‘konzept’,‘oxidativ’,‘methan’,‘methanol’
3	‘prozess’,‘wissenschaft_technisch’,‘technisch_fortschritt’,‘rationalisier’,‘wissenschaft_technisch_fortschritt’	‘funktionalisiert’,‘baustein’,‘verbruckt’,‘chrom’,‘gold’
4	‘schwerpunkt’,‘konzentration’,‘steuerpolit’,‘entwurf’,‘steuerreform’	‘oberflach’,‘adsorption’,‘wasserstoff’,‘wechselwirk’,‘ftir’
5	‘bereich’,‘energie’,‘rationell’,‘brd’,‘konsumgut’	‘the’,‘complex’,‘with’,‘based’,‘catalyst’
6	‘einsatz’,‘erfolgsfaktor’,‘internet’,‘onlin’,‘medi’	‘elektron’,‘clust’,‘zust’,‘fest’,‘spin’
7	‘wandel’,‘osterreich’,‘qualitativ’,‘organisator’,‘inner’	‘olefin’,‘einsatz’,‘stabilisier’,‘homog’,‘ylid’
8	‘mittelstand’,‘intern’,‘berat’,‘modell’,‘unternehmensberat’	‘basis’,‘vorstuf’,‘hinblick’,‘para_phenyl’,‘poly_para’
9	‘industri’,‘effekt’,‘branch’,‘preis’,‘west’	‘platin’,‘komplexbild’,‘cis’,‘stabilitat’,‘phenyl’
10	‘problem’,‘sozialist’,‘beding’,‘volks’,‘aufgab’	‘stereoselektiv’,‘enantioselektiv’,‘enantiomerenrein’,‘aminosaur’,‘diastereoselektiv’
11	‘ddr’,‘kombinat’,‘leitung’,‘sozialismus’,‘nutzung’	‘dynam’,‘synthet’,‘natur’,‘membran’,‘relaxation’
12	‘strategi’,‘ziel’,‘orientiert’,‘unternehmenskris’,‘bewalt’	‘hoh’,‘flussigkristall’,‘niedermolekular’,‘mesog’,‘nemat’
13	‘industriell’,‘aspekt’,‘determinant’,‘organisator’,‘entwicklungsland’	‘ubergangsmetall’,‘phosphan’,‘redox’,‘cyclopentadienyl’,‘fragment’
14	‘extern’,‘prognos’,‘bau’,‘qualitatssicher’,‘steuerungs’	‘for’,‘element’,‘paramet’,‘gallium_indium’,‘aluminium_gallium_indium’
15	‘okonomi’,‘geld’,‘sozial’,‘kritik’,‘okologi’	‘flussigkristallin’,‘amphiphil’,‘monom’,‘phasenverhalt’,‘grenzflach’
16	‘produkt’,‘innovativ’,‘finanz’,‘backed_securiti’,‘rentenversicher’	‘pfeil_recht’,‘eis’,‘typs’,‘eta’,‘mangan’
17	‘polit’,‘institution’,‘quality’,‘histor’,‘islam’	‘bzw’,‘alkaloid’,‘strukturaufklar’,‘pyrrol’,‘cyclisier’
18	‘hintergrund’,‘funktion’,‘okonometr’,‘verander’,‘jung’	‘rhodium’,‘carb’,‘iridium’,‘alkin’,‘zweikern’
19	‘unt’,‘logist’,‘bes_beruck’,‘textil’,‘sektor’	‘elektro’,‘uberbruckt’,‘chromatographi’,‘komplexier’,‘sigma’
20	‘beurteil’,‘anhand’,‘usa’,‘wettbewerbspolit’,‘betracht’	‘unt’,‘phosphor’,‘dimethylamino’,‘phosphoran’,‘nitro’
21	‘forder’,‘mittl’,‘auswahl’,‘massnahm’,‘qualitats’	‘verhalt’,‘cycloaddition’,‘dien’,‘abfang’,‘triazin’
22	‘okolog’,‘nachhalt’,‘sozial’,‘global’,‘umwelt’	‘diel_ald’,‘neutral’,‘hetero_diel’,‘hetero_diel_ald’,‘selektivitat’
23	‘rahm’,‘komplex’,‘nutzung’,‘weiter’,‘beitr’	‘strukturell’,‘kupf’,‘praparativ’,‘oxid’,‘aspekt’
24	‘basis’,‘fuzzy’,‘marktforsch’,‘neuronal_netz’,‘einkaufsstattenwahl’	‘situ’,‘111’,‘adsorption’,‘surfac’,‘non’
25	‘steu’,‘ermittl’,‘kapitalgesellschaft’,‘finanzier’,‘grenzuberschreit’	‘aromat’,‘alkyl’,‘phenol’,‘aliphat’,‘chloriert’
26	‘forschung’,‘betriebs’,‘kost’,‘sicher’,‘kontroll’	‘ausgewahlt’,‘vergleich’,‘ungesattigt’,‘gegenub’,‘substrat’
27	‘bank’,‘roll’,‘kulturell’,‘kund’,‘unternehmenskultur’	‘methyl’,‘total’,‘hydroxy’,‘zugang’,‘est’
28	‘optimier’,‘kommunikation’,‘softwar’,‘mittel’,‘losung’	‘stickstoff’,‘phosphor’,‘schwefel’,‘kohlenstoff’,‘sauerstoff’
29	‘verbesser’,‘qualitat’,‘verwend’,‘neuronal_netz’,‘kunstlich_neuronal’	‘aufbau’,‘messung’,‘druck’,‘temperatur’,‘mpa’
30	‘technisch’,‘informations’,‘rechnergestutzt’,‘darstell’,‘betriebs’	‘iii’,‘oxo’,‘tris’,‘vanadium’,‘chlor’
31	‘struktur’,‘japan’,‘gesellschaft’,‘dimension’,‘alternativ’	‘analyt’,‘modifiziert’,‘hplc’,‘biolog’,‘trennung’
32	‘information’,‘integration’,‘verteilt’,‘heterog’,‘wertorientiert’	‘thermisch’,‘photochem’,‘omega’,‘isomerisier’,‘lamda’
33	‘dynam’,‘optimal’,‘linear’,‘investition’,‘finanzplan’	‘stereo’,‘verwandt’,‘tran’,‘cis’,‘grundlag’
34	‘bezieh’,‘zusammenarbeit’,‘industrieland’,‘nord’,‘kapitalist’	‘modell’,‘einfach’,‘quantenchem’,‘porphyrin’,‘chinon’
35	‘schweiz’,‘wandel’,‘natur’,‘welt’,‘option’	‘metall’,‘modell’,‘chelat’,‘rhenium’,‘haltig’
36	‘uber’,‘zentral’,‘regel’,‘gesetz’,‘plan’	‘dihydro’,‘eta’,‘kenntnis’,‘lambda’,‘sigma’
37	‘sicht’,‘institutionen’,‘betracht’,‘schweizer’,‘wettbewerbsfah’	‘naturstoff’,‘transformation’,‘allyl_substitution’,‘beitr’,‘biolog_aktiv’
38	‘entscheid’,‘computergestutzt’,‘raum’,‘werbung’,‘grenz’	‘verwend’,‘amorph’,‘loslich’,‘kohlenhydrat’,‘materiali’
39	‘risiko’,‘risik’,‘privat’,‘ventur_capital’,‘banking’	‘peptid’,‘konformation’,‘modifizier’,‘zyklisch’,‘racematspalt’
40	‘controlling’,‘umsetz’,‘organisations’,‘operativ’,‘effizient’	‘las’,‘ungewohn’,‘immobilisiert’,‘matrix’,‘studium’
41	‘wirkung’,‘tourismus’,‘mark’,‘stadt’,‘verhaltenswissenschaft’	‘delta’,‘trag’,‘tetra’,‘symmetr’,‘kristallisation’
42	‘markt’,‘industrie’,‘ausgewahlt’,‘fallstudi’,‘transformation’	‘ubergangs’,‘titan’,‘rontgenstrukturanalys’,‘koordination’,‘semiempir’
43	‘hilf’,‘landlich’,‘gebiet’,‘technisch’,‘kennzahl’	‘hilf’,‘infrarot’,‘lichtinduziert’,‘zeitaufgelost’,‘berechn’
44	‘perspektiv’,‘neu’,‘system’,‘regionalpolit’,‘reformvorschlag’	‘gas’,‘massenspektrometr’,‘nachweis’,‘elementar’,‘partiell’
45	‘automobilindustri’,‘netzwerk’,‘kooperation’,‘virtuell’,‘interkulturell’	‘poly’,‘styrol’,‘polystyrol’,‘initiator’,‘copolymerisation’
46	‘servic’,‘financial’,‘engineering’,‘performanc’,‘integration’	‘analoga’,‘festphasen’,‘aufbau’,‘kombinator’,‘strategi’
47	‘arbeit’,‘einflussfaktor’,‘diagnos’,‘grundsatz_ordnungsmass’,‘grundsatz_ordnungsmass_bilanzier’	‘molekul’,‘photo’,‘fluoreszenz’,‘raman’,‘induziert’
48	‘aspekt’,‘ergebnis’,‘land’,‘licht’,‘studi’	‘wassrig’,‘gamma’,‘sio2’,‘al2o3’,‘tio2’
49	‘personal’,‘syst’,‘evaluation’,‘fuhrungskraft’,‘fuhrung’	‘bindung’,‘aktivier’,‘funktionalisier’,‘aktiviert’,‘alkylier’
50	‘staatlich’,‘staat’,‘zusammenhang’,‘land’,‘gesellschaft’	‘optisch’,‘magnet’,‘farbstoff’,‘elektr’,‘schicht’
51	‘makro’,‘fundiert’,‘verhalt’,‘erklar’,‘arbeitsmarkt’	‘amin’,‘amino’,‘ring’,‘aryl’,‘substituent’
52	‘alternativ’,‘losung’,‘geldpolit’,‘finanziell’,‘entscheidungs’	‘molekul’,‘theoret’,‘ion’,‘zeolith’,‘umlager’
53	‘produktion’,‘effektivitat’,‘flexibl’,‘fertig’,‘vorbereit’	‘silicium’,‘kristall’,‘silan’,‘sol_gel’,‘silicat’
54	‘innovation’,‘erfolgreich’,‘fallbeispiel’,‘innovations’,‘organisational’	‘ternar’,‘kristall’,‘lithium’,‘alkali’,‘lanthanoid’
55	‘aufbau’,‘praktisch’,‘rahm’,‘unternehmensfuhr’,‘ansatzpunkt’	‘nickel’,‘koordinations’,‘zink’,‘cobalt’,‘silb’
56	‘marketing’,‘national’,‘einzelhandel’,‘determinant’,‘interaktion’	‘katalyt’,‘mono’,‘aufklar’,‘hydrier’,‘umwandl’
57	‘bestimm’,‘simulation’,‘hilf’,‘system’,‘eignung’	‘cyclisch’,‘umsetz’,‘nucleophil’,‘bzw_beziehungsweis’,‘elektrophil’
58	‘prozess’,‘modellier’,‘unternehmens’,‘dynam’,‘mittel’	‘grupp’,‘element’,‘amid’,‘nebengrupp’,‘moglich’
59	‘regional’,‘studi’,‘rechnungsleg’,‘ifr’,‘bilanzier’	‘oxidation’,‘mechanismus’,‘ruthenium’,‘reduktion’,‘gegenwart’
60	‘gross’,‘markt’,‘operationalisier’,‘bereitstell’,‘erfolgswirk’	‘katalysator’,‘palladium’,‘polymerisation’,‘eth’,‘katalys’
61	‘integriert’,‘unterstutz’,‘technologi’,‘ganzheit’,‘prozessorientiert’	‘optisch_aktiv’,‘pro’,‘baustein’,‘alkohol’,‘katalys’
62	‘einfuhr’,‘business’,‘gruppenarbeit’,‘organisator’,‘produktionsbereich’	‘analys’,‘gebund’,‘optimier’,‘spektr’,‘gaschromatograph’
63	‘dienstleist’,‘relevanz’,‘beschaff’,‘zusammenarbeit’,‘kooperation’	‘wass’,‘syst’,‘thermodynam’,‘mischung’,‘kritisch’
64	‘steuer’,‘konzeptionell’,‘handel’,‘dezentral’,‘handels’	‘addition’,‘versuch’,‘lithium’,‘aldehyd’,‘ungesattigt_ungesattigt’
65	‘rahmenbeding’,‘institutionell’,‘kommunal’,‘bundesland’,‘medizin’	‘wechselwirk’,‘festkorp’,‘hilf’,‘schwach’,‘saur’
66	‘basis’,‘verfahr’,‘entscheidungsorientiert’,‘krankenhaus’,‘energieversorgungs’	‘oligo’,‘sensor’,‘dendrim’,‘kunstlich’,‘potentiell’
67	‘bedeut’,‘entwicklungsland’,‘implikation’,‘wirtschaftspolit’,‘gegenwart’	‘linear’,‘anelliert’,‘thioph’,‘oligom’,‘nichtlinear_optisch’
68	‘region’,‘untersucht’,‘china’,‘strukturwandel’,‘berlin’	‘molekular’,‘anion’,‘experimentell’,‘supramolekular’,‘modellier’
69	‘einfluss’,‘zeitverwend’,‘ausgewahlt’,‘grenz’,‘faktor’	‘protein’,‘dna’,‘wechselwirk’,‘enzymat’,‘inhibitor’
70		‘via’,‘diel_ald’,‘lewis_saur’,‘steroid’,‘intramolekular_diel’
71		‘massenspektrometri’,‘kopplung’,‘icp’,‘prob’,‘direkt’
72		‘struktur’,‘organo’,‘rontgenograph’,‘schwingungs’,‘alkali’
73		‘typ’,‘mechanism’,‘extraktion’,‘chemistry’,‘imidazolin’
74		‘donor’,‘biphenyl’,‘wirt_gast’,‘helical’,‘axial’
75		‘bildung’,‘verfahr’,‘effekt’,‘zerfall’,‘berucksicht’
76		‘verschied’,‘kation’,‘voraussetz’,‘carbonyl’,‘induziert’

