Exploring the Dimensions of Scientific Impact: A Comprehensive Bibliometric Analysis Investigating the Influence of Gender, Mobility, and Open Access

Inaugural dissertation

for the attainment of the title of doctor

in the Faculty of Mathematics and Natural Sciences

at the Heinrich Heine University Düsseldorf

Presented by

Fakhri Momeni

from Esfahan, Iran

August, 2023

Heinrich Heine Universität Düsseldorf

gesis Leibniz-Institut für Sozialwissenschaften

from the Institute for Informatik

at the Heinrich Heine University Düsseldorf

Published by permission of the

Faculty of Mathematics and Natural Sciences at

Heinrich Heine University Düsseldorf

Supervisor: Prof. Dr. Stefan Dietze

Co-supervisor: Prof. Dr. Isabella Peters

Date of the oral examination: 26. October 2023

Statutory Declaration

I herewith formally declare that I have written the submitted dissertation independently. I did not use any outside support except for the quoted literature and other sources mentioned in the thesis.

I clearly marked and separately listed all of the literature and all of the other sources which I employed when producing this academic work, either literally or in content.

I am aware that the violation of this regulation will lead to failure of the thesis.

Düsseldorf, Germany — Fakhri Momeni

2. August 2023

Dedicated to my parents and husband

Acknowledgments

This dissertation is the culmination of my research conducted at the Department of Knowledge Technologies for the Social Sciences (KTS) at GESIS – Leibniz Institute for the Social Sciences in Cologne. I am deeply indebted to Peter Mutschke for his trust in me and for providing me with the incredible opportunity to work at GESIS and engage in research. Although we didn't directly work together, his decision to bring me into the team significantly shaped my journey.

I would also like to extend my sincere appreciation to Dr. Philipp Mayr, who has been an unwavering source of support throughout all phases of my thesis and different research projects. His mentorship and guidance have been instrumental in my professional growth and development.

I am grateful to my first supervisor, Prof. Dietze, for his guidance and valuable input during the early stages of my research. His expertise and constructive feedback have significantly contributed to the quality of my work.

I would also like to express my gratitude to my second supervisor, Prof. Isabella Peters, for her invaluable support and guidance throughout the completion of my thesis. Her extensive knowledge, insightful feedback, and encouragement have been vital in shaping the final outcome of my research. I would like to express my deepest gratitude to my husband for his unwavering support throughout my PhD journey. His constant encouragement, understanding, and belief in my abilities have been invaluable. His presence has been a source of strength, and I am truly grateful for his love and support in every aspect of my life.

Although my parents did not provide direct support, I cannot overlook their profound influ-

ence on shaping me into the hardworking individual I am today. Their resilience, determination, and unwavering commitment to their own work have inspired me to push through challenges and strive for excellence. I owe them a debt of gratitude for instilling in me the values of perseverance and dedication.

Finally, I would like to thank all my friends, colleagues, and fellow researchers who have supported and inspired me along this challenging journey. Their collaboration, discussions, and shared experiences have enriched my research and made this endeavor even more rewarding.

I offer my heartfelt appreciation to all those mentioned above and the countless others who have contributed to my growth as a researcher and individual. Your support and belief in me have been indispensable, and I am truly grateful for the opportunity to have worked with such exceptional individuals.

ABSTRACT

The Science of Science field advances the measurement, evaluation, and prediction of scientific outcomes through the study of extensive scholarly data. For these purposes, bibliometrics is an appropriate approach that studies large volumes of scientific data using mathematical and statistical methods, and is widely used to assess the impact of papers and authors within a specific field or community. However, conducting bibliometric analyses poses several methodological, technical, and informational challenges (e.g., collecting and cleaning data, calculating indicators) which need to be addressed. This thesis aims to tackle some of these challenges and shed light on the factors influencing scientific impact, specifically focusing on open access publishing, international mobility, and influential factors on the h-index. This thesis tackles methodological contributions, such as author disambiguation and co-authorship network analysis, as they provide insights into methodological and informational challenges within bibliometric analysis. Another methodological challenge addressed in this research is the inference of gender for a significant number of authors to obtain gender-related insights. By employing gender inference techniques, the research explores gender as an influential factor in scientific impact, shedding light on potential gender inequalities within the scholarly community. The research employs a bibliometric approach and utilizes mainly Scopus, a comprehensive dataset encompassing various disciplines to make the following contributions:

- We explore the impact of publishing behavior, particularly the adoption of open access practices, on knowledge dissemination and scholarly communication. With this intention, we investigate the impact of journals flipping from closed access to open access publishing models [74]. Changes in publication volumes and citation impact are analyzed, demonstrating an overall increase in publication output and improved citation metrics following the transition to open access. However, the magnitude of

changes varies across scientific disciplines. In another study [76], we utilize a dataset of articles published by Springer Nature and employ correlation and regression analyses to examine the relationship between authors' country affiliations, publishing models, and citation impact. Utilizing machine learning approach, we estimate the publishing model of papers based on different factors. The findings reveal different patterns in authors' choices of publishing models based on income levels, availability of Article Processing Charges waivers, and journal rank. The study highlights potential inequalities in access to open access publishing and its citation advantage.

- We investigate the association between scholars' mobility patterns, socio-demographic characteristics, and their scientific activity and impact. By utilizing network and regression analyses, along with various statistical techniques, we investigate the international mobility of researchers. Furthermore, we conduct a comparative analysis of scientific outcomes, considering factors such as publications, citations, and measures of co-authorship network centrality. The findings reveal gender inequalities in mobility across scientific fields and countries and positive correlations between mobility and scientific success.

- Centered on the prediction of scholars' h-index as a metric of scientific impact, another one of our studies [77] employs machine learning techniques. We examine author, co-authorship, paper, and venue-specific characteristics, in addition to prior impact-based features. The results emphasize the significance of non-prior impact-based features, particularly for early-career scholars in the long term, while also revealing the limited influence of gender on h-index prediction.

The findings of this research hold implications for researchers, academic institutions, and policymakers aiming to advance scientific knowledge and foster equitable practices. By un-

covering the influential factors that shape scientific impact and addressing potential gender disparities, this research contributes to the broader objective of promoting diversity, inclusivity, and excellence within the scholarly community.

Keywords: scientific impact, publishing behavior, open access, international academic mobility, gender analysis, author disambiguation, co-authorship network analysis, bibliometric analysis, machine learning.

Contents

Chapter 1

Introduction

Pursuing scientific inquiry is pivotal in advancing our understanding, spurring innovation, and tackling real-world challenges. As the pursuit of scientific excellence evolves, unraveling the intricate factors that shape scientific impact becomes increasingly crucial. This thesis explores key dimensions influencing scientific success, including open access (OA) publishing, international mobility, gender dynamics, and influential factors on the h-index. By exploring these facets, we strive to gain deeper insights into the complex scientific research ecosystem and contribute to the ongoing discourse surrounding research practices, collaboration, and evaluation. Through a rigorous examination of these dimensions, we hope to provide valuable perspectives and evidence-based recommendations that enhance the effectiveness of research assessment practices.

1.1 Motivation

Understanding the factors contributing to scientific impact is important to researchers, institutions, and policymakers. Motivated by the desire to enhance the effectiveness and efficiency of research assessment practices, this thesis investigates key aspects that influence scientific impact, namely OA publishing, international mobility, gender, and influential factors on the h-index. By leveraging the field of Science of Science, which furthers the measurement, evaluation, and prediction of scientific outcomes relying on big scholarly data [19], this research

aims to gain deeper insights into the complex dynamics that shape scientific impact.

One of the primary methodological challenges in bibliometric analyses is author disambiguation. Distinguishing between authors with similar names or different name variations is essential for accurately attributing publications and understanding individual researchers' contributions. By addressing the challenges associated with disambiguating authors, this research contributes to the reliability and validity of bibliometric analyses, enabling more accurate assessments of scientific impact.

Co-authorship network analysis is essential for understanding collaboration and knowledge exchange among researchers in a field or community. However, analyzing co-authorship networks presents technical challenges, including data collection, network construction, and the application of appropriate centrality measures. The initial phase of this thesis involved conducting network analysis within a small research community. This process deepened the researcher's understanding of centrality measures and their relevance in assessing authors' positions and influence within a collaborative network. By applying network analysis techniques to the study of scientific collaborations, this research expands our understanding of the dynamics and impact of collaboration on scientific outcomes.

Gender reference in bibliometric analysis enables researchers to investigate potential gender disparities in research output, collaboration patterns, and citation impact. However, gender information is often unavailable in bibliometric records, necessitating the inference of gender from external sources. This thesis employs a method introduced by Fariba Karimi et al. [47] to detect gender based on authors' names and images available on the web. By inferring gender information, this research contributes to addressing gender inequalities in science and provides insights into the potential impact of gender on scientific impact. In addition to the methodological and technical challenges, there are informational challenges in bibliometric analyses. The availability and accuracy of data, including publication records,

citation counts, and authors' affiliations, can significantly impact the validity and reliability of the findings. Careful attention to data quality, preprocessing, and validation is crucial to ensure the integrity of the analyses conducted in this thesis.

By addressing the methodological, technical, and informational challenges associated with bibliometric analyses, this research aims to understand the factors influencing scientific impact comprehensively. One of the main motivations behind studying OA publishing is the growing trend towards open science [54, 87, 95] and the accessibility of research findings. Understanding the consequences of journals flipping to OA models is essential to assess their future publication volumes, citation impact, and the overall benefits and challenges associated with this publishing model. Investigating the factors linked to authors' choices in OA publishing helps uncover potential disparities and inequalities in the publication system, providing insights into the motivations and dynamics driving this behavior and its correlation with citation impact. This research contributes to discussions on the role of OA in scholarly publishing.

International mobility of scholars is another important aspect influencing scientific impact. Collaboration and communication between researchers across geographical boundaries can increase knowledge exchange, innovation, and research productivity. By investigating the patterns and outcomes of international mobility, this research seeks to understand the motivations, barriers, and inequalities associated with researchers' mobility. Identifying factors influencing researchers' decisions to pursue international collaborations can provide valuable insights into creating a supportive and collaborative scientific environment. Additionally, analyzing the impact of international mobility on scientific outcomes such as citations, publication volume, and co-authorship can shed light on the benefits and challenges mobile researchers face. This knowledge can inform policies and strategies to foster international collaboration and enhance research's global impact.

The h-index is a widely used metric for evaluating scientific impact, combining productivity measures and citation counts [44]. Understanding the influential factors on the h-index can give researchers and institutions a deeper understanding of the metrics used to assess research impact. By examining the association between academic mobility, OA publishing, and other author and paper-specific features with the future h-index, this research aims to uncover additional factors that contribute to researchers' long-term impact. Machine learning algorithms, a powerful computational tool, are utilized to leverage vast amounts of data and develop predictive models that can accurately forecast the scientific impacts of research. By analyzing various influential factors within the dataset, these algorithms enable a comprehensive examination of the complex dynamics that contribute to scientific impact. The insights gained from this analysis can assist researchers in strategic career planning and institutions in evaluating the effectiveness of their support and development programs.

Overall, the motivation behind this thesis stems from the need to enhance our understanding of the factors that contribute to scientific impact. By investigating OA publishing, international mobility, and influential factors on the h-index, this research aims to provide valuable insights into improving the effectiveness and impact of scientific research. The findings can inform researchers, institutions, and policymakers in their efforts to promote open science, foster international collaboration, and support researchers in achieving long-term impact.

1.2 Research Objectives

This section outlines the research objectives that guided our investigation into scholars' demographic characteristics, scientific behaviors, collaboration patterns, and factors affecting their scientific impact:

1. Open access effect:

(a) Investigating the impact of transitioning to gold OA publishing on journals and authors, considering publication volume, citations, and impact factors.

(b) Examining the association between author-specific factors (such as a country's income level, or one's seniority or gender) and OA publishing, as well as the citation impact for authors from different income levels.

2. International mobility effect:

(a) Analyzing international mobility patterns among researchers and assessing their scientific impact, uncovering motivations, barriers, and inequalities associated with mobility.

(b) Quantifying the relationship between mobility and scientific outcomes, such as citations, publication volume, and co-authorship, while considering authors' characteristics such as gender, country, career stage, and research field.

(c) Determining the role and position of mobile researchers within the collaboration network through network analysis.

3. Influential factors on the researchers' h-index prediction:

(a) Investigating the association between academic mobility, open access publishing, gender, and other author and paper-specific features with the future h-index.

(b) Introducing and examining the impact of novel feature sets (gender, mobility, publishing model) on the future h-index for researchers with varying career backgrounds.

(c) Examining the temporal extent of feature categories' prediction power for the future h-index, considering researchers' seniority.

We aim to comprehensively understand scholars' characteristics, behaviors, and factors influencing their scientific impact by addressing these research objectives.

1.3 Contribution of this Thesis

This thesis used bibliometric data to investigate scholars' demographic characteristics, scientific behaviors (e.g., publishing and citing), and collaboration patterns. Furthermore, we analyzed the factors that influence scientific impact. The main contributions of this thesis are threefold:

- **Open access effect:** To explore this effect, we conducted two studies. In our first study (Section 4.2.1), we examined the outcome of gold OA publishing for journals and authors. For this purpose, we investigated the journals that have transitioned CA to a fully OA model. We compared their publication volume, number of citations, and impact factors before and after flipping. In our second study (Section 4.2.2), we considered author-specific factors, such as their country's income level, seniority, and gender, to examine their association with OA publishing. Additionally, we analyzed the citation impact of authors from countries with different income levels. Finally, we employed machine learning methods to explore the impact of these features on the selection of a publishing model.

- **International mobility effect:** In Chapter 5, we cover the study investigating international mobility and its impact on scholars from several aspects. We began by investigating the mobility patterns among different groups of researchers and analyzed their scientific impact to provide insights into the motivations, barriers, and inequalities associated with mobility. To quantify the relationship between mobility and scientific

outcomes, we extracted various authors' features such as gender, country, career stage, and field of research. Using logistic regression, we determine the probability of being mobile. We also employed Poisson regression to examine the correlation between mobility and scientific outcomes, including received citations, publication volume, and the number of co-authors. To address potential confounding biases and assess the impact of these features on future mobility, we employed a statistical method known as propensity score matching. We compared the scientific outcomes between mobile and non-mobile researchers by controlling for authors' features. Furthermore, we conducted a network analysis and calculated centrality scores of authors to uncover the role and position of mobile researchers within the collaboration network.

- **Influential factors on the h-index:** In Chapter 6, we introduce the study examining the association between academic mobility, open access publishing, gender, and other author and paper-specific features with the future h-index. Our final aim is to find the association between author and paper/venue-specific features with the future h-index. To this end, we made the following contributions:

  - Novel feature sets: We introduced and investigated the effect of different features, namely gender, mobility, and publishing model of papers, on the future h-index for researchers with varying career backgrounds.

  - Feature impact analysis: We employed the machine learning approach to predict the future h-index and analysed the impact of features on the prediction task.

  - Temporal extent of predicting performance: We examined the temporal dimension of the impact of different feature categories to understand the extent of the prediction power for each feature category in the future, considering the seniority of researchers.

For these purposes, we require some informetric techniques and preparatory work, which are introduced in the following:

- **Author disambiguation:** Dealing with synonym problems and common names, especially in the case of Asian names, poses significant challenges in author disambiguation, as it involves multiple individuals who share similar or identical names, potentially leading to confusion and incorrect attribution of publications. In Section 3.1, we outlined our proposed solution to address these challenges and enhance the accuracy of author disambiguation methods by employing a community detection approach within co-authorship networks.

- **Gender detection:** Gender is a significant demographic characteristic that plays a crucial role in addressing gender inequalities in science. However, it is often not directly available in bibliometric data, requiring us to infer it from external sources. In our work, we utilized the method proposed by Karimi et al. [47] to infer gender from authors' names using the Genderize.io API [40] and by analyzing their images on the web using Face++ API [32].

- **Co-author / Citation Network Analysis:** In Section 3.2, we present a case study on co-authorship network analysis, wherein we utilize various centrality measures to assess and determine the authors' positions and roles within the scientific network. By examining measures such as degree centrality, betweenness centrality, and closeness centrality, we gain valuable insights into individual authors' structural significance and influence in the network.

1.4 Structure of the Thesis

This thesis is structured as a cumulative dissertation, incorporating chapters that are predominantly based on previously published papers. Figure 1.1 illustrates the overall structure of our research, highlighting the individual studies conducted and the contributions made in each study.

In Chapter 2, we conduct an extensive literature review to examine and analyze the key findings and insights related to our research topic.

Chapter 3 focuses on our contributions to methodological challenges in bibliometric analysis. Specifically, Section 3.1 presents our study on author disambiguation, while Section 3.2 discusses our research on network analysis in bibliometrics.

Moving to Chapter 4, we delve into our first major contribution, the study of the OA effect. Section 4.2.1 showcases our papers investigating the impact of transitioning journals from the conventional subscription-based model to an OA business model. Additionally, Section 4.2.2 introduces a published paper focusing on the influencing factors of OA publishing.

Chapter 5 centers around our second contribution exploring academic mobility, in which we present our related publication addressing this objective.

Chapter 6 presents our third contribution, a study examining the factors that predict a researcher's h-index.

Chapter 7 serves as the final chapter of this thesis, encompassing a summary of the main findings in Section 7.1, a discussion on the research's limitations, and potential avenues for future research 7.2. The closing section 7.3 offers a personal reflection on the significance and fulfillment derived from contributing to the field of study.

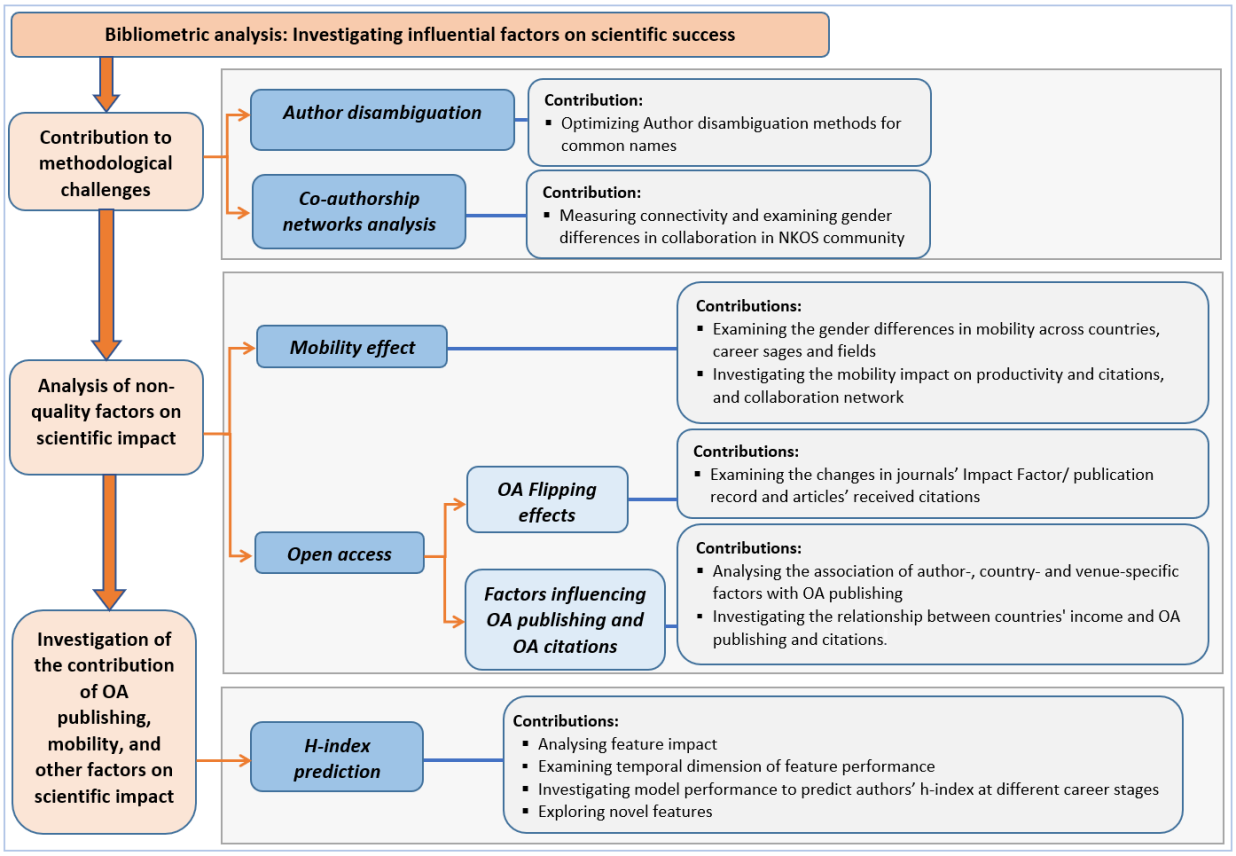

Figure 1.1: The overall structure of the research

1.5 Related Publications

The foundation of this dissertation lies in a collection of previously published papers within the research domain of bibliometrics and informetrics. These papers were published to high-ranked international peer-reviewed journals, conferences, and workshops. The following list enumerates these publications, along with references to their corresponding chapters and sections within this dissertation:

- Fakhri Momeni, Philipp Mayr, and Stefan Dietze (2023). Investigating the contribution of author- and publication-specific features to scholars' h-index prediction. EPJ Data Science, 12(1), 45 [77]

Contributions:

Fakhri Momeni conducted the investigation process of the research goal and formulated it together with Stefan Dietze. She did the computational analyses and visualization of results in tables and figures and prepared the manuscript (90%). Stefan Dietze gave feedback for all steps, particularly implementing the analytical methodologies and polishing the final paper. Philipp Mayer gave feedback on data collection and visualization and reviewed the final draft of the paper.

Dissertation Chapter: 6

Impact factor 2022: 3.6

Status: published.

- Fakhri Momeni, Stefan Dietze, Philipp Mayr, Kristin Biesenbender and Isabella Peters (2023). Which Factors are Associated with Open Access Publishing? A Springer Nature Case Study. Quantitative Science Studies, 4(2), 353-371 [74].

  Contributions: Fakhri Momeni is the main contributor to this paper and formulated the ideas and research goals. She contributed to data preparation and statistical and computational analyses, and data presentation. She prepared the manuscript (90%). Furthermore, she maintained the research data and uploaded the analyzed data and scripts on GitHub for later use. Stefan Dietze gave feedback for the employed machine learning. All co-authors reviewed the paper. Also, Isabella Peters and Philipp Mayr supervised and coordinated the project.

  Dissertation section: 4.2.2

  Impact factor 2022: 6.4

  Status: Published.

- Fakhri Momeni, Fariba Karimi, Philipp Mayr, Isabella Peters and Stefan Dietze (2022). The many facets of academic mobility and its impact on scholars' career. Journal of Informetrics, 16(2), 101280 [76].

  Contributions: Fakhri Momeni formulated the ideas and research goals. She did data engineering and computations in the investigations and applied all statistical and mathematical analyses. she created all figures and tables to visualize the results. Fariba Karimi supported Fakhri Momeni in all steps, including conceptualization, data collection, and methodology. Fakhri Momeni created the original draft and prepared the manuscript (90%). All authors discussed the results and contributed to the final manuscript. The whole work was under the supervision of Stefan Dietze.

  Dissertation chapter: 5

  Impact factor 2022: 3.7

  Status: Published.

- Fakhri Momeni, Philipp Mayr, Nicholas Fraser, and Isabella Peters (2021). 'What happens when a journal converts to Open Access? A bibliometric analysis', Scientometrics, 126(12), 9811-9827 [74].

  Contributions: Fakhri Momeni devised the main conceptual ideas. She prepared the data and performed the analyses. Isabella Peters and Philipp Mayr directed the project and provided critical feedback. Fakhri Momeni wrote mainly the manuscript (80%) with support from Nicholas Fraser. All authors contributed to the final manuscript.

  Dissertation section: 4.2.1

  Impact factor 2021: 3.9

  Status: Published.

- Fariba Karimi, Philipp Mayr, Fakhri Momeni, 'Analyzing the network structure and gender differences among the members of the Networked Knowledge Organization Systems (NKOS) community', International Journal on Digital Libraries (IJDL), 2018 [48].
  Contributions: Fakhri Momeni was involved in preparing data and analyses. Philipp Mayer formulated the main ideas behind the research. Fariba Karimi made valuable contributions to the analyses. All authors collaborated in writing the paper.
  Dissertation section: 3.2
  Status: Published.

- Fakhri Momeni, Philipp Mayer, 'Evaluating Co-authorship Networks in Author Name Disambiguation for Common Names', Research and Advanced Technology for Digital Libraries, 386—391 , 2016. International Conference on Theory and Practice of Digital Libraries (pp. 386-391). Springer [72].
  Contributions: Fakhri Momeni conceived the main idea of this work and performed all computations and experiments. Philipp Mayr provided feedback in evaluating the results. Fakhri Momeni prepared the manuscript (80%) with input from Philipp Mayr.
  Dissertation section: 3.1
  Status: Published.

During her Ph.D. study, Fakhri Momeni collaborated with several other publications aligned with bibliometrics and scientometrics. While these publications are not extensively discussed in this dissertation, they hold relevance and contribute to the broader understanding of these fields. Some of these related publications include:

- Nicholas Fraser, Fakhri Momeni, Philipp Mayr, and Isabella Peters, 'The relationship

between bioRxiv preprints, citations and altmetrics', Quantitative Science Studies, 2020 [38].

Contributions: This study was conducted by Nicholas Fraser. Fakhri Momeni contributed to preparing data and reviewing the paper.

Status: Published.

- Thomas Krämer, Fakhri Momeni, Philipp Mayer, 'Coverage of Author Identifiers in Web of Science and Scopus', Computing Research Repository (CoRR), abs/1703.01319, 2017 [52].

  Contributions: Fakhri Momeni supported Thomas Krämer in analyses and discussed the results with other authors.

  Status: Published as preprint.

- Fakhri Momeni, and Philipp Mayer, 'Analyzing the research output presented at European Networked Knowledge Organization Systems workshops (2000-2015)' NKOS workshop, 2016 [71].

  Contributions: Fakhri Momeni performed the analyses formulated by Philipp Mayr. Also, she contributed to writing the manuscript (50%).

  Status: Published.

- Fakhri Momeni, and Philipp Mayer, 'Using co-authorship networks for author name disambiguation', Digital Libraries (JCDL), Newark, NJ, USA, 19-23 June 2016 [73].

  Contributions: Fakhri Momeni served as the main contributor in this paper, leading the formulation of ideas and research goals. Her significant contributions included data preparation, statistical and computational analyses, and data presentation. Philipp

Mayr provided project supervision, and both authors contributed to the writing of the paper.

Status: Published.

Chapter 2

Related Work

This section reviews the literature and research relevant to our work, highlighting key findings and insights from prior works. We begin in Section 2.1 by discussing information extraction and engineering for scientometric analysis, specifically focusing on gender detection and author disambiguation. Moving forward, we survey selected works contributing to co-authorship network analysis. Additionally, we examine gender differences in collaboration patterns and the impact of gender on co-authorship networks across different academic domains. In the last part of this section, we focus on predictive modeling in scientometrics, specifically predicting scholars' scientific impact. In Section 2.2, our focus shifts to the informetric analysis of scientific impact. We delve into the literature on OACA and studies exploring various factors associated with OA publishing. Additionally, we examine previous research investigating the positive or negative effects of academic mobility and mobility patterns on scientific impact. Lastly, we discuss prior works that have contributed to predicting researchers' h-index using different types of features. Through this comprehensive review, we aim to build upon existing knowledge and provide a solid foundation for our research contributions in this dissertation.

2.1 Methodological Challenges in Informetric Analysis and Modeling

Bibliometric analysis provides meaningful information to quantify researchers' productivity, collaboration patterns, and citation impact. Large-scale bibliometric datasets are mainly employed for these purposes and contain much information about authors, publications, venues, universities, etc. Information Science approaches help us to prepare data and information extraction by cleaning, disambiguating, homogenizing, normalizing, or linking to external resources. These approaches are instrumental in addressing potential issues related to data quality [88]. In addition, due to the large amount of data, many methodologies and predictive models in Informatics have been employed to investigate the researchers' scientific impacts. This section mentions these approaches and studies related to our work.

2.1.1 Information Extraction, Engineering, and Network Analysis for Scientometric Analysis

In scientometric analysis, information extraction involves the process of collecting relevant data from bibliometric databases or external resources. This aspect specifically pertains to gender detection, where algorithms and techniques are used to infer the gender of authors. The engineering aspect in scientometric analysis refers to developing and implementing methods to address challenges such as author disambiguation. Network analysis is fundamental to scientometric studies, specifically co-authorship network analysis. It examines the collaboration patterns and relationships among researchers based on co-authored publications. In this section, we review the related literature and research gaps, aiming to shed light on the advancements and limitations in gender detection, engineering techniques for author

disambiguation, and the crucial role of network analysis, particularly co-authorship network analysis

Gender Detection

To tackle various research questions regarding the scientific impact of researchers, it is crucial to collect their demographic information. Investigating gender differences has been a focal point in numerous studies [24, 50, 59, 119]. However, accurately determining authors' gender solely based on their first name within a bibliometric dataset poses a significant challenge. Consequently, only a few studies have extensively addressed the gender aspects in bibliometric analyses.

The API Genderize.io [40] is an extensively used gender prediction tool that employs algorithms and probabilistic models to estimate individuals' gender based on their names, and it has been applied in numerous studies for gender detection [17, 29, 31, 41]. In the study by Boekhout et al. [17], authors additionally incorporated authors' affiliation country and utilized two other APIs, NameAPI and Gender-API, to determine the gender of six million authors in their analyses. However, using the affiliation country to identify gender assumes that it corresponds to the author's country of origin. This approach has limitations, as affiliation does not always reflect the authors' actual origin, especially in cases of migration or mobility, leading to reduced precision.

Moreover, all the aforementioned APIs predominantly rely on name-based approaches, which exhibit lower precision and recall when dealing with names from Asian nations, notably China [47]. In our work, we adopt the method introduced by Karimi et al. [47], which combines an image-based application (face++ [32]) with the name-based approach (API Genderize.io). This approach enhances precision and recall, particularly for Asian names, surpassing the

performance of relying solely on Genderize.

Author Disambiguation

One of the preliminary steps in evaluating the impact of authors' publications is to identify each author's set of publications. Therefore, author disambiguation was the concern of the informetric studies. Author disambiguation algorithms are divided into supervised and unsupervised approaches. In supervised approaches, a group of disambiguated author names is required as training data to identify patterns for assigning sets of publications to authors. These methods employ similarity measures between publications to determine the publication sets associated with each author [22, 68, 89]. Due to the difficulty of providing appropriate training data for large bibliometric data that represent the patterns and complexities for the entire data [102], most disambiguation algorithms are based on the unsupervised approach. Caron and van Eck [21] employed this approach by grouping names into blocks based on their last name and first initial and clustered publications using rule-based scoring based on bibliometric knowledge. This method has been used for the disambiguation task in many informetric studies [33, 35, 56, 92, 96]. Backes [10] proposed another unsupervised approach that employs an agglomerative clustering algorithm and a probabilistic similarity measure to build authors' publication sets. Tekles and Bornmann [106] evaluated and compared these two methods and some other approaches and presented higher pairwise F1 for the suggested approach by Caron and van Eck [21]. In addition to the disambiguation methods authors use for their analyses, bibliometric datasets may contain some author identifiers (e.g., ORCID iD [83], researcherId [90]). Scopus presented ORCID iD and researcherId in its dataset, but they cover only a small amount of authors. In addition, Scopus has its own author identifier (Id) for almost all authors. Aman [5] used the CVs of 193 Leibniz laureates from Germany to compare their Scopus Id with their CV data and found that 68% of authors have a single

Id. Among the remaining authors with multiple Ids associated with their publications, approximately 97% of these multiple Ids are linked to a single dominant Id, which is the most frequently used identifier across their works. Kawashima and Tomizawa [49] evaluated the accuracy of this identifier and found the precision and recall around 98% and 99%, respectively. However, Krämer et al. [52] discovered that despite the overall high accuracy of the Scopus identifier, it exhibits weaknesses in distinguishing publications associated with common names compared to ORCID iD. To address this issue specifically for common names, we conducted a study utilizing community detection techniques in co-authorship networks to optimize author disambiguation algorithms. We will present it in Section 3.1.

Co-authorship Network Analysis

Co-authorship network analysis has been extensively studied, providing insights into collaboration patterns, network structure, and research impact. We review selected works that have contributed to understanding co-authorship networks. Newman [78] introduced social network analysis concepts and methods for studying collaboration networks in science. He examined characteristics such as degree distribution, clustering coefficient, and network communities. In another paper [79], he investigated collaboration network structures across various fields, revealing patterns such as the small-world phenomenon and preferential attachment. Uzzi et al. [107] employed a network science approach to study collaboration in computer science, identifying key structural features that influence collaboration dynamics and success. Their findings highlight the importance of diversity and interdisciplinarity in fostering innovative research collaborations. Sapmaz et al. [98] conducted a study using social network analysis to investigate collaboration patterns within a specific scientific community. Their analysis reveals central researchers who play crucial roles in connecting different subgroups and facilitating knowledge diffusion. Ilyas et al. [45] provided a com-

prehensive co-authorship network analysis survey overview. They discussed various analysis techniques applied to co-authorship networks and highlighted their potential for understanding research collaborations across domains. Van der Sanden et al. [108] systematically reviewed co-authorship network analysis in the social sciences, synthesizing existing literature and identifying common research themes, methodologies, and challenges. The review emphasizes the importance of considering disciplinary differences and methodological choices. Additionally, some studies [18, 64, 105, 109, 110] specifically addressed gender differences in collaboration patterns. These works analyzed gender disparities in collaboration networks, productivity, citation patterns, and gender homophily in different academic and research domains. Our presented research in Section 3.2 specifically focused on the NKOS community, a prominent group in the field of digital libraries and knowledge organization systems, and provided valuable insights into gender-related dynamics in network analysis.

2.1.2 Predictive Modeling

Predicting scholars' scientific impact has been the topic of many studies [2, 9, 11, 46, 51, 81, 115, 116] since it can lead hiring committees, funding agencies, and research group heads to find researchers with a higher probability of scientific achievements in the future. Penner et al. [84] examined the linear regression model to predict the scholars' future h-index (a widely-used metric for evaluating researchers' productivity and citation counts) and found a strong dependency of career age on the accuracy of the prediction model. Acuna et al. [3] predicted the future h-index using linear regression with elastic net regularization for different scientific disciplines and found a varying range of performance ($\boldsymbol{R^2}$) among disciplines. They examined various features to predict the h-index and suggested a formula based on the regression model containing only the five most important features (number of published articles, the current h-index, years since first publication, number of publications in presti-

gious journals, and the number of distinct journals). Newer studies utilized the regression models based on Machine Learning (ML) approaches such as Support Vector Regression (SVR) [116], Gradient Boosted Regression Trees (GBRT) or Gradient Boosting (GB) [116], Gradient-Boosting Decision Tree (GBDT), Extreme Gradient Boosting (XGBoost) [116], Random Forest (RF) [116], and Neural Networks (NN) [94] to predict the number of citations and h-index. Wu et al. [116] compared various approaches and identified XGBoost as the top-performing method among the ones investigated. Also, some authors considered the prediction task as a classification problem in machine learning and categorized the number of predicting metrics (citation, h-index, or other metrics) into some groups (e.g., high or low) and classified them by employing classification models of machine learning [81, 113]. Nie et al. [80] examined K-Nearest Neighbour (KNN), GBDT, RF, XGBoost, and Support Vector Machine (SVM) algorithms with different feature sets (social, author, venue, and temporal) to detect the rising stars (promising early-career researchers who show potential for significant future impact in their fields) and found the KNN as the best model and author and venue features as the best model feature sets to predict the rising stars. Wang et al. [113] examined the Naiver Bayes, KNN, and RF to identify highly cited papers from bibliometric and altmetric[1] data and used three feature selection techniques to rank the importance of features in the prediction task. They found altmetrics, the scope of knowledge diffusion in the scientific communities, and early-stage citations as the key influential factors on future impact. In conclusion, the prediction of scholars' scientific impact has garnered significant attention in numerous studies in identifying researchers with a higher likelihood of future scientific achievements.

[1] Altmetrics "focuses on the creation, evaluation and use of scholarly metrics derived from the social web"

2.2 Informetric Analysis on Scientific Impact

Informetric analysis is a vital tool for studying scientific impact, offering a quantitative lens through which researchers can examine publication patterns, evaluate publishing models, and predict the future impact of research endeavors. This section provides a comprehensive review of the literature about three key aspects of scientific impact: open access, academic mobility, and the prediction of scientific impact, specifically in terms of the h-index.

2.2.1 Open Access Effect: Open Access Citation Advantages

OA publishing makes scientific literature freely accessible and increases the probability of receiving citations, and it attracted many studies to investigate OACA [27, 36, 61, 63, 65, 66, 87, 103] However, the findings have been mixed with some studies reporting positive impacts [36, 61, 65, 66, 87, 103] while others revealing potential negative effects [8, 27, 63]. These variations in findings can be attributed to different factors such as sample selection, data control, publication types, disciplines, and the specific models of OA implementation. A comprehensive review conducted by Langham-Putrow et al. [55] analyzed 134 studies on this subject and found that most studies (64%) supported the presence of OACA. However, it is important to consider confounding factors such as quality bias[2] or self-selection[3] [69], mandates[4] [39], and self-archiving[117], to accurately assess the true citation impact of OA publishing and differentiate it from other influencing factors. Langham-Putrow et al. [55] mentioned that only 30% of studies on OACA acknowledged the possibility of confounders;

[2]Quality bias refers to the bias that arises when open access articles are perceived to be of higher quality or impact than non-open access articles, potentially distorting research findings.

[3]It refers to authors choosing to publish in OA or CA models based on personal preferences or circumstances, potentially introducing bias in research studies.

[4]a mandate refers to a requirement or policy implemented by a funding agency, institution, or government that mandates or obliges researchers to make their research outputs, such as journal articles or conference papers, freely available to the public.

however, not all controlled them in their analyses. McCabe and Snyder [65] considered these effects in their study, still found 8% increasing in received citations for OA publications. In Section 4.2.1, our research investigates a dataset of journals that have transitioned from a conventional subscription-based model to an OA model. We examine the impact of this transition on scholarly publishing, specifically focusing on changes in article volume, citation impact, and the overall visibility and influence of journals. By analyzing the effects of adopting an OA model, we aim to gain insights into the implications and benefits of OA publishing for both journals and articles. This study contributes to a deeper understanding of the implications and potential advantages of transitioning to an OA publishing model.

2.2.2 Open Access Effect: Open Access Publishing

One of the possible issues making publication freely accessible is transforming the costs of publishing from readers to authors and their funders via Article processing charge (APC). However, this may not be possible for authors who do not have access to financial support for OA publishing. In succession, such authors cannot publish in the OA model (instead they must turn to the CA-model) and cannot profit from the OACA shown for many OA articles. Implementation of OA policies such as mandates, waiver and discount policies, and transformative agreements between publishers and institutions and countries accumulate OA publishing. Simard et al. [101] studied geographic differences in OA publishing and found that low-income countries have the highest OA publishing and citing rates. Robinson-Garcia et al. [93] investigated the OA uptake worldwide across institutions and scientific

fields for different OA models (green[5], gold[6], bronze[7], hybrid[8]) and displayed the highest rate of OA publishing for European and North American countries. Among scientific fields, they found the largest average of OA publishing in Biomedical & Health Science, reported similarly in some other studies [20, 57, 87]). However, their observation of the lowest rate in Social Sciences & Humanities contradicts the results of Larivière and Sugimoto [57] and Piwowar et al. [87], where Social Sciences demonstrated a higher OA rate compared to fields like Engineering & Technology and Chemistry. In a review study by Severin et al. [100] on the uptake of OA publishing, they observed limited consistency in the reported uptake levels across studies, attributing this inconsistency to methodological variations in identifying and measuring OA publishing. It indicates that other features rather than field-specific factors must be studied to find their associations with OA publishing. Our study in Section 4.2.2 examines the association of author- and paper-/venue-specific features with OA publishing. We investigate the relationship between the income level of researchers' affiliation countries and their publication behavior, specifically their preference for OA or CA publishing models. Additionally, we explore the relationship between the income level of researchers' affiliation countries and the citation impact of their publications based on the chosen publishing model. Furthermore, we analyze various factors, including journals, articles, authors, and their countries, to identify the associations with selecting OA or CA business models for publications. By addressing these research questions, we aim to gain valuable insights into the factors influencing the selection of OA or CA business models for publications and deepen our understanding of the impact of publishing behavior across different income levels of researchers' affiliation countries.

[5]Free availability of research articles through self-archiving in institutional or subject repositories

[6]Immediate and unrestricted access to research articles provided by the publisher, usually with a payment of an article processing charge (APC)

[7]A freely available journal article that has no open license

[8]A publishing model that combines both open access and traditional subscription-based access, allowing authors to choose which individual articles they want to make openly accessible while keeping the rest behind a paywall

2.2.3 Mobility Effect

International academic mobility, the subject of extensive research in the Science of Science field [118], refers to the movement of researchers and scholars across borders, playing a crucial role in disseminating and exchanging knowledge. It affects researchers' social capital (networks and relationships that enable collaboration, knowledge exchange, and resource access) by changing their collaboration patterns [15, 16, 112] and can improve their human capital (individual skills and technical knowledge) [12] by accessing new resources. Various studies employed bibliometric data to model researchers' mobility via tracking the affiliations stated in their publications [1, 6, 97, 104, 114] and examined its association with productivity, received citations, and collaboration after mobility. While some studies have shown a decrease in productivity and received citations for researchers after changing their affiliation due to the challenges and difficulties associated with transitioning to a new research environment, known as "adjustment costs" [1, 34], a significant body of research has found a positive effect of mobility on research outcomes [4, 14, 23, 28, 37, 42, 120].

Because of the importance of international academic mobility in knowledge transfer and developing collaboration and scientific communication, many grant programs facilitate it for researchers (e.g., Swiss National Science Foundation [13], Erasmus international mobility [86], University Mobility in Asia and the Pacific [82]). However, research has shown gender inequality in international academic mobility [58, 60, 70, 99], which results in inequality in the scientific promotion. Most studies used a restricted dataset containing a particular group of the population (authors) or restricted analyzed features. For example, previous studies such as Li and Tang [62], Subbotin and Aref [104], Zhao et al. [120] considered only authors from a single country, or Petersen [85] took the sample from a specific scientific field. El-Ouahi et al. [30] investigated the gender differences and career age scientific mobility in the Middle East and North Africa region and found a clear gender gap in mobility for this

region. However, they didn't examine the effect of mobility on the scholars' scientific impact.

In Chapter 5, we undertook a comprehensive global study to explore the impact of gender on international mobility and its influence on scientific outcomes. This investigation allowed us to uncover variations in mobility patterns across countries, scientific fields, and career stages, shedding light on the underlying factors contributing to challenges faced by women in academia. Through our investigation, we aim to explore the relationship between various individual factors, such as country, career stage, field of research, and the mobility of researchers, with a specific focus on understanding potential differences between males and females. Additionally, we seek to examine how different characteristics of mobile researchers are associated with their scientific outcomes. By addressing these research questions we aim to contribute to a deeper understanding of the factors influencing researchers' mobility and its impact on their scientific achievements.

2.2.4 Influential Factors on Predicting the Scientific Impact

One important research topic in the field of Science of Science is predicting the progress and advancement of scientific development. For this purpose, bibliometric datasets provide valuable insights through author-specific attributes (e.g., publication record, collaboration patterns) and paper-specific attributes (e.g., citation counts, publication venue) that have been widely used as predictors in prior studies [2, 7, 43, 94]. Prior scientific impact (current publication record, received citations, and h-index) and their related features (e.g., early citations, changing the publishing behavior or h-index in recent years) have been commonly used as predictors in the past years [11, 94, 115, 116]. These factors simplify the prediction task because they influence the h-index directly. Some other studies examined the relationship between future impact with another author-, co-author, and paper/venue-specific

factors. [81] investigated the association of collaboration patterns and textual content of the author's paper with h-index. They found the collaboration pattern a stronger predictor than the papers' textual content. Dong et al. [26] examined the contribution of factors related to a paper to increase the future h-index and found that topic authority (being highly cited by researchers over a specific domain) and publication venue are critical in determining whether a paper will contribute to increasing the h-index. In their study, Kuppler [53] compared the performance of the prediction model proposed by Weihs and Etzioni [115] across genders and discovered that the models tend to underestimate the future h-index of women to a greater extent than men. Ayaz et al. [9] investigated the role of career age in the prediction task and reported the worse performance for predicting the younger researchers. Their results are reasonable because researchers at the early stage of their careers have a small publication record, and citations and other factors are required to assess and anticipate their scientific impact. In Chapter 6, our research explores the contribution of various factors in predicting scholars' h-index across different career stages. It examines the temporal aspect of prediction performance. We incorporate novel features that have demonstrated associations with scientific impact but have not been utilized for prediction purposes. Furthermore, we investigate the extent to which author- and paper-specific factors contribute to predicting scholars' h-index. Additionally, we assess these predictors' reliability, temporal stability, and applicability in forecasting the h-index across various career stages.

Chapter 3

Contribution to Methodological Challenges in Bibliometrics

3.1 Enhancing Author Disambiguation: A Network Approach for Common Names

3.1.1 Overview

To assess and compare the scientific impact of authors in a bibliometric database, it is crucial to identify their publications accurately. However, manually disambiguating the names of numerous authors within large datasets like Scopus and Web of Science (WOS) is impractical. Consequently, various approaches have been developed to address this challenge, offering solutions with varying levels of coverage and accuracy. One such approach involves researchers self-identifying their publications, which can significantly enhance accuracy. ORCID iD [1] is the author identifier used in this approach, and Scopus has its author identifier for all its indexed authors. In a previous study [52], we assessed the coverage of ORCID iD and Scopus Author ID in the Scopus dataset. Our findings indicated a low coverage of ORCID iD in this dataset (~29%). Furthermore, the study involved comparing the accuracy of these two identifiers by counting the number of publications associated with authors

[1] https://orcid.org/

using ORCID iD and Scopus Identifier. The top 10 authors with the highest publication counts were identified for each identifier system. In the ORCID iD list, both Western and Asian names were present, with the top author having a maximum of 355 publications. Conversely, the Scopus Identifier list exclusively included Asian names, with the top author having 2,338 publications. This indicates a potential limitation or challenge in accurately distinguishing and attributing publications to specific authors with Asian names within the Scopus Identifier system.

In this section, we present our study that introduces an approach capable of enhancing the accuracy of author identifiers for individuals with common names, adaptable to various disambiguation methods. In this approach, we built a network of authors and their publications for authors identified via the rule-based disambiguation approach proposed by [21]. Using community detection, we tried to split up the author identifiers that belong to different communities. The results reveal a noticeable improvement in the accuracy of the author disambiguation method for common names. The following is the paper [72] regarding this study presented at the TPDL conference 2016.

3.1.2 Publication

Title: Evaluating co-authorship networks in author name disambiguation for common names

Authors: Fakhri Momeni and Philipp Mayr

Document Type: Conference paper

Venue: TPDL 2016

DOI: `https://doi.org/10.1007/978-3-319-43997-6_31`

Evaluating Co-Authorship Networks in Author Name Disambiguation for Common Names

Fakhri Momeni and Philipp Mayr

GESIS Leibniz-Institute for the Social Sciences, Cologne, Germany
firstname.lastname@gesis.org

Abstract. With the increasing size of digital libraries it has become a challenge to identify author names correctly. The situation becomes more critical when different persons share the same name (homonym problem) or when the names of authors are presented in several different ways (synonym problem). This paper focuses on homonym names in the computer science bibliography DBLP. The goal of this study is to evaluate a method which uses co-authorship networks and analyze the effect of common names on it. For this purpose we clustered the publications of authors with the same name and measured the effectiveness of the method against a gold standard of manually assigned DBLP records. The results show that despite the good performance of implemented method for most names, we should optimize for common names. Hence community detection was employed to optimize the method. Results prove that the applied method improves the performance for these names.

Key words: Author name homonyms; Co-authorship network; Community detection; Louvain method; Gold standard

1 Introduction

In scholarly digital libraries (DLs) authors are recognized via their publications. It is important for users to know about the author of a particular publication to access possible other publications of this author. For this purpose DLs provide search services using the publication information in their databases. However, when several authors share the same name or authors provide their works with different names DLs need more analysis on author's oeuvres. Many different approaches have been proposed in the field of author name disambiguation. Manual author identification in large DLs is very costly. The consequence is that large sets of ambiguous author names need to be analyzed automatically. In addition the demographic characteristics such as name origin and frequency of names used for authors influence the identification of authors. Therefore, all constrains of the underlying data should be considered to choose the appropriate method for author name disambiguation.

Author assignment method and author grouping method [3] are the two main methods for author name disambiguation. Author assignment method constructs a model that represents the author and assigns proper publications to the

model. It requires former knowledge about the authors. Nguyen and Cao [7] used this method and proposed to link the author names to the matching entities in Wikipedia. The author grouping method clusters the publications on the basis of their properties (co-authors, publication year, keywords, etc.) to assign a group of publications to a certain author. Following this framework, Caron and van Eck [2] applied rule-based scoring to clustered publications. In their approach the authors suppose that there is enough information about authors and their documents. Also, Gurney et al. [4] clustered publications with employing different data fields and integrated a community detection method. Some authors [5],[8],[9] used social networks (mainly co-authorship networks) to cluster publications. Levin and Heuser [5] introduced a set of matching functions based on the social network of authors and measured the strength of connections between the authors. Shin et al. [8] extracted the abstract and author's affiliation from the paper and considered the relation between authors to find similarities between publications. Wang et al. [9] proposed a unified semi-supervised framework to handle the synonym and homonym problem of author names.

In this paper we used an author grouping method (compare [3]) to cluster the publications of a set of random authors with the same name in the DBLP database. Considering the lack of rich bibliographic information in DBLP records, we applied co-authorship network analysis introduced by Levin and Heuser [5] to detect similarities between publications in order to investigate, how the amount of homonym names affects the disambiguation results. In the end, we employed a community detection algorithm (Louvain method) to reduce the effect of common names in our evaluation.

2 Disambiguation Approach

We use the author grouping method in order to assign all publications of each person to a certain group. For this purpose all publications belonging to the same ambiguous author name are categorized into one block. In a next step we compare any pair of publications in each block with each other to find a similarity between them. If we have n blocks and m_i publications in a block i, the number of comparisons for all blocks is:

$$\sum_{i=1}^{n} \frac{m_i(m_i - 1)}{2} \tag{1}$$

The result of each comparison is true or false. The *true* result means that two publications belong to one person and the same cluster. If one of them was compared with another one before and assigned to a cluster, the other one is added to that cluster too. If both of them were compared before and belong to different clusters, two clusters are rebuilt to one cluster. Otherwise a new cluster will be created and two publications are put in new cluster. In the next section we describe how to define the similarity indicator to build the clusters.

The bibliographic information that we can obtain from publications in DBLP is limited mainly to author names (the names of all co-author names are listed),

title and publication venue. We chose the co-author names as our similarity indicator. Therefore we built a network of authors and documents. Figure 1 shows an example of the network. The continuous lines show the links between publications and authors in the network. As it was mentioned before each pair of documents within every block has to be compared.

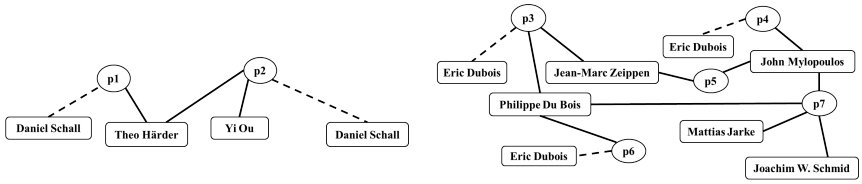

Fig. 1. An example of a co-authorship network

To compare the publications the relations in the network are analyzed. If there is a path between two publications, their distance is defined as the length of the shortest path between them, otherwise it would be infinite. The length of the shortest path is equal to the number of nodes between two nodes. For example the distance between publication p1 and p2 in Figure 1 is 1; the distance between p3 and p4 is 3. The less distance between two publications means that more likely these publications were written by one person. So, the distance between two publications is assumed as the similarity measure. Different thresholds can be considered for the distance. For example, in Figure 1, with the threshold = 1, p1 and p2 are two publications of one person with the name 'Daniel Schall', because they share the same co-author. Accepting the threshold = 3, p3 and p4 belong to same author with the name 'Eric Dubois'. In Section 3 we see the effect of selecting different thresholds on the evaluation results.

3 Evaluation

Gold Standard: In order to evaluate the output of the author disambiguation approach we need a gold standard of disambiguated author names. Many homonym author names in DBLP are disambiguated manually by the DBLP team and are identifiable with an id. For example, 'Wei Li' belongs to 59 different persons: 'Wei Li 0001', 'Wei Li 0002', etc. Thus, the set of publications for each person is recognizable. To build the gold standard [6] we selected these identified author names and compiled all their publications into one set. In our gold standard we provide a list of publications which have at least one disambiguated author name. Asian names, especially Chinese names are the most common names in DBLP and result in many homonym author names. These names are the most problematic names in author disambiguation and should be analyzed in particular. In total 1,578,316 unique author names exist in DBLP. There are 5,408 authors who have an identification number (we mention them as disambiguated authors). These 5,408 authors and their publications form the gold standard. We got these numbers from DBLP, downloaded May/01 2015

from http://dblp.uni-trier.de/xml/. To measure the performance of our method 1,000 disambiguated author names have been randomly selected from the gold standard. In total we have 2,844 different authors and 32,273 publications in our random sample. In the next section we evaluate the performance of our method against the gold standard.

Evaluation Metrics: Bcubed metrics [1] are used to evaluate the quality of the algorithm. These clustering metric satisfy constraints on evaluation the clustering tasks [1] such as cluster homogeneity and cluster completeness. Therefore we applied them to evaluate our method. For this purpose Bcubed precision and recall are computed for each publication.

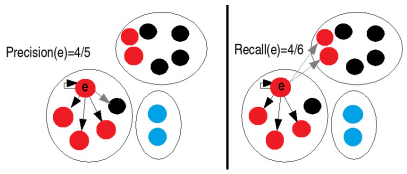

Fig. 2. Example BCubed precision and recall adapted from [1]

Figure 2 shows an example how the precision and recall of one publication of an author are computed by BCubed metrics. In this Figure assume the circles in red, black and blue as the publications belonging to three different authors and our algorithm categorized them to three groups. The publication precision measures how many publications in its group belong to its author. The publication recall measures how many publications from its author appear in its group. Bcubed F as the combination of Bcubed precision and recall is computed as follows:

$$\frac{1}{\alpha(\frac{1}{P}) + (1-\alpha)(\frac{1}{R})} \tag{2}$$

being P and R Bcubed precision and recall and being α and (1- α) the relative weight of each metric (We assumed $\alpha = 0.5$). Bcubed precision, recall and F-measure were computed for every publication in any block. Then we consider their average as the Bcubed precision, recall and F of the block.

4 Results and Discussion

We clustered the publications with regard to the distance between them. For choosing the threshold we have checked the distances larger than 3, which results in a very low precision. Then we chose the threshold equal to 1 and 3. For the distance less than threshold (1 or 3), we assign two publications in the same cluster. The results of the evaluations for two thresholds are demonstrated in Table 1. The results in Table 1 indicate that our co-author networks method performs well on the dataset and it can be utilized as author identification approach. No effort was made to define and compare against an external baseline. Comparing the results for two thresholds (1 and 3) we can conclude that using

Table 1. Mean values of BCubed metrics for 1,000 blocks

	BCubed precision	BCubed recall	BCubed F
Threshold=1	0.98	0.74	0.79
Threshold=3	0.94	0.81	0.82

threshold=3 provides us with the better balance between precision and recall and a higher F (slightly better BCubed recall of 0.81 and F of 0.82). We can shows that with the increasing number of publications in the blocks, the efficiency of our algorithm decreases, especially for threshold=3. We can conclude that although using threshold=3 results the better performance generally, it is less efficient than using threshold=1 for common names. The reason is that common names enhance the probability of being authors with the same name in the same area of research activity and increase the likelihood of detecting the shared co-author for different researchers with the same name. Furthermore, it is more likely that these authors have co-authors with similar common names. This results in a higher probability of ambiguous co-authors and wrong connections between publications. Therefore, we should be more cautious to use the co-author of co-author as the similarity measure for these cases and will verify them more deeply. To remove the wrong connections that link two groups of publications from different authors community detection is a good solution. *Community detection* aims at grouping nodes in accordance with the relationships among them to form strongly linked subgraphs from the entire graph. Hence, we applied a community detection algorithm to optimize the results (threshold=3) for the common names. We chose a subset of the author's names which have more than 200 publications (totally 28 names) in our DBLP dataset. To detect communities in the network we utilized the Louvain method in Pajek. This method maximizes the modularity of network. Single refinement is selected and the resolution parameter was set to 1. Because the less distance between publications increases the probability of being the same author, we gave the weight to connections. For the distances equal to 1 and 2 have weights with values 2 and 1 respectively. Table 2 shows that community detection improved the results for the most repeated names in our sample.

Table 2. BCubed metrics for author names with more than 200 publications, thr.=3

	BCubed precision	BCubed recall	BCubed F
Before optimization	0.46	0.87	0.45
After optimization	0.79	0.61	0.58

5 Conclusions and Future Work

In this paper we implemented a method to identify authors with the same name based on co-authorship networks in DBLP. The results showed that although

co-author networks have a substantial impact on author name disambiguation, but common names decrease the performance of our method and should be optimized in an extra step. For this reason, we implemented the community detection method which showed an improvement for highly frequent common names. Our approach can be applied to disambiguate author names in DBLP. In this way we create the network and link the publications automatically, then apply the community detection to find the suspicious connections and check them manually if they are a wrong connection. In this case, they will be removed from the network and increase the performance of algorithm. So, the speed of automatic disambiguating and the accuracy of manual checking can be combined. Our approach improves the disambiguation of common names, but this is not sufficient. To get better results we need to optimize the parameters such as resolution in the community detection method for different numbers of publications per name. We could also investigate the effect of changing the weights of links between publications depending on their distances. Because this method is based on co-author network, it is limited to multi-author papers. Therefore a multi-aspect indicator is required for single author papers. We can use the titles of publications to extract keywords and add this information to calculate similarity measures.

6 Acknowledgment

This work was funded by BMBF (Federal Ministry of Education and Research, Germany) under grant number 01PQ13001, the Competence Centre for Bibliometrics. We thank our colleagues at DBLP who helped with generating the testbed [6].

References

1. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Inf. Retr. 12(4), 461–486 (2009)
2. Caron, E., van Eck, N.J.: Large scale author name disambiguation using rule-based scoring and clustering (2014)
3. Ferreira, A.A., Gonçalves, M.A., Laender, A.H.F.: A brief survey of automatic methods for author name disambiguation. SIGMOD Record 41(2), 15–26 (2012)
4. Gurney, T., Horlings, E., den Besselaar, P.V.: Author disambiguation using multi-aspect similarity indicators. Scientometrics 91(2), 435–449 (2012)
5. Levin, F.H., Heuser, C.A.: Evaluating the use of social networks in author name disambiguation in digital libraries. JIDM 1(2), 183–198 (June 2010)
6. Momeni, F., Mayr, P.: An Open Testbed for Author Name Disambiguation Evaluation (2016), `http://dx.doi.org/10.7802/1234`
7. Nguyen, H.T., Cao, T.H.: Named entity disambiguation: A hybrid statistical and rule-based incremental approach. In: ASWC 2008
8. Shin, D., Kim, T., Jung, H., Choi, J.: Automatic method for author name disambiguation using social networks (2010)
9. Wang, P., Zhao, J., Huang, K., Xu, B.: A unified semi-supervised framework for author disambiguation in academic social network. In: DEXA 2014

3.2 Exploring Network Connectivity and Gender Dynamics in Academic Contexts

3.2.1 Overview

Network analysis in bibliometrics investigates the relationships between publications, citations, and authors. We can employ centrality measures to examine the authors' and publications' power and influence in the bibliometric networks. We started experimenting with network analysis in [71] by building a co-authorship network from papers belonging to a community of authors who participated in European*Networked Knowledge Organization Systems* (NKOS) workshops and measured their centrality scores. In the following paper, we extended the data to European and US NKOS workshops and published the results in the International Journal on Digital Libraries. We described the basic centrality measures required to investigate authors' properties in the networks and used them in our next study [75] in Chapter 5. We also investigated gender differences in collaboration networks and found that homophily is higher among women, which contributes to widening inequalities [91].

3.2.2 Publication

Title: Analyzing the network structure and gender differences among the members of the Networked Knowledge Organization Systems (NKOS) community

Authors: Fariba Karimi, Philipp Mayr, and Fakhri Momeni

Document Type: Journal paper

Venue: International Journal on Digital Libraries

DOI: https://doi.org/10.1007/s00799-018-0243-0

International Journal on Digital Libraries (2019) 20:231–239
https://doi.org/10.1007/s00799-018-0243-0

Analyzing the network structure and gender differences among the members of the Networked Knowledge Organization Systems (NKOS) community

Fariba Karimi[1] · **Philipp Mayr**[1] · **Fakhri Momeni**[1]

Received: 4 October 2017 / Revised: 4 May 2018 / Accepted: 8 May 2018 / Published online: 14 June 2018

Abstract
In this paper, we analyze a major part of the research output of the Networked Knowledge Organization Systems (NKOS) community in the period 2000–2016 from a network analytical perspective. We focus on the papers presented at the European and US NKOS workshops and in addition four special issues on NKOS in the last 16 years. For this purpose, we have generated an open dataset, the "NKOS bibliography" which covers the bibliographic information of the research output. We analyze the co-authorship network of this community which results in 123 papers with a sum of 256 distinct authors. We use standard network analytic measures such as degree, betweenness and closeness centrality to describe the co-authorship network of the NKOS dataset. First, we investigate global properties of the network over time. Second, we analyze the centrality of the authors in the NKOS network. Lastly, we investigate gender differences in collaboration behavior in this community. Our results show that apart from differences in centrality measures of the scholars, they have higher tendency to collaborate with those in the same institution or the same geographic proximity. We also find that homophily is higher among women in this community. Apart from small differences in closeness and clustering among men and women, we do not find any significant dissimilarities with respect to other centralities.

Keywords NKOS workshops · Network analysis · Co-authorship networks · Gender · Homophily · Collaboration

1 Introduction

The Networked Knowledge Organization Systems (NKOS)[1] community in Europe and in the USA has held a long-running series of annual workshops at the European Conference on Digital Libraries (ECDL), latterly renamed as the International Conference on Theory and Practice of Digital Libraries (TPDL), the Joint Conference on Digital Libraries (JCDL) and some other scattered events. The NKOS workshops in the USA have started in 1997/1998 organized by Linda Hill, Gail Hodge, Ron Davies and others. Slightly later, the first NKOS workshop was organized in Europe at ECDL 2000 in Lisbon (Portugal) by Martin Doerr, Traugott Koch, Douglas Tudhope and Repke de Vries.

Typically, recent advances in Knowledge Organization Systems (KOS) have been reported at the annual NKOS workshops, including for example the Simple Knowledge Organization System (SKOS) W3C standard, the ISO 25964 thesauri standard, the CIDOC Conceptual Reference Model (CRM), Linked Data applications, KOS-based recommender systems, KOS mapping techniques, KOS registries and metadata, social tagging, user-centered issues and many other topics[2]. Special issues on Networked Knowledge Organization Systems were published in the Journal of Digital Information in 2001 [8] and 2004 [24], in the New Review of Hypermedia and Multimedia in 2006 [25] and recently in the International Journal of Digital Libraries in 2016 [14].

✉ Philipp Mayr
philipp.mayr@gesis.org

Fariba Karimi
fariba.karimi@gesis.org

Fakhri Momeni
fakhri.momeni@gesis.org

[1] GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany

[1] For an introduction of KOS and NKOS and recent applications see [8,14].

[2] Comprehensive review articles on KOS and NKOS topics were published in [9,26].

Recently, the NKOS workshop activities have accelerated again, e.g., with two European NKOS workshops in 2016 at the TPDL and Dublin Core conference and a revival of the US NKOS activities in 2017. In addition, the last two NKOS workshops at TPDL resulted in formal conference proceedings published as CEUR Workshop Proceedings [15,16].

The motivation of this paper is to analyze and visualize the collaboration network of the NKOS community. We are focusing here on the informal part of this output, i.e., the paper presentations given at the past NKOS workshops. The specialty of this research output is that these research papers are typically not published in journals or conference proceedings. These papers appear just as oral presentations at the workshop and are documented as such on the corresponding websites. To cover this informal research output, we collected presentation information from the workshop agendas. To analyze the co-authorship network of this community, we restrict our analysis to papers which have been authored by a minimum of two authors. This results in 123 papers with a sum of 256 distinct authors. It is important to state that practices at the NKOS workshops in the USA and Europe are different. In the USA, NKOS workshops were previously not based on open call for papers, but contributions are rather collected via inviting speakers. This practice explains the relatively low ratio of co-authorship in the US workshop series. From the beginning, in Europe, the NKOS workshops were based on accepting academic papers and resulted in an open call for papers and a subsequent peer review of submitted paper abstracts.

In the following, we report about the network structure and gender differences among the members of the NKOS community as we could recall from the past European and US workshop agendas and published special issues.

This paper is an extended version of the paper "Analyzing the research output presented at European Networked Knowledge Organization Systems workshops (2000–2015)" [18] presented at the 15th NKOS workshop at the TPDL conference 2016. In [18], we focused on the European workshops and special issues. Meanwhile, we have extended the dataset and included the US NKOS workshops and some other scattered NKOS events. In this way, the paper enables a more comprehensive overview of the international NKOS research community. To the best of our knowledge, this paper is the first attempt to analyze the co-authorship network of NKOS in great details.

In the following sections we describe the underlying dataset (Sect. 2), perform network analysis (Sect. 3), highlight some results of our analysis (Sect. 4) and conclude our paper (Sect. 5).

2 NKOS workshop bibliography dataset

For our analysis, we have compiled an open dataset derived from the "NKOS bibliography"[3]. The NKOS bibliography has been started in 2016 [18] and covers bibliographic information of all research papers presented at the past NKOS workshops. Editing and organizing activities (incl. the introductions) at the workshops have not been covered in our dataset. Journal papers published in four special issues on NKOS which were edited by members of the NKOS community in the same period were added. These journal papers are the only formal publications in our analysis. In the end, we manually disambiguate the author names of all papers. The bibliography is stored in single bibtex files (one bibtex file for each venue).

To this date, the NKOS bibliography covers:

- sixteen European NKOS workshops from 2000 to 2016. In total 16 workshop agendas: ECDL 2000, 2003–2010, TPDL 2011–2016, Dublin Core 2016,
- eight US NKOS workshop agendas: JCDL 2000–2003, 2005 and NKOS-CENDI 2008–2009, 2012,
- four special issues on NKOS [8,14,24,25], and
- two scattered NKOS workshops at ISKO-UK 2011 and ICADL 2015.

For the analysis in this paper, we compiled all research presentations at NKOS workshops and papers published in special issues. We restrict our analysis to papers which were authored by a minimum of two authors. This restriction reduces the content of the dataset, e.g., the ECDL NKOS workshop from 2000 is missing in Table 1 because all papers were single author papers. In total, this results in a dataset of 123 papers with a sum of 256 distinct authors (see Table 1)[4].

3 Network analysis of the NKOS community

In order to analyze the collaboration of the NKOS community, we build a network of all authors at the workshops and special issues and compute various centrality measures for each author. A link in this network represents two authors who wrote a paper together. Therefore, if we have n_p number of papers and a paper i has m_i authors, the total number of pairs (links) E is

$$E = \sum_{i=1}^{n_p} \frac{m_i(m_i - 1)}{2} \qquad if \quad m_i \geq 1 \tag{1}$$

[3] The NKOS workshop bibliography is maintained in the following repository: https://github.com/PhilippMayr/NKOS-bibliography.

[4] The data for this subset are available under https://github.com/PhilippMayr/NKOS-bibliography/tree/master/publications/ijdl17.

Table 1 Overview of all NKOS papers sorted by years. In general, the community shows a high average clustering in many years indicating that there are many triangles in the network

Year	Nr. of papers	Nr. of authors	Nr. of links	Avg. clustering
2001	4	9	6	0.37
2002	3	10	13	0.8
2003	5	12	9	0.4
2004	13	39	47	0.65
2005	7	22	26	0.81
2006	11	33	39	0.73
2007	4	15	24	1.0
2008	7	15	9	0.2
2009	10	34	60	0.68
2010	8	21	19	0.61
2011	8	32	59	0.80
2012	6	26	56	0.92
2013	5	18	31	0.86
2014	6	16	13	0.85
2015	9	24	23	0.58
2016	17	60	114	0.75

If two authors published more than one paper together, we give weights to the link equivalent to the number of times they collaborated in different papers. Thus, the resulting network is a weighted undirected graph.

In this paper, first, we investigate global properties of the network over time. Second, we analyze the centrality of the authors in this network. Lastly, we investigate gender differences in the collaboration behavior in this community.

4 Results

Figure 1 demonstrates the overall NKOS co-authorship network. In this view, each author has at least one co-author. The node color represents the gender; purple for men and orange for women. This network contains 44 components. From the network illustrated in this figure, we selected the largest component that is represented in Fig. 3. One hundred and seven authors (41% of all authors) are connected in this component. The NKOS co-authorship network in the "NKOS bibliography" is a typical co-authorship network with one relatively large component, some smaller components and many isolated co-authorships or triples.

Figure 2 shows the degree distribution for this network. Despite being a rather small network, the degree distribution follows a similar trend as a power-law degree distribution that has been observed in other co-authorship networks [1,11].

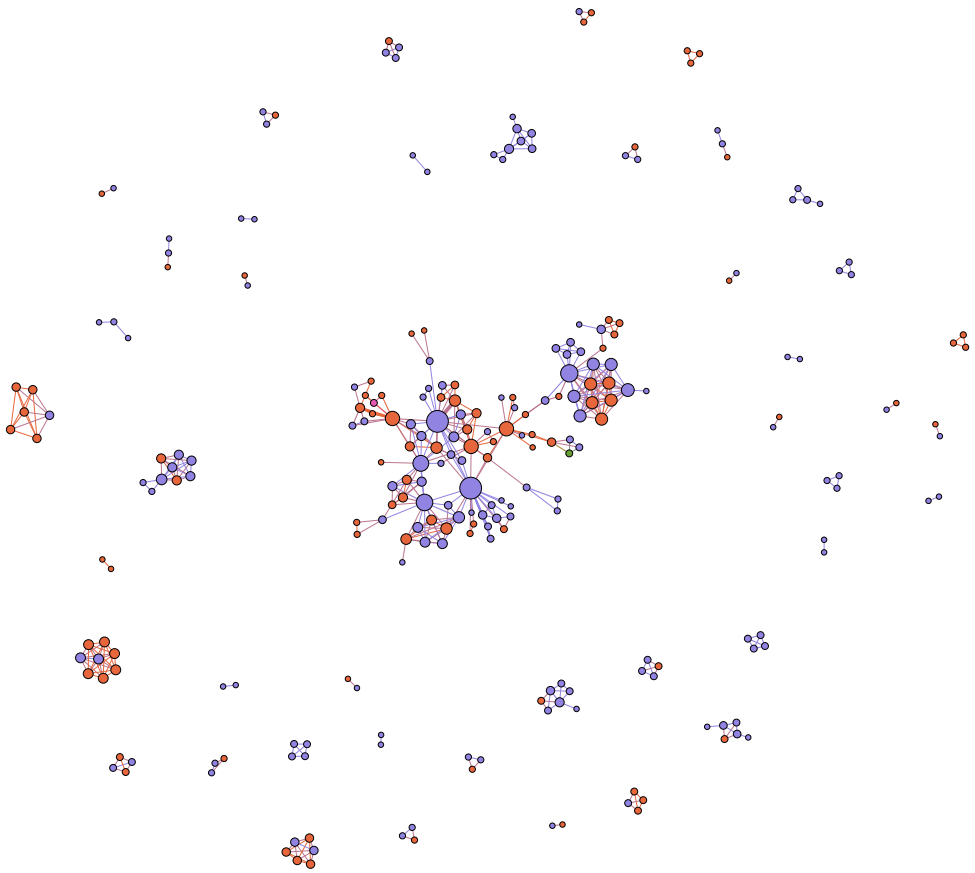

Fig. 1 Co-authorship network of the NKOS community. In general, the network is sparse and contains 44 isolated components. The largest connected component (the cluster in the middle) contains 107 nodes. Nodes are colored based on their gender. Purple nodes represent men, and orange nodes represent women (color figure online)

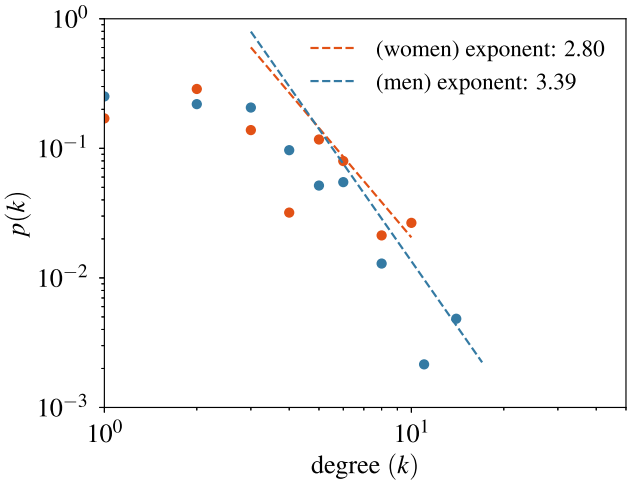

Fig. 2 Degree distribution of the NKOS network. Blue and orange colors indicate the distribution for men and women, respectively. Although the network is small, it exhibits power-law degree distribution (color figure online)

In Fig. 3, the largest connected component, we see that scientists tend to forge intra-institutional collaborations [6]. Good examples are the clusters from Johannes Keizer (FAO), Antoine Isaac (Vrije Universiteit Amsterdam/Europeana) and Philipp Mayr (GESIS). A large fraction of their co-authors are affiliated with the same institution. Also a tendency to select those co-authors who are in geographic proximity is visible in Fig. 3. For example, Douglas Tudhope (University of South Wales, UK) has a larger fraction of UK-affiliated co-authors.

4.1 Node centralities

To detect the influence of authors on information exchange, we calculate various measures of centrality, namely, degree centrality, betweenness centrality and closeness centrality of the authors. Here, we only focus on the largest connected component (LCC) in order to have a robust comparison.