Source: Own data.

Appendix B. Logistic regression model and average marginal effects (AME)

Variable	Coefficient	AME (sample)
Same first name (reference = missing)		
Same first name = T	2.353*** (0.02)	0.080*** (0.00)
Same class =T (reference = F)	2.482*** (0.03)	0.092*** (0.00)
Diff. in publication year	-0.057*** (0.00)	-0.001*** (0.00)
Diff. in no. of citations	-0.000* (0.00)	-0.000* (0.00)
Thematic similarity title	0.539*** (0.03)	0.011*** (0.00)
Thematic similarity abstract	0.719*** (0.03)	0.015*** (0.00)
Jaccard distance title	-18.038*** (0.15)	-0.372*** (0.00)
Institution string similarity	6.427*** (0.05)	0.133*** (0.00)
Diff. in no. of coauthors	-0.093*** (0.003)	-0.002*** (0.00)
First name count 1	-0.000*** (0.00)	-0.000*** (0.00)
Block set size	0.000*** (0.00)	0.000 (0.00)
First name count group 1	-0.000*** (0.00)	-0.000*** (0.00)
Same keyword	-1.135 (0.96)	-0.018 (0.01)
Ratio Block Last	0.748*** (0.06)	0.015*** (0.00)
Same second initial (reference = F)		
Same second initial = missing	2.837*** (0.07)	0.035*** (0.00)
Same second initial = T	4.805*** (0.07)	0.105*** (0.00)
Same country (reference = F)		
Same country = missing	1.563*** (0.02)	0.030*** (0.00)
Same country = T	0.338*** (0.02)	0.005*** (0.00)
Same coauthors	2.968** (0.25)	0.123*** (0.00)
Author position difference	0.028*** (0.00)	0.001*** (0.00)
Publication Year 1	-0.053*** (0.00)	-0.001*** (0.00)
Publication Year 2	-0.053*** (0.00)	-0.001*** (0.00)
Second initial count set 1	0.000*** (0.00)	0.000*** (0.00)
Intercept	224.77*** (3.21)	

Note: Standard Errors in parentheses; significance levels: *p<0.01, **p<0.05, ***p<0.01. AME were calculated for a random sample of 10,000 observations.

Source: Own data and calculation.

Appendix C. Definition of the explanatory variables

Variable name	Definition
Regional origin	
birthplace_east	Dummy 1: birthplace in eastern Germany, 0: birthplace in western Germany
workplace_east	Dummy 1: workplace in eastern Germany, 0: workplace in western Germany
workplace_berlin	Dummy 1: workplace in Berlin, 0: workplace elsewhere
university_east	Dummy 1: university in eastern Germany, 0: university elsewhere
Individual characteristics	
Age	Continuous variable
Sex	Dummy 1: female, 0:male
Work experience	Continuous variable, measured in days/100 up until one year before graduation
Move_region	Change between region of university and region of employer in t+5 after obtaining PhD, dummy 1: yes, 0: no
Occupational characteristics	
Sector	9 sectors, dummy 1: yes, 0: no
Requirement level of the job task	According to German classification of occupations (KldB2010); dummy 1: low skills, 2: medium skills, 3: specialist skills, 4: expert skills
Region type	Agglomeration, urbanized region, rural region, dummy 1: yes, 0: no
Scientific characteristics	
Discipline	Scientific disciplines as classified by the DNB; 1: architecture, 2: history, 3: computer science, 4: engineering, 5: arts and music, 6: mathematics and statistics, 7: sciences, 8: philosophy, 9: political science, 10: psychology, 11: education, 12: law and administration, 13: social sciences, 14: sports, 15: languages and linguistics, 16: theology, 17: economics and business
Other variables	
Year	Dummy 1: yes, 0: no for the years 2000-2015

Source: Own compilation.