Degree centrality is the most straightforward measure of centrality that depicts the importance of nodes in terms of total number of unique links. The authors with high degree

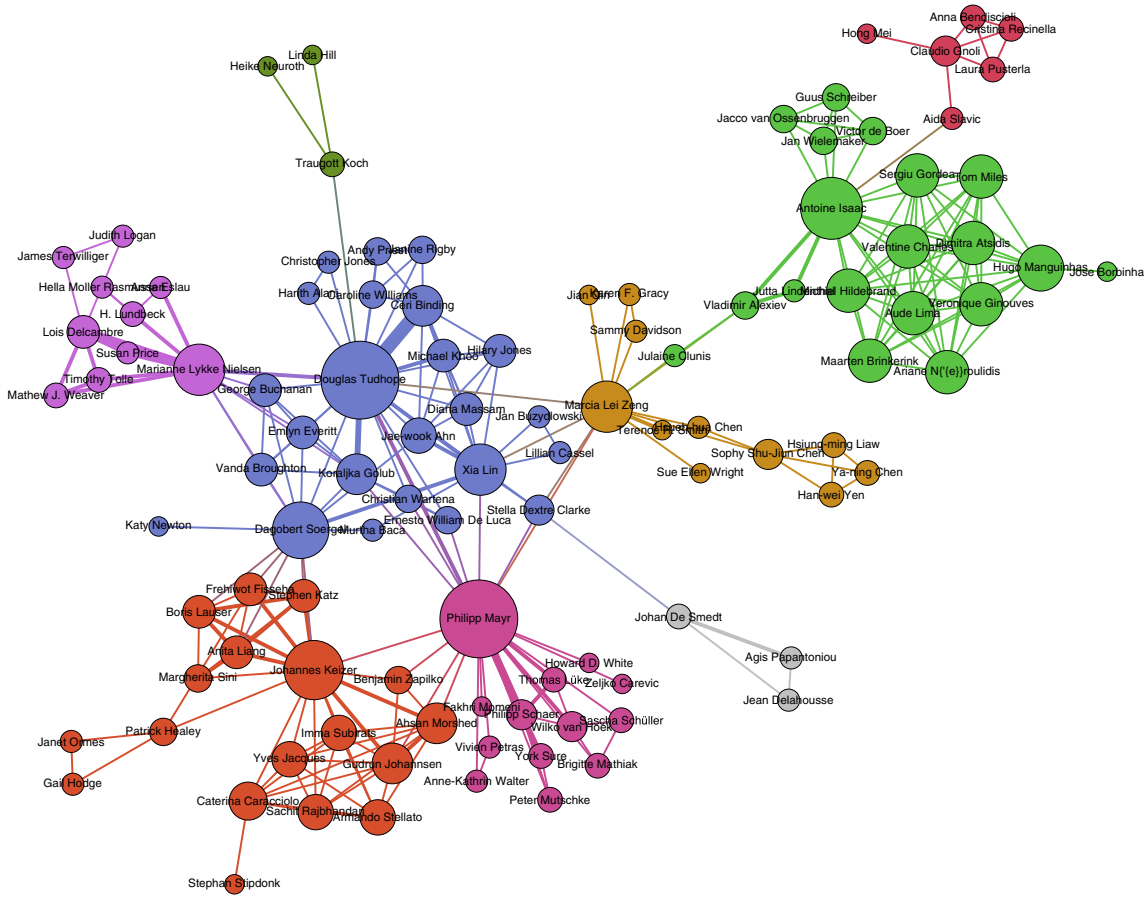

Fig. 3 Largest component in the NKOS co-authorship network. The network is clustered into 9 clusters using the Louvain clustering method [2]. Nodes are colored based on their cluster, and the size of the node represents the node's degree. Clusters are shaped based on the location of the groups and collaboration among their members. The majority of the scholars in the largest component are based in Europe (color figure online)

centrality have established a wide collaboration with many different scholars.

Betweenness centrality indicates the fraction of the shortest paths between all pairs of nodes that pass through a node. The betweenness of a node indicates the node's ability to funnel the flow in the network [20]. In this network, the author with a high betweenness has a large influence in transferring the information from one part of the network to another.

Closeness centrality indicates how close scholars are from others. Mathematically, it is sum of all the shortest paths between a node to all other nodes [7]. If a shortest path between node u to v is $d(u, v)$ and the total number of nodes in the graph is denoted by N, closeness centrality of the node u is defined as follows:

$$c(u) = \frac{N-1}{\sum_{v}^{N-1} d(u, v)} \tag{2}$$

where $N-1$ in the nominator normalizes the measure so that it becomes size independent. Scholars with high closeness centrality are on average closer to other nodes in the network.

Figure 4 shows the comparison of centrality measures for top 15 authors in the largest connected component. It is interesting to note that author centrality ranks may vary depending on the type of the centrality measures. For example, even though H. Manguinhas has a relatively high degree centrality, this author does not appear in the top closeness or betweenness rank. A closer look at the author's location in the graph 3 shows that this author is embedded in the light green cluster with high clustering and few connectivity with other clusters.

Comparing closeness centrality and betweenness centrality also shows interesting results. Although some authors have a high closeness to other scholars, they may not have a high betweenness centrality. For example, K. Golub has a relatively high closeness centrality due to the special location of the author in connection with many other authors from different clusters. However, this author does not have a relatively high betweenness centrality because her network position does not allow to connect to other further distanced clusters. In contrast, author A. Slavic does not have a high degree or a high closeness centrality, but this author has a high betweenness centrality due to connecting an almost isolated red cluster to the rest of the network. The same is true for T. Koch. It is important to note that while scholars with higher closeness centrality are on average closer to other scholars and thus can access novel ideas more frequently, authors with high betweenness centrality play a crucial role in transferring the knowledge in the community [10].

4.2 Structural holes and bridges

Weak ties play a crucial role in networks as they connect disconnected clusters and act as bridges in networks. The structural hole idea first coined by sociologist Ronald Burt suggests that nodes can act as a mediator between two or more closely connected clusters. This is in particular important since novel ideas or information need to pass from these gatekeepers to transfer to other parts of the network. Here, we measure the effective size of a node based on the concept of redundancy. A person's ego network has redundancy to the extent to which her neighbors are connected to each other as well. In a simple graph, the effective size of a node u, $e(u)$, can be expressed as:

$$e(u) = n - \frac{2t}{n} \tag{3}$$

where t is the number of the total ties in the egocentric network (excluding those ties to the ego) and n is the number of total nodes in the egocentric network (excluding the ego). The effective size can vary from 1 to the total number of links in the ego [3]. The higher the effective size, the more effective a node is in terms of being a bridge.

Figure 5 displays the top 15 ranked authors with respect to their effective size. The ranking suggests that in this community, nodes with a high degree (hubs) also act as bridges between the clusters; thus, they can transfer novel ideas among their peers.

4.3 Gender differences in the co-authorship network

To infer the gender of the scholars, we use the state-of-the-art approach by combining the results of the first names and Google images of the scholars with their full names [13]. For the remaining unidentified names or names with initials, we manually check the author's online profile based on the title of their papers. Our complete network consists of 97 (38%) women and 157 (62%) men and 2 unidentified names. Compared to other scientific communities and in particular in science and engineering fields, this community shows a higher percentage of active women [11]. The share of women and men in the largest connected component also shows an interesting effect. We find 46 women and 59 men in the LCC which means women occupy 43% of the nodes in this component.

Homophily In the first step, we measure homophily in this network. There are various ways to define homophily. Here, we use two well-defined measures. The first measure of homophily was proposed by Newman that computes the Pearson correlation between attributes when corrected by what we would expect from a node's degree [19]. The homophily varies between -1 (disassortativity) to $+1$ (complete assortativity). We find that gender assortativity in this community is 0.1. This means that there is a positive tendency among scholars in this community to collaborate with the same gender. One can observe the gender homophily from Fig. 1.

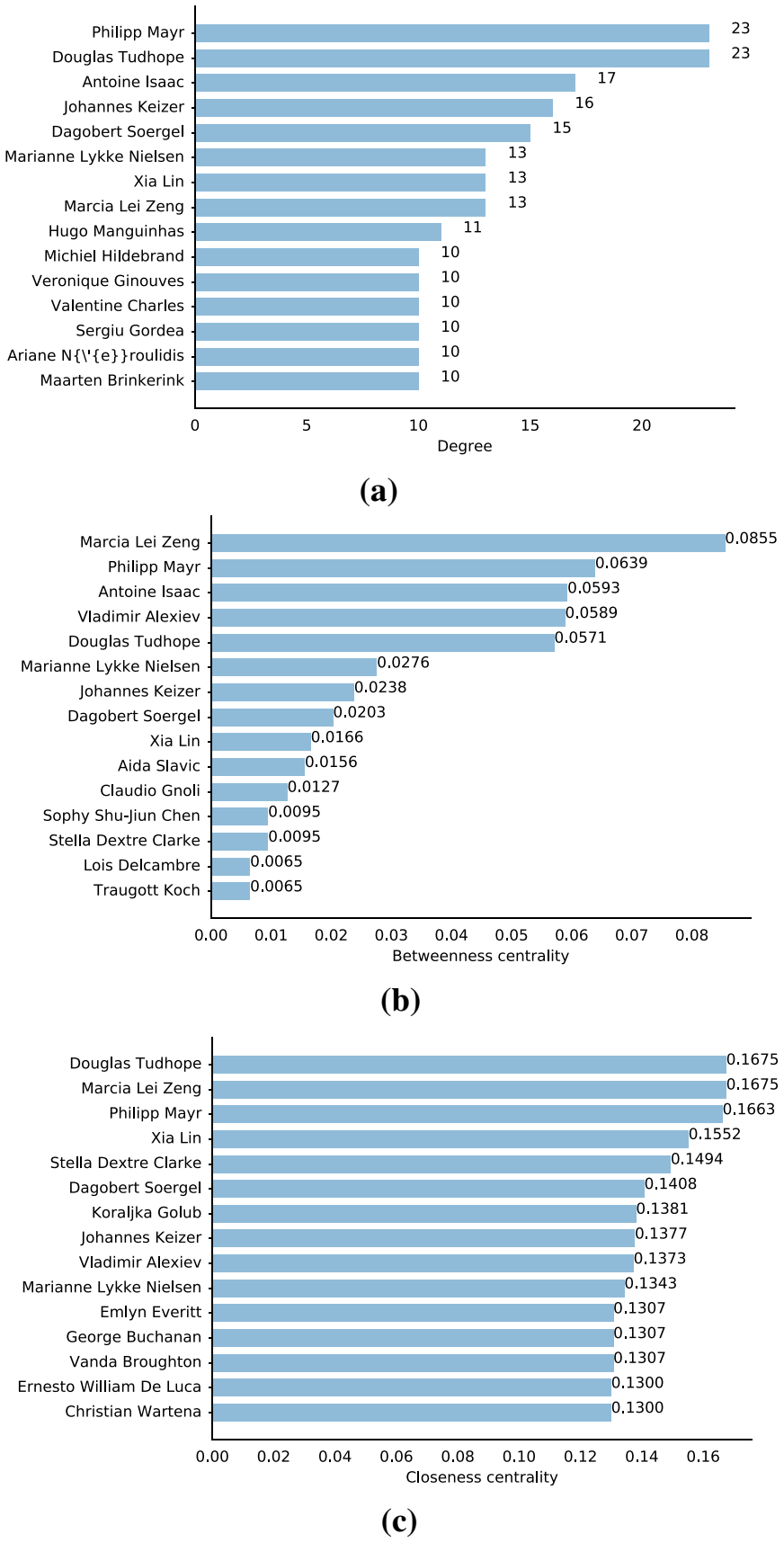

Fig. 4 Top 15 authors with the highest **a** degree centrality, **b** betweenness centrality and **c** closeness centrality

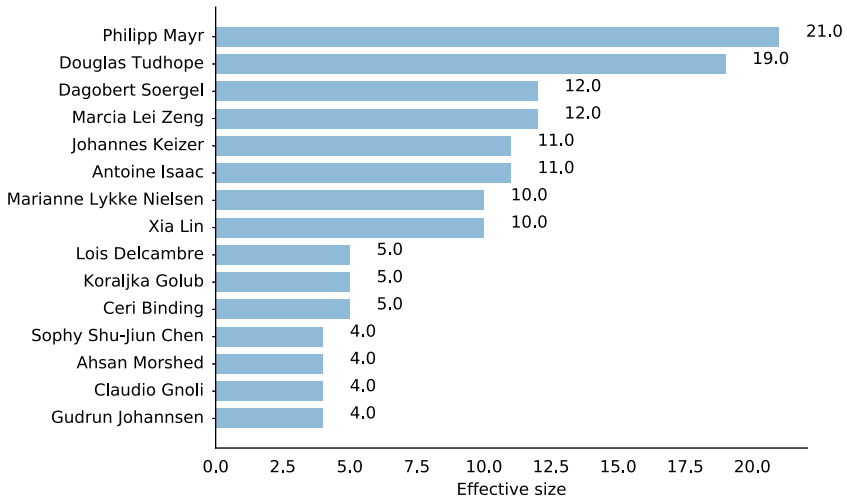

Fig. 5 The top 15 scholars with the highest effective size. The effective size indicates the ability of a node to connect otherwise disconnected nodes and therefore the node can act as a weak tie or bridge

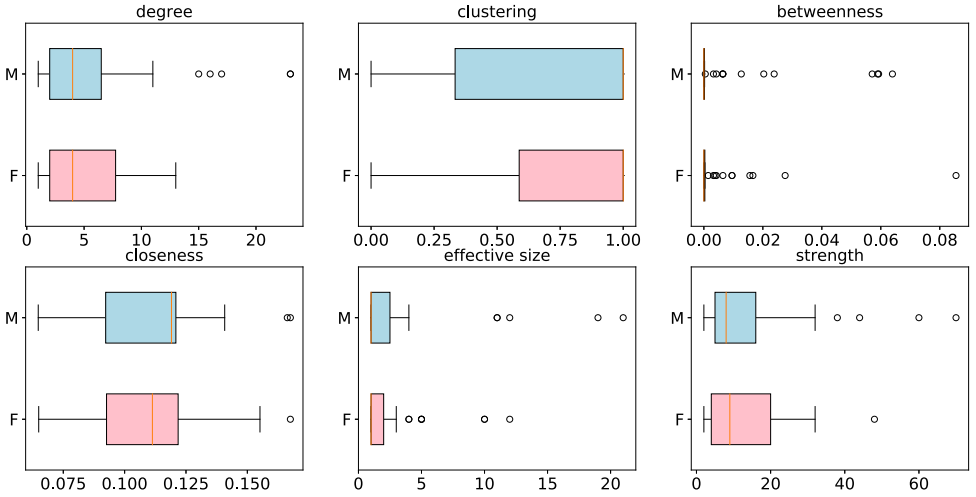

Fig. 6 Box plots indicating median and quartiles of network properties for male and female scholars in the largest connected component. Median is similar for the majority of the node characteristics except for closeness centrality that is higher for men. With regard to degree centrality, there are more outliers among men with a high degree. For clustering, women have higher clustering on average than men. Men also show outliers with higher effective size and strength compared to women

Although the assortativity measure captures the overall homophily in the network, it does not provide additional insights whether or not the nature of homophily is symmetric or asymmetric. Indeed, we have shown previously that asymmetric homophily can impact the degree centrality of the nodes and in particular a minority group in networks [12]. To capture the asymmetric nature of the homophily, we take a simple approach first proposed by Coleman (1958). In this case, we measure the probability of links that exist between two scholars of the same gender. Let us denote the probability of links that exist among women as p_{ww} and among men as p_{mm}. To compare groups of different sizes, the probabilities are compared with group sizes and normalized by the maximum values. If the fraction of women is denoted by f_w and of men by f_m, the Coleman index for women is:

$$C_w = \frac{p_{ww} - f_w}{1 - f_w} \tag{4}$$

A similar definition will apply for men. The maximum value for the Coleman homophily index is 1. When applying this index to our network we get $C_w = -0.12$ for women and $C_m = -0.42$ for men. These results suggest that the homophily among women is higher than the homophily among men in this network. Similar findings were also found in other co-authorship networks [11].

Network characteristics and gender differences Next, we measure the network characteristics among men and women in the largest connected component. We use six measures of networks similar to the previous section. We also include the strength of the node as the sum of all weighted links.

Figure 6 shows box plots comparing network measures for men and women. Overall, the median and quartiles for degree and betweenness are the same for men and women. Women show a higher tendency for higher clustering compared to men. Men show a higher median for closeness centrality compared to women. In addition, there is a higher number

of outliers among men in terms of the degree, effective size and strength compared to women.

5 Conclusion

In this paper, we have analyzed the collaborative research of authors and their connectivity for the special case of NKOS workshop activities including four special issues on NKOS. The results highlight the most active and central scholars in this community. We found differences among centrality measures of the scholars which indicate that scholars play a different role in their collaboration network. We also found the most influential scholars who act as bridges between the clusters. We found 9 clusters in the largest component that show that scholars have a higher tendency to collaborate with those in the same institution or the same geographic proximity [6]. Our analyses show that the NKOS community is rather successful in bringing researchers from different domains together in recent years.

The NKOS co-authorship network consists of 38% women in total, and the share of women in the largest connected component is 43%. The network shows positive gender homophily, and the homophily among women is higher compared to men. We found on average that men have a higher closeness centrality compared to women. In addition, women have a slightly higher clustering compared to men. Apart from these differences, we did not find any significant dissimilarities between men and women with respect to their centralities.

This study has some limitations. First of all, we have included only research paper presentations. Editing and organizing activities at the workshops, which have an enormous impact on the visibility and connectivity of researchers, have not been covered in our dataset. This leads to artifacts, e.g., Traugott Koch,[5] a long-term organizer of the NKOS workshops and editor of the early JoDI special issues on NKOS, is not covered very well in our dataset and the network.

Second, many influential papers (e.g., [9,26]) and standardization activities (e.g., the W3C Recommendation for SKOS [17]) presented and discussed at NKOS events and published after the NKOS workshops are missing. This fact is reducing the representativeness and completeness of the network.

Third, we have not included bibliometric data to complete our analysis. This is because most of the NKOS workshop activities (presentations) are not formally cited or even mentioned in scientific papers. In difference to the workshop output, the few journal papers in the special issues on NKOS are cited. Some works (e.g., [4,5,21–23]) are cited well in the literature. So adding citation data would be a next reasonable step to complete the dataset.

[5] Traugott Koch was a central protagonist and networker of the US and European NKOS community. He retired and left the NKOS community in 2012.

6 Future work

We are planning to extend the analysis of the NKOS network. In this way, we first plan to complement the dataset with other NKOS research output. We also plan to analyze the development of topics in the titles and abstracts of the presentations and papers. Combining network analytic measures with bibliometric analysis (e.g., co-citations, bibliographic coupling) would complement our preliminary observations and advance our understanding of the role of gender and other attributes in scientific collaboration. We invite people to contribute to our open dataset.

Acknowledgements We thank our colleague Marcia Lei Zeng (Kent State University) who provided us with internal information about the US NKOS workshops. This work was partly funded by DFG, Grant No. SU 647/19-1; the "Opening Scholarly Communication in the Social Sciences" (OSCOSS) project at GESIS.

References

1. Barabási, A.L.: Scale-free networks: a decade and beyond. Science **325**(5939), 412–413 (2009)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J Stat Mech Theory Exp **2008**(10), P10008 (2008)
3. Burt, R.S.: Structural holes and good ideas. Am J Sociol **110**(2), 349–399 (2004)
4. Cranefield, S.: Networked knowledge representation and exchange using UML and RDF. J. Dig. Inf. (2001). https://journals.tdl.org/jodi/index.php/jodi/article/view/30
5. Doerr, M.: Semantic problems of thesaurus mapping. J. Dig. Inf. (2001). https://journals.tdl.org/jodi/index.php/jodi/article/view/31
6. Evans, T.S., Lambiotte, R., Panzarasa, P.: Community structure and patterns of scientific collaboration in business and management. Scientometrics **89**(1), 381–396 (2011). https://doi.org/10.1007/s11192-011-0439-1
7. Freeman, L.C.: Centrality in social networks conceptual clarification. Soc. Netw. **1**(3), 215–239 (1978)
8. Hill, L., Koch, T.: Networked Knowledge Organization Systems: introduction to a special issue. J. Dig. Inf. **1**(8) (2001). https://journals.tdl.org/jodi/index.php/jodi/article/view/32/33
9. Hodge, G.: Systems of knowledge organization for digital libraries: beyond traditional authority files (2000). https://www.clir.org/pubs/reports/pub91/pub91.pdf
10. Iyer, S., Killingback, T., Sundaram, B., Wang, Z.: Attack robustness and centrality of complex networks. PloS ONE **8**(4), e59613 (2013)
11. Jadidi, M., Karimi, F., Wagner, C.: Gender disparities in science? dropout, productivity, collaborations and success of male and female computer scientists (2017). arXiv preprint arXiv:1704.05801

12. Karimi, F., Génois, M., Wagner, C., Singer, P., Strohmaier, M.: Visibility of minorities in social networks (2017). arXiv preprint arXiv:1702.00150
13. Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., Strohmaier, M.: Inferring gender from names on the web: A comparative evaluation of gender detection methods. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 53–54. International World Wide Web Conferences Steering Committee (2016)
14. Mayr, P., Tudhope, D., Clarke, S.D., Zeng, M.L., Lin, X.: Recent applications of Knowledge Organization Systems: introduction to a special issue. Int. J. Dig. Libr. **17**(1), 1–4 (2016). https://doi.org/10.1007/s00799-015-0167-x
15. Mayr, P., Tudhope, D., Golub, K., Wartena, C., De Luca, E.W.: Proceedings of the 15th European Networked Knowledge Organization Systems (NKOS) Workshop. CEUR-WS.org (2016). http://ceur-ws.org/Vol-1676/
16. Mayr, P., Tudhope, D., Golub, K., Wartena, C., De Luca, E.W.: Proceedings of the 17th European Networked Knowledge Organization Systems (NKOS) Workshop. CEUR-WS.org (2017). http://ceur-ws.org/Vol-1937/
17. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System Reference (2009). https://www.w3.org/TR/skos-reference/
18. Momeni, F., Mayr, P.: Analyzing the research output presented at European Networked Knowledge Organization Systems Workshops (2000–2015). In: Proceedings of the 15th European Networked Knowledge Organization Systems Workshop (NKOS 2016). pp. 7–14. CEUR-WS.org, Hannover, Germany (2016). http://ceur-ws.org/Vol-1676/paper1.pdf
19. Newman, M.E.: Assortative mixing in networks. Phys. Rev. Lett. **89**(20), 208701 (2002)
20. Opsahl, T., Agneessens, F., Skvoretz, J.: Node centrality in weighted networks: generalizing degree and shortest paths. Soc. Netw. **32**(3), 245–251 (2010). https://doi.org/10.1016/j.socnet.2010.03.006
21. Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S.: Reengineering thesauri for new applications: the agrovoc example. J. Dig. Inf. (2004). https://journals.tdl.org/jodi/index.php/jodi/article/view/112
22. Trant, J.: with the participants in the steve.museum project: exploring the potential for social tagging and folksonomy in art museums: Proof of concept. New Rev. Hypermed. Multimed. (2006). http://www.tandfonline.com/doi/abs/10.1080/13614560600802940
23. Tudhope, D., Alani, H., Jones, C.: Augmenting thesaurus relationships: possibilities for retrieval. J. Dig. Inf. (2001). https://journals.tdl.org/jodi/index.php/jodi/article/view/181/160
24. Tudhope, D., Koch, T.: New applications of knowledge organization systems: introduction to a special issue. J. Dig. Inf. **4**(4) (2004). https://journals.tdl.org/jodi/index.php/jodi/article/view/109/108
25. Tudhope, D., Lykke Nielsen, M.: Introduction to Knowledge Organization Systems and Services. New Rev. Hypermed. Multimed. **12**(1), 3–9 (2006)
26. Zeng, M.L., Chan, L.M.: Trends and issues in establishing interoperability among knowledge organization systems. J. Am. Soc. Inf. Sci. Technol. **55**(3), 377–395 (2004)

Chapter 4

Investigating Impact and Factors of Open Access Publishing

4.1 Overview

In Section 4.2.1, we introduce the study [74], which aims to describe the transformation process from traditional to OA publishing in more detail from a bibliometric aspect. To this end, we tracked changes in articles' number, received citations, and impact factor of journals after flipping from CA to the OA publishing model. By studying these changes, we gain a deeper understanding of the issues and conflicts related to OA publishing.

In Section 4.2.2, we present our study [76] that examines the factors related to OA publishing. Specifically, we investigated the relationship between countries' income levels and the adoption of OA publishing, both in terms of publication and citation patterns. We analyzed various factors at the author, paper / venue, and country levels to determine which models authors use to publish their articles. To this end, we employed correlation analysis and machine learning prediction models to assess the association between these factors and OA publishing.

4.2 Publications

4.2.1 Impacts of Flipping a Journal to Open Access

Title: What happens when a journal converts to open access? A bibliometric analysis

Authors: Fakhri Momeni, Philipp Mayr, Nicholas Fraser, and Isabella Peters

Document Type: Journal paper

Venue: Scientometrics

DOI: https://doi.org/10.1007/s11192-021-03972-5

Scientometrics (2021) 126:9811–9827
https://doi.org/10.1007/s11192-021-03972-5

What happens when a journal converts to open access? A bibliometric analysis

Fakhri Momeni[1] · **Philipp Mayr**[1,2] · **Nicholas Fraser**[3] · **Isabella Peters**[3]

Received: 15 June 2020 / Accepted: 25 March 2021 / Published online: 26 April 2021

Abstract

In recent years, increased stakeholder pressure to transition research to Open Access has led to many journals converting, or 'flipping', from a closed access (CA) to an open access (OA) publishing model. Changing the publishing model can influence the decision of authors to submit their papers to a journal, and increased article accessibility may influence citation behaviour. In this paper we aimed to understand how flipping a journal to an OA model influences the journal's future publication volumes and citation impact. We analysed two independent sets of journals that had flipped to an OA model, one from the Directory of Open Access Journals (DOAJ) and one from the Open Access Directory (OAD), and compared their development with two respective control groups of similar journals. For bibliometric analyses, journals were matched to the Scopus database. We assessed changes in the number of articles published over time, as well as two citation metrics at the journal and article level: the normalised impact factor (IF) and the average relative citations (ARC), respectively. Our results show that overall, journals that flipped to an OA model increased their publication output compared to journals that remained closed. Mean normalised IF and ARC also generally increased following the flip to an OA model, at a greater rate than was observed in the control groups. However, the changes appear to vary largely by scientific discipline. Overall, these results indicate that flipping to an OA publishing model can bring positive changes to a journal.

Keywords Open access publishing · Scholarly communication · Citation analysis · Scholarly journals · Journal publishing models

✉ Fakhri Momeni
fakhri.momeni@gesis.org

✉ Philipp Mayr
philipp.mayr@gesis.org; mayr@informatik.uni-goettingen.de

✉ Nicholas Fraser
N.Fraser@zbw.eu

Isabella Peters
I.Peters@zbw.eu

[1] GESIS Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany

[2] Institute of Computer Science, University of Göttingen, Göttingen, Germany

[3] ZBW Leibniz Information Centre for Economics, Düsternbrooker Weg 120, 24105 Kiel, Germany

Introduction

For hundreds of years, the closed-access (CA) model has been the traditional publishing model, where journal articles are published behind a "paywall" that can be removed through the payment of subscription fees to the publisher, most commonly by academic libraries or research funders. Over the past three decades, the growth of the Internet and resulting opportunities for low-cost distribution of digital content have led to a revolution in scholarly publishing (Björk, 2017; Laakso et al., 2011). In the midst of these changes, a new business model for publishers of scholarly journals has emerged besides the traditional model: an open-access (OA) model, where journal articles are made freely-available to all readers, and the publication costs are borne by third-parties, such as authors, institutions, societies or funders. These publishing models may not be mutually exclusive (e.g. a CA journal may still allow certain articles to be published under OA licenses, usually referred to as "Hybrid" publishing), and may not remain static over time; a journal may convert from a CA model to an OA model or vice versa, processes commonly termed as "flipping" or "reverse flipping", respectively (Solomon et al., 2016; Matthias et al., 2019). Recent quantitative studies found e.g. that more than 50% of the newer articles indexed by Web of Science are freely available in "some form" of OA via Google Scholar (Martin-Martin et al., 2018). As a result, the speed of adoption of OA is increasing constantly. Hobert et al. (2020) can show this trend (OA uptake) in a large-scale study for German universities and non-university research institutions in the period 2010–2018. They found out that 45% of all considered articles in the observed period 2010–2018 were openly available at the time of analysis in one form of OA (Green OA, Gold OA and other OA variants). Hobert et al. showed for Germany that subject-specific repositories are still the most prevalent OA type, but fully OA journals are steadily increasing in the analysed period.

Journal flipping in itself is not a new concept: Peter Suber previously noted that "*Subscription journals have been converting or "flipping" to open access (OA) for about as long as OA has been an option*" (in Solomon et al., 2016). However, the topic has received more attention in recent years due to increased funder pressure to accelerate the transition to OA, for example through Plan S in Europe (https://www.coalition-s.org/), driven in part by the increasingly unsustainable economic costs of the subscription model (Schimmer et al., 2015; Tennant et al., 2016). For publishers who intend to transition from a CA to an OA business model there is an urgent need to understand the journal flipping process and its consequences. A clear concern of these publishers is to find an alternative stable stream of income to subscription fees. OA journal revenue streams are commonly associated with Article Processing Charges (APCs), whereby authors, institutions or funders pay fees to a publisher on a per-published-article basis. Björk (2012) demonstrated that this author-pays model in hybrid journals is unpopular with authors, whilst Peterson et al. (2013) argued that the cost of APC in this model is often too much for many authors, and publishers try to solve this problem in different ways such as fee waiver policies, subsidizing academic publishing directly without profiteering intermediaries, etc. However, according to the public journal dataset from the Directory of Open Access Journals (DOAJ),[1] only 4,021 of 14,741 (27.2%) of journals charge APCs (data accessed on 10th June 2020); the remainder may be supported, for example, by individual societies or library presses. Even so, according to

[1] https://doaj.org/, public metadata dump available at https://doaj.org/public-data-dump.

Solomon et al. (2016), transitioning a journal to an OA model for those societies with low numbers of publications can be expensive.

Predicting how a change in the business model will affect the long-term viability of a flipped journal is of immense importance to those responsible for journal management, thus in-depth, longitudinal bibliometric studies can help to inform decision making of publishers, and their assessment of chances and risks of flipping their journals (see e.g. Perianes-Rodríguez & Olmeda-Gómez, 2019). Such bibliometric studies may focus on multiple aspects of publishing behaviour, such as changes in publishing volume, which is itself a function of submission volumes and editorial selection processes, as well as changes in article impact, which may be a proxy for the future "attractiveness" of a journal to researchers. This study addresses both of these aspects, building on work presented in Momeni et al. (2019) but substantially expanding its scope, in terms of the data sources of flipped journals and the bibliometric data analysed (from Web of Science to Scopus). Moreover, in this study we included a comparison of flipped journals and journals from the same disciplines that still publish under the CA model (as suggested recently by Bautista-Puig et al., 2020). We aim to answer the following research questions:

(1) Do journals flipping from a CA to an OA model experience a positive/negative change in the volume of articles published?
(2) Do journals flipping from a CA to an OA model experience a positive/negative change in their Impact Factor?
(3) Do articles in journals flipping from a CA to an OA model experience a positive/negative change in their individual citation impact?

An important point to note, is that this study focuses only on journals that have flipped from a CA to an OA model, whilst retaining the same journal name. Over the past years a number of journal "declarations of independence"[2] have resulted in the resignation of editors from a CA journal to form a new OA journal at an alternative publisher (e.g. the editorial board of *Journal of Informetrics*, published by Elsevier, transitioned to a new journal called *Quantitative Science Studies*, published by MIT Press; Waltman et al., 2020). Although these transitions are closely related to the concept of "flipping", they not only concern the journal name, but also involve a change in journal venue and potential attractiveness, which may make the direct effects of transitioning from CA to OA difficult to distinguish. In our study we considered just journals which kept the same journal name after flipping.

Related work

Studies on journal flipping from a bibliometric perspective

Relatively few studies have been carried out that systematically research the effects flipping has on a journal's publication output and impact. One of the earliest studies from Solomon et al. (2013) documented the growth of OA journals, their articles and normalized citation rates between 1999 and 2010, whilst also controlling for whether the journal had

[2] http://oad.simmons.edu/oadwiki/Journal_declarations_of_independence.

been launched as an OA journal, or flipped to an OA journal at a later point. The authors combined data from Scopus and DOAJ, and manually reviewed the public websites of all journals included in their matched dataset (N = 2012), finding that of these journals, 1,064 were flipped from a CA to an OA model, whilst 931 journals were "born-OA" (17 were undetermined). According to the data from Solomon et al. (2013), the number of flipped OA journals peaked in 2005, and since then decreased year-on-year; in 2012 less than 20 journals were discovered that had flipped from a CA to an OA model. In terms of citations, the authors compared longitudinal trends in Source Normalized Impact per Paper (SNIP), a citation metric that accounts for field-specific differences in citation. They find that overall citation rates for flipped OA journals were approximately 50% lower than those from CA journals, but this relationship did not change greatly over time. Conversely, born-OA journals experienced a strong growth in SNIP between the years 2003–2005, eventually reaching a plateau with citation rates almost at the same levels of CA journals.

In another study, Busch (2014a, b) investigated the response of the Impact Factor (IF) of six journals which were transferred from CA models at other publishers to the OA model of BioMed Central between the years 2006 and 2011. IFs were compared to the median IF of journals from the same Web of Science subject category. Four of the six journals experienced a sizeable increase in IF following the flip to OA—for example the *Journal of Cardiovascular Magnetic Resonance* increased its IF from 1.87 in the year prior to flipping, to 4.33 in the year after flipping, a ~ 130% increase. For the remaining two of the journals, IFs remained relatively static or even fell following the flip, although the author notes that the goal of these journals for the years in question was to increase their publishing volume, which may have led to less selective editorial decisions; both journals in fact accepted around 50% more submissions in the post-flip years than pre-flip. These results must, however, be interpreted carefully; not only did the journals flip from a CA to an OA model, but they also transferred to a new publisher (although keeping their old name) which may have had an important effect on the journal's visibility.

As well as converting from a CA model to an OA model, some journals may also convert in the opposite direction, from an OA to a CA model, a process that has been termed "reverse flipping" (Matthias et al., 2019). The study of Matthias et al. (2019) investigated the publication and citation behaviour of 152 journals that were identified as having reverse-flipped from 2005 onwards. Interestingly, 62% of those journals had initially been CA journals and flipped to an OA model, before flipping back to a CA model. The authors also found that publication volumes and citation metrics changed little in the two years before or after the reverse flip, although some individual journals encountered large variability. Reasons for reverse flipping may in part be due to a lack of success with the OA model, for reasons such as financial sustainability or low article volumes, although 69% of reverse flips were related to a change in publisher and thus the journal may have simply adopted the prevalent publication model of the new publisher.

A more recent study by Bautista-Puig et al. (2020) follows a similar methodology to this study. The authors used data combined from DOAJ (N = 119 journals) and the Open Access Directory (OAD)[3] (N = 100 journals), who host a community-maintained list of journals that have flipped from a CA (referred to by OAD as "TA", for Toll Access) to an OA publishing model.[4] The authors compared post-flip and pre-flip bibliometric indicators including publishing volumes and normalised citation rates, against two distinct control

[3] http://oad.simmons.edu/oadwiki/Main_Page.

[4] http://oad.simmons.edu/oadwiki/Journals_that_converted_from_TA_to_OA.

groups: a standard control group, as well as a "tailor-made" control group accounting for a journal's national orientation. The authors found evidence of an OA citation advantage: DOAJ journals increased their normalised IF by ~ 50% at 4-years post-flipping, compared to just ~ 10% in the standard control group, whilst OAD journals increased their normalised IF by ~ 35% in the same time interval, compared to ~ 15% in the standard control group. However, the authors found no evidence for an OA publication advantage: for all groups, the journals experienced an increase in publishing volumes in the range of 10–20%. The authors also assessed changes in the affiliation countries of publishing and citing authors after a journal had flipped. They found that overall, the share of authors from high-income countries declined after a journal flipped to an OA model, although a similar effect was also found in the respective control groups.

The present study is an extension of the previous study of Momeni et al. (2019). In the previous study, we used a list of flipped journals available from OAD, as also used by Bautista-Puig et al. (2020). The list of journals was matched to journals contained in the Web of Science (N = 171) to determine the effects on publication volume and two citation metrics, one at the journal level (IF) and one at the article level (average of relative citations; ARC), of flipping a journal to OA. These initial results showed that flipping a journal mostly had positive effects on a journal's IF, but conversely had no strong effect on the citation impact at the level of individual articles. We also observed a small decline in the number of articles that were published by a journal after flipping to an OA model. Whilst these initial findings were interesting, they also came with several limitations: (1) we only considered a small sample size of journals from a single source, (2) we did not consider the relevant journal and article metrics with respect to any form of control group, thus we could not interpret whether these changes deviated from global publishing and citation patterns, and (3) we did not consider how publication and citation behaviour might vary in different scientific communities. We therefore attempt to address these limitations in the current study, by increasing our sample size with the addition of a new list of journals from DOAJ, by generating a control group for comparing to the group of flipped journals, and conducting analysis at the level of scientific disciplines.

The OA citation advantage

In our study we also aim to report on changes in citation impact resulting from a journal flipping from a CA to an OA model, both at the journal level and article level. A number of studies have already attempted to study the relationship between OA and citation impact, with most evidence pointing towards an open access citation advantage (OACA) for OA articles over CA articles (Lewis, 2018; McKiernan, 2016; Ottaviani, 2016; Piwowar & Vision, 2013; Sotudeh et al., 2015; Swan, 2010). The Scholarly Publishing and Academic Resources Coalition (SPARC) maintained a repository of 70 studies investigating the OACA[5] until 2015; of these, 46 (65.7%) found a citation advantage, whilst only 17 (24%) found no advantage (the remaining 7 records were inconclusive). A subsequent large-scale study of 3.3 million articles published between 2007 and 2009 by Archambault et al. (2016) found that OA papers received ~ 23% higher citation impact overall than the global average citation rate, although the effect was stronger in Green OA (i.e. OA articles made available through open repositories) forms than Gold OA (OA articles published in

[5] https://bit.ly/SPARC-OACA.

fully OA journals). These findings were echoed in a study by Piwowar, (2018), who found that OA articles receive on average ~ 18% more citations than CA articles, but again this advantage was driven primarily by Green OA, whilst Gold OA was found to have slightly lower citation rates than the global average. Whilst many of these studies note a strong correlation between OA and citation rates, it is important to note that these findings do not necessarily imply causation, as citations may be influenced by a number of additional structural and author-specific factors (Tahamtan et al., 2016). Other studies based on randomized control trials (Davis, 2011) have also reported conflicting results, indicating that methodologies taking into account multiple factors are necessary to understand the exact mechanism driving higher citation rates of OA articles.

Data and methods

Groups of flipped journals

For this study we compiled groups of flipped journals from two main sources: DOAJ and OAD. DOAJ is a directory of ~ 14,700 OA journals, maintained by the Infrastructure Services for Open Access (IS4OA). Journals must apply for indexing in DOAJ and meet a set of basic quality control and transparency criteria to be included. DOAJ provides access to metadata of all indexed journals, which includes a field containing the first calendar year that a complete volume of the journal provided OA to the full text of all articles (herein referred to as "flipping year"). Note that journal metadata in DOAJ is provided by the publishers directly and is thus not "verified" by any third party. As Sotudeh and Horri (2007) and Bautista-Puig et al. (2020) have shown this can often lead to inaccurate data, e.g., in terms of flipping date. To build a group of flipped journals, we extracted details of all journals in DOAJ as well as the flipping year. This group of journals was compared to the *Zeitschriftendatenbank* ("Journal database"; ZDB[6]), a database of high-quality journals and other periodicals maintained by the *Staatsbibliothek zu Berlin* ("State Library of Berlin"). An advantage of using the ZDB is that they maintain a "first issued year" field for each contained journal, and thus by comparing this year with the flipping year field from DOAJ, we can discover journals that were previously a CA journal and then changed to an OA model (i.e. we exclude any journals that were initiated as OA journals). For bibliometric analysis, this group of journals was then matched to journals contained in Scopus via matching of journal names (case-insensitive) and ISSNs. We therefore have only considered journals that had the same names and ISSNs before and after the flip. Access to Scopus was provided via the German Competence Centre for Bibliometrics,[7] who maintain an in-house, quality-controlled version of the Scopus database. To follow common standards of bibliometric studies, we applied a number of filters to journals matched between the datasets, namely that the journal must have flipped between 2001 and 2013, that there must be more than 5 years distance between the first issued year and the flipping year, and the journals must have published citable articles in every year for the 4 years prior to and following the flipping year. This resulted in a final group of 234 flipped journals from DOAJ.

[6] https://www.zeitschriftendatenbank.de.

[7] http://www.forschungsinfo.de/Bibliometrie/en/.

The second group of journals was derived from OAD, a wiki where the OA community can create and support simple factual lists about open access to science and scholarship, hosted by the School of Library and Information Science at Simmons College. OAD contains a community-maintained list of journals that have flipped from CA to OA. We manually retrieved the full list of journals as well as their flipping years from the public web page. Annotations on the website described whether the journal had flipped to a full OA or a hybrid OA model—in this study we only retained journals that flipped to a full OA model. The group of journals were matched to journals indexed in the Scopus via matching of journal names. Just journals with citable articles in all four years around the flipped year, and with flipping years between 2001 and 2013 were included in the study. The final OAD-group contains 87 journals.

Our two compiled groups of journals from DOAJ and OAD have 12 journals in common. In the following, we will treat these two groups as independent journal groups.[8]

Control groups

For comparative analysis we defined a control group of CA journals for each of the two groups of flipped journals. The control journals were designed to be similar to our flipped journals in terms of discipline, number of published articles and IF in the year of journal flipping. We first defined a candidate list of CA journals, which were obtained from data in Unpaywall, a service that finds OA versions of journal articles and also provides open access to metadata relating to journal publishing models. We used the metadata fields "journal_is_oa" and "article_is_oa" to generate a list of CA journals that do not contain any OA articles (i.e. where journal_is_oa = FALSE and article_is_oa = FALSE for all articles within a journal). These journals were matched with journals contained in Scopus on the basis of shared journal titles (case-insensitive), and then for each journal in our groups of flipped journals, the top 20 percent of CA journals were taken from the same discipline, with the smallest difference in number of published articles in the flipping year. Lastly, from this group of journals with similar volume of articles, a single control journal was selected with the smallest difference in calculated IF to the flipped journal in the flipping year. For the journals with multiple disciplines, a control journal was selected from each individual discipline of the flipped journal. With this method, we have generated two separate control groups, one for the list of flipped DOAJ journals and one for the flipped OAD journals. However, these two control groups may entail common journals.

Bibliometric indicators used in study

In this study, we investigate the effect of flipping from CA to OA through a descriptive analysis of the timeline of CA to OA conversions, the change in the number of articles published by the flipped and control journals over time, as well as two metrics of citations at the article and journal level: the average relative citations (ARC) and normalised impact factor (IF), respectively. An explanation of the latter two metrics is given in the following paragraphs.

[8] We have published our two journal groups with the control group journals on the following page (https://github.com/momenifi/flipped-journals).

Table 1 Fields and disciplines as provided by Scopus

Fields	Disciplines
Physical Sciences	Chemical Engineering
	Chemistry
	Computer Science
	Earth and Planetary Sciences
	Energy
	Engineering
	Environmental Science
	Material Science
	Mathematics
	Physics and Astronomy
	Multidisciplinary
Health Sciences	Medicine
	Nursing
	Veterinary
	Dentistry
	Health Professions
	Multidisciplinary
Social Sciences	Arts and Humanities
	Business, Management and Accounting
	Decision Sciences
	Economics, Econometrics and Finance
	Psychology
	Social Sciences
	Multidisciplinary
Life Sciences	Agricultural and Biological Sciences
	Biochemistry, Genetics and Molecular Biology
	Immunology and Microbiology
	Neuroscience
	Pharmacology, Toxicology and Pharmaceutics
	Multidisciplinary

ARC is calculated, by first calculating a relative citation (RC) count for each individual article published within our flipped journal and control journal datasets, normalised to account for different citation patterns across disciplines. For this calculation we only included articles with the "type" property of "Article" or "Review", as contained within Scopus. The RC of a paper is calculated for each year by computing the sum of citations gained by the individual article, divided by the average number of citations of all papers across its discipline(s) published in the same year. We use a citation window of three years. An RC value above 1 means that a paper is cited more frequently than the average citation level for all papers in that discipline, and vice versa. To calculate the citation performance of a group of papers relative to papers in the same discipline and publication year, we simply calculate the arithmetic mean of the RC of all papers in the group, referred to as the average of relative citations (ARC).

For each journal in our dataset, we calculated the two years IF following a similar methodology to that commonly associated with the Journal Citation Reports, produced

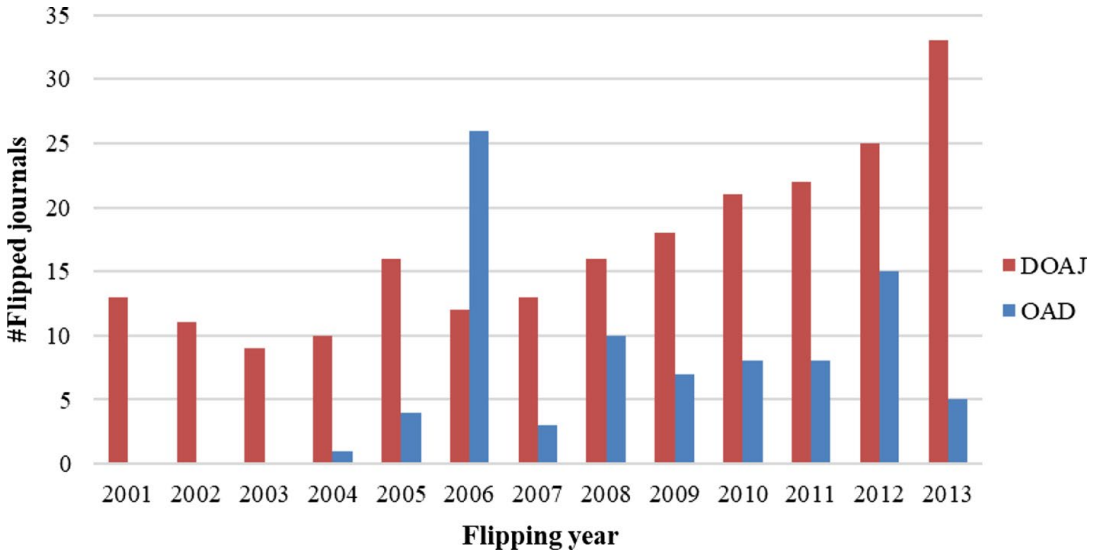

Fig. 1 Distribution of flipped DOAJ and OAD journals by year. In this study we considered only journals with flipping years between 2001 and 2013

by Clarivate Analytics.[9] Based on this definition, the IF is defined as all citations to the journal in the current year to items published in the previous two years, divided by the total number of citable items (these comprise articles, reviews, and proceedings papers) published in the journal in the previous two years. In order to compare IFs between different disciplines, we conducted an additional normalisation step using the rescaling method introduced by Radicchi et al. (2008). So the citation rate for each individual article used in the IF calculation was rescaled by dividing by the arithmetic mean of the citation rate of all articles in its discipline.

To calculate RC and normalized IF across disciplines we used the "All Science Journal Classification" (ASJC) classification system[10] of Scopus which has four fields (called 'subject areas' in Scopus) and 27 disciplines (called 'subject area classifications' in Scopus; see Table 1). In this classification system, journals can have more than one category, therefore we considered the mean citation rate of all articles in all disciplines of which the journal belongs to.

Results

In the following we will present the results of our descriptive analysis.

Analysis of the flipping year

IFs are calculated based on citations earned by articles published in the two past years, thus we expect to observe the impact of converting to OA at least one year after the flip.

[9] https://clarivate.com/webofsciencegroup/essays/impact-factor/.

[10] https://service.elsevier.com/app/answers/detail/a_id/14882/supporthub/scopus/~/what-are-the-most-frequent-subject-area-categories-and-classifications-used-in/.

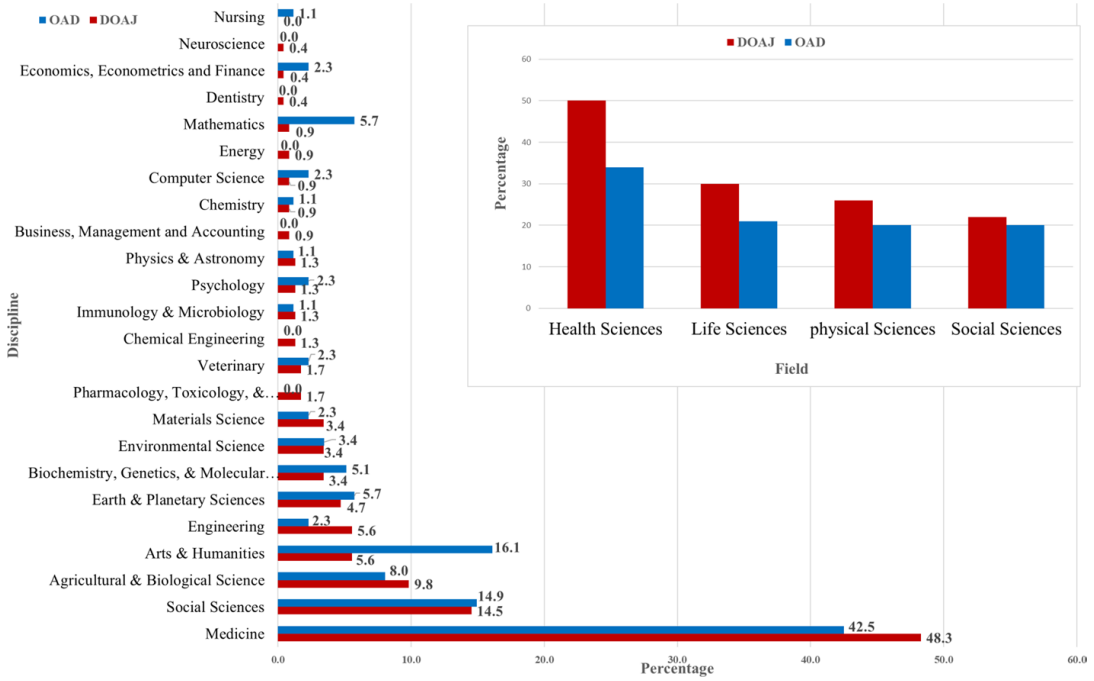

Fig. 2 Proportion of flipped DOAJ and OAD journals per discipline and field. Note that a journal can belong to more than one discipline; thus, percentages do not sum up to 100%. Only disciplines that included at least one journal are shown

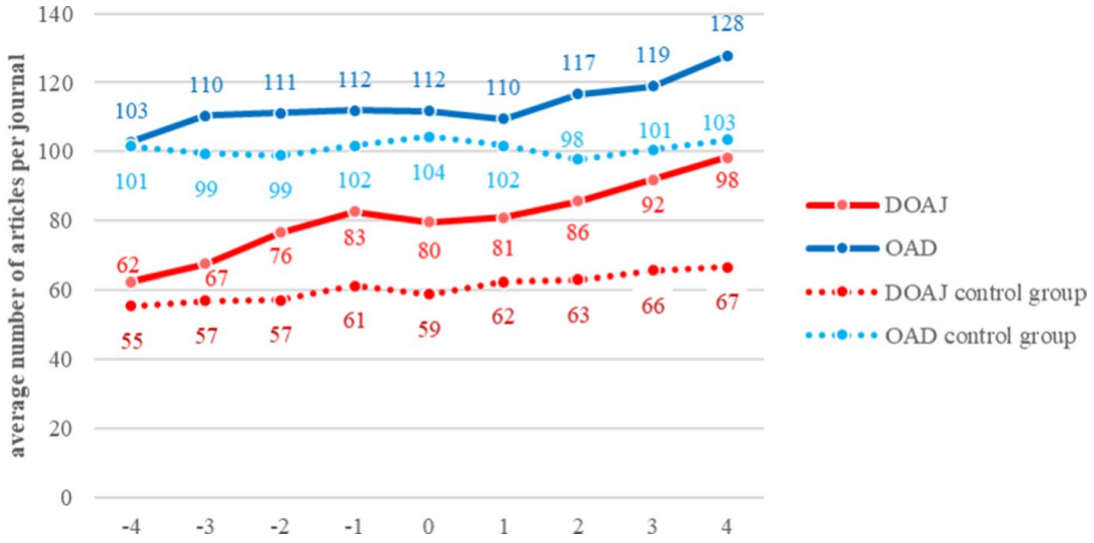

Fig. 3 Yearly average number of articles for flipped journals four years before and after flipping. The x-axis refers to the year with respect to the flipping year: 0 represents the year of flipping, positive values the years following the flip, and negative values the years preceding the flip

Due to the journal review time, e.g. when a journal flips, newly submitted articles will take several months to proceed through the review process. Therefore for the OAD-group (in the case of having the month of flipping) we considered the following year as the flipping point for journals which were flipped in the fourth quarter to ensure that articles reflect the OA model under which they were submitted. Figure 1 shows the distribution of years

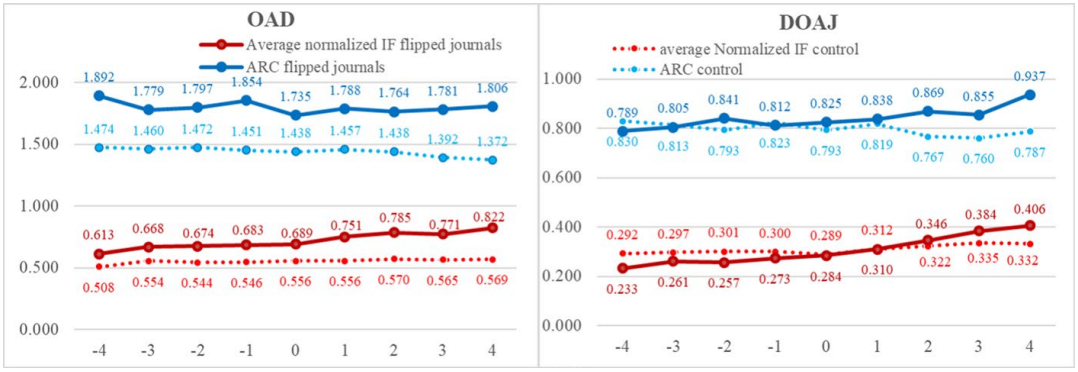

Fig. 4 Mean normalised IF (blue lines) and ARC (red lines) of OAD (left) and DOAJ (right) journals considered in this study. Solid lines represent the group of flipped journals, dotted lines the respective control group. (Color figure online)

in which journals in the two datasets flipped. We observe a peak in the number of flipped journals in 2006, as well as a long-term steady increase in the number of journals that have converted to OA across all years. The peak in 2006 for OAD is caused by a large number of journal conversions carried out by two major publishers: *CSIC Consejo Superior de Investigaciones Cientificas* and *Hindawi*. The peak for the DOAJ journals in 2013 has different publishers and is not dominated by specific publishers. Figure 2 shows the distribution of flipped journals across fields and disciplines. The majority of journals are categorised into the disciplines 'Medicine', 'Social Sciences', 'Agricultural and Biological Sciences' and 'Arts & Humanities'. However, the two groups differ with regard to disciplines included: the OAD-group seems to include a considerably higher amount of journals from 'Arts & Humanities' and 'Mathematics' than the DOAJ-group.

Journal publishing volumes

We assessed the evolution of publishing volumes for journals that flipped from a CA to an OA model, for 4 years prior to and 4 years following the year of the flipping (see Fig. 3). For each group of flipped journals (i.e. DOAJ or OAD) we calculated the mean number of articles published per journal per year and compared these numbers to the control group. Independent from the flipping process, we can observe a general difference between the two groups of flipped journals and their respective control groups regarding the number of published articles: the OAD-group of flipped journals publishes more articles on average than the control group, whereas the DOAJ-group of flipped journals publishes considerably fewer articles on average than the control group.

The number of articles published in the flipping year ranged from 1 in *Journal Hungarian Geographical Bulletin*, to 3,807 in *Journal of Acta Crystallographica Section E*. In general, we observe a small but steady increase in the number of articles published by journals following the flip to an OA model, which continues for the entire 4-year period of our analysis. For DOAJ flipped journals, the mean number of published articles increased from 80 articles in the flipping year, to 98 articles 4 years after flipping, an increase of 22.5%. In contrast, the DOAJ control group only increased from 59 to 67 articles on average, an increase of 14%. For OAD flipped journals the number increased from 112 articles in the flipping year, to 128 articles 4 years after flipping, an increase of 14.3%, whilst the

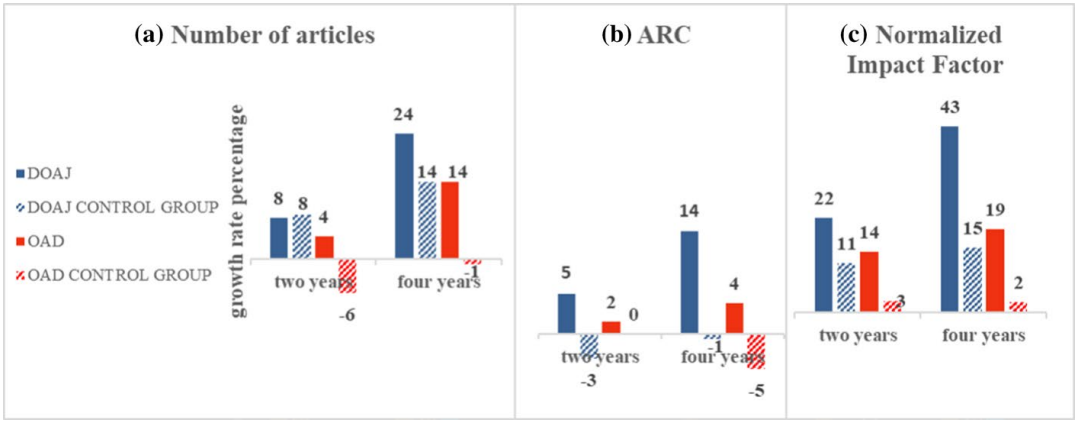

Fig. 5 Growth rates, calculated as the percentage growth between the flipping year and measurement year, for a) number of articles, b) ARC) and c) normalized impact factor. Growth rates were measured for both groups of flipped journals (DOAJ and OAD), as well as their respective control groups, at two and four years post-flipping

control group decreased from 104 to 103 articles in the same period, a decrease of 1%. Thus, for both groups of journals the mean increase in publishing volumes for flipped journals exceeded the increase in publishing volumes for journals that remained CA.

For OAD flipped journals, the post-flip increase in publishing volume appears to be insensitive to general long-term trends, as in all three years prior to flipping the number of published articles per journal remained relatively static at ~ 111 articles per year, and only began to increase prominently at 2 years post-flipping. For DOAJ, the interpretation is less clear – in general the number of published articles increased in the 4-year period prior to flipping, but the trend is characterised by a decline in the number of published articles in the year immediately preceding the flip.

Article and journal level citation metrics

Figure 4 shows the mean normalised IF (red line) and ARC (blue line) for journals and articles, respectively, in our dataset for the four years before and after flipping. The ranges of IF and ARC in the year of flipping are from 0 to ~ 5.55 and 0 to 60.8 respectively. The left panel shows the values for journals and articles within the OAD flipped journals (solid line) and respective control group (dashed line), and the right panel the same for DOAJ journals. To also more clearly demonstrate changes in ARC and normalized IF at specific time points following flipping, we also calculate growth rates of each metric at two and four years post-flipping, relative to the values in the flipping year (see Fig. 5). For OAD flipped journals, we observe no major difference in ARC before or after flipping, as values remain relatively stable in the range from ~ 1.8 to ~ 1.9. In terms of normalized IF, we observe a small increase for OAD journals, from 0.69 in the flipping year to 0.82 at 4 years after flipping (an increase of 19%). However, this increase is relatively small and not significantly greater than the interannual variability that we observe in either the OAD flipped journals or the respective control group.

In the DOAJ group, we observe clearer temporal trends in ARC and normalized IF which may, at least in part, be attributed to the flipping of the journal. ARC increases from 0.83 to 0.94 between the flipping year and 4 years after the flipping year (an increase of 14%), whilst normalized IF increases from 0.29 to 0.41(an increase of 43%). These values

Table 2 Two-year and four-year growth rates in numbers of published articles, normalized IF and ARC for CA and flipped journals, by scientific field. Numbers in the left column refer to the number of journals within each field, for DOAJ and OAD journals, respectively

#journals per field		growth rate (%) two years after flip				growth rate (%) four years after flip			
		DOAJ	DOAJ Control group	OAD	OAD Control group	DOAJ	DOAJ Control group	OAD	OAD Control group
Health Sciences DOAJ: 117 OAD: 40	#artic	10	9	7	1	22	13	15	9
	Av. IF	22	12	26	2	44	13	41	4
	ARC	6	−6	1	−3	14	−7	2	−10
Social Sciences DOAJ: 49 OAD: 25	#artic	−7	2	−3	2	43	4	7	-2
	Av. IF	36	22	0	15	46	31	8	5
	ARC	−4	7	−1	−14	−5	4	16	−11
Life Sciences DOAJ: 71 OAD: 23	#artic	4	18	−3	−7	16	36	13	−6
	Av. IF	16	2	15	−2	43	7	29	0
	ARC	8	−7	2	0	20	−4	3	−4
Physical Sciences DOAJ: 61 OAD: 24	#artic	9	3	−2	−14	24	11	8	−5
	Av. IF	29	12	−5	−2	48	15	0	2
	ARC	7	−7	5	5	11	0	3	3

are both higher than those observed for the control group, which decreased ARC by 1% and increased normalized IF by 15%.

Variability between scientific fields

The effect of flipping a journal to OA on ARC and IF may be manifested differently across different scientific fields. To investigate these possible differences, we additionally grouped journals by field and compared changes in ARC and normalized IF between the time of flipping, and two years and four years post-flipping. Results for each field are shown in Table 2.

In general, we observe a strong variability between the different bibliometric dimensions under study (i.e. number of articles published, normalised IF and ARC) and between each field, suggesting that the effect of flipping a journal differs strongly between different fields and included disciplines, respectively. For Health Sciences (117 journals in DOAJ and 40 journals in OAD), for example, growth rates of all dimensions were positive at two and four years after flipping for both sets of flipped journals, and higher than values observed in the control groups. Conversely, in the Social Sciences (49 journals in DOAJ, 25 journals in OAD), growth rates in the number of articles published are negative at two years after flipping, but become positive, and for DOAJ far greater than the growth rates of the control group, at four years after flipping, indicating that the effect of flipping, at least in terms of article volume, takes a longer time to diffuse in the Social Sciences.

Conclusion

We have presented one of few studies on journals which flipped from a CA to an OA model and its effect on journal publication volumes, article- and journal-level citations metrics and how these compare to journals which still pursue the CA model. The literature reporting studies on flipped journals shows that journals' IFs usually increase after flipping (Bautista-Puig et al., 2020; Busch, 2014a). Our results agree with these previous findings, but show that whilst IF and ARC increase generally in the years following flipping, they vary greatly across scientific fields. Previous studies found that the effect of the OA model on received citations is field specific (Björk & Solomon, 2012; Li et al., 2018). One reason for the higher advantage by some disciplines is probably the lack of available OA journals at the same quality level for those disciplines. Of course, this effect is accelerated by the general relation between the quality of articles and journals and received citations which is not discipline-dependent. For example, Gargouri et al. (2010) found a greater OA advantage for articles published in journals with higher impact factors. Moreover, the amount of charged APCs may be a factor influencing the number of citations. Björk and Solomon (2012) showed that the average number of citations for OA journals with an APC model is higher than for those without an APC model. Zhang et al. (2020) found that 'Life Sciences' charge the highest APCs followed by 'Health Sciences' and 'Physical Sciences', while the 'Social Sciences' charge the lowest APCs.

We also observe that a higher number of articles are published after flipping, pointing to either a higher tendency amongst authors to submit to OA journals, which complements the research by Rowley et al. (2017), or higher acceptance rates by journals (e.g. because of lack of space-wise restrictions for online-only publications). Here again we saw different trends across fields for both groups of flipped journals. The willingness to submit to OA journals with APCs is related to the amount of available funding (e.g. from institutions, universities, or governments). Zhang et al. (2020) reported that authors from the 'Social Sciences' show a lower willingness-to-pay for APCs because of less financial support whereas authors from 'Health and Life Sciences' are able and willing to spend more on OA publishing because of more financial resources available.

Our study as well as related work have shown that flipping to an OA publishing model can positively affect the number of published articles as well as journal and article citation indicators. However, journals that flip to OA are confronted with a complex net of interrelated factors that determine success or failure of the flipping procedure. More in-depth studies are needed to control for the various factors affecting journal success.

Limitations and future work

This study has a number of limitations, which can be built upon and improved in future work. Most importantly, the study has a relatively small sample size, with only 234 journals considered from DOAJ, and 87 journals from OAD. It is therefore not clear how representative this sample is of the total number of journals that have flipped from CA to OA models – but it is almost certainly not a complete list of the entirety of flipped journals. Thus, more advanced methods for identifying journals that have flipped from CA to OA models (e.g. by utilising data from large-scale aggregators of OA information such as Unpaywall or CORE) may help to generate a more complete picture in future.

However, it should be noted that Unpaywall data might be a source of error also affecting our study (regarding the construction of the control groups). Akbaritabar and Stahlschmidt (2019) have studied Unpaywall and showed that 13% of publications that Unpaywall classified as OA was classified as CA in Crossref. Here more work is needed to better determine the OA/CA-status of articles.

Another limitation is the lack of data on submissions to flipped journals which we assume to better reflect the willingness of authors to publish in an OA journal. The results of our analyses are only based on the number of accepted articles which may have also increased due to changes in editorial policies, amongst others.

Although we used article and journal level citation indicators to increase the precision of comparisons of groups of flipped journals with CA journals we didn't exclude outliers at article level (e.g. the merit or quality of individual articles which might draw attention and higher impact among the community) and journal level (impact factor) which could affect the measures. Future work will be more sensible to such issues (e.g. by removing outliers from analyses).

A number of additional factors are important to consider, when using bibliometric indicators to understand the development of a flipped journal over time. For example, it is important to consider the exact business model that is used by a flipped journal – some journals may use an APC-driven model, whilst others may be supported by individual societies or library presses. These different models bring different economic challenges, as highlighted by Matthias et al. (2019) who found that a large percentage of journals that flipped to an OA model eventually flipped back to a CA model, in part for monetary reasons. These economic pressures may also cause downstream changes on editorial decisions, not least because APC-driven revenues are closely tied to journal acceptance rates. A related factor is that of the publisher itself – in this study we considered changes at the journal level, but did not consider how a change in the publisher may also accompany a change in business model. Different publishers bring differences in platform quality and visibility, and these may also have an effect, for example, on the willingness of authors to publish their work with a journal. Bautista-Puig et al. (2020) also investigated how countries of publishing and citing authors change before and after a flip, which is important for understanding exactly who is supporting new OA models and what effect that may have on bibliometric indicators and publishing behaviour. In addition, long-term changes in the support of institutions and funders, as well as increasing pressure to transition to OA models will mean that the findings presented here will evolve over time. Future work should therefore focus on trying to understand the complicated interactions between these different factors.

It is important that quantitative bibliometrics, such as the results presented here, also involve the views of stakeholders such as publishers, funders, libraries and societies. Therefore, future work should also be complemented with more qualitative information from interviews with these stakeholders, to reveal their attitudes towards journal flipping and OA, their expectations regarding journal quality and indicators as well as their motivation to change the publication model.

Acknowledgement This work is supported by BMBF project OASE, grant number 01PU17005A. We are thankful to Masoud Davari for his assistance with preparing and validating the data. We thank Najko Jahn and the anonymous reviewers for helpful comments on the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Akbaritabar, A., & Stahlschmidt, S. (2019). Applying Crossref and Unpaywall information to identify gold, hidden gold, hybrid and delayed Open Access publications in the KB publication corpus. https://osf.io/sdzft/download

Archambault, É., Côté, G., Struck, B., & Voorons, M. (2016). Research impact of paywalled versus open access papers. https://digitalcommons.unl.edu/scholcom/29/

Bautista-Puig, N., Lopez-Illescas, C., de Moya-Anegon, F., Guerrero-Bote, V., & Moed, H. F. (2020). Do journals flipping to gold open access show an OA citation or publication advantage? *Scientometrics*. https://doi.org/10.1007/s11192-020-03546-x.

Björk, B. C. (2012). The hybrid model for open access publication of scholarly articles: A failed experiment? *JASIST, 63*(8), 1496–1504.

Björk, B. C., & Solomon, D. (2012). Open access versus subscription journals: a comparison of scientific impact. *BMC medicine, 10*(1), 1–10.

Björk, B.-C. (2017). Scholarly journal publishing in transition- from restricted to open access. *Electronic Markets, 27*(2), 101–109. https://doi.org/10.1007/s12525-017-0249-2.

Busch, S. (2014a). "The careers of converts – how a transfer to BioMed Central affects the Impact Factors of established journals" BioMed Central (http://blogs.biomedcentral.com/bmcblog/2014/01/15/the-careers-of-converts-how-a-transfer-to-biomed-central-affects-the-impact-factors-of-established-journals/).

Busch, S. (2014b). "The Impact Factor of Journals Converting from Subscription to Open Access." BioMed Central (https://blogs.biomedcentral.com/bmcblog/2014/11/06/the-impact-factor-of-journals-converting-from-subscription-to-open-access/).

Davis, P. M. (2011). Open access, readership, citations: a randomized controlled trial of scientific journal publishing. *The FASEB Journal, 25*(7), 2129–2134.

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., & Harnad, S. (2010). Self-selected or mandated, open access increases citation impact for higher quality research. *PLoS ONE, 5*(10), e13636.

Hobert, A., Jahn, N., Mayr, P., Schmidt, B., & Taubert, N. (2020). Open Access Uptake in Germany Adoption in a diverse research landscape. https://doi.org/10.5281/zenodo.3892951

Laakso, M., Welling, P., Bukvova, H., Nyman, L., Björk, B.-C., & Hedlund, T. (2011). The Development of Open Access Journal Publishing from 1993 to 2009. *PLoS ONE, 6*(6), e20961. https://doi.org/10.1371/journal.pone.0020961.

Lewis, C. L. (2018). The Open Access Citation Advantage: Does It Exist and What Does It Mean for Libraries? *Information Technology and Libraries, 37*(3), 50–65. https://doi.org/10.6017/ital.v37i3.10604.

Li, Y., Wu, C., Yan, E., & Li, K. (2018). Will open access increase journal CiteScores? An empirical investigation over multiple disciplines. *PLoS ONE, 13*(8), e0201885.

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics, 12*(3), 819–841. https://doi.org/10.1016/j.joi.2018.06.012.

Matthias, L., Jahn, N., & Laakso, M. (2019). The Two-Way Street of Open Access Journal Publishing: Flip It and Reverse It. *Publications, 7*(2), 23.

McKiernan, E. C., et al. (2016). How open science helps researchers succeed. *eLife*. https://doi.org/10.7554/eLife.16800.001.

Momeni, F., Mayr, P., Fraser, N., & Peters, I. (2019). From closed to open access: A case study of flipped journals. Proceedings of the 17th International Conference on Scientometrics & Informetrics (ISSI 2019), 1270–1275.

Ottaviani, J. (2016). The post-embargo open access citation advantage: it exists (probably), it's modest (usually), and the rich get richer (of course). *PLoS ONE, 11*(8), e0159614.

Perianes-Rodríguez, A., & Olmeda-Gómez, C. (2019). Effects of journal choice on the visibility of scientific publications: A comparison between subscription-based and full Open Access models. *Scientometrics, 121*(3), 1737–1752. https://doi.org/10.1007/s11192-019-03265-y.

Peterson, A. T., Emmett, A., & Greenberg, M. L. (2013). Open Access and the Author-Pays Problem: Assuring Access for Readers and Authors in the Global Academic Community. *Journal of Librarianship and Scholarly Communication, 1*(3), 1064. https://doi.org/10.7710/2162-3309.1064.

Piwowar, H., et al. (2018). The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ, 6*, e4375. https://doi.org/10.7717/peerj.4375.

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ, 1*, e175. https://doi.org/10.7717/peerj.175.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences, 105*(45), 17268–17272.

Rowley, J., Johnson, F., Sbaffi, L., Frass, W., & Devine, E. (2017). Academics' behaviors and attitudes towards open access publishing in scholarly journals. *Journal of the Association for Information Science and Technology, 68*(5), 1201–1211.

Schimmer, R., Geschuhn, K. K., & Vogler, A. (2015). Disrupting the subscription journals' business model for the necessary large-scale transformation to open access. Technical report. https://doi.org/10.17617/1.3

Solomon, D. J., Laakso, M., & Björk, B.-C. (2013). A longitudinal comparison of citation rates and growth among open access journals. *Journal of Informetrics, 7*(3), 642–650. https://doi.org/10.1016/j.joi.2013.03.008.

Solomon, D. J., Laakso, M., & Björk, B. C. (2016) Converting scholarly journals to open access: a review of approaches and experiences [Internet]. tTechnical report. Harvard Library Office for Scholarly Communication.

Sotudeh, H., & Horri, A. (2007). The citation performance of open access journals: A disciplinary investigation of citation distribution models. *Journal of the American Society for Information Science and Technology, 58*(13), 2145–2156.

Sotudeh, H., Ghasempour, Z., & Yaghtin, M. (2015). The citation advantage of author-pays model: the case of Springer and Elsevier OA journals. *Scientometrics, 104*(2), 581–608.

Swan, A. (2010). The Open Access citation advantage: Studies and results to date. https://doi.org/10.1002/leap.1056.

Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics, 107*(3), 1195–1225. https://doi.org/10.1007/s11192-016-1889-2.

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., & Hartgerink, C. H. (2016). The academic, economic and societal impacts of Open Access an evidence-based review. *F1000Research, 5*, 632.

Waltman, L., Larivière, V., Milojević, S., & Sugimoto, C. R. (2020). Opening science: The rebirth of a scholarly journal. *Quantitative Science Studies, 1*(1), 1–3. https://doi.org/10.1162/qss_e_00025.

Zhang, X., Grebel, T., & Budzinski, O. (2020). The prices of open access publishing: The composition of APC across different fields of sciences (No. 145). Ilmenau Economics Discussion Papers. http://hdl.handle.net/10419/225259

4.2.2 Factors Associated with Open Access Publishing

Title: Which Factors are Associated with Open Access Publishing? A Springer Nature Case Study

Authors: Fakhri Momeni, Stefan Dietze, Philipp Mayr, Kristin Biesenbender, and Isabella Peters

Document Type: Journal paper

Venue: Quantitative Science Studies

DOI: https://doi.org/10.1162/qss_a_00253

***

Momeni, F., Dietze, S., Mayr, P., Biesenbender, K., and Peters, I. (2023). Which Factors are associated with Open Access Publishing? A Springer Nature Case Study. Quantitative Science Studies. Advance Publication. https://doi.org/10.1162/qss_a_00253

Which Factors are associated with Open Access Publishing? A Springer Nature Case Study

Fakhri Momeni[1*], Stefan Dietze[1,3], Philipp Mayr[1], Kristin Biesenbender[2] and Isabella Peters[2]

[1*]KTS, GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne, 50667, Germany.
[2*]Web Science, ZBW – Leibniz Information Centre for Economics, Düsternbrooker Weg 120, Kiel, 24105, Germany.
[3*]Computer Sciences, Heinrich-Heine-University Düsseldorf, Universitätsstr, Düsseldorf, 40225, Germany.

*Corresponding author(s). E-mail(s): Fakhri.Momeni@gesis.org;
Contributing authors: Stefan.Dietze@gesis.org;
Philipp.Mayr@gesis.org; k.biesenbender@zbw.eu; i.peters@zbw.eu;

Abstract

Open Access (OA) facilitates access to articles. But, authors or funders often must pay the publishing costs preventing authors who do not receive financial support from participating in OA publishing and citation advantage for OA articles. OA may exacerbate existing inequalities in the publication system rather than overcome them. To investigate this, we studied 522,411 articles published by Springer Nature. Employing correlation and regression analyses, we describe the relationship between authors affiliated with countries from different income levels, their choice of publishing model, and the citation impact of their papers. A machine learning classification method helped us to explore the importance of different features in predicting the publishing model. The results show that authors eligible for APC waivers publish more in gold-OA journals than others. In contrast, authors eligible for an APC discount have the lowest ratio of OA publications, leading to the assumption that this discount insufficiently motivates authors to publish in gold-OA journals. We found a strong correlation between the journal rank and the publishing model in gold-OA journals, whereas the OA

option is mostly avoided in hybrid journals. Also, results show that the countries' income level, seniority, and experience with OA publications are the most predictive factors for OA publishing in hybrid journals.