Appendix D. Descriptive statistics for dependent and explanatory variables

Variable	No. of observations	No. of observations = 1	Mean	Std. dev.	Min.	Max.
Dependent variables						
Wage above contribution assessment ceiling in t+5	2,758	1,051	0.38	0.49	0	1
Skill requirement level of the job	2,758		3.62	0.68	1	4
Of which: Low skills	2,758	11				
Medium skills	2,758	287				
Specialist skills	2,578	428				
Expert skills	2,758	2,016	0.73	0.44	0	1
Explanatory variables						
Regional origin						
birthplace_east	2,758	670	0.24	0.42	0	1
workplace_east	2,758	637	0.23	0.42	0	1
workplace_berlin	2,758	91	0.03	0.17	0	1
university_east	2,758	918	0.33	0.47	0	1
Individual characteristics						
Age	2,758		36.79	3.59	22	62
Age^2	2,758		1,367.10	289.6	484	3,844
Of which: aged 22-32	2,758	96				
aged 33-35	2,758	1,067				
aged 36-38	2,758	980				
aged 39-42	2,758	430				
aged 43-46	2,758	125				
aged 47-62	2,758	60				
Sex	2,758	682	0.24	0.43	0	1
Work experience			19.04	11.13	0	103.03
Of which: work exp. <=3.91	2,758	219				
work exp. >3.91;<=7.56	2,758	75				
work exp. >7.56;<=10.94	2,758	209				
work exp. >10.94	2,758	2,255				
Move_region	2,758	1,829	0.66	0.47	0	1
Occupational characteristics						
Sector						
Agriculture, forestry and horticulture	2,758	7				
Production, processing	2,758	555				
Construction, architecture	2,758	24				
Natural science, geography, computer science	2,758	827				
Transport, traffic, security	2,758	20				
Commercial services, wholesale and trade	2,758	71				
Business administration, accounting, law	2,758	589				
Health, social, education	2,758	547				
Humanities, culture, arts, media	2,758	102				
Skill requirement level of the job						
Low skills	2,758	11				
Medium skills	2,758	287				
Specialist skills	2,758	428				
Expert skills	2,758	2,016				
Region type						
Agglomerations	2,758	1,673				
Urbanized regions	2,758	801				
Rural regions	2,578	284				
Scientific characteristics						
Discipline (double counts possible)						
Architecture	2,758	12				

History	2,758	16				
Computer science	2,758	97				
Engineering	2,758	400				
Arts and music	2,758	27				
Mathematics and statistics	2,758	108				
Sciences	2,758	1,870				
Philosophy	2,758	30				
Political science	2,758	10				
Psychology	2,758	43				
Education	2,758	20				
Law and administration	2,758	20				
Social sciences	2,758	15				
Sports	2,758	3				
Languages and linguistics	2,758	30				
Theology	2,758	2				
Economics and business	2,758	121				
Other characteristics						
Year of employment spell	2,758		2011.70	3.25	2000	2015
2000-2005	148					
2006-2010	789					
2011-2015	1,895					

Note: The table shows the descriptive statistics for the initial dataset in the regressions. Due to Singularities not all of the observations were used in the regressions.

Sources: IIPED data set, own research; own calculation.

Appendix E. Average marginal effects for holding an expert job in t+5

Variable	dy/dx	std. error	z-score	p- value
Main independent variables				
birthplace_east	0.069	0.023	0.30	0.764
workplace_east	0.023	0.025	0.92	0.358
workplace_berlin	-0.007	0.042	-0.16	0.870
university_east	-0.033	0.043	-1.48	0.140
Individual characteristics				
Age	0.032	0.022	1.27	0.206
Age^2	-0.000	0.022	-1.23	0.217
Sex	-0.044**	0.018	-2.43	0.015
Work experience	0.000	0.001	-1.09	0.274
Move_region	0.032**	0.016	1.98	0.047
Occupational characteristics				
Sector				
Production, processing	0.056	0.155	0.36	0.717
Construction, architecture	0.111	0.171	0.65	0.517
Natural science, computer science, geography	0.131	0.154	0.85	0.394
Transport, traffic, security	-0.450**	0.183	-2.45	0.014
Commercial services, wholesale/trade, tourism	-0.523**	0.160	-3.25	0.001
Business admin., accounting, law, administration	-0.342**	0.155	-2.21	0.027
Health, social, education	0.214	0.153	1.39	0.165
Humanities, culture, arts, media	-0.235	0.162	-1.45	0.148
Region type				
reference=agglomerations				
rural regions	-0.026	0.026	-0.99	0.322
urbanized regions	0.008	0.016	0.49	0.625
Year of employment spell				
2001	not estimable			
2002	-0.004	0.137	-0.03	0.978
2003	0.156	0.131	1.19	0.233
2004	0.092	0.120	0.77	0.441
2005	0.062	0.121	0.51	0.609
2006	0.089	0.117	0.76	0.447
2007	0.055	0.116	0.47	0.637
2008	0.028	0.117	0.24	0.812
2009	0.051	0.116	0.44	0.660
2010	0.029	0.116	0.26	0.797
2011	0.101	0.116	0.87	0.386
2012	0.139	0.115	1.20	0.230
2013	0.146	0.114	1.26	0.206
2014	0.090	0.115	0.79	0.431
2015	0.096	0.115	0.84	0.403
Discipline characteristics				
Architecture	-0.069	0.114	-0.61	0.542
History	0.026	0.117	0.22	0.823
Computer science	-0.061	0.090	-0.68	0.494
Engineering	0.068	0.079	0.87	0.386
Arts and music	0.023	0.102	0.23	0.816
Mathematics and statistics	0.037	0.089	0.41	0.680
Sciences	0.020	0.079	0.25	0.800
Philosophy	-0.124	0.133	-0.92	0.355
Political science	-0.005	0.133	-0.04	0.968
Psychology	0.256	0.137	1.86	0.062
Education	0.049	0.165	0.30	0.766
Law and administration	0.200	0.113	1.76	0.078
Social sciences	-0.121	0.112	-1.09	0.277
Sports	0.121	0.197	0.61	0.539
Languages and linguistics	0.053	0.111	0.47	0.635
Theology	omitted			
Economics and business	0.037	0.082	0.45	0.651
Number of observations =2,733				

Note: Delta method, Model VCE: robust, dy/dx for factor levels is the discrete change from the base level. */**/*** indicates statistical significance at the 10/5/1% level, respectively.

Sources: IIPED data set, own research; own calculation.

Appendix F. Average marginal effects for exceeding the contribution assessment ceiling in t+5

Variable	dy/dx	std. error	z-score	p- value
Main independent variables				
birthplace_east	-0.001	0.027	-0.04	0.966
workplace_east	-0.203***	0.026	-7.65	0.000
workplace_berlin	-0.069	0.044	-1.58	0.115
university_east	-0.021	0.026	-0.82	0.410
Individual characteristics				
Age	-0.092***	0.026	-350	0.000
Age^2	0.001***	0.0003	3.21	0.001
Sex	-0.176***	0.019	-8.89	0.000
Work experience	0.004***	0.001	3.77	0.000
Move_region	0.058	0.019	3.00	0.003
Occupational characteristics				
Sector	not estimable			
Region type				
reference=agglomerations				
rural regions	0.078**	0.030	-1.53	0.012
urbanized regions	-0.029	0.019	2.52	0.125
Skill requirement level of the job				
reference=low skills				
medium skills	0.160**	0.096	1.66	0.097
specialist skills	0.248**	0.095	2.59	0.010
expert skills	0.240**	0.094	2.56	0.010
Year of employment spell				
2001	0.079	0.23	0.23	0.820
2002	0.175	0.75	0.75	0.453
2003	-0.279	-1.24	-1.24	0.215
2004	-0.139	-0.63	-0.63	0.528
2005	-0.118	-0.55	-0.55	0.582
2006	-0.082	-0.39	-0.39	0.697
2007	-0.167	-0.79	-0.79	0.428
2008	-0.083	-0.39	-0.39	0.694
2009	-0.145	-0.69	-0.69	0.492
2010	-0.089	-0.43	-0.43	0.670
2011	-0.153	-0.73	-0.73	0.467
2012	-0.169	-0.81	-0.81	0.420
2013	-0.019	0.93	-0.93	0.351
2014	-0.245	-1.17	-1.17	0.241
2015	0.075	-0.36	-0.36	0.720
Discipline characteristics				
Architecture	-0.240	0.171	-1.41	0.160
History	-0.358**	0.177	-2.02	0.044
Computer science	-0.001	0.102	-0.01	0.991
Engineering	-0.007	0.092	-0.08	0.936
Arts and music	-0.107	0.132	-0.81	0.418
Mathematics and statistics	-0.115	0.102	-1.12	0.261
Sciences	-0.101	0.093	-1.08	0.279
Philosophy	0.053	0.133	0.40	0.687
Political science	-0.188	0.166	-1.13	0.258
Psychology	-0.150	0.139	-1.07	0.282
Education	-0.122	0.162	-0.75	0.452
Law and administration	-0.120	0.130	-0.92	0.358
Social sciences	-0.392*	0.226	-1.73	0.083
Sports	-0.048	0.254	-0.19	0.847
Languages and linguistics	-0.47**	0.225	-2.13	0.034
Theology	omitted			
Economics and business	0.012	0.099	0.13	0.899
Number of observations =2733				

Note: Delta method, Model VCE: robust, dy/dx for factor levels is the discrete change from the base level. */**/*** indicates statistical significance at the 10/5/1% level, respectively.

Sources: IIPED data set, own research; own calculation

References

Abbott, A., Cyranoski, D., Jones, N., Maher, B., Schiermeier, Q., & Van Noorden, R. (2010, June 17). Metrics: Do metrics matter? *Nature*, Vol. 465, pp. 860–862. https://doi.org/10.1038/465860a

Acemoglu, D. (1995). Reward structures and the allocation of talent. *European Economic Review*, *39*(1), 17–33.