Keywords: APC policies, bibliometrics, open access, citation impact, machine learning

1 Introduction

The unrestricted availability of Open Access (OA) publications is linked to the goal of granting all interested parties free access to scientific knowledge and ensuring greater equality of access (Munafò et al., 2017). This view is strongly related to the consumers of scholarly knowledge, who then would not have to pay for access. However, when taking the authors of those articles into account, they are affected by OA in two different ways: a) when choosing a publication model for an article and b) when receiving citations (and along with its reputation) for articles that have been published via a certain model (usually described as citation advantage, see e.g., Langham-Putrow, Bakker, and Riegelman (2021)). Those two aspects of OA may introduce significant biases and inequity into the scholarly publication and reputation system since they may restrict participation in OA in particular ways (Bahlai et al., 2019).

First, the OA publishing model generally shifts the publishing costs from readers to authors or their institutions and funders by introducing article processing charges (APCs). This can be a severe constraint for those authors who cannot afford these costs or do not receive any financial support. To overcome this issue, most publishers have implemented an APC waiver/discount policy for authors from, e.g., low-income countries (Lawson, 2015). However, it is open how the different options for OA publishing and waivers/discounts are considered and adopted by researchers with various characteristics such as their countries' income level, but also their seniority and gender – factors which are also often associated with the decision to publish OA (Iyandemye & Thomas, 2019; Olejniczak & Wilson, 2020; Simard, Ghiasi, Mongeon, & Larivière, 2021; Smith et al., 2022; Zhu, 2017). Rouhi, Beard, and Brundy (2022) discussed the waiver issues from the perspectives of the publisher, institutions, and developing countries. They mentioned the potential unfairness authors are confronted with, which may be caused by APC-based models. They argued that waiver programs have yet to address this problem successfully. They suggested that meeting the equity standard requires a cross-functional approach involving publishers, funders, research institutions, individual researchers, libraries, and service providers.

To accommodate OA publishing costs, three funding options have emerged over time. First, Diamond OA journals are funded by public institutions such as libraries, which enable free reading and publishing for all researchers. Second, transformative agreements between public institutions and publishers have

been introduced that include reading and publishing contracts and which are also funded by the institutions. In this case, there are no direct fees for authors, but their institutions pay for the APCs as part of a consortium. Access to publishing and access to publications is limited to participating organizations only. Thirdly, APCs could also be paid by the authors or their institutions themselves. The first option leads to Gold OA at the journal level. Transformative agreements allow authors to publish in either gold OA or hybrid (which – for a fee – allow publishing individual articles as an OA-variant) journals. The third option is often associated with hybrid journals. All other publishing models for journals usually require funding via subscriptions, resulting in closed-access articles (CA) that can only be read after paying the article or journal fee.

The publishing model is also strongly associated with the visibility of authors and articles. For many researchers, it makes a difference where, i.e., in which journals they publish (e.g., considering discipline-specific journal rankings). If they want to be noticed by others and/or seek promotion, it can be crucial to publish in reputable journals, especially for early career researchers. And to achieve this, not only financial hurdles and APCs have to be overcome, but, for example, English language skills and technical skills are needed, as well as institutions that can help with legal advice or infrastructure support. Against this background, researchers have to decide which publishing model to choose and whether OA is not only an altruistic but feasible option at all.

The second possible source of bias and inequity is related to the paying for access case: It has been shown already that articles published as OA-variants are more visible, leading to higher citation counts and altmetrics (Evans & Reimer, 2009; Fraser, Momeni, Mayr, & Peters, 2020; Lewis, 2018; McKiernan et al., 2016; Ottaviani, 2016). Moreover, the Matthew effect shows that researchers who are already well-known and widely cited receive even more citations (Farys & Wolbring, 2021) – which directly affects rewards for publication in prestigious journals, for prominence, and citations. For researchers, publications play a central role in their daily practice and the reputation system in which they operate. Publications enable researchers to build on the body of knowledge and refer to those findings by citing the publications (which accumulate reputation in this way). Hence, access to publications is crucial for the progress of science and building of reputation – which both can be impeded by a lack of access to OA publishing options and the risk of CA-articles not being cited as frequently as OA articles.

From that, we hypothesize that researchers with better access to financial resources have better access to publications – both in terms of access to read openly and in terms of access to publish openly. Associated with that may be an even stronger citation advantage for those researchers (usually WEIRD: Western, educated, industrialized, rich, and democratic; (Henrich, Heine, & Norenzayan, 2010)) with extensive OA-publishing options. As such, OA may carry the risk of perpetuating already existing inequalities rather than resolving such marginalization in the scholarly communication system (Fox et al., 2021).

2 Related work

Related work also indicates a strong association between economic factors, OA, and citation advantages. The scientific output of countries is associated with their economic evolution because scientific progress needs governments' financial support. Samimi (2011) used a Granger Causality Test to examine the causal relationship between scientific output and GDP in 176 countries and found a two-way positive relationship between them. King (2004) compared published papers and their citation impacts across countries and found that only 31 countries contributed to 98% of the world's highly cited papers and that the remaining 161 countries contributed less than 2%.

Open Access publishing is also highly influenced by the authors' country of affiliation since it determines APC waiver/discount policies or the availability of transformative agreements with publishers. Some publishers offer general waivers or have a discount policy for all of their journals for eligible authors, and the country's income level mainly determines eligibility. Lawson (2015) has studied the waiver policy of the 32 most prominent publishers and found that 68% of them grant APC waivers. Simard et al. (2021) found that low-income countries publish and cite OA more than upper-middle and high-income countries. The positive correlation between OA citing and publishing is 1.3 times weaker for high-income countries than other countries. Similarly, Iyandemye and Thomas (2019) showed that biomedicine researchers from low-income countries have the highest percentage in OA publishing. Smith et al. (2022) reported the proportionately fewer OA articles published in Elsevier's journals for low-income countries, despite their eligibility for APC waivers.

Olejniczak and Wilson (2020) studied the articles published by faculty members at research universities in the United States and found that in the United States, male and senior authors are more likely to publish in OA form. Zhu (2017) conducted a survey with over 1800 researchers at 12 Russell Group universities[1] to find the differences in OA publishing regarding discipline, seniority, and gender. Their results revealed disciplinary differences in OA publishing (Medical and Life Scientists are most likely to publish in Gold OA journals), more tendency toward OA publishing for senior authors, and across genders for men.

The journal rank is a decisive factor in submitting the article in addition to its business model. Schroter, Tite, and Smith (2005) conducted a survey study with 28 international authors who submitted to the BMJ and found that for authors, the journal's ranking is more important than the availability of OA.

Many studies have investigated the OA citation outcome, and most found a citation advantage for OA articles (Evans & Reimer, 2009; Fraser et al., 2020; Lewis, 2018; McKiernan et al., 2016; Ottaviani, 2016). However, regarding biases (e.g., quality bias, self-selecting, mandating, self-archiving), different sampling and controlling data makes it difficult to conclude that receiving more citations is only the effect of OA. Momeni, Mayr, Fraser, and Peters (2021)

[1] https://russellgroup.ac.uk/about/our-universities/

studied the citation impact of flipping journals from CA to OA and generally found a slightly higher growth in receiving citations compared to journals in the same discipline and the impact factor's range. However, they didn't observe this trend in all scientific fields. Momeni, Mayr, and Dietze (2022) examined the correlation between different factors and the future authors' h-index and found a positive but weak correlation coefficient between them.

One issue which is often discussed together with OA publishing and APCs is the problem of predatory publishing. Predatory publishers take advantage of the OA movement but work against the good scientific practice. Ross-Hellauer et al. (2021) did a systematic review to study the threat to equity in science via open science implementations. They concluded that less well-resourced researchers, researchers from non-English-speaking countries, and early-career researchers are particularly affected by the 'predatory publishing' problem.

3 Research questions

We conduct our study on the association between publishing models, the economic background of researchers, and other author-specific and structural factors along three major research questions:

RQ1: What is the relationship between the income level of researchers' affiliation countries and their publication behavior (do they prefer OA or CA)?

RQ2: What is the relationship between the income level of researchers' affiliation countries and their publication behavior (OA or CA) with their citation impact?

To answer these questions, we categorize corresponding authors based on the income level of their affiliation country and compare the access status of articles they have published and their citation impact. Whereas the first two RQs are rather descriptive and aim at quantifying the extent to which access to publish openly and access to read openly (and along with it to make them easier/more likely to cite) are related to the economic background of authors, the third RQ takes a variety of factors into account that have been shown to be strongly associated with tendencies to publish OA (Iyandemye & Thomas, 2019; Olejniczak & Wilson, 2020; Simard et al., 2021; Smith et al., 2022; Zhu, 2017).

RQ3: What factors (e.g., journals, articles, authors, or their countries) are associated with selecting the business model of publications (OA against CA)?

Here we aim to give a detailed view of associating factors with OA publishing using correlation, regression, and machine learning analyses. To this end, structural features, such as APC waivers, are considered besides author-specific properties, such as gender or years of publishing activity (see Table 2). We will also look closely at the different access forms to publications such as Gold OA, Hybrid, and Closed Access. Concerning the level of journals, the relationships between journal rankings, APCs, and research fields (Health Sciences, Life Sciences, Physical Sciences, Social Sciences, and multiple fields) will be examined. In addition, possible country-related influencing factors will be

investigated, such as countries' income level, transformation agreements' existence, or opportunities for researchers to obtain APC discounts or waivers. At the journal article level, the ratio of OA to CA citations in an article and the number of authors involved are examined. Other author-specific influencing factors can be gender and age, the ratio of OA to CA publications in the past, or even the proportion of international co-authors.

4 Data and methodology

To conduct our study, information on the business model, author characteristics, and article impact are needed, and several approaches and databases must be linked to receive a complete dataset.

4.1 Data selection

For the business model of journals (OA, Hybrid, CA) it is possible to crawl the information from the journal's or publisher's website or to look up sources such as the Directory of Open Access Journals (DOAJ) and Unpaywall, which both include OA information. But information about the history of the business model of journals is rarely available. In recent years, many journals have converted (flipped) from closed access to open access and vice versa, but often there is not enough information about the exact date of starting with the new access model. The Open Access Directory (OAD), a wiki hosted by the School of Library and Information Science at Simmons University[2], is the only resource containing a list of a few flipped journals and the date of flipping. The open-access start date of journals was available in the DOAJ dataset until 2020. Bautista-Puig, Lopez-Illescas, de Moya-Anegon, Guerrero-Bote, and Moed (2020) and Momeni et al. (2021) used OAD and DOAJ for their studies about flipping journals. Unfortunately, DOAJ stopped collecting that information by now: "As time progressed, open access models became more complicated ... It has become harder to find the right answer to that seemingly simple question: when did open access start for this journal?"[3]. Matthias, Jahn, and Laakso (2019) employed different snapshots of datasets that have the open access status (Scopus, DOAJ, Ulrichsweb, publishers' website, etc.) and some other resources to find out the reverse flip (converting from OA back to CA) and verified them manually. For the bibliometric analyses related to open access, it is necessary to know about the access status of journals for the period in which we study the effect of OA. Obtaining information more coherently requires looking into different journals' business models and harmonizing them to make them comparable. In addition, every publisher has its own rules for APC exemptions to foster publishing in OA format. For example, eligibility for APC waivers for publishing in Elsevier's journals is based on the

[2] http://oad.simmons.edu/oadwiki/Main_Page

[3] https://blog.doaj.org/2021/02/05/why-did-we-stop-collecting-and-showing-the-open-access-start-date-for-journals/

'Research4Life program'[4] and for Springer Nature based on 'World bank classification'. Various transformative agreements with publishers and the period of their contracts are other influential factors that should be considered in studying the publishing behavior of each publisher separately.

Due to these varying APC-related rules for different publishers, we focused on one major publisher. To analyse papers for various disciplines and countries, we chose Springer Nature, the largest publisher of academic journals (more than 2,900 journals[5]) with worldwide authors from various disciplines, which provides us with a large amount of data and data diversity for more accurate results. Also, compared to Elsevier, the second most prominent publisher of scholarly journals (above 2,700 journals [6]), this publisher has a higher OA update (Sotudeh, Ghasempour, & Yaghtin, 2015; Sullo, 2016), resulting in fewer data skewness.

We downloaded the list of journals and their access status from the snapshot from the year 2019 which is available on the publisher's website[7]. Three publishing models exist for these Springer Nature (SN) journals: Gold Open Access, Hybrid (with the open access option: Open Choice), and Closed Access. Figure 1 displays the distribution of journals and their publishing models.

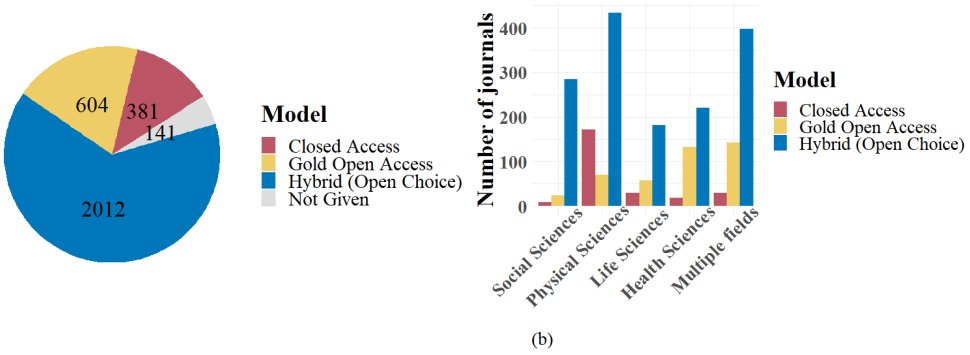

Fig. 1 Distribution of Springer Nature's journals by (a) publishing model and (b) field and publishing model.

For the bibliometric analyses, we employed Scopus[8]. We matched the list of SN journals with journals in Scopus via title and ISSN. From 3,138 SN journals, we could match 2,757 journals, which we used for further analyses. Because of the problems regarding journals' flipping mentioned above, we limited our data to two years, 2017 and 2018, to reduce the errors related to detecting the journals' and articles' business model. It resulted in 522,411 articles.

[4] https://www.research4life.org/access/eligibility/
[5] https://www.springernature.com/gp/librarians/products/journals/springer-journals
[6] https://www.elsevier.com/about/this-is-elsevier
[7] https://www.springernature.com/gp/open-research/journals-books/journals
[8] The in-house Scopus database maintained by the German Competence Centre for Bibliometrics (Scopus-KB), 2021 version

To detect the publishing model of articles in hybrid journals, we employed Unpaywall[9] (the snapshot of 2019), a service to find the available version of articles. We can obtain the publishing model of articles in hybrid journals from metadata in this dataset.

We obtained the APC amount in dollars for 1,741 hybrid journals and 297 gold OA journals from the website of Springer Nature[10]. There was no fixed APC for 147 gold OA journals (only 5% of investigated articles belong to these journals), and we had to visit their website to obtain the exact amount for these journals. Therefore we replaced the APC amount for these journals with null values (empty) and excluded them from the data for the classification task.

To detect the gender status of authors, we utilized a combined name and image-based approach introduced by Karimi, Wagner, Lemmerich, Jadidi, and Strohmaier (2016), which categorizes the gender into male and female. Based on this method, we tried detecting gender using the API *Genderize.io* [11]. For those names that the API couldn't identify the gender of, we looked for names on the web. We detected their gender using image-based recognition algorithms, which increases the recall and accuracy compared to *Genderize.io* (Karimi et al., 2016). We acknowledge that the person's gender is not a binary variable. Considering the social dimensions, more gender identities could not be identified with this approach, and that is left out for the analysis. Using Scopus author ID, we found 381,074 unique corresponding authors for the investigated articles, and 10,614 authors (about 3%) had only initials or no first name, and we couldn't detect their gender.

Overall, we identified the gender status for 49% of them. Therefore, we excluded 254,044 articles (about 49%) that we couldn't detect the gender status of their corresponding author from data in the regression analysis and classification task. One possible reason for a low rate of identifying gender is the large percentage of authors affiliated with Asian countries (136,591 above 35%)[12] and probably originally from these countries. Previous studies tested gender detection tools for authors with different nationalities and found them less effective for Asian names (Karimi et al., 2016; Santamaría & Mihaljević, 2018). Table 1 shows the number and percentage of OA and CA publications belonging to the corresponding authors with a gender status across scientific fields. The percentage of detected gender of authors for OA publications is 4% more than for CA publications.

4.2 Features and definitions

To investigate the factors that are associated with higher rates of OA publishing, we defined some features presented in Table 2. Figure 3 presents an

[9] https://unpaywall.org/
[10] https://www.springernature.com/de/open-research/journals-books/journals
[11] https://genderize.io/
[12] Authors from Armenia, Azerbaijan, Georgia, Kazakhstan, Russia, and Turkey, which belong to both Asia and Europe, are not included in this list.

Table 1 Number and proportion of articles among scientific fields and publishing model that we detected the gender status of their corresponding author.

	Publishing Model	
	CA model (percentage)	**OA model (percentage)**
Health Sciences	31,642 (53%)	20,534 (49%)
Life Sciences	23,011 (54%)	10,032 (57%)
Physical Sciences	74,742 (48%)	9,927 (50%)
Social Sciences	9,210 (40%)	2,020 (41%)
Multiple fields	38,507 (52%)	48,742 (58%)
Total	177,112 (50%)	91,255 (54%)

overview of data collection and preparation steps. The final analysed data is available on Git repository [13].

To compare the publishing and citation behavior across countries, we classified countries by income based on the World Bank classification[14] into four groups: low, lower-middle, upper-middle and high-income economies. The income level of a country has been evaluated every year and its history is available[15]. From 218 listed countries by the World Bank, we excluded 20 countries with different income levels from 2015 to 2018. Springer Nature offers APC waiver and discount to those articles with the corresponding author from low and lower-middle-income countries (classified by the World Bank), respectively[16].

From the website *Transformative Agreement Registry* provided by ESAC[17] we found three organizations with an open access agreement with this publisher during the investigated years 2017 and 2018 (KEMOE/FWF in Austria, Max Planck Society in Germany and Bibsam consortium in Sweden) and two organizations (VSNU-UKB in Netherlands and FinELib consortium in Finland) in 2018. We obtained the list of involved institutions in the agreement by asking KEMOE/FWF, Bibsam, and FinELib organizations. The list of participating institutions via VSNU-UK was available on the website of SN [18]. We assumed that the publications with the corresponding author affiliated with institutions included in the transformative agreement are free of APC charges. To find Max Planck institutions, we used disambiguated institutional addresses for German institutions (Rimmert, Schwechheimer, & Winterhager, 2017) available on Scopus-KB. We manually looked up the participating institutions for the rest of the four countries. We found 12,323 articles and used them to set the feature 'OA_agreement' value.

Figure 2 represents the number of articles published in Springer Nature where their corresponding author is affiliated with a country with the respective income group. Sixty-seven articles had a corresponding author with

[13] https://github.com/momenifi/open_access_springer_nature
[14] https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups
[15] http://databank.worldbank.org/data/download/site-content/OGHIST.xlsx
[16] https://www.springernature.com/gp/open-research/policies/journal-policies/apc-waiver-countries
[17] https://esac-initiative.org/about/transformative-agreements/agreement-registry/
[18] https://resource-cms.springernature.com/springer-cms/rest/v1/content/19371608/data/v3

multiple affiliation countries and we excluded them from the analyses. Publication distribution by countries and their income level is available on GitHub[19].

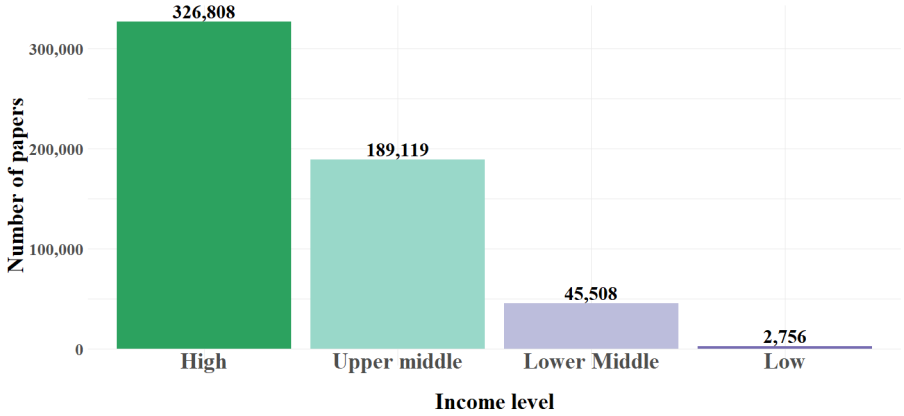

Fig. 2 Number of papers published by Springer Nature grouped by income level of countries.

We needed to identify authors and their publications to obtain the ratio of authors' previous OA publications. Scopus Author Id enabled us to get each author's published article list. For the variable Country_income, we consider average GDP per capita in 2017 and 2018 obtained from the world bank group[20]. We used the year of the first publication of authors indexed in Scopus to calculate their career age as a measurement of seniority.

To evaluate and rank the quality of journals, we employed the journal's H-index, which Hodge and Lacasse (2011) suggested as a better measurement for ranking journals than the 5-year impact factor in social science and that has been used in previous studies (Barner, Holosko, & Thyer, 2014; Xia, 2012). We calculated the H-index of all journals in Scopus classified in 27 subject categories[21] within the years 2011 and 2016.

4.3 Methodology

4.3.1 Normalizing the citation impact

To evaluate and compare the citation impact at the article and journal level among different subject areas, we should normalize them because of varying citation patterns across scientific disciplines and fields. To normalize the journal's H-index across categories, we computed the Percentile Rank (PR) of each journal (inspired by Bornmann and Mutz (2014)) in its category. This method gives the journals within a category a rank between 0 (lowest H-index) to 100 (highest H-index). In this approach, journals with the same H-index have the

[19] https://github.com/momenifi/open_access_springer_nature/blob/main/publications_country_distribution.csv
[20] https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
[21] https://service.elsevier.com/app/answers/detail/a_id/14882/supporthub/scopus/related/1/

Table 2 Features used to study the associated factors with OA publishing.

Feature type	**Feature**	**Description**
Journal	journal_ranking	H-index ranking of the journal in the related discipline (for multidisciplinary journals, the average ranking among disciplines).
	journal_APC	The cost of APC to publish OA in the journal (US-Dollar).
	field	Field of journal (If the journal has more than one field, the value is *'multiple fields'*).
	Health Sciences	
	Life Sciences	
	Physical Sciences	
	Social Sciences	
	multiple fields	
Country	country_income	Income level (GDP per capita) of the country in which the corresponding author is affiliated.
	OA_agreement	If the corresponding author's country of affiliation has an OA agreement with the publisher, it equals 1, otherwise 0.
	discount_eligible	If the corresponding author's country of affiliation belongs to the lower-middle income group, it equals 1, otherwise 0.
	waiver_eligible	If the corresponding author's country of affiliation belongs to the low-income group, it equals 1, otherwise 0.
Paper	OA_cite	ratio of citing OA against CA in this paper
	authors_count	number of authors
Author*	gender	for females equals 1 and for male 0.
	age	years since first publication
	OA_publish	ratio of OA publications against CA in the past (number of previous OA publications divided by the number of CA publications)
	international_coauthors	proportion of international co-authors** to all co-authors in this paper

* Corresponding author

** An international co-author is a co-author who has a different affiliation country than the corresponding author.

same rank. Therefore, this normalization method is an advantage in case of skewed distributions. If the journal belongs to more than one category, we used the weighted PR (Bornmann & Williams, 2020). Based on this approach, weighted PR (wPR) will be calculated using the formula:

$$wPR = \frac{PR_{sc1} * n_{sc1} + PR_{sc2} * n_{sc2} + ... + PR_{sci} * n_{sci}}{n_{sc1} + n_{sc2} + ... + n_{sci}} \quad (1)$$

whereby, *sci* is the ith subject category that the journal belongs to and n_{sci} is the number of journals in this subject category, and PR_{sci} is PR of the journal in it.

We employed a similar normalizing approach to present the citation impact of articles. Because the citation count is confounded by time since publication, we consider the citations during a time window of two years since the publication, as in previous studies (Jannot, Agoritsas, Gayet-Ageron, & Perneger, 2013; Piwowar et al., 2018). Next, we categorized the articles into groups with the same subject category and publishing year and ranked them from 0 to

***

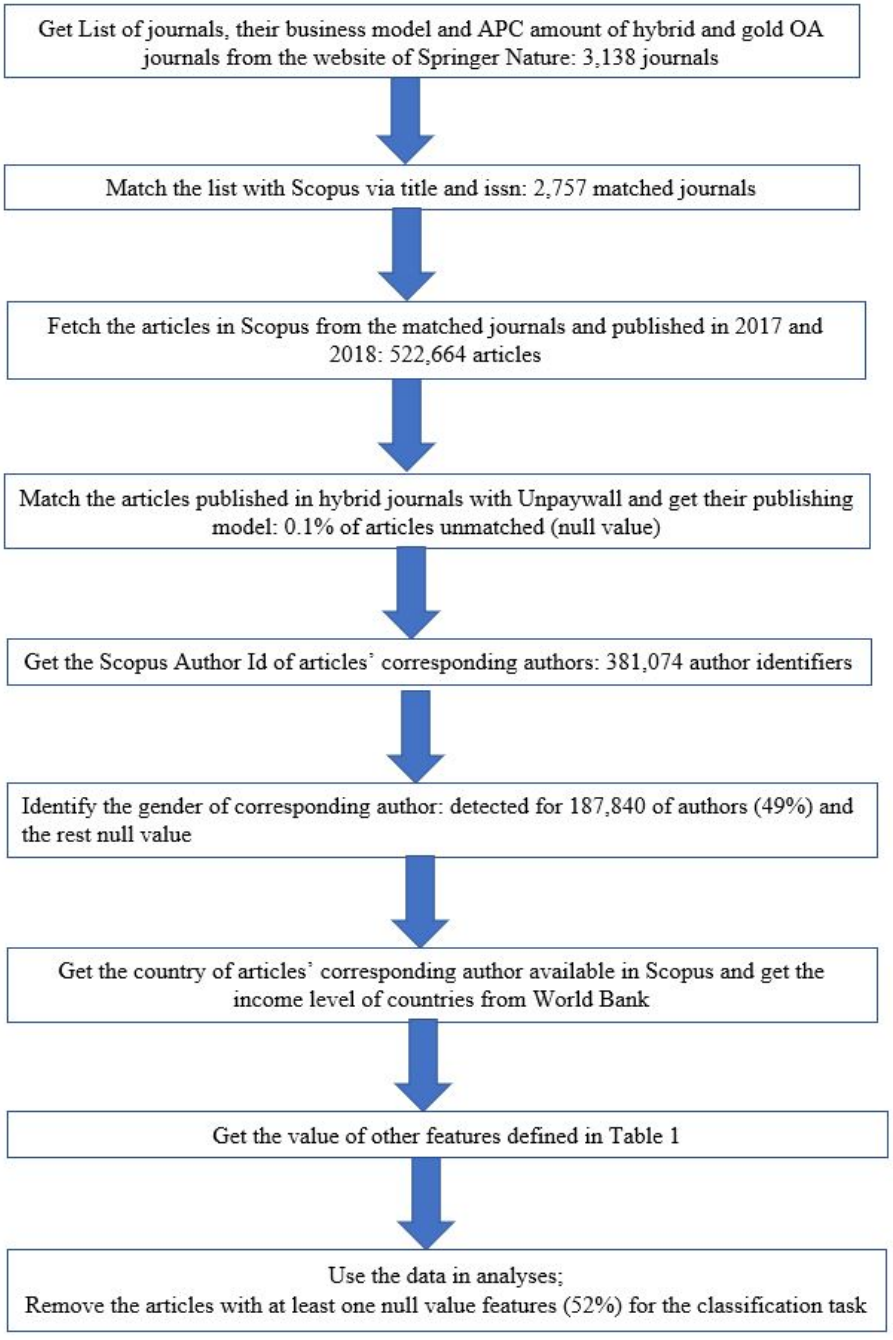

Fig. 3 Flow chart of data collection and preparation process.

100 based on received citations. We define a percentile rank of 50 (citation's median) as a threshold for highly cited articles. An article is highly cited if its rank is above 50% of PR in its group, meaning that it has received more citations than half of the articles in the same subject category and publishing

year. For articles belonging to multiple subject categories, we used wPR mentioned in Equation 1, where *sci* is the ith subject category of the article and n_{sci} is the number of articles in this subject category, and PR_{sci} is PR of the article in it.

4.3.2 Correlation analysis

To find the association between OA publishing and any feature defined in Table 2 we conducted a correlation analysis. The first variable in calculating the correlation is OA publishing, a dichotomous variable (a case of categorical variable). To assess the association with *field*, which is a categorical variable, we selected *Cramer's V* coefficient. Cramer's V is based on the chi-squared test and measures the strength of association between two variables. Its value ranges from 0 (no association) to 1 (complete association). The association with binary variables (OA_agreement, discount_eligible, waiver_eligible, gender) was examined with *Phi* coefficient (Ekström, 2011). This correlation coefficient ranges from -1 to +1 and shows the strength of the positive or negative correlation between two dichotomous variables. To measure the association with other numerical or continuous variables, we applied the Point-Biserial Correlation Coefficient, which is used instead of the Pearson correlation when a variable is dichotomous (LeBlanc & Cox, 2017) and can range from -1 to +1.

4.3.3 Regression analysis

We used multivariate logistic regression to find the relationship between various variables (defined in Table2) and OA publishing. It is a common method for modeling the relationship between the dichotomous dependent variable and multiple independent variables. It allows us to understand the association of the dependent variable with an independent variable in the presence of other independent variables in the data.

4.3.4 Classification method

We employed a machine learning method to estimate the likelihood of choosing the publishing model. To this end, we categorized the publishing model of articles into two groups, OA and CA. Then, we utilized the value of defined features in Table 2 to predict the publishing model. This process is a classification task in machine learning.

To estimate the publishing model of articles, we use a supervised machine learning method, random forest (RF), a common tool for classification tasks (Behr, Giese, Theune, et al., 2020; Kumar, Mukhopadhyay, Gupta, Handa, & Shukla, 2019; Roy, Chopra, Lee, Spampinato, & Mohammadi-ivatlood, 2020; Yamak, Saunier, & Vercouter, 2016). We utilize this tool for binary classification (OA=1 or CA=0) and use the features introduced in Table 2 as independent variables. We implement the algorithm for hybrid journals in which authors can choose their paper's business model. We used k-Fold cross-validation (k=10) procedure to train and test the model.

Due to the skewed distribution in the target variable (91% CA and 9% OA publishing), we balance them by re-sampling data via *SMOTE* (Synthetic Minority Over-sampling Technique), which was proven to be a suitable method to handle a class imbalance problem (Spelmen & Porkodi, 2018).

5 Results

In this section, first, we present some descriptive statistics about the publishing model of articles across four country groups and address RQ1. Next, we display their differences in terms of citation impact among different models to answer RQ2. Then we focus on RQ3 and present the correlation coefficient between the publishing model and features defined in Table 2 and multivariate logistic regression to show the relationship between variables. Also, we demonstrate the performance of estimating the publishing model of articles in hybrid journals and the importance of defined features in the estimation task to reveal the influential factors in selecting the OA model for publishing.

5.1 Countries' income level of corresponding authors and their publishing model

Figure 4 shows the distribution of articles categorized by publishing model and the country income level of the corresponding authors. Authors with affiliations in countries with the lowest income level and eligible for the APC waiver have the highest proportion of gold OA publications. In contrast to this, authors from lower-middle-income countries who are eligible for the APC discount have the lowest percentage in gold OA publishing.

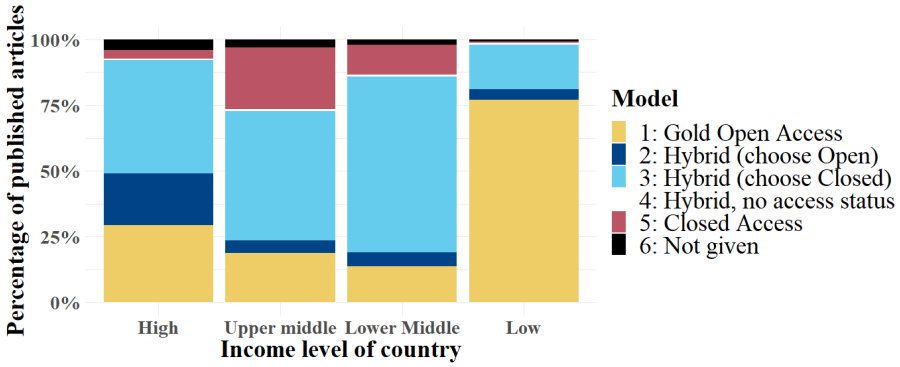

Fig. 4 Distribution of articles published in journals with three publishing models across four groups of countries. The access status of hybrid articles has been identified from Unpaywall (cases 2 and 3). For case 4 (Hybrid, no access status), we couldn't find hybrid journals' articles in Unpaywall.

5.2 Countries' income level of corresponding authors and their citation impact

Figure 5 shows the ratio of highly cited articles with different publishing models across country groups for the investigated articles. Generally, we observe a higher percentage of highly cited papers for corresponding authors from countries with higher income levels.

The ratio of highly cited articles among all countries for gold and hybrid OA models is higher than in other models. Also, this ratio is higher for gold OA articles and indicates the better citation impact of articles published in gold OA journals. The only exception is for countries with low-income levels, with more highly cited papers in the hybrid OA model. Compared to CA journals, journals in hybrid CA have more highly cited articles, except for countries with a high-income level.

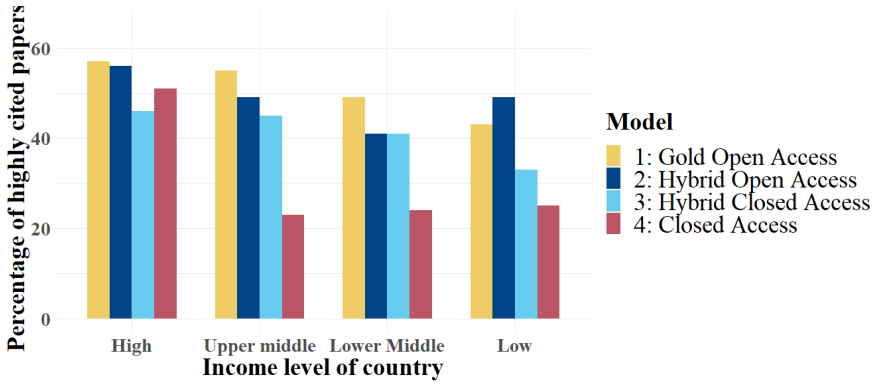

Fig. 5 Percentage of highly cited papers published in different models. Hybrid Open Access / Closed Access belongs to articles published as OA/CA in hybrid journals.

5.3 Influential factors on the publishing model

First, we conducted a correlation analysis to find the associations between OA publishing and features. Table 3 shows the correlation coefficient between the publishing model (if open access is equal to 1 otherwise 0) and features in Table 2. We separated the data into two sets, set 1 for articles published in OA or CA journals (non-hybrid journals) and set 2 for articles in hybrid journals. Set 1 reveals the association of discount and waiver policies with OA publishing, while optional OA publishing for hybrid journals in set 2 displays more author-specific factors related to OA publishing. The weak negative correlation with *gender* demonstrates that the tendency toward gold OA publishing for women is slightly more than for men, which disagrees with previous findings (Olejniczak & Wilson, 2020; Zhu, 2017). As we observed the lowest proportion of OA publishing for countries with a lower-middle-income level in Figure 4, the negative correlation for *discount_eligible* (also positive value for *waiver_eligible*) in Table 3 points out that the discount policies are insufficient

to motivate the authors from these countries for gold OA publishing. Table 4 displays the relationship between the publishing model and features in Table 3 by considering all features in multivariate logistic regression. The results confirm the negative/ positive correlation calculated in correlation analysis, except the positive correlation between *discount_eligible* and the publishing model is inconsistent with the result in the correlation coefficient. The highest Odds Ratios for Social Sciences among fields in Table 4 reveal the highest proportion of OA publishing in this field. for *field* indicate that among scientific fields, those authors having an article with a multidisciplinary subject are more likely to choose a gold OA. This field has experienced a dramatic since 2009 Liu and Li (2018). The strong positive correlation between *journal_ranking* and the publishing model for the first set suggests that the journal's rank is the dominant factor in choosing a gold OA journal to publish. Therefore, we estimate the publishing model for articles in set 2 (hybrid journals) to discover other feature categories rather than journal-specific factors influencing the authors' decision for an OA option. Moreover, the optional choice of the OA model in hybrid journals better reveals characteristics leading to the OA model.

Table 3 Correlation coefficient between independent variables and the target variable. The value of the target equal to 1 (0) means the paper has been published in the OA (CA) model.

		Correlation Coefficient	
Feature	**Correlation Test**	**Set 1 (non-hybrid)**	**Set 2 (hybrid)**
journal_ranking	Point-Biserial	0.70	0.07
journal_APC	Point-Biserial	-	0.10
field	Cramer's V	0.69	0.09
country_income	Point-Biserial	0.28	0.16
OA_agreement	Phi	0.08	0.30
discount_eligible	Phi	-0.08	-
waiver_eligible	Phi	0.06	-
OA_cite	Point-Biserial	0.42	0.13
authors_count	Point-Biserial	0.09	0.07
gender	Phi	-0.08	-0.01
age	Point-Biserial	-0.08	0.02
OA_publish	Point-Biserial	0.46	0.41
international_coauthors	Point-Biserial	0.17	0.11
Sample Size:		192,498	329,913

Table 5 shows the performance of the RF classifier for the second set (hybrid journals). Figure 6 displays the *permutation importance* of features employed to predict the publishing model implemented for this set. The permutation importance of a feature shows a decrease in the model performance when the feature's value is randomly shuffled while the values of other predictors remain unchanged. A higher value for a feature shows more predictive power in the proposed model. The highest importance values for *country_income*, and *age* in Figure 6 indicate that the most significant factors in selecting an OA model are the income level of countries and seniority. The lowest value for the variable *gender* presents that gender has a lower impact on the authors' decision for

Table 4 The results of Logistic regression. The target variable is the publishing model and is equal to 1 for OA and 0 for CA publishing. The outputs are Odds Ratio, exp(β). (1-exp(β)) shows the percentage change of the target variable per unit increase in an independent variable. So, the Odds Ratio greater/less than one displays a positive/negative correlation between variables.

	Set 1		Set 2	
	Odds Ratio	95% CI	Odds Ratio	95% CI
Intercept	0.002^{***}(-72.4)	0.001 to 0.002	0.00^{***}(-87.7)	0.00 to 0.00
Independent Variables				
journal_ranking	1.98^{***}(10.38)	1.74 to 2.25	110.7^{***}(86.5)	99.5 to 100.23
journal_APC	1.00^{***}(8.05)	1.0001 to 1.0002	-	-
field				
Health Sciences	reference	reference	reference	reference
Life Sciences	1.01(0.31)	0.94 to 1.08	0.67^{***}(-9.55)	0.62 to 0.73
Physical Sciences	0.97(-0.91)	0.91 to 1.07	0.20^{***}(-44.29)	0.18 to 0.21
Social Sciences	1.90^{***}(13.81)	1.73 to 2.08	3.49^{***}(12.2)	2.86 to 4.27
multiple fields	1.25^{***}(8.5)	1.19 to 1.32	3.4^{***}(30.87)	3.17 to 3.71
country_income	1.00^{***}(33.88)	1.000 to 1.000	1.000^{***}(16.18)	1.00 to 1.00
OA_agreement	14.9^{***}(65.07)	13.78 to 16.22	0.93(-0.78)	0.78 to 1.11
discount_eligible	-	-	1.7^{***}(9.17)	1.52 to 1.90
waiver_eligible	-	-	20.19^{***}(5.53)	8.29 to 77.5
OA_cite	0.55^{***}(-12.97)	0.500 to 0.600	1.55^{***}(8.4)	1.39 to 1.71
authors_count	1.003(0.80)	0.99 to 1.01	1.17^{***}(33.15)	1.16 to 1.18
gender	0.94^{**}(-2.8)	0.90 to 0.98	0.93^{*}(-2.5)	0.88 to 0.98
age	1.05^{***}(29.63)	1.05 to 1.1.054	0.97^{***}(-15.36)	0.96 to 0.98
OA_publish	196.79^{***}(105.65)	178.46 to 217.09	23.86^{***}(50.58)	21.1 to 26.99
international_coauthors	1.17^{***}(18.21)	1.15 to 1.19	1.03(1.34)	0.99 to 1.06
McFadden's Pseudo R^2	0.25		0.60	
Sample Size	96,674		162,773	
significant codes: . $p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ z-values of coefficients in parentheses CI: Confidence Interval				

Table 5 performance of predicting the publishing model of papers with random forest method.

Classification	**OA**	**CA**
Precision	0.85	0.94
Recall	0.95	0.83
F1score	0.89	0.88
Accuracy	0.92	

the OA model compared to other factors. OA_agreement is one of the weakest features in predicting the publishing model, and the correlation analysis also shows a weak correlation between them. One possible reason for the weak effect is that only 2.3% of papers have been involved in transformative agreements. In addition, the income level of countries is the most important feature, and regarding the positive correlation of this feature with OA publishing, it is more likely for authors from high-income countries (even without a transformative agreement) to publish in the OA model. This may also smooth the association of the agreement with OA publishing.

6 Conclusion and discussion

This work presents a detailed study of the relationship between author-specific and structural factors (e.g., income level of authors' affiliation country), OA

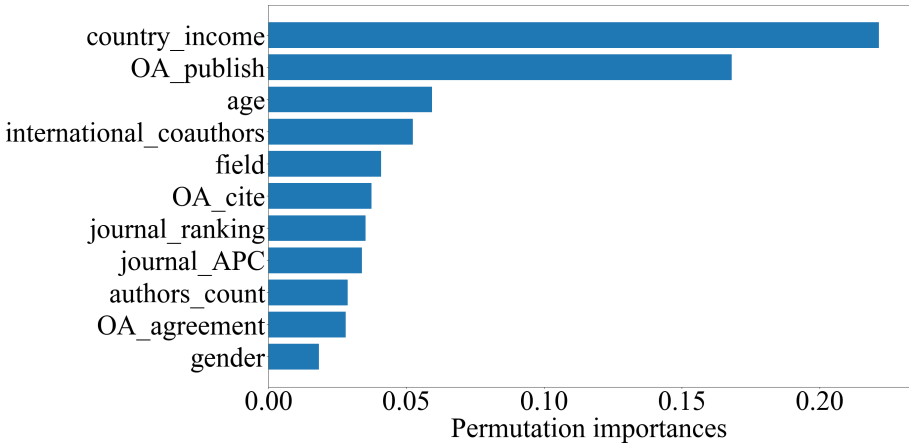

Fig. 6 Permutation importance of features employed to predict the publishing model of papers with random forest method for the articles published in hybrid journals.

publishing, and OA citation advantage. First, we investigated the relationship between the income level of countries and OA publishing for articles published by Springer Nature in the years 2017 and 2018. We found that authors from lower-middle-income countries with the eligibility to use APC discounts have a lower proportion of gold OA publications in all published papers by this publisher compared to other countries. It indicates that discounted APC is still too much for these authors to pay for a gold OA model and agrees with the statement of (Rouhi et al., 2022), who pointed out that waiver and discount issues couldn't bring author equity in reading and publishing. In contrast, this proportion of authors from countries with a low-income level who receive APC waivers is higher than authors from other countries. This result conflicts with the study's results by Smith et al. (2022), which found fewer OA papers proportions published by Elsevier for these countries compared to others. The reason can be stricter conditions, which this publisher considers for waiver eligibility.

We examined the citation impact of these articles and compared the percentage of highly cited papers among the publishing models and the income levels of the corresponding authors' countries. For all countries, the OA model in gold OA or hybrid has the highest percentage of highly cited papers. Also, the results demonstrate a higher proportion of highly cited articles for countries with higher income levels. Although it displays more citation impact for OA models, it can result from confounding factors such as self-selection and quality biases (Gargouri et al., 2010). Also, examining the preprint and green OA publishing (if the article has been published in the CA model, but a free version is available in a repository outside of the publisher's website) effect will result in more accurate analyses (Fraser et al., 2020; Wang, Glänzel, & Chen, 2020).

We conducted correlation, regression, and machine learning analyses to find more characteristics (e.g., author, journal, paper) related to OA publishing.

The results of the correlation analysis displayed the strength of positive/negative correlation between the publishing model and every feature defined in Table 2. Using regression analysis, we examined the association of each factor while accounting for other factors. The results reinforced the correlation outcomes. The only conflict between these two methods was the negative correlation between *discount_egibility* with OA publishing in correlation analysis, but positive in regression evaluation. In addition, we estimated the publishing model of articles (OA or CA) using a random forest-based machine learning approach and examined the impact of each feature on the estimation task. The results show that the country's income and more experiences in OA rather than CA publishing are the most influential factors in estimating the publishing model. We discovered that the tendency toward OA publishing was slightly higher for women, but it was a less important feature than other features in estimating the OA model.

7 Limitations and future work

One obvious limitation of this study is that we included articles from just one publisher, Springer Nature. Authors' publishing behavior may differ among articles published by other publishers, which limits the generalizability of the results of our study.

We obtained the access status of journals in 2019 based on the list published on Springer Nature's website (the same for the access status at the article level from Unpaywall). Some journals may have flipped from CA to OA (Momeni et al., 2021) or vice versa, and we did not detect them, which can cause errors in results. Furthermore, we did not control the correctness of external data (Springer nature and Unpaywall). The accuracy of these data affects the results' precision. We identified the gender of 49% authors and removed 49% of articles without gender status corresponding authors in regression and machine learning analyses. In addition, 2% of the data have been removed because of the null value in other features (e.g., journals' APC). Because the gender detection approach doesn't work well for Asian names, especially Chinese ones, we have a lower proportion of these authors with gender status in the dataset, which also creates biases in our analyses.

For future work, we can consider other publishers to examine how the different APC policies among publishers impact OA publishing. Also, controlling for articles' language in the analyses encourages future studies. **Springer Nature** is an international publisher and publishes mostly articles in English[22], and articles in other languages are underrepresented in this study. considering other publishers with non-English content and the articles' language in the analyses can reveal the role of languages in publishing international OA articles and citation advantages.

[22] https://support.springernature.com/en/support/solutions/articles/6000219817-are-any-of-your-titles-available-in-other-languages-

***

8 Declarations

Author contributions

Fakhri Momeni: Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Resources; Writing - Original Draft; Writing - Review & Editing; Visualization.
Kristin Biesenbender: Conceptualization; Resources; Writing - Review & Editing.
Philipp Mayr: Writing - Review & Editing; Project administration; Funding acquisition.
Stefan Dietze: Supervision; Methodology; Writing - Review & Editing
Isabella Peters: Supervision; Writing - Review & Editing; Project administration; Funding acquisition;

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The dataset analysed during the current study and codes are available on https://github.com/momenifi/open_access_springer_nature.git.

Funding information

This work is financially supported by BMBF project OASE, grant number 01PU17005A. We acknowledge the support of the German Competence Center for Bibliometrics (grant: 01PQ17001) for maintaining the used dataset for the analyses.

References

Bahlai, C., Bartlett, L.J., Burgio, K.R., Fournier, A.M., Keiser, C.N., Poisot, T., Whitney, K.S. (2019). Open science isn't always open to all scientists. *American Scientist*, *107*(2), 78–82.

https://doi.org/10.1511/2019.107.2.78

Barner, J.R., Holosko, M.J., Thyer, B.A. (2014). American social work and psychology faculty members' scholarly productivity: A controlled comparison of citation impact using the h-index. *The British Journal of Social Work*, *44*(8), 2448–2458.

https://doi.org/10.1093/bjsw/bct161

Bautista-Puig, N., Lopez-Illescas, C., de Moya-Anegon, F., Guerrero-Bote, V., Moed, H.F. (2020). Do journals flipping to gold open access show an oa citation or publication advantage? *Scientometrics*, *124*(3), 2551–2575.

https://doi.org/10.1007/s11192-020-03546-x

Behr, A., Giese, M., Theune, K., et al. (2020). Early prediction of university dropouts–a random forest approach. *Jahrbücher für Nationalökonomie und Statistik*, *240*(6), 743–789.

Bornmann, L., & Mutz, R. (2014). From p100 to p100': A new citation-rank approach. *Journal of the Association for Information Science and Technology*, *65*(9), 1939–1943.

https://doi.org/10.1002/asi.23152

Bornmann, L., & Williams, R. (2020). An evaluation of percentile measures of citation impact, and a proposal for making them better. *Scientometrics*, *124*(2), 1457–1478.

https://doi.org/10.1007/s11192-020-03512-7

Ekström, J. (2011). The phi-coefficient, the tetrachoric correlation coefficient, and the pearson-yule debate.

https://doi.org/10.1016/j.jkss.2012.10.002

Evans, J.A., & Reimer, J. (2009). Open access and global participation in science. *Science*, *323*(5917), 1025–1025.

https://doi.org/10.1126/science.1154562

Farys, R., & Wolbring, T. (2021). Matthew effects in science and the serial diffusion of ideas: Testing old ideas with new methods. *Quantitative Science Studies*, *2*(2), 505–526.

https://doi.org/10.1162/qss_a_00129

Fox, J., Pearce, K.E., Massanari, A.L., Riles, J.M., Szulc, Ł., Ranjit, Y.S., ... others (2021). Open science, closed doors? countering marginalization through an agenda for ethical, inclusive research in communication. *Journal of Communication*, *71*(5), 764–784.

https://doi.org/10.1093/joc/jqab029

Fraser, N., Momeni, F., Mayr, P., Peters, I. (2020). The relationship between biorxiv preprints, citations and altmetrics. *Quantitative Science Studies*, *1*(2), 618–638.

https://doi.org/10.1162/qss_a_00043

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., Harnad, S. (2010). Self-selected or mandated, open access increases citation impact for higher quality research. *PloS one*, *5*(10), e13636.

https://doi.org/10.1371/journal.pone.0013636

Henrich, J., Heine, S.J., Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83.

https://doi.org/10.1017/S0140525X0999152X

Hodge, D.R., & Lacasse, J.R. (2011). Evaluating journal quality: Is the h-index a better measure than impact factors? *Research on Social Work Practice*, *21*(2), 222–230.

https://doi.org/10.1177/1049731510369141

Iyandemye, J., & Thomas, M.P. (2019). Low income countries have the highest percentages of open access publication: A systematic computational analysis of the biomedical literature. *PLoS One*, *14*(7), e0220229.

https://doi.org/10.1371/journal.pone.0220229

Jannot, A.-S., Agoritsas, T., Gayet-Ageron, A., Perneger, T.V. (2013). Citation bias favoring statistically significant studies was present in medical research. *Journal of clinical epidemiology*, *66*(3), 296–301.

https://doi.org/10.1016/j.jclinepi.2012.09.015

Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., Strohmaier, M. (2016). Inferring gender from names on the web: A comparative evaluation of gender detection methods. *Proceedings of the 25th international conference companion on world wide web* (pp. 53–54). https://doi.org/10.1145/2872518.2889385

King, D.A. (2004). The scientific impact of nations. *Nature*, *430*(6997), 311–316.

https://doi.org/10.1038/430311a

Kumar, N., Mukhopadhyay, S., Gupta, M., Handa, A., Shukla, S.K. (2019). Malware classification using early stage behavioral analysis. *2019 14th asia joint conference on information security (asiajcis)* (pp. 16–23).

Langham-Putrow, A., Bakker, C., Riegelman, A. (2021). Is the open access citation advantage real? a systematic review of the citation of open access and subscription-based articles. *PloS one*, *16*(6), e0253129.

https://doi.org/10.1371/journal.pone.0253129

Lawson, S. (2015). Fee waivers for open access journals. *Publications*, *3*(3), 155–167.

https://doi.org/10.3390/publications3030155

LeBlanc, V., & Cox, M. (2017). Interpretation of the point-biserial correlation coefficient in the context of a school examination. *Tutor. Quant. Methods Psychol*, *13*, 46–56.

https://doi.org/10.20982/tqmp.13.1.p046

Lewis, C.L. (2018). The open access citation advantage: Does it exist and what does it mean for libraries? *Information technology and libraries*, *37*(3), 50–65.

https://doi.org/10.6017/ital.v37i3.10604

Liu, W., & Li, Y. (2018). Open access publications in sciences and social sciences: A comparative analysis. *Learned Publishing*, *31*(2), 107–119.

https://doi.org/10.1002/leap.1114

Matthias, L., Jahn, N., Laakso, M. (2019). The two-way street of open access journal publishing: flip it and reverse it. *Publications*, *7*(2), 23.

https://doi.org/10.3390/publications7020023

McKiernan, E.C., Bourne, P.E., Brown, C.T., Buck, S., Kenall, A., Lin, J., ... others (2016). Point of view: How open science helps researchers succeed. *elife*, *5*, e16800.

https://doi.org/10.7554/eLife.16800

Momeni, F., Mayr, P., Dietze, S. (2022). Investigating the contribution of authors' and papers' characteristics to predict the scholars' h-index. *arXiv preprint arXiv:2207.09655*.

https://doi.org/10.48550/arXiv.2207.09655

Momeni, F., Mayr, P., Fraser, N., Peters, I. (2021). What happens when a journal converts to open access? a bibliometric analysis. *Scientometrics*, 1–17.

https://doi.org/10.1007/s11192-021-03972-5

Munafò, M.R., Nosek, B.A., Bishop, D.V., Button, K.S., Chambers, C.D., Percie du Sert, N., ... Ioannidis, J. (2017). A manifesto for reproducible science. *Nature human behaviour*, *1*(1), 1–9.

https://doi.org/10.1038/s41562-016-0021

Olejniczak, A.J., & Wilson, M.J. (2020). Who's writing open access (oa) articles? characteristics of oa authors at ph. d.-granting institutions in the united states. *Quantitative science studies*, *1*(4), 1429–1450.

Ottaviani, J. (2016). The post-embargo open access citation advantage: it exists (probably), it's modest (usually), and the rich get richer (of course). *PLoS One*, *11*(8), e0159614.

https://doi.org/10.1371/journal.pone.0165166

Piwowar, H., Priem, J., Larivière, V., Alperin, J.P., Matthias, L., Norlander, B., ... Haustein, S. (2018). The state of oa: a large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, *6*, e4375.

https://doi.org/10.7717/peerj.4375

Rimmert, C., Schwechheimer, H., Winterhager, M. (2017). Disambiguation of author addresses in bibliometric databases-technical report.

Ross-Hellauer, T., Reichmann, S., Cole, N.L., Fessl, A., Klebel, T., Pontika, N. (2021). Dynamics of cumulative advantage and threats to equity in open science: a scoping review. *Royal Society Open Science*, *9*(1), 211032.

Rouhi, S., Beard, R., Brundy, C. (2022). Left in the cold: the failure of apc waiver programs to provide author equity. *Science Editor*, 5–13.

https://doi.org/10.36591/SE-D-4501-5

Roy, S.S., Chopra, R., Lee, K.C., Spampinato, C., Mohammadi-ivatlood, B. (2020). Random forest, gradient boosted machines and deep neural network for stock price forecasting: a comparative analysis on south korean companies. *International Journal of Ad Hoc and Ubiquitous Computing*, *33*(1), 62–71.

Samimi, A.J. (2011). Scientific output and gdp: Evidence from countries around the world. *Journal of Education and Vocational Research*, *2*(2), 38–41.

https://doi.org/10.22610/jevr.v2i2.23

Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, *4*, e156.

https://doi.org/10.7717/peerj-cs.156

Schroter, S., Tite, L., Smith, R. (2005). Perceptions of open access publishing: interviews with journal authors. *BMJ*, *330*(7494), 756.

https://doi.org/10.1136/bmj.38359.695220.82

Simard, M.-A., Ghiasi, G., Mongeon, P., Larivière, V. (2021). Geographic differences in the uptake of open access. *18th international conference on scientometrics and informetrics conference, issi 2021.* Retrieved from https://issi2021.org/proceedings/

Smith, A.C., Merz, L., Borden, J.B., Gulick, C.K., Kshirsagar, A.R., Bruna, E.M. (2022, 02). Assessing the effect of article processing charges on the geographic diversity of authors using Elsevier's "Mirror Journal" system. *Quantitative Science Studies*, *2*(4), 1123-1143.

https://doi.org/10.1162/qss_a_00157

Sotudeh, H., Ghasempour, Z., Yaghtin, M. (2015). The citation advantage of author-pays model: the case of springer and elsevier oa journals. *Scientometrics*, *104*(2), 581–608.

https://doi.org/10.1007/s11192-015-1607-5

Spelmen, V.S., & Porkodi, R. (2018). A review on handling imbalanced data. *2018 international conference on current trends towards converging technologies (icctct)* (pp. 1–11).

Sullo, E. (2016). Open access papers have a greater citation advantage in the author-pays model compared to toll access papers in springer and

***

elsevier open access journals. *Evidence Based Library and Information Practice*, *11*(1).

http://dx.doi.org/10.18438/B84W67

Wang, Z., Glänzel, W., Chen, Y. (2020). The impact of preprints in library and information science: an analysis of citations, usage and social attention indicators. *Scientometrics*, *125*(2), 1403–1423.

https://doi.org/10.1007/s11192-020-03612-4

Xia, J. (2012). Positioning open access journals in a lis journal ranking. *College & Research Libraries*, *73*(2), 134–145.

https://doi.org/10.5860/crl-234

Yamak, Z., Saunier, J., Vercouter, L. (2016). Detection of multiple identity manipulation in collaborative projects. *Proceedings of the 25th international conference companion on world wide web* (pp. 955–960).

Zhu, Y. (2017). Who support open access publishing? gender, discipline, seniority and other factors associated with academics' oa practice. *Scientometrics*, *111*(2), 557–579.

https://doi.org/10.1007/s11192-017-2316-z

Chapter 5

Impact of International Academic Mobility on Researchers' Career

5.1 Overview

International academic mobility plays a significant role in the globalization of science and is a key aspect of the institutional strategy of universities and research organizations in many countries. In this study [75], we investigated gender inequalities in mobility programs across countries, career stages, and scientific fields at a large scale to tackle mobility problems. Using statistical methods and regression analyses, we also examined mobility's impact on researchers in terms of productivity and received citations and collaboration. A global co-authorship network has been built to analyse the social aspects of mobility and the role of mobile researchers in scientific communities.

5.2 Publication

Title: The many facets of academic mobility and its impact on scholars' career

Authors: Fakhri Momeni, Fariba Karimi, Philipp Mayr, Isabella Peters, and Stefan Dietze

Document Type: Journal paper

Venue: Journal of Informetrics

DOI: https://doi.org/10.1016/j.joi.2022.101280

Journal of Informetrics 16 (2022) 101280

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

The many facets of academic mobility and its impact on scholars' career

Fakhri Momeni[a,*], Fariba Karimi[b], Philipp Mayr[a], Isabella Peters[c], Stefan Dietze[a,d]

[a] GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany
[b] Complexity Science Hub Vienna, Josefstädter Straße 39,1080 Vienna, Austria
[c] ZBW – Leibniz Information Centre for Economics, Düsternbrooker Weg 120, 24105 Kiel, Germany
[d] Heinrich-Heine-University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany

ARTICLE INFO

Keywords:
International academic mobility
Gender inequality
Co-authorship network
Bibliometrics
Scientific performance
Scientific success

ABSTRACT

International mobility in academia can enhance the human and social capital of researchers and consequently their scientific outcome. However, there is still a very limited understanding of the different mobility patterns among scholars with various socio-demographic characteristics. By studying these differences, we can detect inequalities in access to scholarly networks across borders, which can cause disparities in scientific advancement. The aim of this study is twofold. First, we investigate to what extent individuals' factors (e.g., country, career stage, and field of research) associate with the mobility of male and female researchers. Second, we explore the relationship between mobility and scientific activity and impact. For this purpose, we used a bibliometric approach to track the mobility of authors. To compare the researchers' scientific outcomes, we considered the number of publications and received citations as indicators, as well as the number of unique co-authors in all their publications. We also analyzed the co-authorship network of researchers and compared centrality measures of "mobile" and "non-mobile" researchers. Results show that researchers from North America and Sub-Saharan Africa, particularly female ones, have the lowest, respectively, highest tendency towards international mobility. Having international co-authors increases the probability of international movement. Our findings uncover gender inequality in international mobility across scientific fields and countries. Across genders, researchers in the Physical sciences have the most and in the Social sciences the least rate of mobility. We observed more mobility for Social scientists at the advanced career stage, while researchers in other fields prefer to move at earlier career stages. Also, we found a positive correlation between mobility and scientific outcomes, but no apparent difference between females and males. Indeed, researchers who have started mobility at the advanced career stages had a better scientific outcome. Comparing the centrality of mobile and non-mobile researchers in the co-authorship networks reveals a higher social capital advantage for mobile researchers.

1. Introduction

Scientific progress is the result of a collaborative process that involves researchers across the world and international collaboration. For that, mobility of researchers is important, since it fosters communication, collaboration and knowledge transfer between researchers – all factors which are considered crucial for scientific progress as well as for success and performance of researchers.

* Corresponding author.
E-mail addresses: fakhri.momeni@t-online.de (F. Momeni), karimi@csh.ac.at (F. Karimi), philipp.mayr@gesis.org (P. Mayr), I.Peters@zbw.eu (I. Peters), stefan.dietze@hhu.de (S. Dietze).

https://doi.org/10.1016/j.joi.2022.101280
Received 29 April 2021; Received in revised form 2 March 2022; Accepted 14 March 2022
Available online 28 March 2022

Table 1
Investigated features in studies that used a bibliometric approach to study international mobility of authors.

Study	Investigated feature				Restriction in author selection
	Gender	Field	Country	Career stage	
Aman (2018a)	×	√	×	×	Authors with German affiliation
Petersen (2018)	×	×	√	×	Physics researchers
Robinson-Garcia et al. (2019)	×	×	√	×	-
Subbotin and Aref (2021)	√	√	×	×	Russian Authors
El-Ouahi et al. (2021)	√	×	√	√	Authors with an affiliation in Middle East and North Africa region
Our study	√	√	√	√	-

However, research has shown that extent and distribution of mobility is gender-dependent (El-Ouahi et al., 2021; Jöns, 2011; Leemann, 2010; Ryazanova & McNamara, 2019). For example, although participation in science of researchers that have been identified as females has seen a significant advancement during the last years (Carr et al., 2015; Ovseiko et al., 2017), the vast majority of female researchers are not mobile (Jöns, 2011).

Gender inequality in academic international mobility needs to be tackled in many societies, because of the existing troubles to go abroad for women. Depending on the family situations during the career life, women can face more barriers to being mobile at any career stage (Jöns, 2011). Identifying differences in mobility among various countries, scientific fields and the career stages can help to better explore the root of problems women deal with.

Studying the mobility patterns among different societies enhances our understanding of the researchers' motivations and restrictions to move internationally. Someone moves to a new country due to experience working with colleagues and researchers in another environment, while another one moves away from problems in the current country (e.g., gender inequality, lack of labor in a particular field, financial and political grounds, etc.), which may as well be a potential barrier to going aboard. On the other hand, the chance of obtaining an appropriate research position abroad is not the same for all groups, which affects the decision of researchers and mobility directions. Some fields are more in demand in a particular country and academic positions are offered to international or inexperienced researchers regardless of gender, whereas, in some other fields or countries, positions are restricted to skilled or male researchers. In addition, the comparative analyses of mobility effects on researchers across various groups and different career phases, reveal the importance of mobility between countries. By applying a bibliometric approach, we can track the mobility of researchers through their publications. Some studies have already investigated international academic mobility in a similar way (Aman, 2018a; El-Ouahi et al., 2021; Petersen, 2018; Robinson-Garcia et al., 2019; Subbotin & Aref, 2021). Table 1 summarizes the notable related studies in terms of academic mobility and the features they considered as well as the scope of analyzed data. No prior work exists that investigates the role of gender in the context of mobility and scientific impact on a global scale. The main contribution of this paper is to comparatively study the mobility pattern of different genders, i.e., women and men, and their scientific outcome among scientific fields and countries across the stages of career development. This will reveal the extent of gender inequality in different societies. We define international academic mobility as changes in the scholars' country of affiliation over time. In the following, *'mobility'* refers to '*international academic mobility'*.

In this paper, we aim to answer two main research questions:

1. To what extent individuals' factors (e.g., country, career stage, and field of research) do associate with the mobility of researchers and how do they differ for males and females?
2. How do different characteristics of mobile researchers correlate with scientific outcomes of researchers?

For the first research question we investigate the role of gender, scientific field, country of origin, and international collaboration on the likelihood of becoming mobile and compare the mobility pattern of two genders, i.e., male and female, across countries and scientific fields and through three career stages. For the second research question, we use the number of publications, received citations and number of unique co-authors of researchers and examine the relationship between mobility and these indicators. We also analyze the co-authorship networks to present the differences between centrality measures of mobile and non-mobile researchers.

Our study shows the following novel aspects; firstly, the scale and broad coverage of the used dataset from various fields and countries, that contributes to the generalizability of the results. Second, inferring the gender of scholars with high accuracy on scale that enables comparative analyses between two genders. Third, tracking the mobility of scholars over time using scholarly publications and considering the frequency of movements in analyses that leads to more comprehensive results. Lastly, by applying centrality measures in large-scale collaboration networks at the individual level, we compare the position and role of mobile with non-mobile researchers in these networks.

2. Related work

In the academic world, communication and collaboration between scientists are crucial to individual and scientific success. International mobility can connect scholars from different countries with various scientific backgrounds and along with it may enhance knowledge exchange. It associates with both human capital (refers to the knowledge and experience of individuals (De Cleyn et al., 2015)) and social capital (as the resources available to individuals and groups through membership in social networks Villalonga-Olives & Kawachi, 2015) of researchers. Thus, it can affect the researchers' scientific impact positively by sharing, exchanging knowl-

edge and obtaining other opportunities to enhance individuals' skills or negatively by disconnecting from local co-authors and having difficulties making connections with new colleagues because of different languages and cultural backgrounds (Almansour, 2015; Caniglia et al., 2017). Also, at the country level, a researcher with the experience of staying abroad can act as a hub which mediates between different countries and, overall, increases collaboration between both countries (brain circulation) (Saxenian, 2007). However poor countries suffer from losing talents who migrate and labor shortage (brain drain) (Arrieta et al., 2017).

2.1. Mobility and field of research

Epistemic characteristics of fields influence decisions of researchers for national or international mobility (Laudel & Bielick, 2019). In some fields, the human capital of researchers is more transferable, but some others are more specific for a country. For example, Bäker (2015) reports that researchers in disciplines with both quantitative and qualitative research methods (pluralistic) are more likely to lose their human capital after changing affiliation, because of the diversity in research approaches. Aman (2020) measured the knowledge transmission among mobile and non-mobile researchers and discovered the highest knowledge transmission in "Earth and Planetary Sciences" and "Neurosciences". Depending on the size and domain of used data, prior studies have mentioned different proportions of mobility across disciplines. Cañibano et al. (2011) analysed a set of 10,000 PhD holders in Spain from *Scientific Information System of Andalusia* dataset (SICA) and reported the most international mobility in "social sciences" and "science and technology of health" as the least mobile discipline. In contrast, Subbotin and Aref (2021) found that Russian scientists have the most and least international mobility in Physical Sciences and Social Sciences, respectively.

2.2. Mobility and gender

Gender inequality in science is more obvious in mobility, due to barriers for women to go abroad. It can decrease their visibility and scientific impact, as Kong et al. (2021) explained that women suffer from citation inequality due to first-mover advantage of men. However, Bozeman and Gaughan (2011) found no evidence that men or women adopt a "nationalist" strategy (wishing collaborators from one's own nation or shared language) in collaboration. Prior studies have shown that women are less likely to have international mobility (El-Ouahi et al., 2021; Jöns, 2011; Leemann, 2010; Ryazanova & McNamara, 2019). This varies by many factors such as discipline, career stage and country of origin. Bhandari (2017) showed a lower percentage of internationally mobile female researchers in STEM disciplines and the results of a study by Jöns (2011) report a less international mobility of women in natural sciences. Jayachandran (2015) showed that many poor countries favor men in mobility than women due to cultural norms. There are some mobility programs around the world that prioritize women. For example, women are over-represented by Erasmus mobility program (Böttcher et al., 2016; De Benedictis & Leoni, 2020). Jöns (2011) found that at the earlier career stages, male and female students are equally internationally mobile, but at advanced career stages flexibility of women to go abroad decreases much more than those of their male colleagues. However, Leemann (2010) revealed that the probability to move abroad decreases with age for both genders.

2.3. Mobility and academic impact

Academic mobility influences the co-authorship pattern that impacts quality and quantity of scientific productivity. These effects differ between disciplines with varying characteristics. Halevi et al. (2016) analysed the data of 100 top authors in seven disciplines and showed that for some disciplines, country mobility has a negative effect on productivity and received citations, while for others it has a positive or no effect. Bäker (2015) analysed the impact of changing affiliation for economics as a less pluralistic discipline (e.g., only quantitative research methods) and management as a more pluralistic discipline (e.g., quantitative and qualitative research methods) and report a worse effect for the most pluralistic disciplines in the short-term, because researchers in those disciplines are more likely to lose human capital due to variety of approaches in the new institutions. Petersen (2018) reported an increase in co-author diversity as the effect of mobility for physics scientists. Wang et al. (2019) examined the change in collaboration patterns of mobile researchers and found an increase in domestic collaboration but at the cost of decreasing international collaboration. Also, Bernard et al. (2021) showed the reduced likelihood of collaboration with previous co-authors after mobility.

Also, the time of moving is a significant factor influencing academic outcomes. Zhao et al. (2020) found that the productivity of researchers who move to China at an earlier career stage is higher than those who move at a later stage. However, Bauder (2020) reported that mobility can lead to the loss of national social capital that negatively affects early-career researchers in particular. Furthermore, the results of a study by Ryazanova and McNamara (2019) indicate a negative effect of international mobility at the first postdoctoral researcher (postdoc) job on research productivity, however they found that an international movement between year 2 and 7 of a postdoc is better than later.

2.4. Approaches, data resources and investigated features

Many studies utilized qualitative approaches to investigate academic mobility.. The major drawback in qualitative analysis is mainly the small size of data as well as bias problems (Bäker, 2015; Bauder et al., 2017; Bedenlier, 2018; Cohen et al., 2020; Laudel & Bielick, 2019; Leung, 2017; Morano-Foadi, 2005; Nikunen & Lempiäinen, 2020; Schaer et al., 2017). Other studies with quantitative approaches employed resources such as CV (Cañibano et al., 2008; Laudel & Bielick, 2019; Li & Tang, 2019; Youtie et al., 2013; Zhao et al., 2020) and bibliometric data of researchers (Aman, 2018a; Chinchilla-Rodríguez et al., 2018;

El-Ouahi et al., 2021; Petersen, 2018; Robinson-Garcia et al., 2019; Subbotin & Aref, 2021) to track their movements with larger data sample sizes. Table 1 shows these studies with a bibliometric approach and the investigated features as well as the set of selected authors. The most of these studies used a restricted set of authors or features. For example, Petersen (2018) analysed the *American Physical Society* (APS) dataset which covers publications in the domain of physics. Subbotin and Aref (2021) employed the Scopus dataset for analysing the international migration of researchers who have published with a Russian affiliation address, by discipline. El-Ouahi et al. (2021) investigated the international mobility for countries in the Middle East and North Africa region from Web of Science dataset. In this study, they compared the Gender ratio of migrants for the countries in this region. Among all these, only the study by Robinson-Garcia et al. (2019) covered all authors in Scopus from various countries and classified mobile authors into three groups (migrant, directional travelers and non-directional travelers). They did a comparative analysis for these mobility classes at the country level and compared the scientific outcome of mobile authors with those of non-mobile authors. Our study includes the authors from different countries too, but we rank the mobility of authors according to the frequency of changing their affiliated countries, which leads to a more detailed analysis of the extent of mobility. Gender, field, career stage and network centralities of researchers are other distinctive aspects of our study that enable us to discover the disparities and issues in mobility in different societies and scientific communities.

3. Data and methods

3.1. Data sources

The in-house Scopus database maintained by the German Competence Centre for Bibliometrics (Scopus-KB), 2020 version, is used as the main resource of analyses. We utilized publications indexed in Scopus to study the international mobility of scholars. In order to identify authors, we used Scopus author ID which enable us to track the international mobility of authors (Aman, 2018b). Kawashima and Tomizawa (2015) estimated the accuracy of Scopus author ID using KAKEN database (largest funding database in Japan) and found a very high precision (99%) and recall (98%).

For detecting the gender status, we apply a combined name and image-based approach introduced by Karimi et al. (2016). They tested the accuracy of this method in their paper with a sample of 693 male and 723 female names. The ground truth consists of a manually labelled random sample of academics, their full names, institutions, countries, and their gender. This method (combination of first names, family names, and images) has a general f-score of 93% which is higher than other existing gender inference methods and is more robust for different nationalities. The only exception is for Asian names, especially Chinese names, where this method has low accuracy. Therefore, we try to eliminate those ambiguous names to increase the accuracy of results for these countries. From 32,110,580 identifiers in Scopus, 7,956,823 had no first name or just initial among their publications that any gender detection methods would not be able to infer genders. For the remaining 24,153,757 identifiers, our gender inference method was able to infer the gender of 8,592,307 (~35%) names that could be identified.

We acknowledge that gender is a non-binary identity. For our purposes, and due to the lack of more fine-grained gender information, we consider it as binary in this work. The term “gender” doesn’t refer to the sex of the authors, nor the gender that the authors identify themselves with. We refer to gender as the general societal convention in assigning first names to individuals in combination with what machine learning face recognition algorithms identify as female or male (Karimi et al., 2016). Hence, this work can only be a starting point for more detailed analyses of the role of gender on the mobility of researchers.

To detect the disciplines of authors we used the “All Science Journal Classification” (ASJC) system of Scopus[1] which contains 27 subject categories. Next, we classify these disciplines to four main fields according to the Scopus classification.

The field with most publications is considered as the *main field* of the author. About 1.5% of the authors had more than one most popular field and we excluded these authors from the analyses.

3.2. Career stages

Several approaches to stratify researchers into career stages have been proposed and discussed. The major challenge lies in the individual situations of career progression, which is highly dependent on many factors (e.g., discipline, faculty, and career interruption). Although not optimal, this is why most approaches work with fixed time periods for every career stage. Li et al. (2019), for example, defined researchers as ‘junior’ within three years after their first publication. In contrast, Bäker (2015) recommended 4-6 years for this phase of career.

We agree with Bazeley (2003) and Bosanquet et al. (2017) who considered a period of five years as the minimum for the early career stage and therefore also adopted the approach presented by Mascarenhas et al. (2017) who used the following calculations for the three career stages:

Early career stage (s_1): years from the first publication year to 4 years after it.

Mid-career stage (s_2): years between 5 and 9 years after the first publication year.

Late career stage (s_3): years more than 10 years after the first publication year.

[1] More information: https://service.elsevier.com/app/answers/detail/a_id/14882/supporthub/scopus/~/what-are-the-most-frequent-subject-area-categories-and-classifications-used-in/ Accessed 14 Sep. 2021.

Table 2
Two examples calculating the mobility score. Author A has a mobility score of 5 and author B has a mobility score of 4.