Agan, A., & Starr, S. (2018). Ban the box, criminal records, and racial discrimination: A field experiment. *Quarterly Journal of Economics*, *133*(1), 191–235. https://doi.org/10.1093/qje/qjx028

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, *80*(1), 123–129. https://doi.org/10.1016/S0165-1765(03)00032-6

Aigner, D. J., & Cain, G. G. (1977). Statistical Theories of Discrimination in Labor Markets. *ILR Review*, *30*(2), 175–187. https://doi.org/10.1177/001979397703000204

AlShebli, B., Makovi, K., & Rahwan, T. (2020). The association between early career informal mentorship in academic collaborations and junior author performance. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19723-8

Andersen, J. P., & Hammarfelt, B. (2011). Price revisited: On the growth of dissertations in eight research fields. *Scientometrics*, *88*(2), 371–383. https://doi.org/10.1007/s11192-011-0408-8

Antoni, M., Ganzer, A., & Vom Berge, P. (2016). *Sample of Integrated Labour Market Biographies (SIAB) 1975-2014*.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., … Zhu, M. (2013). A practical algorithm for topic modelling with provable guarantees. In S. Dasgupta & S. McAllester (Eds.), *Proceedings of the 30th international conference on machine learning, Volume 28 of proceedings of machine learning research* (pp. 280–288). Atlanta: PLMR.

Arrow, K. J. (1973). The Theory of Discrimination. In O. Aschenfelter & A. Rees (Eds.), *Discrimination in Labor Markets*. Princeton: Princeton University Press.

Auer, W., Fichtl, A., Hener, T., Piopiunik, M., & Rainer, H. (2017). *Bildungsrenditen und nichtmonetäre Erträge der wissenschaftlichen Qualifizierung (Begleitstudie B8) Studien im Rahmen des Bundesberichts Wissenschaftlicher Nachwuchs (BuWiN) 2017*.

Auriol, L., Misu, M., & Freeman, R. A. (2013). Careers of Doctorate Holders: Analysis of Labour Market and Mobility Indicators. In *OECD Science, Technology and Industry Working Papers*. https://doi.org/10.1787/5k43nxgs289w-en

Azoulay, P., Liu, C. C., & Stuart, T. E. (2017). Social influence given (Partially) deliberate matching: Career imprints in the creation of academic entrepreneurs. *American Journal of Sociology*, *122*(4), 1223–1271. https://doi.org/10.1086/689890

Azoulay, P., Stuart, T., & Wang, Y. (2014). Matthew: Effect or fable? *Management Science*, *60*(1), 92–109. https://doi.org/10.1287/mnsc.2013.1755

Bangani, S. (2018). The impact of electronic theses and dissertations: a study of the institutional repository of a university in South Africa. *Scientometrics*, *115*(1), 131–151. https://doi.org/10.1007/s11192-018-2657-2

Barnes, B. J., & Austin, A. E. (2009). The role of doctoral advisors: A look at advising from the advisor's perspective. *Innovative Higher Education*, *33*(5), 297–315. https://doi.org/10.1007/s10755-008-9084-x

Baruffaldi, S., Visentin, F., & Conti, A. (2016). The productivity of science & engineering PhD students hired from supervisors' networks. *Research Policy*, *45*(4), 785–796. https://doi.org/10.1016/j.respol.2015.12.006

Becker, G. (1957). *The economics of discrimination*. Chicago: University of Chicago Press.

Belitz-Demiriz, H., Voigt, D., & Gries, S. (1990). *Die Sozialstruktur der promovierten Intelligenz in der DDR und in der Bundesrepublik Deutschland 1950-1982*. Bochum: Brockmeyer.

Bender, K. A., & Heywood, J. S. (2009). Educational mismatch among Ph.D.s: determinants and consequences. In R. B. Freeman & D. L. Goroff (Eds.), *Science and Engineering Careers in the United States: An Analysis of Markets and Employment* (pp. 229–255). Chicago: University of Chicago Press.

Bender, K., & Heywood, J. (2011). Educational mismatch and the careers of scientists. *Education Economics*, *19*(3), 253–274. https://doi.org/10.1080/09645292.2011.577555

Bergmann, B. R. (1971). The Effect on White Incomes of Discrimination in Employment. *Journal of Political Economy*, *79*(2), 294–313. https://doi.org/10.1086/259744

Bertrand, M., & Duflo, E. (2016). *Field Experiments on Discrimination*. Retrieved from https://econpapers.repec.org/RePEc:nbr:nberwo:22014

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, *94*(4), 991–1013. https://doi.org/10.1257/0002828042002561

Best, H. (2005). Cadres into managers: structural changes of East German economic elites before and after reunification. *Historical Social Research*, *30*(2), 6–24.

Bettinger, E. P., Long, B. T., Ehrenberg, R., Jacob, B., & Murnane, R. (2005). Do faculty serve as role models? the impact of instructor gender on female students. *American Economic Review*, *95*(2), 152–157. https://doi.org/10.1257/000282805774670149

Bhattacharya, J., & Packalen, M. (2020). Stagnation and Scientific Incentives. *Nber Working Paper Series*, *53*(8), 1689–1699. https://doi.org/10.3386/w26752

Birkmaier, D., & Wohlrabe, K. (2014). The Matthew effect in economics reconsidered. *Journal of Informetrics*, *8*(4), 880–889. https://doi.org/10.1016/j.joi.2014.08.005

Blakely, T., & Salmond, C. (2002). Probabilistic record linkage and a method to calculate the positive predictive value. *Great Britain International Journal of Epidemiology*, *31*(6), 1246–1252.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1), 17–35. https://doi.org/10.1214/07-AOAS114

Blei, D. M., Ng, A. Y., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blickenstaff, J. C. (2005). Women and science careers: Leaky pipeline or gender filter? *Gender and Education*, *17*(4), 369–386. https://doi.org/10.1080/09540250500145072

Blinder, A. S. (1973). Wage Discrimination: Reduced Form and Structural Estimates. *The Journal of Human Resources*, *8*(4), 436. https://doi.org/10.2307/144855

Bluhm, M., & Jacobs, O. (2016). Wer beherrscht den Osten? Ostdeutsche Eliten ein Vierteljahrhundert nach der deutschen Wiedervereinigung. Retrieved January 24, 2019, from https://www.mdr.de/heute-im-osten/wer-beherrscht-den-osten-studie-100.html

Bol, T., De Vaan, M., & Van De Rijt, A. (2018). The Matthew effect in science funding. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(19), 4887–4890. https://doi.org/10.1073/pnas.1719557115

Boll, C., Leppin, J. S., & Schömann, K. (2016). Who is overeducated and why? Probit and dynamic mixed multinomial logit analyses of vertical mismatch in East and West Germany. *Education Economics*, *24*(6), 639–662. https://doi.org/10.1080/09645292.2016.1158787

Bornmann, L., & Enders, J. (2001). *Karriere mit Doktortitel?: Ausbildung, Berufsverlauf und Berufserfolg von Promovierten*. München: Campus Verlag.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222. https://doi.org/10.1002/asi.23329

Bourdieu, P. (1984). *Homo academicus*. Paris: Editions de Minuit.

Bourdieu, P. (1988). *Homo academicus*. Stanford: Stanford University Press.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Büchel, F., & van Ham, M. (2003). Overeducation, regional labor markets, and spatial flexibility. *Journal of Urban Economics*, *53*(3), 482–493. https://doi.org/10.1016/S0094-1190(03)00008-1

Buchmueller, T. C., Dominitz, J., & Lee Hansen, W. (1999). Graduate training and the early career productivity of Ph.D. economists. *Economics of Education Review*, *18*(1), 65–77. https://doi.org/10.1016/s0272-7757(98)00019-3

Buenstorf, G., & Geissler, M. (2014). Like Doktorvater, like son? Tracing role model learning in the evolution of German laser research. *Jahrbucher Fur Nationalokonomie Und Statistik*, *234*(2–3), 158–184. https://doi.org/10.1515/jbnst-2014-2-305

Buenstorf, G., & Heinisch, D. P. (2020). When do firms get ideas from hiring PhDs? *Research Policy*, *49*(3), 103913. https://doi.org/10.1016/j.respol.2019.103913

Busch, O., & Weigert, B. (2010). Where have all the graduates gone? Internal cross-state migration of graduates in Germany 1984-2004. *Annals of Regional Science*, *44*(3), 559–572. https://doi.org/10.1007/s00168-008-0274-3

Canaan, S., & Mouganie, P. (2019). Female Science Advisors and the STEM Gender Gap. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3396119

Canal Domínguez, J. F., & Rodríguez Gutiérrez, C. (2013). Wage differences among

Ph.D.s by area of knowledge: Are science areas better paid than humanities and social ones? The Spanish case. *Journal of Education and Work*, *26*(2), 187–218. https://doi.org/10.1080/13639080.2011.638623

Carlsson, M., & Eriksson, S. (2019). Age discrimination in hiring decisions: Evidence from a field experiment in the labor market. *Labour Economics*, *59*, 173–183. https://doi.org/10.1016/j.labeco.2019.03.002

Caron, E., & van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. *Proceedings of the Science and Technology Indicators Conference 2014 Leiden*.

Carroll, D., & Tani, M. (2013). Over-education of recent higher education graduates: New Australian panel evidence. *Economics of Education Review*, *32*(1), 207–218. https://doi.org/10.1016/j.econedurev.2012.10.002

Clarivate Analytics. (2021). Web of Science Core Collection Help. Retrieved February 15, 2021, from https://images.webofknowledge.com/images/help/WOS/hp_full_record.html

Cole, J. R., & Zuckerman, H. (1984). *The productivity puzzle. Advances in Motivation and Achievement. Women in Science.* Greenwich: JAI Press.