Author A			Author B		
Publishing year	Affiliation country	score	Publishing year	Affiliation country	score
2002	USA	0	2000	Japan	0
				China	1
				Singapore	1
2004	Germany	1	2005	Japan	0
	Canada	1		China	0
	France	1			
2007	USA	1	2006	Australia	1
	Canada	0		Japan	0
2008	USA	0	2008	China	1
	Germany	1			
Sum of scores		5			4

We did not find a clear definition of advanced career stages (middle and late) in other studies. The only study from Ponjuan et al. (2011) defined 4 to 5 years for mid-career and more than five years for late career stages for pre-tenure faculty members, which is close to our thresholds for these two career phases.

To further increase comparability and homogeneity among the studied researchers (and to disregard career paths that show too much variance) as mentioned above, we excluded authors who published less than one publication per three years.

3.3. Mobility detection

In Scopus, affiliation information as well as the country of affiliation of authors are separately available for each publication. We utilized the country information of the affiliation to track the mobility of authors. Since the authors' affiliation is provided for each publication, we can track the changes of affiliations over time.

Mobility in this study is defined as having a co-affiliation (affiliated with more than one country in the same publication) or multiple affiliations (affiliated with at least two countries in two papers) (Chinchilla-Rodríguez et al., 2017; Petersen, 2018). Therefore, an author with one affiliation country through the author's publications is considered as non-mobile.

The *origin country* of the author is the country of author's affiliation on the first publication.

Mobility score is applied to measure the frequency of mobility. To calculate the mobility score of the author, we sort the lists of affiliation countries based on publishing years. Next, we compare the affiliation countries for each year to those from previous publishing year and assign one score for each country in the current year that doesn't exist in the list of previous year. Then, the sum of scores across all publishing years will be assigned as the mobility score of an author. For the first publishing year with non-empty list of countries, the number of unique countries except the first country which is the origin country, will be considered as the score for that year. Table 2 shows two examples of calculating the mobility score.

We select those authors for our analyses who have a Scopus author ID, gender status (male or female), and at least early and mid-career stage publications (10 years career age). We consider active authors who published at least one-third of their career age (e.g., an author having the first publication in 2001 and last publication in 2013 has a career age of 12 and should have at least four publications). To count the number of received citations, we apply a three years citation window after the publication year. To ensure that we count the received citations equally for all publications, we include all publications until 2016 and assume it as the last publication year for those authors published after this year. In addition, to have more authors with the late-career stage, we include the authors with the first publication year until 2002. Since most authors have their first publication from 1996, we exclude the authors with the first year before this year. By applying all these filters, we extract a list of 1,184,355 authors.

3.4. Region and Income level of countries

We use annual Gross Domestic Product (*GDP) per capita*, Purchasing Power Parity (PPP) (current international $) and region of countries from the World Bank[2] in the analyses. The average GDP per capita from 1996 to 2016 is considered for each country.

3.5. Mobility outcome

Similar to Barabási and Musciotto (2019) we define two concepts to assess the impact of mobility on the research outcome: performance and success. According to this definition "Performance is about individual effort, while success is a collective quantity capturing community's acknowledgment of effort and performance". Therefore, we evaluate the performance by *mean publication per year* (*PPY*) (by dividing the number of publications by career age) and success by *mean citation per publication* (*CPP*) (by dividing the sum of citations by number of publications). To calculate CPP, we consider the citations received until three years after the publication

[2] https://data.worldbank.org/indicator/NY.GDP.MKTP.PP.CD Accessed 14 Sep. 2021.

year (to account for differences in the age of publications). Finally, we use *Mean unique co-authors per publication* (*COPP*) (number of unique co-authors among all publications divided by number of publications) to measure the impact of mobility on co-author diversity.

While first/corresponding authors are considered to have made the major contribution to the paper, all publications of researchers would affect their career through the accumulation of citations and h-index. On the other hand, most researchers have their first-authored papers at the earlier years of their career life and their position in publications moves from first to the last author while progressing through career stages (Gingras et al., 2008; Way et al., 2017). Therefore, for calculating PPY, CPP, and COPP, we consider all publications that an author has co-authored and don't differentiate between various authorship positions.

3.6. Co-authorship network

We measure the social capital of authors by analysing the co-authorship network. We generate the co-authorship network for each discipline based on the publication records. In this network, every author is one node, and each edge represents a co-authorship activity between two authors. The number of shared publications between authors indicates the weight of the edges. The structure of the network changes and evolves over the years. Because of that, for each discipline and year, we create a network. We assume any two authors (nodes) are connected in a given year if they have at least one co-authorship in the past five years. Therefore, those nodes and edges related to publications older than five years are not included in the network of that year. Also, nodes without any connection are removed from the networks. The network analyses are performed with the *igraph*[3] library in Python.

Degree, closeness and betweenness are three well-known centrality measures employed in most previous works (Abbasi et al., 2011; Chinchilla-Rodríguez et al., 2017; Karimi et al., 2019; Li et al., 2013; Servia-Rodríguez et al., 2015) that analysed a co-authorship network. We use degree and closeness, and disregard betweenness because of its high calculation complexity and time expenditure for such large-scale collaboration networks on the author level.

Co-authorship networks consist of many communities in which co-authors are connected via their publications. *Clustering coefficient* is another centrality metric utilized in this study to observe how the collaboration patterns in communities modify when authors change their communities via mobility, as it was used by Abbasi et al. (2011).

Collaborating with top researchers is a motivation for many researchers to go abroad. We apply *coreness* centrality to examine whether they have better access to top authors.

We calculate these centrality measures for any node (author) in a given network with N nodes and E edges:

- *Degree:* Number of ties that the node has with other nodes (Freeman, 1978):

$$d_i = \sum_j a_{ij}$$

where a_{ij} is an element of the adjacency matrix and indicates the existence or non-existence of a link between node$i$ and node$j$.

- *Closeness*: Average length of the shortest path between the node and all other nodes in the graph (Freeman, 1978). It indicates how close an author is to all other authors in the network. Because this centrality is a global centrality and it is affected by the number of nodes in the network, we use the normalized closeness by multiplying raw closeness by total number of nodes except the one $(n-1)$ (Mohammadamin et al., 2017):

$$C_i = \frac{n-1}{\sum_j e_{ij}}$$

where n is the number of nodes and e_{ij} is the number of links in the shortest path from node i to node j.

- *Coreness*: It represents how well a node is connected to other important nodes and also with periphery nodes in the network (Saxena & Iyengar, 2020). A k-core of k indicates that a node is connected to a subset of nodes that have at least k degree or higher.
- *Clustering coefficient:* The probability that the adjacent nodes of a node are connected. In co-authorship network, clustering indicates how likely it is that two co-authors of a given author are also co-authors (Zare-Farashbandi et al., 2014).

4. Results

First, we will present some descriptive statistics about the analysed authors and their characteristics. Next, we will show the results of two regression analyses. The first model displays the factors influencing mobility (Table 4) and the second model indicates the impact of mobility on scholars' career (Table 5).

Also, we will utilize propensity score matching (PSM), for causal inference, to examine how mobility influences the scientific outcome. By comparing the centrality measures of mobile and non-mobile authors in the co-authorship networks, we will try to explain their position and role in these scientific networks.

[3] https://igraph.org/python/ Accessed 14.9.2021.

Table 3
Number of analysed authors and proportion of mobile and non-mobile among gender and scientific fields. The percentages identify the proportion of non-mobile or mobile researchers to all researchers.

		Non mobile, Mobility Score = 0 (percentage)	Mobile, mobility score> = 1 (percentage)	Total
Health Sciences	Women	83,107 **(65%)**	43,916 **(35%)**	127,023
	Men	161,941 **(58%)**	119,508 **(42%)**	281,449
	Total	245,048 **(60%)**	163,424 **(40%)**	408,472
Life Sciences	Women	55,132 **(61%)**	34,915 **(39%)**	90,047
	Men	100,478 **(55%)**	83,067 **(45%)**	183,545
	Total	155,610 **(57%)**	117,982 **(43%)**	273,592
Physical Sciences	Women	48,396 **(60%)**	32,634 **(40%)**	81,030
	Men	187,143 **(54%)**	157,812 **(46%)**	344,955
	Total	235,539 **(56%)**	190,447 **(44%)**	425,985
Social Sciences	Women	16,288 **(72%)**	6,364 **(28%)**	22,652
	Men	34,681 **(65%)**	18,973 **(35%)**	53,654
	Total	50,969 **(67%)**	25,337 **(33%)**	76,306
Total	Women	202,923 **(63%)**	117,829 **(37%)**	320,752
	Men	484,243 **(56%)**	379,360 **(44%)**	863,603
	Total	687,166 **(58%)**	497,189 **(42%)**	1,184,355

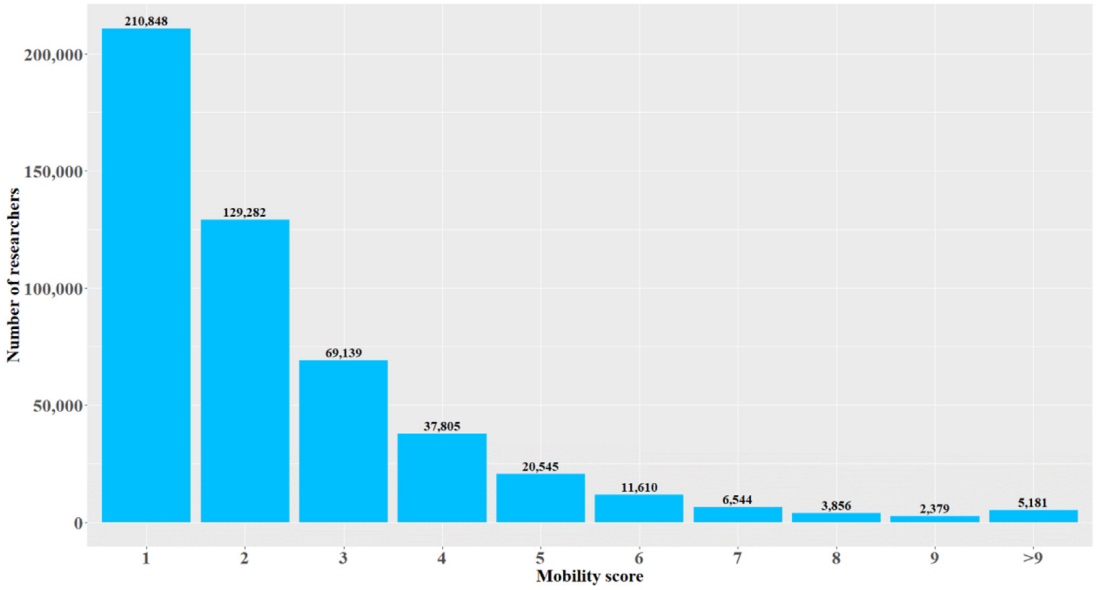

Fig. 1. Distribution of mobile researchers based on their mobility score. The biggest group shows researchers with a mobility score of one and the number of researchers decreases for higher mobility scores.

4.1. Descriptive statistics

Table 3 shows the number of included male and female researchers in this study and the proportion of international mobility per discipline. Comparing the proportion of mobile to non-mobile researchers for each gender shows that women have been less mobile (37%) than men (44%) overall.

Fig. 1 represents the distribution of mobile researchers based on their mobility score. The mobility score has a range from 1 to 58 and it has a skewed distribution. Most mobile researchers (~42%) have only one movement over their career life.

4.2. Mobility differences across countries, disciplines, and career stages

Fig. 2 (a) shows that in all scientific fields, women are underrepresented in international mobility, especially in Physical Sciences women have least participation. This agrees with the result of study by Bhandari (2017) and can be the result of gender inequality in this field (Wang & Degol, 2017; Miyake et al., 2010). From Fig. 2 (b) we observe that in all fields the proportion of internationally mobile female to all female researchers is less than for male researchers. Physical Sciences and Life Sciences have the most mobile

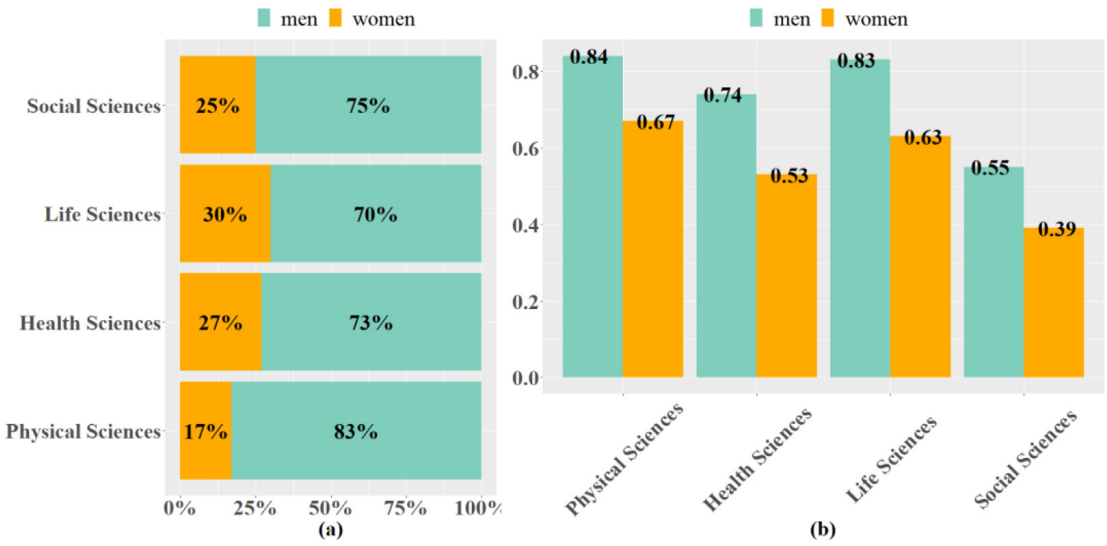

Fig. 2. (a) The share of male and female mobile researchers across fields, Physical Sciences has the lowest percentage of mobile women among all fields. (b) Ratio of male/female mobile researchers to non-mobile male/female researchers across fields, Physical Sciences and Social Sciences have the highest and lowest participation in mobility for both genders, respectively.

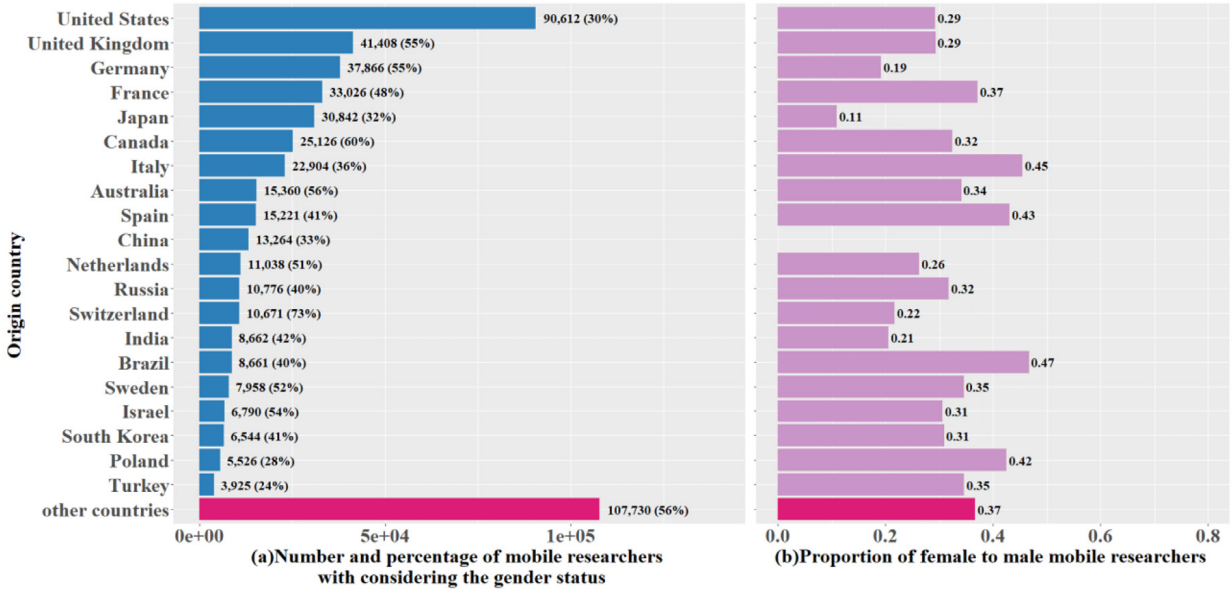

Fig. 3. (a) Distribution of mobile researchers by origin country grouped by 20 top countries with the largest number of mobile researchers and remained countries. The numbers show the number of mobile researchers and percentages in parentheses represent the percentage of mobile to all researchers (b) The proportion of female to male mobile researchers among countries, for all countries, the proportion is less than one and it means in general women participate in international mobility less than men. Note: given the fact that gender detection is very weak for Chinese names and we have tried to eliminate those names as much as possible from the raw data, we present no proportion for this country in part (b).

researchers for both genders. This complements the results of study by Aman (2020), which found knowledge transmitting by researchers in these fields are relatively high. The results in this figure agree with the results of prior studies (Bauder, 2020; Jöns, 2011; Leemann, 2010; Ryazanova & McNamara, 2019).

Fig. 3 displays 20 top origin countries of researchers with the number and percentage of mobile researchers (Fig. 3 (a)) and the share of women among mobile researchers (Fig. 3 (b)). Fig. 3. (a) reveals the lowest and highest percentage of mobile researchers for Turkey and Switzerland, respectively. From Fig. 3(b), we observe that women from Japan have the lowest proportion in mobility among all countries. Interestingly Brazil, one of the BRICS countries, has the highest ratio of female mobile researchers. To show what bias can yield the gender detection method we build another dataset and apply all filters mentioned in the section 'Data and methods' except filtering for gender status. This dataset involves 1,878,545 authors. We compare Fig. 3 (a) with the result for this dataset and present it in Appendix A.

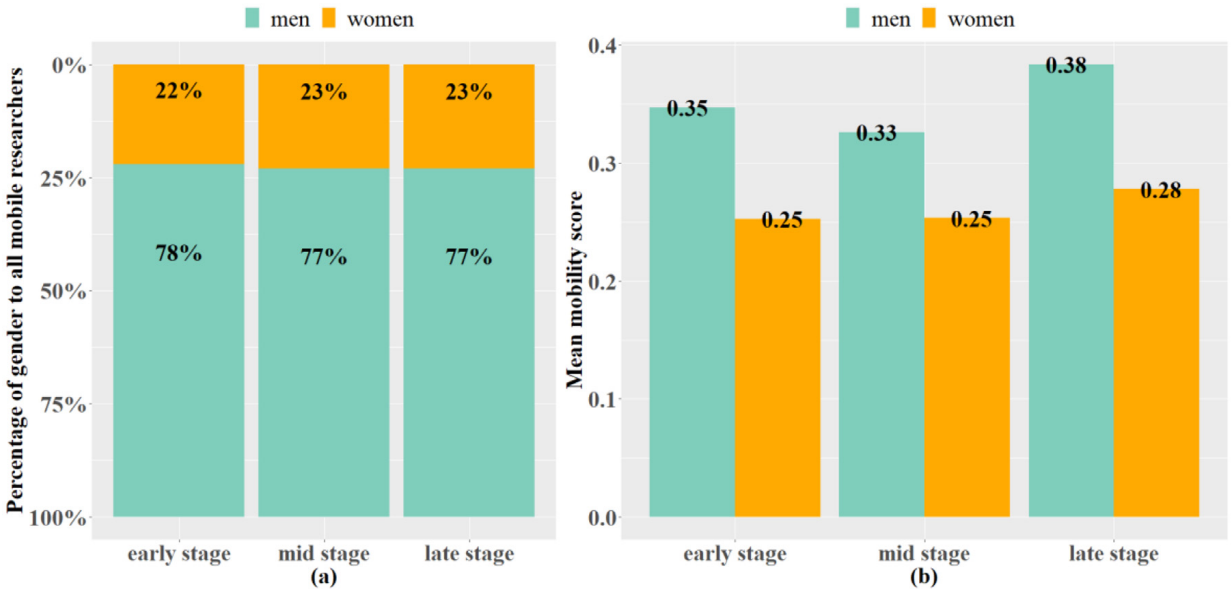

Fig. 4. (a) Percentage of mobile men and women to all mobile researchers in three career stages. (b) Average mobility score of men and women in different career stages, the percentage numbers are the percentage change of mobility score from early to current career stage.

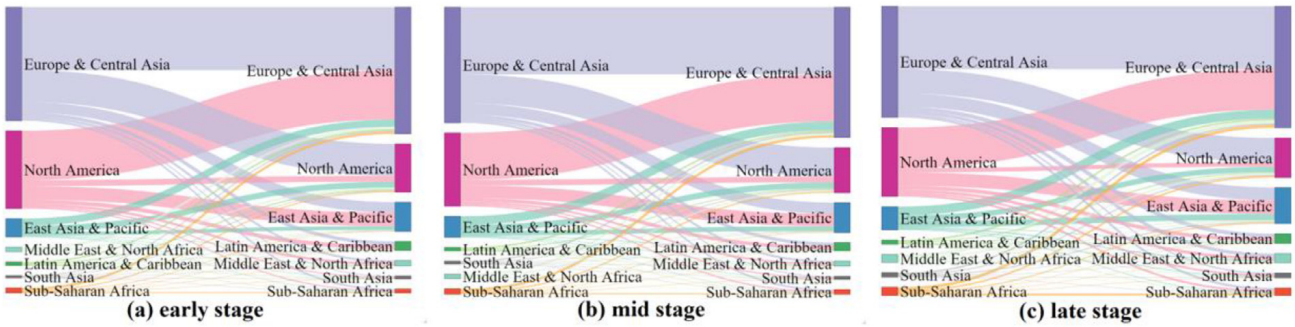

Fig. 5. Movement between and within world regions. Each movement is a changing of the country affiliation for an author at a particular career stage. The right and left sides in each stage are regions before and after movement, respectively. Most movements happen between Europe & Central Asia for all career stages, but the tendency to move to this region has decreased in the late stage. Also, countries in this region are the most popular destination for authors from other regions. Movement from and to two regions "Middle East & North Africa" and Sub-Saharan Africa" have increased in the late career stage.

We calculate the mean mobility score of mobile researchers for both genders at each career stage. Fig. 4 (a) shows the proportion of mobile men and women is stable through career stages and just one percent point decrease for men at mid and late career stages. Also, Fig. 4 (b) indicates the highest mobility score in late-career stage for both genders, which disagrees with previous work by Leemann (2010) who found that probability to go abroad reduces by each age year for both men and women which is related to having family and children. After a decrease in movement in the mid-career stage for men, their mobility score grows again in the late career stages.

To display the flow of international mobility, we counted each movement among all authors and their career stages from one country to another. Fig. 5 displays the aggregated results of the flow of mobility between continents and across career stages. These results show a higher tendency to move to "Europe & Central Asia" and "North America" in the early stage which slowly inclines to other regions in the late stage.

4.3. Factors associated with mobility

In order to understand which socio-demographic characteristics associate with mobility and how mobility correlates with scholars' academic performance, we utilize two regression models. The first model shows the factors that relate to mobility at mid and later career stages considering the history of researchers at earlier career stages. Poisson regression is used for this model. Table 4 reports exponentiated coefficient, $\exp(\beta)$ of independent variables. In this model, values greater (less) than unity indicate positive

Table 4
The results of Poisson regression. The dependent variable is the mobility score at career stage s_i where i = 2,3. The outputs are Odd ratio, (exp(β)). (1-exp(ß)) shows the percentage change of dependent variable per unit increase in an independent variable, therefore numbers greater/less than one indicate a positive/negative correlation between variables. For interpreting the interaction between variables, we should multiply odd ratio of the related interaction to the odd ratio of both variables. For example, the odds for "Latin America & Caribbean and female" equals to 0.996*1.56*0.75 = 1.16 and the value more than one means that the likelihood of mobility for female researchers from Latin America & Caribbean is higher than males from North America.

	exp(β) at s_1	exp(β) at s_2	exp(β) at s_3
Intercept	0.26***(−280)	0.02***(−236)	0.06***(−247.7)
Independent variables			
Having international co-author at career stage $s_i - 1$	——	7.9***(137.3)	3.53***(130)
Gender:			
Male	Reference	Reference	Reference
Female	0.75***(−25.2)	0.84***(−35.6)	0.81***(−13.3)
Region of origin country (Average GDP per Capita):			
North America (43,207)	Reference	Reference	Reference
Latin America & Caribbean (13,110)	1.56***(43.7)	1.46***(28.01)	1.33***(16.6)
Europa & Central Asia (30,223)	1.77***(119.1)	1.50***(68.5)	1.26***(31.5)
Sub-Saharan Africa (6,197)	2.00***(37.5)	2.17***(36.12)	2.35***(30.7)
Middle East & North Africa (21,981)	1.57***(38.8)	1.54***(28.55)	1.53***(23.4)
South Asia (3,390)	1.20***(12.9)	1.74***(36.1)	1.49***(19.8)
East Asia & Pacific (26,783)	1.11***(16.4)	1.32***(44)	0.90***(−10.0)
Interaction between Region of origin country and Gender:			
North America and male	Reference	Reference	Reference
Latin America & Caribbean and female	0.87***(−6.9)	0.93** (−2.6)	0.89***(−4.2)
Europa & Central Asia and female	0.996 (−0.37)	1.01 (0.7)	0.87***(−8.8)
Sub-Saharan Africa and female	1.21***(4.60)	1.11* (2.06)	1.08 (1.42)
Middle East & North Africa and female	0.89***(−3.9)	0.99 (-0.37)	0.86***(−3.7)
South Asia and female	1.04 (1.14)	1.07 . (1.7)	1.00 (0.1)
East Asia & Pacific and female	1.32***(20.1)	1.27***(13.4)	1.26***(10.84)
Field: Physical Sciences	Reference	Reference	Reference
Life Sciences	0.93***(−15.6)	1.07***(11.75)	0.83***(−21.7)
Health Sciences	0.93***(−18.1)	0.94***(−11.48)	1.01 (1.5)
Social Sciences	0.47***(−68.0)	0.87***(−12)	1.45***(33.19)
Interaction between Field and Gender:			
Physical Sciences and male	Reference	Reference	Reference
Life Sciences and female	0.86***(−15.4)	0.88** (−2.9)	0.97 . (5.2)
Health Sciences and female	0.998 (−0.16)	0.96***(−8.6)	1.08***(−1.74)
Social Sciences and female	0.96 . (−1.9)	0.90***(−3.96)	0.96 . (−1.74)
Pseudo R^2	0.03	0.07	0.05
N	1,183,662	919,692	784,857

Significant codes: p<0.1,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
z-values of coefficients in parentheses.

(negative) correlation with the dependent variable. To avoid confounding effects in this regression, we select only those researchers at career stage s_i who were non-mobile at the past career stages. Thus, we can observe how those independent variables are related to mobility of researchers. Results show that having international co-authors at the previous career stage is the most significant factor in increasing the probability of international mobility. This complements the research by Bauder (2020) showing that international social capital facilitates international mobility. Besides, we observe that at the earlier career stages, the tendency for mobility in Social Sciences is less than in other fields, but social scientists are most likely to be mobile at the late-career stage. Researchers from the North America region have the highest GDP per capita and lowest probability of mobility across regions, respectively. Among all regions, Sub-Saharan Africa with a relative low GDP per capita has the most engaged researchers in mobility. Also, Chinchilla-Rodríguez et al. (2021) reported the highest participation rate of international collaboration for this region. Interaction between gender and two other independent variables in this table, indicates the extent of gender inequality in mobility for different regions or scientific fields. For example, regarding the interaction between gender and region of origin, females from Sub-Saharan Africa are more likely to be mobile than those from other regions. Given the low value of R^2, our results suggest that in general it is not easy to predict the determinants of mobility. We denote that the very low p-value in regression results can be affected by the size of the sample. Hence, its representativeness for significance of statistical results may suffer in large-N settings. Therefore, only relying on low p-value is not sufficient to support the hypotheses. To reduce the p-value problem Lin et al. (2013) and Khalilzadeh and Tasci (2017) suggested some solutions (e.g., presenting effect size, reporting confidence intervals and using charts). According to the

Table 5
OLS regression to estimate PPY, CPP and COPP. Dependent variables are log-transformed, therefore exponentiated coefficient of independent variables are presented. (1-exp(ß)) shows the percentage change of dependent variable per unit increase in an independent variable, therefore numbers greater/less than one indicate a positive/negative correlation between variables.

	PPY		*CPP*		*COPP*	
	Men	Women	Men	Women	Men	Women
	exp(β)	exp(β)	exp(β)	exp(β)	exp(β)	exp(β)
Intercept	1.43***(223.1)	1.44***(125.7)	7.57***(1310)	9.68***(846.5)	1.68***324.4)	1.77***(193.3)
Independent variables:						
Mobility score	1.15***(210.4)	1.15***(98.8)	1.06***(85.5)	1.05***(37.3)	1.05***(70.1)	1.06***(38.03)
***Field*:**						
Physical Sciences	Reference	Reference	Reference	Reference	Reference	Reference
Health Sciences	0.99 (-0.45)	0.87***(-40.7)	1.5***(210.7)	1.47***(122.7)	2.24***(388.8)	2.25***(231.3)
Life Sciences	0.87***(−57.1)	0.74***(−80.8)	2.2***(351)	2.09***(217.0)	1.74***(236.7)	1.77***(151.1)
Social Sciences	0.59***(−136.9)	0.59***(−89.1)	0.84***(−46.2)	0.92***(−16.5)	0.45***(−201.31)	0.55***(−100.25)
Career stage of first mobility:						
Non-mobile	Reference	Reference	Reference	Reference	Reference	Reference
Early stage	1.37***(105.7)	1.24***(42.7)	1.18***(58.9)	1.13***(26.7)	0.92***(−26.1)	0.96***(−8.33)
Mid-stage	1.33***(89.6)	1.22***(38.6)	1.27***(78.0)	1.19***(35.7)	0.98***(−6.2)	0.99*** (−2.5)
Late stage	1.41***(100.9)	1.36***(56.2)	1.26***(71.4)	1.18***(33.3)	1.04***(11.7)	1.05*(8.4)
R-square	0.20	0.16	0.18	0.17	0.24	0.24
Residual standard error	0.82	0.77	0.78	0.70	0.81	0.78
N	863,595	320,744	860,83	320,114	859,568	319,782

Significant codes: <* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
z-values in parentheses.

proposed recommendations by Lin et al. (2013), we report the coefficients and their confidence intervals for Table 4 and Table 5 in the Appendix B.

4.4. Impact of mobility on scholars' career

The second model demonstrates the relationship between mobility and co-author diversity, productivity, and citation. We chose ordinary least squares (OLS) regression for this purpose. PPY, CPP and COPP are dependent variables in this model. To reduce the residual standard error of results, we used the log transformation of dependent variables.

Table 5 shows the regression results for men and women. Again, the results show exponentiated coefficient, exp(β) of independent variables. We observe a similar effect of mobility for both genders. The results show that mobility has better outcomes in terms of PPY, CPP for all mobile researchers and greater COPP for those who start mobility at the late career stage.

To estimate the effect of mobility by accounting for the covariates, we use propensity score matching, which calculates the causal effects of the treatment (mobility). To this end, we selected the 1:1 nearest-neighbor matching method to pair mobile with non-mobile authors that share similar characteristics and close research attributes at the stage before mobility. Then we compare the scientific outcomes regarding PPY, CPP, and COPP for these two groups. We select nine covariances for matching. To control the effect of the starting stage of mobility, we divide mobile researchers according to the career stage in which they start to be mobile and paired them separately with non-mobile researchers.

Table 6 shows these covariates and Standardized Mean Difference (SMD) to assess the covariance balance before and after matching. Stuart et al. (2013) recommended threshold equals to 0.1 for asserting covariance balance between two groups. All SMDs are less than the threshold after matching and it shows that all covariates were balanced. Table 7 reports the result of t-tests which reveals higher PPY and CPP for mobile researchers.

4.5. Position of mobile researchers in the co-authorship networks

In this section, we will compare the centrality scores of mobile and non-mobile researchers in the co-authorship network. Because the structure of co-authorship networks is dynamic (it changes yearly) and the position of authors in the network may change by their career development, we compare the centrality scores of researchers at the same career stage and for each year separately. To this end, we chose the researchers who started publishing at the same year and tracked their centrality scores for the future years. We represent the results of analyses for authors who started publishing in the arbitrary year (2000), but it is generalizable to other years.

We divided the data into non-mobile and mobile groups. From the mobile group, we selected authors who have mobility at the early or mid-career stage. To calculate the yearly centrality score of each group, first, we computed the average centrality of each author across networks which she/he belongs to them and considered it as her/his centrality score. Then we calculated the average score among researchers.

Table 6
The Balance report before and after matching treatment (mobile) and control (non-mobile) groups; mobile researchers with different stage of stating mobility have been paired separately with non-mobile researchers through nine covariances.

First mobility stage	s_1		s_2		s_3	
	SMD Before matching	SMD After matching	SMD Before matching	SMD After matching	SMD Before matching	SMD After matching
Covariance:						
Gender	0.16	0.008	0.10	0.012	0.1	0.006
Career age	0.39	0.036	0.28	0.007	0.50	0.008
Region	0.33	0.041	0.22	0.019	0.16	0.013
GPD per capita of the first affiliation country	0.05	0.064	0.07	0.016	0.02	0.020
Field	0.20	0.021	0.13	0.005	0.08	0.018
Having an international co-author	0.65	0.001	0.6	<0.001	0.58	0.004
CPP at the previous stage and before mobility	-	-	0.14	0.001	13.6	0.016
PPY at the previous stage and before mobility	-	-	0.36	0.003	1.85	0.006
COPP at the previous stage and before mobility	-	-	0.17	0.006	0.17	0.02
Sample sizes:						
Control (non-mobile)	679,456	260,169	679,456	133,420	637,446	90,922
Treated (mobile)	262,294	260,169	133,829	133,420	90,977	90,922
Paired matched	-	260,169	-	133,420	-	90,922

Table 7
The result of paired-samples t-test. Positive Mean diff. shows a higher outcome (PPY, CPP and COPP) for mobile researchers. PPY and CPP for mobile researchers are higher than non-mobile researchers regardless of the stage of starting mobility. Mobile researchers have the better outcome in terms of COPP than non-mobile researchers, only if they start mobility at the late career stage.

First mobility stage	s_1			s_2			s_3		
	Mean Diff.	SE	t (*p*-value)	Mean Diff.	SE	t (*p*-value)	Mean Diff.	SE	t (*p*-value)
PPY	1.93	0.01	198.1 (***)	0.79	0.01	83.5 (***)	0.71	0.01	69.9 (***)
CPP	2.91	0.01	94.4 (***)	2.5	0.04	54.6 (***)	1.8	0.05	32.9 (***)
COPP	−0.68	0.06	−11.5 (***)	−0.23	0.07	−3.1 (***)	0.13	0.1	1.2 (***)
n	260,169			133,420			90,922		

Significant codes: p<0.1,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

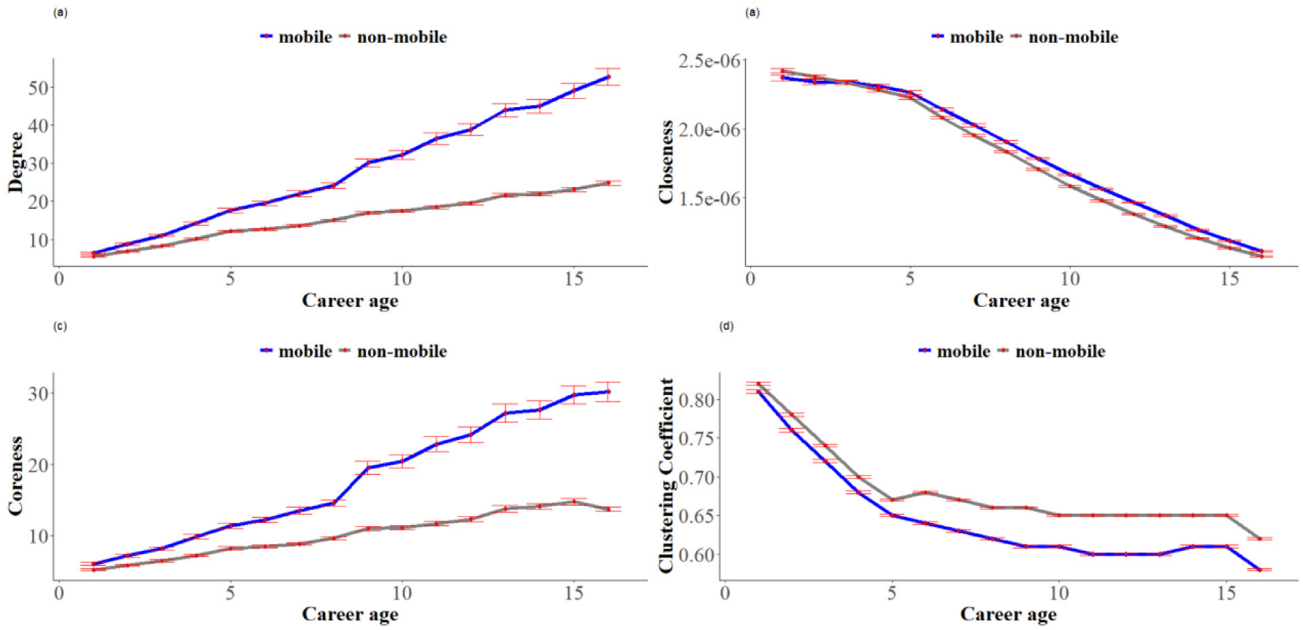

Fig. 6. Centrality measures of mobile and non-mobile researchers. Number of observations are 18,734 and 29,122 for mobile and non-mobile groups respectively. Error bars show standard errors. Career age has a range of 1 to 16 (first publication year and 15 years after that).

Fig. 6 shows yearly centrality scores of selected researchers. For mobile researchers, we observe a growing trend in degree and coreness by increasing the career age. These two observations suggest that not only mobility increases the social capital of the scholars (degree), but also it helps to position the scholars into the core of the community which would allow them to get better access to highly influential people (Kitsak et al., 2010).

In contrast, the lower clustering coefficient of mobile researchers indicates that by increasing their degree, they diversify their co-authors and belong to various communities and thus they act more as bridges between the communities.