Cyranoski, D., Gilbert, N., Ledford, H., Nayar, A., & Yahia, M. (2011). Education: The PhD factory. *Nature*, *472*(7343), 276–279. https://doi.org/10.1038/472276a

D'Angelo, C. A., & van Eck, N. J. (2020). Collecting large-scale publication data at the level of individual researchers: a practical proposal for author name disambiguation. *Scientometrics*, *123*(2), 883–907. https://doi.org/10.1007/s11192-020-03410-y

Dahrendorf, R. (1965). *Gesellschaft und Demokratie in Deutschland*. München: Piper.

Dasgupta, N., & Stout, J. G. (2014). Girls and Women in Science, Technology, Engineering, and Mathematics. *Policy Insights from the Behavioral and Brain Sciences*, *1*(1), 21–29. https://doi.org/10.1177/2372732214549471

Dasgupta, P., & David, P. A. (1994). Toward a new economics of science. *Research Policy*, *23*(5), 487–521. https://doi.org/10.1016/0048-7333(94)01002-1

Davis, K., & Moore, W. E. (1945). Some Principles of Stratification. *American Sociological Review*, *10*(2), 242. https://doi.org/10.2307/2085643

De Bellis, N. (2009). *Bibliometrics and citation analysis : from the Science citation index to cybermetrics*. Scarecrow Press.

De Carvalho, A., Ferreira, A. A., Laender, A. H. F., & Gonçalves, M. A. (2011). Incremental Unsupervised Name Disambiguation in Cleaned Digital Libraries. *Journal of Information and Data Management*, *2*(3), 289–289. Retrieved from https://periodicos.ufmg.br/index.php/jidm/article/view/128

de Solla Price, D. J. (1963). *Little Science, Big Science*. New York: Columbia University Press.

de Solla Price, D. J. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*(5), 292–306. https://doi.org/10.1002/asi.4630270505

Dee, T., Dee, & Thomas. (2004). Are there civic returns to education? *Journal of Public Economics, 88*(9–10), 1697–1720. Retrieved from https://econpapers.repec.org/RePEc:eee:pubeco:v:88:y:2004:i:9-10:p:1697-1720

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum Likelihood from Incomplete

Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Destatis. (2019). *Bildung und Kultur: Promovierendenstatistik*. Retrieved from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Publikationen/Downloads-Hochschulen/promovierendenstatistik-5213501197004.pdf?__blob=publicationFile

Deutsche Demokratische Republik. (1968). *Verordnung über die akademischen Grade vom 06.11.1968*.

Deutsche Nationalbibliothek. (2019). The German National Library in brief. Retrieved September 5, 2019, from https://www.dnb.de/EN/Ueber-uns/Portraet/portraet_node.html

Di Paolo, A., & Mañé, F. (2016). Misusing our talent? Overeducation, overskilling and skill underutilisation among Spanish PhD graduates. *The Economic and Labour Relations Review*, *27*(4), 432–452. https://doi.org/10.1177/1035304616657479

Dolton, P. J., & Silles, M. A. (2008). The effects of over-education on earnings in the graduate labour market. *Economics of Education Review*, *27*(2), 125–139. https://doi.org/10.1016/j.econedurev.2006.08.008

Dolton, P., & Vignoles, A. (2000). The incidence and effects of overeducation in the U.K. graduate labour market. *Economics of Education Review*, *19*(2), 179–198. Retrieved from www.elsevier.com/locate/econedurev

Enamorado, T., & Fifield, B. (2019). Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records. *American Political Science Review*, *113*, 353–371. https://doi.org/10.1017/S0003055418000783

Enamorado, T., Fifield, B., & Imai, K. (2020). *Package "fastLink": Fast Probabilistic Record Linkage with Missing Data*. Retrieved from http://imai.fas.harvard.edu/research/linkage.html

Engelage, S., & Schubert, F. (2009). PhD and Career – Do academics with a doctoral degree in Switzerland find adequate jobs? *Zeitschrift Fur Arbeitsmarktforschung*, *42*(3), 213–233. https://doi.org/10.1007/s12651-009-0017-7

Enserink, M. (2009). Scientific publishing. Are you ready to become a number? *Science*, *323*(5922), 1662–1664. https://doi.org/10.1126/science.323.5922.1662

Falk, S., & Küpper, H.-U. (2013). Verbessert der Doktortitel die Karrierechancen von Hochschulabsolventen? *Beiträge Zur Hochschulforschung*, *35*(1), 58–77.

Fan, X., Wang, J., Pu, X., Zhou, L., & Lv, B. (2011). On graph-based name disambiguation. *Journal of Data and Information Quality*, *2*(2), 1–23. https://doi.org/10.1145/1891879.1891883

Feeney, M. K., & Bozeman, B. (2008). Mentoring and network ties. *Human Relations*, *61*(12), 1651–1676. https://doi.org/10.1177/0018726708098081

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Ferreira, A. A., Veloso, A., Gonçalves, M. A., & Laender, A. H. F. (2014). Self-training author name disambiguation for information scarce scenarios. *Journal of the Association for Information Science and Technology*, *65*(6), 1257–1278. https://doi.org/10.1002/asi.22992

Fisher, A. J., Mendoza-Denton, R., Patt, C., Young, I., Eppig, A., Garrell, R. L., … Richards, M. A. (2019). Structure and belonging: Pathways to success for underrepresented minority and women PhD students in STEM fields. *PLOS ONE, 14*(1), e0209279. https://doi.org/10.1371/journal.pone.0209279

Fox, M. F., & Stephan, P. E. (2001). Careers of young scientists: Preferences, prospects and realities by gender and field. *Social Studies of Science, 31*(1), 109–122. https://doi.org/10.1177/030631201031001006

Frey, B. S. (2008). Evaluitis — eine neue Krankheit. In H. Matthies & S. Dagmar (Eds.), *Wissenschaft unter Beobachtung* (pp. 125–140). https://doi.org/10.1007/978-3-531-90863-2_8

Fuchs, M., Rauscher, C., & Weyh, A. (2014). Lohnhöhe und Lohnwachstum: Die regionalen Unterschiede in Deutschland sind groß. *IAB-Kurzbericht, 2014*(17).

Fuchs, M., & Rehs, A. (2019). Erwerbsbiographien ost-und westdeutscher Promovierter nach der Wiedervereinigung: Gleiche Qualifikation, gleiche Karriereverläufe? *ifo Dresden berichtet, 26*(06), 17-22.

Fuchs, M., & Rehs, A. (2020). Career paths of PhD holders in eastern and western Germany Same qualification, same labor market outcomes? *IAB-Discussion Paper, 2020*(1). Retrieved from http://doku.iab.de/discussionpapers/2020/dp0120.pdf

Fudickar, R., Hottenrott, H., & Lawson, C. (2018). What's the price of academic consulting? Effects of public and private sector consulting on academic research. *Industrial and Corporate Change, 27*(4), 699–722. https://doi.org/10.1093/icc/dty007

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178*(4060), 471–479. https://doi.org/10.1126/science.178.4060.471

Gaule, P., & Piacentini, M. (2018). An advisor like me? Advisor gender and post-graduate careers in science. *Research Policy, 47*(4), 805–813. https://doi.org/10.1016/j.respol.2018.02.011

Geißler, R. (2014). *Die Sozialstruktur Deutschlands*. https://doi.org/10.1007/978-3-531-19151-5

Gilman, H. (1965). Economic discrimination and unemployment. *American Economic Review, 55*(5), 1077–1096.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics, 56*(3), 357–367. https://doi.org/10.1023/A:1022378804087

Glänzel, W., & Schubert, A. (2006). Analysing Scientific Networks Through Co-Authorship. In H. F. Moed, U. Schmoch, & W. Glänzel (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 257–276). https://doi.org/10.1007/1-4020-2755-9_12

Granato, N., Haas, A., Hamann, S., & Niebuhr, A. (2010). Arbeitskräftemobilität in Deutschland – Qualifikationsspezifische Befunde regionaler Wanderungs- und Pendlerströme. *Raumforschung Und Raumordnung, 67*(1), 21–33. https://doi.org/10.1007/bf03183140

Grave, B. S., & Goerlitz, K. (2012). Wage differentials by field of study - the case of German university graduates. *Education Economics, 20*(3), 284–302. https://doi.org/10.1080/09645292.2012.680549

Gray, D. E., & Goregaokar, H. (2010). Choosing an executive coach: The influence of gender on the coach-coachee matching process. *Management Learning, 41*(5), 525–544. https://doi.org/10.1177/1350507610371608

Gruhn, W., & Lauterbach, G. (1977). Die Organisation der Forschung in der DDR. In

Institut für Gesellschaft und Wissenschaft (pp. 127–213). Erlangen: Campus Verlag.

Guenther, K. (1989). *Das Bildungswesen der Deutschen Demokratischen Republik: Gemeinschaftsarbeit der Akademie der Pädagogischen Wissenschaften*. Berlin: Volk und Wissen.

Gupta, B. M., Kumar, S., & Aggarwal, B. S. (1999). A comparision of productivity of male and female scientists of CSIR. *Scientometrics*, *45*(2), 269–289. https://doi.org/10.1007/BF02458437

Gutmann, G. (1979). Employment problems under socialism. *Intereconomics*, *14*(2), 96–100. https://doi.org/10.1007/BF02930205

Hachmeister, C.-D. (2019). *Im Blickpunkt: Promotionen als Indikator für die Leistung von Hochschulen Auswertung von Daten des Statistischen Bundesamtes und des CHE Rankings 2019/20*. Retrieved from https://www.che.de/wp-content/uploads/upload/Im_Blickpunkt_Promotionen_2019.pdf

Hagen, N. T. (2010). Deconstructing doctoral dissertations: How many papers does it take to make a PhD? *Scientometrics*, *85*(2), 567–579. https://doi.org/10.1007/s11192-010-0214-8

Hahn, E. (2009). Publikationsverhalten in der Chemie. Publikationsverhalten in unterschiedlichen wissenschaftlichen Disziplinen: Beiträge zur Beurteilung von Forschungsleistungen. Retrieved August 1, 2020, from https://www.humboldt-foundation.de/pls/web/docs/F13905/12_disk_papier_publikationsverhalten2_kompr.pdf.

Hartmann, M., & Kopp, J. (2001). Elite selection by means of education or by means of social origin? Doctorate, social origin and the recruitment of the german business elite. *Kolner Zeitschrift Fur Soziologie Und Sozialpsychologie*, *53*(3), 436–466. https://doi.org/10.1007/s11577-001-0074-6

Haustein, S., & Larivière, V. (2015). The use of bibliometrics for assessing research: Possibilities, limitations and adverse effects. In I. Welpe, U. Wollersheim, S. Ringelhan, & M. Osterloh (Eds.), *Incentives and Performance: Governance of Research Organizations* (pp. 121–139). https://doi.org/10.1007/978-3-319-09785-5_8

Heckman, J. J., & Moktan, S. (2020). Publishing and promotion in economics: The tyranny of the top five. *Journal of Economic Literature*, *58*(2), 419–470. https://doi.org/10.1257/JEL.20191574

Heineck, G., & Matthes, B. (2012). Zahlt sich der Doktortitel aus? Eine Analyse zu monetären und nicht-monetären Renditen der Promotion. In N. Huber, A. Schelling, & S. Hornbostel (Eds.), *Der Doktortitel zwischen Status und Qualifikation. , IFQ-Working Paper No. 12* (pp. 85–89).

Heinisch, D. P., & Buenstorf, G. (2018). The next generation (plus one): an analysis of doctoral students’ academic fecundity based on a novel approach to advisor identification. *Scientometrics*, *117*(1), 351–380. https://doi.org/10.1007/s11192-018-2840-5

Heinisch, D. P., Koenig, J., & Otto, A. (2020). A supervised machine learning approach to trace doctorate recipients’ employment trajectories. *Quantitative Science Studies*, *1*(1), 94–116. https://doi.org/10.1162/qss_a_00001

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, *44*(2), 193–

215. https://doi.org/10.1007/BF02457380

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

Hilmer, C., & Hilmer, M. (2007). Women helping women, men helping women? Same-gender mentoring, initial job placements, and early career publishing success for economics PhDs. *American Economic Review, 97*(2), 422–426. https://doi.org/10.1257/aer.97.2.422

Hirsch, J. E. (2005). 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), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Horta, H., Cattaneo, M., & Meoli, M. (2018). PhD funding as a determinant of PhD and career research performance. *Studies in Higher Education, 43*(3), 542–570. https://doi.org/10.1080/03075079.2016.1185406

Horta, H., & Santos, J. M. (2016). The Impact of Publishing During PhD Studies on Career Research Publication, Visibility, and Collaborations. *Research in Higher Education, 57*(1), 28–50. https://doi.org/10.1007/s11162-015-9380-0

Hsieh, C.-T., Hurst, E., Jones, C. I., & Klenow, P. J. (2019). The Allocation of Talent and U.S. Economic Growth. *Econometrica, 87*(5), 1439–1474. https://doi.org/10.3982/ecta11427

Huaco, G. A. (1966). The functionalist theory of stratification: Two decades of controversy. *Inquiry (United Kingdom), 9*(1–4), 215–240. https://doi.org/10.1080/00201746608601459

Hussain, I., & Asghar, S. (2017). A survey of author name disambiguation techniques: 2010–2016. *The Knowledge Engineering Review, 32*. https://doi.org/10.1017/s0269888917000182

Igami, M. Z., Bressiani, J., & Mugnaini, R. (2014). A new model to identify the productivity of theses in terms of articles using co-word analysis. *Journal of Scientometric Research, 3*(1), 3. https://doi.org/10.4103/2320-0057.143660

Jaksztat, S. (2014). Bildungsherkunft und Promotionen: Wie beeinflusst das elterliche Bildungsniveau den Übergang in die Promotionsphase? *Zeitschrift Fur Soziologie, 43*(4), 286–301. https://doi.org/10.1515/zfsoz-2014-0404

Jaksztat, S. (2017). Geschlecht und wissenschaftliche Produktivität: Erklären Elternschaft und wissenschaftliches Sozialkapital Produktivitätsunterschiede während der Promotionsphase? *Zeitschrift Fur Soziologie, 46*(5), 347–361. https://doi.org/10.1515/zfsoz-2017-1019

Jinha, A. (2010). Article 50 million: An estimate of the number of scholarly articles in existence. *Learned Publishing, 23*(3), 258–263. https://doi.org/10.1087/20100308

Karahan, F., Guvenen, F., Ozkan, S., & Song, J. (2015). What Do Data on Millions of U.S. Workers Reveal About Life-Cycle Earnings Risk? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2563279

Kehm, B. M. (2006). Doctoral education in Europe and North America: a comparative analysis. *Wenner Gren International Series, 83*, 67–78.

Kim, J. (2019). A fast and integrative algorithm for clustering performance evaluation in author name disambiguation. *Scientometrics, 120*(2), 661–681. https://doi.org/10.1007/s11192-019-03143-7

Kim, J., & Kim, J. (2018). The impact of imbalanced training data on machine learning for author name disambiguation. *Scientometrics, 117*(1), 511–526.

https://doi.org/10.1007/s11192-018-2865-9

Kim, K., Rohatgi, S., & Lee Giles, C. (2019). Hybrid deep pairwise classification for author name disambiguation. *International Conference on Information and Knowledge Management, Proceedings*, 2369–2372. https://doi.org/10.1145/3357384.3358153

Kim, S., Hansen, D., & Helps, R. (2018). Computing research in the academy: insights from theses and dissertations. *Scientometrics*, *114*(1), 135–158. https://doi.org/10.1007/s11192-017-2572-y

Koenig, J. (2019). Leaving on a high note? Postdoctoral academic employment and future non-academic career prospects. *Paper Presented at Druid Conference 2019*. Retrieved from https://conference.druid.dk/acc_papers/ogoaxil6awenxf0t5fpt359znos0m9.pdf

Kolloch, E. (2001). Abwicklung und Neuaufbau der wirtschaftswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin zwischen November 1989 und Dezember 1993. In F. Theißen (Ed.), *Zwischen Plan und Pleite. Erlebnisberichte aus der Arbeitswelt der DDR*. Bühlau Verlag.

Konsortium Bundesbericht Wissenschaftlicher Nachwuchs. (2017). *Bundesbericht Wissenschaftlicher Nachwuchs 2017*. Bielefeld: W. Bertelsmann.

Kousha, K., & Thelwall, M. (2019). Can Google Scholar and Mendeley help to assess the scholarly impacts of dissertations? *Journal of Informetrics*, *13*(2), 467–484. https://doi.org/10.1016/j.joi.2019.02.009

Kousha, K., & Thelwall, M. (2020). Google Books, Scopus, Microsoft Academic and Mendeley for impact assessment of doctoral dissertations: A multidisciplinary analysis of the UK. *Quantitative Science Studies*, *1*(2), 1–26. https://doi.org/10.1162/qss_a_00042

Kracke, N., Reichelt, M., & Vicari, B. (2018). Wage Losses Due to Overqualification: The Role of Formal Degrees and Occupational Skills. *Social Indicators Research*, *139*(3), 1085–1108. https://doi.org/10.1007/s11205-017-1744-8

Kuhn, T. S. (2012). *The structure of scientific revolutions*. Chicago: University of Chicago press.

Landhuis, E. (2016). Scientific literature: Information overload. *Nature*, *535*(7612), 457–458. https://doi.org/10.1038/nj7612-457a

Larivière, V. (2012). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. *Scientometrics*, *90*(2), 463–481. https://doi.org/10.1007/s11192-011-0495-6

Larivière, V. (2013). PhD students' excellence scholarships and their relationship with research productivity, scientific impact, and degree completion. *Canadian Journal of Higher Education*, *43*(2), 27–41. Retrieved from https://journals.sfu.ca/cjhe/index.php/cjhe/article/view/2270

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature*, *504*(7479), 211–213. https://doi.org/10.1038/504211a

Larsen, P. O., & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, *84*(3), 575–603. https://doi.org/10.1007/s11192-010-0202-z

Leahey, E. (2006). Gernder differences in productivity: Research Specialization as a Missing Link. *Gender and Soceity*, *20*(6), 754–780.

https://doi.org/10.1177/0891243206293030

Leibbrandt, A., & List, J. A. (2015). Do women avoid salary negotiations? Evidence from a large-scale natural field experiment. *Management Science, 61*(9), 2016–2024. https://doi.org/10.1287/mnsc.2014.1994

Leininger, W. (2008). Publikationsverhalten in den Wirtschaftswissenschaften. In *Publikationsverhalten in unterschiedlichen wissenschaftlichen Disziplinen: Beiträge zur Beurteilung von Forschungsleistungen* (pp. 39–40). Alexander von Humboldt Foundation.

Leuven, E., & Oosterbek, H. (2011). Overeducation and mismatch in the labor market. In E. A. Hanushek, S. Machin, & L. Wössmann (Eds.), *Handbook of the Economics of Education* (pp. 283–326). Amsterdam: Elsevier.

Levin, S. G., & Stephan, P. E. (1991). Research Productivity Over the Life Cycle: Evidence for Academic Scientists. *American Economic Review, 81*(1), 114–132.

Levitt, S. D. (2004). Testing theories of discrimination: Evidence from weakest link. *Journal of Law and Economics, 47*(2), 431–452. https://doi.org/10.1086/425591

Leydesdorff, Loet; Milojevič, S. (2015). Scientometrics. In J. D. Wright (Ed.), *International Encyclopedia of Social and Behavioral Sciences* (2nd ed., pp. 322–327). Amsterdam: Elsevier.

Lippens, L., Baert, S., Ghekiere, A., Verhaeghe, P.-P., & Derous, E. (2020). Is labour market discrimination against ethnic minorities better explained by taste or statistics? A systematic review of the empirical evidence. *GLO Discussion Paper Series*. Retrieved from https://ideas.repec.org/p/zbw/glodps/615.html

List, J. A. (2004). The nature and extent of discrimination in the marketplace: Evidence from the field. *Quarterly Journal of Economics*, *119*(1), 49–89. https://doi.org/10.1162/003355304772839524

Liu, W., Hu, G., & Tang, L. (2018). Missing author address information in Web of Science - An explorative study. *Journal of Informetrics*, *12*(3), 985–997. https://doi.org/10.1016/j.joi.2018.07.008

Lotka, A. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*(12), 317–323.

Louppe, G., Al-Natsheh, H. T., Susik, M., & Maguire, E. J. (2016). Ethnicity sensitive author disambiguation using semi-supervised learning. *Communications in Computer and Information Science*, *649*, 272–287. https://doi.org/10.1007/978-3-319-45880-9_21

Lukas, J., & Reinhard, D. (2016, May 19). Wer beherrscht den Osten? *Die Zeit*. Retrieved from https://www.zeit.de/2016/22/ostdeutschland-2016-macht-einfluss-mdr-doku

Mann, R. (1979). Internationale Wissenschaftsbeziehungen. In Institut für Gesellschaft und Wissenschaft (Ed.), *Das Wissenschaftssystem in der DDR*. Berlin: Campus Verlag.

Marx, K., & Engels, F. (1848). *The Communist Manifesto*. Retrieved from http://la.utexas.edu/users/hcleaver/368/368CommunistManifestoPtItable.pdf

May, R. M. (1997). The scientific wealth of nations. *Science*, *275*(5301), 793–796. https://doi.org/10.1126/science.275.5301.793

McGuinness, S. (2006). Overeducation in the labour market. *Journal of Economic Surveys*, *20*(3), 387–418. https://doi.org/10.1111/j.0950-0804.2006.00284.x

Medoff, M. H. (2006). Evidence of a Harvard and Chicago Matthew Effect. *Journal of Economic Methodology*, *13*(4), 485–506. https://doi.org/10.1080/13501780601049079

Merga, M. K., Mason, S., & Morris, J. E. (2020). ‘What do I even call this?’ Challenges and possibilities of undertaking a thesis by publication. *Journal of Further and Higher Education*, *44*(9), 1245–1261. https://doi.org/10.1080/0309877X.2019.1671964

Mertens, A., & Röbken, H. (2013). Does a doctoral degree pay off? An empirical analysis of rates of return of German doctorate holders. *Higher Education*, *66*(2), 217–231. https://doi.org/10.1007/s10734-012-9600-x

Merton, R. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago: University of Chicago Press.

Merton, R. K. (1957). Priorities in Scientific Discovery: A Chapter in the Sociology of Science. *American Sociological Review*, *22*(6), 635. https://doi.org/10.2307/2089193

Merton, R. K. (1968). The matthew effect in science. *Science*, *159*(3810), 56–62. https://doi.org/10.1126/science.159.3810.56

Meske, W. (2004). *From System Transformation to European Integration. Science and technology in Central and Eastern Europe at the beginning of the 21st century*. Münster: LIT Verlag.

Meyer, M. (2011). Bibliometrics, stylized facts and the way ahead: How to build good social simulation models of science? *JASSS*, *14*(4). https://doi.org/10.18564/jasss.1824

Millar, M. M. (2013). Interdisciplinary research and the early career: The effect of interdisciplinary dissertation research on career placement and publication productivity of doctoral graduates in the sciences. *Research Policy*, *42*(5), 1152–1164. https://doi.org/10.1016/j.respol.2013.02.004

Mincer, J. (1958). Investment in Human Capital and Personal Income Distribution. *Journal of Political Economy*, *66*(4), 281–302. https://doi.org/10.1086/258055

Mincer, J., & Polachek, S. (1974). Family Investments in Human Capital: Earnings of Women. *Journal of Political Economy*, *82*(2, Part 2), S76–S108. https://doi.org/10.1086/260293

Möller, C. (2015). *Herkunft zählt (fast) immer. Soziale Ungleichheiten unter Universitätsprofessorinnen und –professoren*. Weinheim, Basel: Beltz Juventa.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Morichika, N., & Shibayama, S. (2016). Use of dissertation data in science policy research. *Scientometrics*, *108*(1), 221–241. https://doi.org/10.1007/s11192-016-1962-x

Nelson, R. R., & Winter, S. G. (1982). *An Evolutionary Theory of Economic Change* (Belknap Press of Harvard University Press, Ed.). Cambridge MA and London.

Neumark, D., & Gardecki, R. (1998). Women helping women? Role model and mentoring effects on female Ph.D. students in economics. *Journal of Human Resources*, *33*(1), 220–246. https://doi.org/10.2307/146320

Newman, M. (2018). *Networks* (2.). Oxford University Press.

Noe, R. A. (1988). Women and Mentoring: A Review and Research Agenda. *Academy of Management Review*, *13*(1), 65–78. https://doi.org/10.5465/amr.1988.4306784

Oaxaca, R. (2001). Economics of Discrimination. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 3756–3762). https://doi.org/10.1016/b0-08-043076-7/02292-0

Oaxaca, Ronald. (1973). Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, *14*(3), 693. https://doi.org/10.2307/2525981

OECD. (2019). OECD work on careers of doctorate holders. Retrieved December 12, 2020, from https://www.oecd.org/innovation/inno/careers-of-doctorate-holders.htm

On, B. W., Lee, I., & Lee, D. (2012). Scalable clustering methods for the name disambiguation problem. *Knowledge and Information Systems*, *31*(1), 129–151. https://doi.org/10.1007/s10115-011-0397-1

Onodera, N., Iwasawa, M., Midorikawa, N., Yoshikane, F., Amano, K., Ootani, Y., … Yamazaki, S. (2011). A method for eliminating articles by homonymous authors from the large number of articles retrieved by author search. *Journal of the American Society for Information Science and Technology*, *62*(4), 677–690. https://doi.org/10.1002/asi.21491

Ooms, J. (2018). cld3: Google’s Compact Language Detector 3. Retrieved February 7, 2019, from https://cran.r-project.org/web/packages/cld3/cld3.pdf

Pasternack, P. (2007). *Forschungslandkarte Ostdeutschland*. Halle-Wittenberg: Institut für Hochschulforschung: HoF.

Paulus, W., & Matthes, B. (2013). The German classification of occupations 2010: structure, coding and conversion table. *FDZ-Methodenreport*, *2013*(8).

Peisert, H., & Framheim, G. (1994). *Das Hochschulsystem in der Bundesrepublik Deutschland: Struktur und Entwicklungstendenzen*. Bad Honnef: Bock.

Pezzoni, M., Mairesse, J., Stephan, P., & Lane, J. (2016). Gender and the publication output of graduate students: A case study. *PLoS ONE*, *11*(1). https://doi.org/10.1371/journal.pone.0145146

Phelps, E. S. (1972). The Statistical Theory of Racism and Sexism. *American Economic Review*, *62*(4), 659–661. Retrieved from https://www.researchgate.net/publication/4728049

Polanyi, M. (2015). *Personal knowledge: towards a post-critical philosophy*. Chicago: University of Chicago Press.

Prpić, K. (2002). Gender and productivity differentials in science. *Scientometrics*, *55*(1), 27–58. https://doi.org/10.1023/A:1016046819457

Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. H. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(41), 10870–10875. https://doi.org/10.1073/pnas.1706255114

Ramage, D., Dumais, S., & Liebling, D. (2010). Characterizing microblogs with topic models. *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*.

Rehs, A. (2020a). A structural topic model approach to scientific reorientation of economics and chemistry after German reunification. *Scientometrics*. https://doi.org/10.1007/s11192-020-03640-0

Rehs, A. (2020b). Dataset: A structural topic model approach to scientific re-orientation of economics and chemistry after German reunification. https://doi.org/10.5281/zenodo.3895119

Rehs, A. (2021). A supervised machine learning approach to author disambiguation in the Web of Science. *Journal of Informetrics*, *15*(3).

Rehs, A. (2021). The scientific productivity of German PhD graduates: A machine learning-based author name disambiguation and record linkage approach. *Proceedings of the 18th conference of Scientometrics & Informetrics*, 1531-1533.

Rehs, A. (2021). Protégé-advisor gender-pairings in academic survival and productivity of German PhD graduates. *Proceedings of the 18th Conference on Scientometrics and Informetrics*, 955-967.

Reichelt, M., & Vicari, B. (2014). Ausbildungsinadäquate Beschäftigung in Deutschland: Im Osten sind vor allem Ältere für ihre Tätigkeit formal überqualifiziert. *IAB-Kurzbericht*, *2014*(25).

Rimmert, C., Schwechheimer, H., & Winterhager, M. (2017). *Disambiguation of author addresses in bibliometric databases-technical report*. Retrieved from https://pub.uni-bielefeld.de/download/2914944/2914947/DisambiguationOfAuthorAddressesInBibliometricDatabases.pdf

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003. https://doi.org/10.1080/01621459.2016.1141684

Roberts, M. E., Stewart, B. M., & Dustin, T. (2019). Journal of Statistical Software stm: R Package for Structural Topic Models. *Journal of Statistical Software*, *91*(2). https://doi.org/10.18637/jss.v000.i00

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

Romer, P. (1990). Endogenous Technological Change. *Journal of Political Economy*, *98*(5), S71-102. Retrieved from https://econpapers.repec.org/RePEc:ucp:jpolec:v:98:y:1990:i:5:p:s71-102

Romer, P. (1994). The Origins of Endogenous Growth. *Journal of Economic Perspectives*, *8*(1), 3–22. https://doi.org/10.1257/JEP.8.1.3

Rossen, A., Boll, C., & Wolf, A. (2019). Patterns of Overeducation in Europe: The Role of Field of Study. *IZA Journal of Labor Policy*, *9*(1). https://doi.org/10.2478/izajolp-2019-0003

Rossiter, M. W. (1993). The Matthew Matilda Effect in Science. *Social Studies of Science*, *23*(2), 325–341. https://doi.org/10.1177/030631293023002004

Rosvall, M., & Bergstrom, C. T. (2007). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(4), 1118–1123. https://doi.org/10.1073/pnas.0706851105

Rukwid, O. (2012). Deutschland, Grenzen der Bildungsexpansion? Ausbildungsinadäquate Beschäftigung von Ausbildungs- und Hochschulabsolventen in Deutschland. *Schriftenreihe Des Promotionsschwerpunkts Globalisierung Und Beschäftigung*, *37*.

Sabatier, M., Carrere, M., & Mangematin, V. (2006). Profiles of academic activities and careers: Does gender matter? An analysis based on french life scientist CVs. *Journal of Technology Transfer*, *31*(3), 311–324. https://doi.org/10.1007/s10961-006-7203-3

Salheiser, A. (2012). Socialist and post-socialist functional elites in East Germany. *Historical Social Research*, Vol. 37, pp. 123–138. https://doi.org/10.12759/hsr.37.2012.2.123-138

Sayers, A., Ben-Shlomo, Y., Blom, A. W., & Steele, F. (2016). Probabilistic record linkage. *International Journal of Epidemiology*, *45*(3), 954–964. https://doi.org/10.1093/ije/dyv322

Schmoch, U., & Schulze, N. (2010). Performance and structures on the German science system in an international comparison 2009 with a special focus on East Germany. *Studien Zum Deutschen Innovationssystem*. Retrieved from https://ideas.repec.org/p/zbw/efisdi/82010.html

Schnabel, C. (2016). United, Yet Apart? A Note on Persistent Labour Market Differences between Western and Eastern Germany. *Jahrbucher Fur Nationalokonomie Und Statistik*, *236*(2), 157–179. https://doi.org/10.1515/jbnst-2015-1012

Schulz, J. (2016). Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analyses. *Scientometrics*, *107*(3), 1283–1298. https://doi.org/10.1007/s11192-016-1892-7

Sen, A. (1995). *Inequality Reexamined*. https://doi.org/10.1093/0198289286.001.0001

Shin, D., Kim, T., Choi, J., & Kim, J. (2014). Author name disambiguation using a graph model with node splitting and merging based on bibliographic information. *Scientometrics*, *100*(1), 15–50. https://doi.org/10.1007/s11192-014-1289-4

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, *87*(3), 355–374. https://doi.org/10.2307/1882010

Statistisches Bundesamt. (2020). Frauenanteile nach akademischer Laufbahn. Retrieved October 29, 2020, from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Tabellen/frauenanteile-akademischelaufbahn.html

Steeg, M. van der, Wiel, K. van der, & Wouterse, B. (2014). Individual returns to a PhD education in the Netherlands: income differences between Masters and PhDs. *CPB Discussion Paper*, (276).

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

Stephan, P. E., Sumell, A. J., Black, G. C., & Adams, J. D. (2004). Doctoral education and economic development: The flow of new Ph.D.s to industry. *Economic Development Quarterly*, *18*(2), 151–167. https://doi.org/10.1177/0891242403262019

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63*(9), 1820–1833. https://doi.org/10.1002/asi.22695

Stüber, H. (2016). Berufsspezifische Lebensentgelte: Qualifikation zahlt sich aus. *IAB-Kurzbericht*, *2016*(7).

Sturm, J.-E., & Ursprung, H. W. (2017). The Handelsblatt Rankings 2.0: Research Rankings for the Economics Profession in Austria, Germany, and Switzerland.

German Economic Review, *18*(4), 492–515. https://doi.org/10.1111/geer.12145

Talburt, J. R. (2011). Entity Resolution and Information Quality. In *Entity Resolution and Information Quality*. https://doi.org/10.1016/C2009-0-63396-1

Tang, J., Fong, A. C. M., Wang, B., & Zhang, J. (2011). A Unified Probabilistic Framework for Name Disambiguation in Digital Library. *IEEE Transactions on Knowledge and Data Engineering*, *24*(6), 975–987. https://doi.org/10.1109/TKDE.2011.13

Teichert, C., Niebuhr, A., Otto, A., & Rossen, A. (2018). Graduate migration in Germany - new evidence from an event history analysis. *IAB Discussion Paper*. Retrieved from https://ideas.repec.org/p/iab/iabdpa/201803.html

Tekles, A., & Bornmann, L. (2019). Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. *17th International Conference on Scientometrics and Informetrics, ISSI 2019 - Proceedings*, *2*, 1548–1559. Retrieved from http://arxiv.org/abs/1904.12746

Thijssen, L., Coenders, M., & Lancee, B. (2021). Is there evidence for statistical discrimination against ethnic minorities in hiring? Evidence from a cross-national field experiment. *Social Science Research*, *93*, 102482. https://doi.org/10.1016/j.ssresearch.2020.102482

Tol, R. S. J. (2009). The matthew effect defined and tested for the 100 most prolific economists. *Journal of the American Society for Information Science and Technology*, *60*(2), 420–426. https://doi.org/10.1002/asi.20968

Tol, R. S. J. (2013). The Matthew effect for cohorts of economists. *Journal of Informetrics*, *7*(2), 522–527. https://doi.org/10.1016/j.joi.2013.02.001

Torvik, V. I., & Smalheiser, N. R. (2009). Author Name Disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data*, *3*(3). https://doi.org/10.1145/1552303.1552304

Torvik, V. I., Weeber, M., Swanson, D. R., & Smalheiser, N. R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology*, *56*(2), 140–158. https://doi.org/10.1002/asi.20105

Treeratpituk, P., & Giles, C. L. (2009). Disambiguating authors in academic publications using random forests. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 39–48. https://doi.org/10.1145/1555400.1555408

Tregellas, J. R., Smucny, J., Rojas, D. C., & Legget, K. T. (2018). Predicting academic career outcomes by predoctoral publication record. *PeerJ*, *2018*(10). https://doi.org/10.7717/peerj.5707

Tutz, G., & Schmid, M. (2016). *Modeling Discrete Time-to-Event Data*. https://doi.org/10.1007/978-3-319-28158-2

United Nations. (2015). *Concepts of Inequality*. Retrieved from https://www.un.org/en/development/desa/policy/wess/wess_dev_issues/dsp_policy_01.pdf

van de Schoot, R. (2020). Intro to Discrete-Time Survival Analysis in R. Retrieved February 18, 2021, from https://www.rensvandeschoot.com/tutorials/discrete-time-survival/

Volkskammer der DDR. (1976). *Verfassung der Deutschen Demokratischen Republik vom 6. April 1968 in der Fassung des Gesetzes zur Ergänzung und Änderung der Verfassung der Deutschen Demokratischen Republik vom 7. Oktober 1974.* Berlin:

Staatsverlag der Deutschen Demokratischen Republik.

Waldinger, F. (2010). Quality matters: The expulsion of professors and the consequences for PhD student outcomes in Nazi Germany. *Journal of Political Economy*, *118*(4), 787–831. https://doi.org/10.1086/655976

Weingart, P, Strate, J., & Winterhager, M. (1991). *Bibliometrisches Profil der DDR. Bericht an den Stifterverband für die Deutsche Wissenschaft und den Wissenschaftsrat, Universitätsschwerpunkt Wissenschaftsforschung*.

Weingart, Peter. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, *62*(1), 117–131. https://doi.org/10.1007/s11192-005-0007-7

Wollgast, S. (2001). *Zur Geschichte des Promotionswesens in Deutschland*. Bergisch Gladbach: Dr. Frank Graetz Verlag.

Wu, H., Li, B., Pei, Y., & He, J. (2014). Unsupervised author disambiguation using Dempster–Shafer theory. *Scientometrics*, *101*(3), 1955–1972. https://doi.org/10.1007/s11192-014-1283-x