We note that the relationship between mobility and scholars' careers that we observed in Table 5 could be to some extent explained by the collaboration networks (Jadidi et al., 2018) as we see in Fig. 6. Given the significant influence of mobility on centrality and position of researchers in their collaboration network, this also confirms the importance of mobility in advancing one's career.

5. Discussion and conclusion

In this paper, we have identified and explored many facets of academic mobility of males and females that relate to 1) mobility differences across nationalities, disciplines, and career stages, 2) the international mobility flow 3) the factors that associate with mobility, 4) relationship between mobility and scholars' publications and citations, 5) different position of mobile and non-mobile scholars in their collaboration network.

Our findings regarding the first research question revealed that in the world of academia, international mobility of women is less than men, which is consistent with the past studies (Jöns, 2011; Leemann, 2010; Myers & Griffin, 2019; Toader & Dahinden, 2018) We observe that the mobility score of both genders has the lowest level in the early and mid-career stage, which can be related to the time that families should focus more on preschool children than going abroad. The results demonstrate that gender inequality in international mobility exists for all scientific fields and women are underrepresented particularly in Physical Sciences. Physical Sciences and Life Sciences have the most mobile researchers for both genders, which agrees with the findings by

Subbotin and Aref (2021). Our findings show that mobility of researchers decreases with increasing the career age for both genders which is in line with the results of the study by Leeman (2010) The regression result of the first model pointed to the importance of having international collaboration for mobility in future. Also, we observed various tendencies for mobility across geographic regions, scientific fields, and career stages. Researchers from the Sub-Saharan Africa region with relatively low-income levels are most likely to move. Females from this region have the highest probability to be mobile as well. It seems that getting better funding opportunities is a motivation for researchers to go abroad, as Hunter et al. (2009) found that top scholars head to countries with high R&D spending levels. From the results, authors in the Social Sciences have the least and most probability for mobility in the mid and late-career stages, respectively. Receiving postdoctoral positions in this field might be hard for international researchers and a reason for low participation in the mid-career stage. Perhaps these researchers should gain academic experiences at the earlier career stage in their own country to increase their chance of receiving an international position.

Regarding the second research question, the results of our second regression analysis demonstrated the relationship between mobility with the performance, success, and number of unique co-authors of researchers. We compared the outcomes of mobility for men and women and found no clear difference between them. We observed that although mobility improves the performance and success of researchers regardless of the stage of starting mobility. We used PSM to draw the causal reference and minimize the confounding bias in examining the effect of mobility. By matching mobile with non-mobile researchers with the similar characteristics, we compared their scientific outcome and found the positive impact of mobility on scientific performance and success.

Finally, the co-authorship networks of mobile and non-mobile researchers reveals that mobile researchers have a more diverse social capital and better access to influential scholars in their network compared to non-mobile researchers.

The results carry potential insights for policy-makers concerned with the issues and inequalities in this area to provide fair opportunities at the proper phase of the academic career for all researchers who desire to communicate and collaborate with their international colleagues.

6. Limitations and future work

This study has some limitations that should be noted. We used bibliometric data to indicate the mobility and career age of the authors. We used Scopus author ID to associate authors with their publications. Although this author ID system has high precision to assign the set of articles to a particular author ID, an author ID may not cover the complete article set of an author and result in multiple author IDs for one author in Scopus (Moed et al., 2013). It causes problems in tracking the affiliation of authors whose works have been split into multiple author IDs and may underestimate the mobility of authors. Besides, for publications that aren't indexed in Scopus, this approach can lead to errors in specifying the country of origin, mobility score, and researcher's career stage. Next, we have assumed a fixed period of early and mid-career stages for all authors. These can vary depending on the study field, career interruption, or type of affiliate organization. Although a comprehensive discussion of the few related works is included in Section 'Data & Methods', which argues in favor of fixed periods, future work would need to consider the robustness of this approach. Also, we didn't distinguish between academic researchers and those outsides academia (industry or government researchers) who have varying average number of publications and career advancement. In addition, this approach cannot indicate temporary mobilities, such as research visits, that the host countries are not considered as the author's affiliation. The collaboration pattern of these may differ from other mobile researchers. For example, they are less likely to lose co-authors from their country of origin. Moreover, our analyses don't contain authors for whom we could not detect their gender status, because the applied method has some weaknesses for common names especially Chinese names.

In this study we built the collaboration networks for each discipline separately. This might affect the network position of scholars who work on interdisciplinary topics. In the future, more attention is needed to study the influence of mobility and collaboration networks on researchers with interdisciplinary backgrounds. Also, we didn't control for the size of the network or communities that researchers belong to (one can start academic life in an organisation with high density and interaction between authors, others in a much more isolated one) and that may impact the future career perspectives.

In this paper, we considered all publications without regarding the position of authors in the paper. By this means we make statements about all effects of collaboration not only 'prestigious effects' that lead to becoming a first author or a last author. In future, it would be interesting to consider the influence of mobility on the authorship position.

In this study we didn't consider destination countries of researchers and different kinds of mobility (e. g. immigrants, returnees). Including the different individual-level mobility introduced by Robinson-Garcia et al. (2019) in analyses leads to more comprehensive results and give us a better knowledge about the motivation of movement, advantages, and disadvantages of mobility for the origin and destination countries.

We looked at the mobility of authors at the country level. Investigating the mobility in scientific fields and its relation to geographic mobility will give us a better understanding of the impact of mobility on knowledge transfer between scientific fields. Also, it would be interesting to analyse the co-authorship networks among different kinds of mobile researchers to discover the differences in collaboration pattern and their impact on other researchers.

During the COVID-19 pandemic, international mobility has been limited and many countries have reduced international academic positions. In the meantime, telecollaboration has become more popular, which makes international collaboration easier. Investigating the impact of the pandemic on the scholars' career and comparing collaboration patterns before and after the pandemic, will give us an understanding of the importance of physical mobility. Perhaps virtual mobility can be an alternative for scientists.

Declaration of Competing Interest

We declare we have no competing interests.

CRediT authorship contribution statement

Fakhri Momeni: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization. **Fariba Karimi:** Conceptualization, Resources, Writing – review & editing. **Philipp Mayr:** Writing – review & editing, Project administration, Funding acquisition. **Isabella Peters:** Writing – review & editing, Project administration, Funding acquisition. **Stefan Dietze:** Supervision, Writing – review & editing.

Acknowledgments

This work is supported by BMBF project OASE, grant number 01PU17005A. We acknowledge the support of the German Competence Center for Bibliometrics (grant: 01PQ17001). We are thankful to Nina Smirnova for her aid with analysing data. We thank Dr. Matthias Raddant and Dr. Anne-Kathrin Stroppe for helpful comments.

Appendix A

Figure A

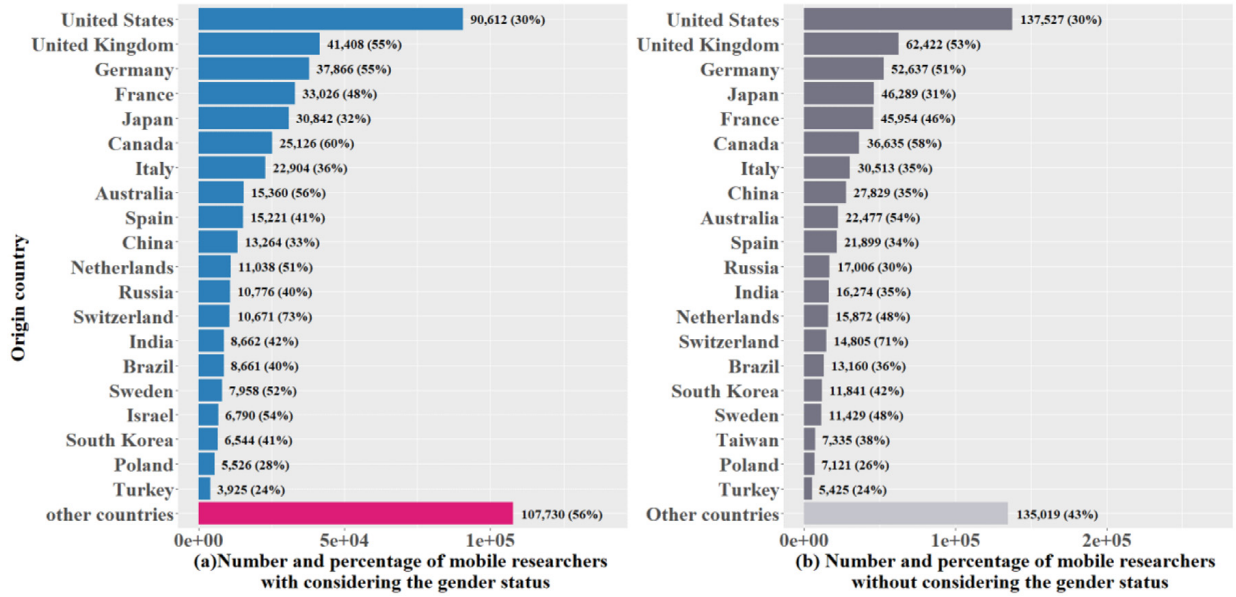

Fig. A. Comparison of distribution of mobile researchers by origin country for two samples: (a) The sample with gender status and (b) The sample without gender status. Some Asian countries such as China, South Korea and India have a lower ranking in the sample with gender status.

Appendix B

Table B1 and B2

Table B1
Coefficients and confidence intervals for the Poisson regression presented in Table 4.

	s_1			s_2			s_3		
	β*	2.5% CI**	97.5% CI***	β	2.5% CI	97.5% CI	β	2.5% CI	97.5% CI
Intercept	−1.32	−1,33	−1,32	−3.7	−3.77	−3.71	−2.8	−2.86	−2.81
independent variables									
Gender:									
Male	Reference			Reference			Reference		
female	−0.28	−0.30	−0.21	−0.17	−0.21	−0.14	−0.21	−0.25	-0.18
Having international co-author at career stage $s_i - 1$	——			2.1	2.04	2.1	1.26	1.24	1.28
Region of origin country (Average GDP per Capita):									
North America (43,207)	Reference			Reference			Reference		
Latin America & Caribbean (13,110)	0.44	0.43	0.47	0.37	0.35	0.40	0.28	0.24	0.31
Europa & Central Asia (30,223)	0.57	0.56	0.58	0.40	0.39	0.41	0.23	0.21	0.24
Sub-Saharan Africa (6,197)	0.69	0.65	0.73	0.77	0.72	0.82	0.85	0.80	0.91
Middle East & North Africa (21,981)	0.45	0.42	0.47	0.43	0.40	0.46	0.42	0.39	0.46
South Asia (3,390)	0.18	0.15	0.21	0.55	0.52	0.59	0.40	0.36	0.44
East Asia & Pacific (26,783)	0.10	0.10	0.12	0.28	0.26	0.29	−0.1	−0.12	−0.08
Interaction between Region of origin country and Gender:									
North America and male	Reference			Reference			Reference		
Latin America & Caribbean and female	−0.14	−0.18	−0.01	−0.07	−0.12	−0.02	−0.12	−0.19	−0.07
Europa & Central Asia and female	−0.003	−0.02	0.01	0.00	−0.02	0.03	−0.13	−0.16	−0.10
Sub-Saharan Africa and female	0.18	0.10	0.26	0.10	0.00	0.2	0.08	−0.03	0.19
Middle East & North Africa and female	−0.11	−0.16	−0.06	−0.01	−0.08	0.05	−0.14	−0.22	−0.07
South Asia and female	0.04	−0.03	0.1	0.07	−0.01	0.15	0.004	−0.09	0.09
East Asia & Pacific and female	0.28	0.25	0.30	0.24	0.20	0.27		0.18	0.26
Field : Physical Sciences	Reference			Reference			Reference		
Life Sciences	−0.08	−0.09	−0.07	0.08	0.07	0.09	−0.19	−0.20	−0.17
Health Sciences	−0.08	−0.08	−0.07	−0.03	−0.05	−0.02	0.01	0.00	0.02
Social Sciences	−0.7	−0.8	−0.70	−0.11	−0.13	−0.08	0.37	0.35	0.39
Interaction between Field and Gender:									
Physical Sciences and male	Reference			Reference			Reference		
Life Sciences and female	−0.001	−0.02	0.17	−0.04	−0.07	−0.01	−0.03	−0.07	0.003
Health Sciences and female	−0.15	−0.17	−0.13	−0.12	−0.15	−0.09	0.08	0.05	0.11
Social Sciences and female	−0.04	−0.1	0.0	−0.10	−0.15	−0.05	−0.04	−0.08	0.00

* Coefficient
** Coefficient at confidence interval: 2.5%
*** Coefficient at confidence interval: 97.5%

Table B2
Coefficients and confidence intervals for the OLS regression presented in Table 5.

	PPY						CPP						COPP					
	Men			Women			Men			Women			Men			Women		
	β*	2.5% CI**	97.5% CI***	β	2.5% CI	97.5% CI	β	2.5% CI	97.5% CI	β	2.5% CI	97.5% CI	β	2.5% CI	97.5% CI	β	2.5% CI	97.5% CI
Intercept	0.36	0.36	0.36	0.37	0.36	0.37	2.02	2.02	2.03	2.27	2.26	2.27	0.52	0.52	0.52	0.57	0.57	0.58
independent variables:																		
mobility score	0.15	0.14	0.14	0.14	0.14	0.14	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06
***Field*:**																		
Physical Sciences	Reference			Reference			Reference			Reference			Reference			Reference		
Health Sciences	−0.001	−0.005	0.003	−0.14	−0.15	−0.13	0.42	0.41	0.42	0.39	0.38	0.39	0.81	0.80	0.81	0.80	0.8	0.81
Life Sciences	−0.13	−0.14	−0.13	−0.30	−0.31	−0.29	0.79	0.79	0.80	0.74	0.73	0.74	0.56	0.55	0.56	0.56	0.56	0.58
Social Sciences	−0.52	−0.53	−0.51	−0.51	−0.53	−0.50	−0.17	−0.18	−0.16	−0.09	−0.1	−0.07	−0.78	−0.79	−0.78	−0.6	−0.61	−0.58
Career stage of first mobility:																		
Non-mobile	Reference			Reference			Reference			Reference			Reference			Reference		
Early stage	0.31	0.31	0.32	0.22	0.21	0.23	0.17	0.16	0.17	0.12	0.11	0.13	−0.08	−0.08	−0.07	0.04	−0.05	−0.03
Mid-stage	0.28	0.28	0.29	0.20	0.19	0.21	0.24	0.23	0.24	0.17	0.16	0.18	−0.04	−0.03	-0.05	−0.01	−0.002	−0.003
Late stage	0.34	0.34	0.35	0.31	0.30	0.32	0.23	0.23	0.24	0.17	0.16	0.18	0.04	0.03	0.05	0.05	0.04	0.06

*Coefficient
**Coefficient at confidence interval: 2.5%
***Coefficient at confidence interval: 97.5%

References

Abbasi, A., Hossain, L., Uddin, S., & Rasmussen, K.J. (2011). Evolutionary dynamics of scientific collaboration networks: Multi-levels and cross-time analysis. *Scientometrics, 89*, 687–710.

Almansour, S. (2015). The challenges of international collaboration: Perspectives from Princess Nourah Bint Abdulrahman University. *Cogent Education, 2*, Article 1118201.

Aman, V. (2018a). A new bibliometric approach to measure knowledge transfer of internationally mobile scientists. *Scientometrics, 117*, 227–247. 10.1007/s11192-018-2864-x.

Aman, V. (2018b). Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates. *Scientometrics, 117*, 705–720. 10.1007/s11192-018-2895-3.

Aman, V. (2020). Transfer of knowledge through international scientific mobility: Introduction of a network-based bibliometric approach to study different knowledge types. *Quantitative Science Studies*, 1–17. 10.1162/qss_a_00028.

Arrieta, O.A., Pammolli, F., & Petersen, A.M. (2017). Quantifying the negative impact of brain drain on the integration of European science. *Science Advances, 3*, Article e1602232.

Bäker, A. (2015). Non-tenured post-doctoral researchers' job mobility and research output: An analysis of the role of research discipline, department size, and coauthors. *Research Policy, 44*, 634–650. 10.1016/j.respol.2014.12.012.

Barabási, A.L., & Musciotto, F. (2019). Science of success: An introduction. *Computational Social Science and Complex Systems, 203*, 57. 10.3254/190005.

Bauder, H. (2020). *International mobility and social capital in the academic field*. Minerva. 10.1007/s11024-020-09401-w.

Bauder, H., Hannan, C.-A., & Lujan, O. (2017). International experience in the academic field: knowledge production, symbolic capital, and mobility fetishism. *Population, Space and Place, 23*, e2040. 10.1002/psp.2040.

Bazeley, P. (2003). Defining'early career'in research. *Higher Education, 45*, 257–279.

Bedenlier, S. (2018). The impact of my work would be greater here than there': Implications of the international mobility of colombian academics. *Research in Comparative and International Education, 13*, 378–396. 10.1177/1745499918784681.

Bernard, M., Bernela, B., & Ferru, M. (2021). Does the geographical mobility of scientists shape their collaboration network? A panel approach of chemists' careers. *Papers in Regional Science, 100*, 79–99.

Bhandari, R. (2017). Women on the move: The gender dimensions of academic mobility. *Women on the move: The gender dimensions of academic mobility*. New, York, NY: Institute of International Education Retrieved from https://www.iie.org/-/ .

Bosanquet, A., Mailey, A., Matthews, K.E., & Lodge, J.M. (2017). Redefining 'early career'in academia: A collective narrative approach. *Higher Education Research & Development, 36*, 890–902. 10.1080/07294360.2016.1263934.

Böttcher, L., Araújo, N.A., Nagler, J., Mendes, J.F., Helbing, D., & Herrmann, H.J. (2016). Gender gap in the ERASMUS mobility program. *PLoS One, 11*, Article 0149514. 10.1371/journal.pone.0149514.

Bozeman, B., & Gaughan, M. (2011). How do men and women differ in research collaborations? An analysis of the collaborative motives and strategies of academic researchers. *Research Policy, 40*, 1393–1402.

Cañibano, C., Otamendi, F.J., & Solís, F. (2011). International temporary mobility of researchers: a cross-discipline study. *Scientometrics, 89*, 653–675. 10.1007/s11192-011-0462-2.

Cañibano, C., Otamendi, J., Andújar, I., & others (2008). Measuring and assessing researcher mobility from CV analysis: The case of the Ramón y Cajal programme in Spain. *Research Evaluation, 17*, 17–31. 10.3152/095820208X292797.

Caniglia, G., Luederitz, C., Groß, M., Muhr, M., John, B., Keeler, L.W., … Lang, D. (2017). Transnational collaboration for sustainability in higher education: Lessons from a systematic review. *Journal of Cleaner Production, 168*, 764–779. 10.1016/j.jclepro.2017.07.256.

Carr, P.L., Gunn, C.M., Kaplan, S.A., Raj, A., & Freund, K.M. (2015). Inadequate progress for women in academic medicine: Findings from the national faculty study. *Journal of Women's Health, 24*, 190–199.

Chinchilla-Rodríguez, Z., Bu, Y., Robinson-García, N., Costas, R., & Sugimoto, C.R. (2018). Travel bans and scientific mobility: utility of asymmetry and affinity indexes to inform science policy. *Scientometrics, 116*, 569–590. 10.1007/s11192-018-2738-2.

Chinchilla-Rodríguez, Z., Liu, J., & Bu, Y. (2021). Patterns of knowledge diffusion via research collaboration on a global level. In Proceedings of the 18th International Conference on Scientometrics and Informetrics (ISSI 2021)*, (S. 269-280)*.

Chinchilla-Rodríguez, Z., Miao, L., Murray, D., Robinson-García, N., Costas, R., & Sugimoto, C.R. (2017). Networks of international collaboration and mobility: a comparative study. *Networks of international collaboration and mobility: A comparative study*.

Cohen, S., Hanna, P., Higham, J., Hopkins, D., & Orchiston, C. (2020). Gender discourses in academic mobility. *Gender, Work & Organization, 27*, 149–165. 10.1111/gwao.12413.

De Benedictis, L., & Leoni, S. (2020). Gender bias in the Erasmus students network. *Gender bias in the Erasmus students network*.

De Cleyn, S.H., Braet, J., & Klofsten, M. (2015). How human capital interacts with the early development of academic spin-offs. *International Entrepreneurship and Management Journal, 11*, 599–621. 10.1007/s11365-013-0294-z.

El-Ouahi, J., Robinson-García, N., & Costas, R. (2021). Analysing scientific mobility and collaboration in the middle East and North Africa. *Quantitative Science Studies, 2*(3), 1023–1047. 10.1162/qss_a_00149.

Freeman, L.C. (1978). Centrality in social networks conceptual clarification. *Social Networks, 1*, 215–239. 10.1016/0378-8733(78)90021-7.

Gingras, Y., Lariviere, V., Macaluso, B., & Robitaille, J.-P. (2008). The effects of aging on researchers' publication and citation patterns. *PloS One, 3*, e4048.

Halevi, G., Moed, H.F., & Bar-Ilan, J. (2016). Researchers' mobility, productivity and impact: Case of top producing authors in seven disciplines. *Publishing Research Quarterly, 32*, 22–37. 10.1007/s12109-015-9437-0.

Hunter, R.S., Oswald, A.J., & Charlton, B.G. (2009). The elite brain drain. *The Economic Journal, 119*, F231–F251. 10.1111/j.1468-0297.2009.02274.x.

Jadidi, M., Karimi, F., Lietz, H., & Wagner, C. (2018). Gender disparities in science? Dropout, productivity, collaborations and success of male and female computer scientists. *Advances in Complex Systems, 21*, Article 1750011.

Jayachandran, S. (2015). The roots of gender inequality in developing countries. *Economics, 7*, 63–88. 10.1146/annurev-economics-080614-115404.

Jöns, H. (2011). Transnational academic mobility and gender. *Globalisation, Societies and Education, 9*, 183–209. 10.1080/14767724.2011.577199.

Karimi, F., Mayr, P., & Momeni, F. (2019). Analyzing the network structure and gender differences among the members of the Networked Knowledge Organization Systems (NKOS) community. *International Journal on Digital Libraries, 20*, 231–239.

Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., & Strohmaier, M. (2016). Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th international conference companion on world wide web* (pp. 53–54). 10.1145/2872518.2889385.

Kawashima, H., & Tomizawa, H. (2015). Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. *Scientometrics, 103*, 1061–1071.

Khalilzadeh, J., & Tasci, A.D. (2017). Large sample size, significance level, and the effect size: Solutions to perils of using big data for academic research. *Tourism Management, 62*, 89–96.

Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., & Makse, H.A. (2010). Identification of influential spreaders in complex networks. *Nature Physics, 6*, 888–893. 10.1038/nphys1746.

Kong, H., Martin-Gutierrez, S., & Karimi, F. (2021). First-mover advantage explains gender disparities in physics citations. *arXiv preprint arXiv:2110.02815*.

Laudel, G., & Bielick, J. (2019). How do field-specific research practices affect mobility decisions of early career researchers? *Research Policy, 48*, Article 103800. 10.1016/j.respol.2019.05.009.

Leemann, R.J. (2010). Gender inequalities in transnational academic mobility and the ideal type of academic entrepreneur. *Discourse: Studies in the Cultural Politics of Education, 31*, 605–625. 10.1080/01596306.2010.516942.

Leung, M.W. (2017). Social mobility via academic mobility: Reconfigurations in class and gender identities among Asian scholars in the global north. *Journal of Ethnic and Migration Studies, 43*, 2704–2719. 10.1080/1369183X.2017.1314595.

Li, E.Y., Liao, C.H., & Yen, H.R. (2013). Co-authorship networks and research impact: A social capital perspective. *Research Policy, 42*, 1515–1530.
Li, F., & Tang, L. (2019). When international mobility meets local connections: Evidence from China. *Science and Public Policy, 46*, 518–529. 10.1093/scipol/scz004.
Li, W., Aste, T., Caccioli, F., & Livan, G. (2019). Early coauthorship with top scientists predicts success in academic careers. *Nature Communications, 10*, 1–9.
Lin, M., Lucas, H.C., Jr, & Shmueli, G (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research, 24*, 906–917.
Mascarenhas, A., Moore, J.E., Tricco, A.C., Hamid, J., Daly, C., Bain, J., & Straus, S.E. (2017). Perceptions and experiences of a gender gap at a Canadian research institute and potential strategies to mitigate this gap: a sequential mixed-methods study. *CMAJ Open, 5*, 144. 10.9778/cmajo.20160114.
Miyake, A., Kost-Smith, L.E., Finkelstein, N.D., Pollock, S.J., Cohen, G.L., & Ito, T.A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science, 330*, 1234–1237. 10.1126/science.1195996.
Moed, H.F., Aisati, M., & Plume, A. (2013). Studying scientific migration in Scopus. *Scientometrics, 94*, 929–942.
Mohammadamin, E., Ali, R.V., & Abrizah, A. (2017). Co-authorship network of scientometrics research collaboration. *Malaysian Journal of Library & Information Science, 17*, 73–93.
Morano-Foadi, S. (2005). Scientific mobility, career progression, and excellence in the european research area1. *International Migration, 43*, 133–162. 10.1111/j.1468-2435.2005.00344.x.
Myers, R.M., & Griffin, A.L. (2019). The geography of gender inequality in international higher education. *Journal of Studies in International Education, 23*, 429–450. 10.1177/1028315318803763.
Nikunen, M., & Lempiäinen, K. (2020). Gendered strategies of mobility and academic career. *Gender and Education, 32*, 554–571. 10.1080/09540253.2018.1533917.
Ovseiko, P.V., Chapple, A., Edmunds, L.D., & Ziebland, S. (2017). Advancing gender equality through the Athena SWAN Charter for Women in Science: an exploratory study of women's and men's perceptions. *Health Research Policy and Systems, 15*, 1–13.
Petersen, A.M. (2018). Multiscale impact of researcher mobility. *Journal of the Royal Society Interface, 15*, Article 20180580. 10.1098/rsif.2018.0580.
Ponjuan, L., Conley, V.M., & Trower, C. (2011). Career stage differences in pre-tenure track faculty perceptions of professional and personal relationships with colleagues. *The Journal of Higher Education, 82*, 319–346.
Robinson-Garcia, N., Sugimoto, C.R., Murray, D., Yegros-Yegros, A., Larivière, V., & Costas, R. (2019). The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics, 13*, 50–63. 10.1016/j.joi.2018.11.002.
Ryazanova, O., & McNamara, P. (2019). Choices and consequences: Impact of mobility on research-career capital and promotion in business schools. *Academy of Management Learning & Education, 18*, 186–212. 10.5465/amle.2017.0389.
Saxena, A., & Iyengar, S. (2020). *Centrality measures in complex networks: a survey arXiv preprint arXiv:2011.07190*.
Saxenian, A. (2007). *The new argonauts: Regional advantage in a global economy*. Harvard University Press.
Schaer, M., Dahinden, J., & Toader, A. (2017). Transnational mobility among early-career academics: gendered aspects of negotiations and arrangements within heterosexual couples. *Journal of Ethnic and Migration Studies, 43*, 1292–1307. 10.1080/1369183X.2017.1300254.
Servia-Rodríguez, S., Noulas, A., Mascolo, C., Fernández-Vilas, A., & Díaz-Redondo, R.P. (2015). The evolution of your success lies at the centre of your co-authorship network. *PloS One, 10*, Article e0114302.
Stuart, E.A., Lee, B.K., & Leacy, F.P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology, 66*, S84–S90.
Subbotin, A., & Aref, S. (2021). Brain drain and brain gain in Russia: Analyzing international migration of researchers by discipline using scopus bibliometric data 1996-2020. *Scientometrics*, 7875–7900. 10.1007/s11192-021-04091-x.
Toader, A., & Dahinden, J. (2018). Family configurations and arrangements in the transnational mobility of early-career academics: Does gender make twice the difference? *Migration Letters, 15*, 67–84. 10.33182/ml.v15i1.339.
Villalonga-Olives, E., & Kawachi, I. (2015). The measurement of social capital. *Gaceta Sanitaria, 29*, 62–64. 10.1016/j.gaceta.2014.09.006.
Wang, J., Hooi, R., Li, A.X., & Chou, M.H. (2019). Collaboration patterns of mobile academics: The impact of international mobility. *Science and Public Policy, 46*, 450–462. 10.1093/scipol/scy073.
Wang, M.-T., & Degol, J.L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review, 29*, 119–140. 10.1007/s10648-015-9355-x.
Way, S.F., Morgan, A.C., Clauset, A., & Larremore, D.B. (2017). The misleading narrative of the canonical faculty productivity trajectory. *Proceedings of the National Academy of Sciences, 114*, E9216–E9223.
Youtie, J., Rogers, J., Heinze, T., Shapira, P., & Tang, L. (2013). Career-based influences on scientific recognition in the United States and Europe: Longitudinal evidence from curriculum vitae data. *Research Policy, 42*, 1341–1355.
Zare-Farashbandi, F., Geraei, E., & Siamaki, S. (2014). Study of co-authorship network of papers in the Journal of Research in Medical Sciences using social network analysis. *Journal of Research in Medical Sciences: the Official Journal of Isfahan University of Medical Sciences, 19*, 41.
Zhao, Z., Bu, Y., Kang, L., Min, C., Bian, Y., Tang, L., & Li, J. (2020). An investigation of the relationship between scientists' mobility to/from China and their research performance. *Journal of Informetrics, 14*, Article 101037. 10.1016/j.joi.2020.101037.

Chapter 6

Exploring Influential Factors on Researchers' h-Index Prediction

6.1 Overview

In the previous two chapters, we discussed two factors: academic mobility, and OA publishing, that are associated with scientific impact. Our analysis revealed a positive correlation between these factors and the number of received citations. Additionally, we conducted gender-specific analyses within this context. In this paper [77], we considered these three factors, along with other author- and publication-specific features, to examine their collective contribution in predicting scholars' h-index. To achieve this, we categorized researchers into three groups based on their career stage (junior, mid-level, and senior) and employed a machine learning approach to predict their h-index. We then compared the model's performance across different career stages and future time windows.

6.2 Publication

Title: Investigating the contribution of author- and publication-specific features to scholars' h-index prediction

Authors: Fakhri Momeni, Philipp Mayr, and Stefan Dietze

Document Type: Journal paper

Venue: EPJ Data Science

DOI: https://doi.org/10.1140/epjds/s13688-023-00421-6

Momeni et al. *EPJ Data Science* (2023) 12:45
https://doi.org/10.1140/epjds/s13688-023-00421-6

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Investigating the contribution of author- and publication-specific features to scholars' h-index prediction

Fakhri Momeni[1*], Philipp Mayr[1] and Stefan Dietze[1,2]

*Correspondence:
fakhri.momeni@t-online.de
[1]GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany
Full list of author information is available at the end of the article

Abstract

Evaluation of researchers' output is vital for hiring committees and funding bodies, and it is usually measured via their scientific productivity, citations, or a combined metric such as the h-index. Assessing young researchers is more critical because it takes a while to get citations and increment of h-index. Hence, predicting the h-index can help to discover the researchers' scientific impact. In addition, identifying the influential factors to predict the scientific impact is helpful for researchers and their organizations seeking solutions to improve it. This study investigates the effect of the author, paper/venue-specific features on the future h-index. For this purpose, we used a machine learning approach to predict the h-index and feature analysis techniques to advance the understanding of feature impact. Utilizing the bibliometric data in Scopus, we defined and extracted two main groups of features. The first relates to prior scientific impact, and we name it 'prior impact-based features' and includes the number of publications, received citations, and h-index. The second group is 'non-prior impact-based features' and contains the features related to author, co-authorship, paper, and venue characteristics. We explored their importance in predicting researchers' h-index in three career phases. Also, we examined the temporal dimension of predicting performance for different feature categories to find out which features are more reliable for long- and short-term prediction. We referred to the gender of the authors to examine the role of this author's characteristics in the prediction task. Our findings showed that gender has a very slight effect in predicting the h-index. Although the results demonstrate better performance for the models containing prior impact-based features for all researchers' groups in the near future, we found that non-prior impact-based features are more robust predictors for younger scholars in the long term. Also, prior impact-based features lose their power to predict more than other features in the long term.

Keywords: h-index prediction; Feature importance; Academic mobility; Machine learning; Open access publishing

1 Introduction

Predicting scientific impact helps to anticipate the career trajectories of researchers and reveal mechanisms of the scientific process that influence future impact, which has al-

Springer

ways been a concern of individual researchers, universities, recruitment committees, and funding agencies. Also, it can reveal factors influencing the future outcome and propose path-ways to young researchers on how to improve future impact and their organizations for more support.

Scientific productivity and received citations are the basis for many evaluation metrics (e.g., h-index [1], g-index [2], h_s-index [3]). The h-index is the most common metric which evaluates the scholars' scientific impact since it measures researchers' productivity and citation impact and has a leading role in hiring and funding decisions. Therefore, predicting this metric is crucial for these purposes. The shorter publication record, received citations, and h-index (prior impact-based features) simplify the h-index prediction task because these features reflect the scholar's impact. Since more senior scholars have a distinguished research profile, predicting their h-index is easier. Assessing the future impact is more pivotal for young scholars than seniors because prior impact-based features are less available for junior researchers as they have a shorter data history. The prediction task will be more complicated for rising stars (who have a lower research profile at the beginning of their career compared to other authors in the same career stage but may become prominent contributors in the future [4]), and we need non-prior impact-based features to evaluate their impact in the long term. Although previous studies demonstrated high accuracy by employing prior impact-based features [5–7], they displayed a substantial decline in the performance of predicting the h-index in the distant future. We hypothesise that publication/citation-based features may be efficient short-term predictors, but other feature categories may be more efficient in predicting long-term impact.

To address these limitations and improve the accuracy of h-index prediction, this study takes a comprehensive approach by investigating a wide array of features and feature combinations. We consider traditional publication/citation-based features and explore other feature categories that may play a role in predicting long-term impact. Our primary objective is to gain a deeper understanding of feature contributions to the h-index prediction task for researchers at different career stages. Our investigation involves analyzing various features and feature combinations in the context of h-index prediction. Drawing from prior research associating specific features with productivity and received citations, we examine how these attributes contribute to researchers' future h-index. To accomplish this, we leverage a machine learning approach to predict the h-index for the upcoming ten years and conduct an extensive feature analysis. To assess the temporal stability of our predictions, we implement our method on three distinct groups of authors: junior, middle-level, and senior researchers. By comparing the accuracy of different feature combinations within each group, we gain insights into the efficacy of the predictive models over time.

In summary, our study makes three significant contributions to the field:

1. *Feature impact analysis:* We advance the understanding of the impact of different feature categories on various h-index prediction tasks for researchers in different career phases and examine the reliability of these predictions.
2. *Temporal dimension of feature performance:* We investigate the temporal dimension of predictors to advance the understanding of feature performance depending on the time window considered for the future prediction, i.e., to understand which features/categories perform better for long- and short-term prediction regarding their seniority.

3. *Novel features:* We introduce and investigate the effect of non-prior impact-based features, namely gender and academic mobility, on the prediction task to reveal the influential factors on the scientific impact (prior impact-based features that implicitly or explicitly encode citation counts simplify the h-index prediction task dramatically by providing the model with data that directly influences the target metric (h-index)).

2 Related work

To identify the future scientific impact, several studies focus on predicting the citations count for a specific paper [8–12], others tried to predict the impact at the author level with the h-index [5–7, 13]. Among all models and methods presented in these studies to predict the h-index, those that took the number of prior publications, received citations, or the current h-index (prior impact-based features) into consideration achieved the highest performance. Although prior impact-based features are the strongest predictors of future impact, sometimes we need to predict it using the other author, paper, and venue characteristics.

2.1 Features used for the prediction tasks

Many studies employed various properties of papers, venues, authors, and their coauthors to predict the scientific impact. Abrishami and Aliakbary [9] and Bai et al. [8] use time series methods and early citations count to predict the number of citations in the long term. Jiang et al. [10] presented a citation time series approach to predict the citations for newly published papers. They used the paper's topic (via keyword), author reputation, venue prestige, and temporal cues (e.g., increasing network centrality over time) to detect citation signals and convert them into signals for citation time series generation. Nie et al. [14] utilized some features and categorized them into the author (regarding citations and publication), venue, social (coauthor), and temporal (average citation increment of the author and coauthors within two years) features and examined their importance in predicting academic rising star. Ayaz et al. [5] and Weihs and Etzioni [6] used the number of current publications, citations, or h-index with other features to predict the future h-index and both presented models with $R^2 = 0.93$. Wu et al. [7] included related indicators to these features, such as changes in citations and h-index over the last two years to the predictors' list and demonstrated a model with a higher precision $R^2 = 0.97$. Further studies focused on other feature types rather than prior impact-based features to identify the influential factors on the scientific impact of researchers. For example, McCarty et al. [15] investigated the relationship between some characteristics of the coauthor network and the h-index. Their results showed the significance of coauthors' productivity via collaborating with many authors and their impact on predicting the h-index. Nikolentzos et al. [13] extracted two types of features, papers' textual content and graph features (related to collaboration patterns), and found that graph features alone are more robust predictors. Dong et al. [16] studied the contribution of a publication to the author's h-index and found that topical authority and publication venues are the most predictive features in the absence of citation-related features of prior publications. Otherwise, they reported citation count as the most decisive factor in predicting the future h-index. Jiang et al. [10] found that certain features, such as the author's reputation, are more predictive than others. Therefore, they applied trainable weights to preserve the unequal contribution of different kinds of features. Ayaz et al. [5] reported the career age, number of high-quality

papers, and number of publications in distinct journals as the most compelling feature in predicting the h-index after prior impact-based features. They observed a lower performance for younger researchers and concluded the investigated features are insufficient to predict their h-index and a need to evaluate future features for better prediction.

Wu et al. [7] investigated the stability of predictive models for long-term prediction (ten future years) and compared their method with state-of-the-art [5, 6, 16]. They used time series features (the history of the h-index) and more impact-based features in their analyses, which are less valuable to predict the future impact of young researchers. They found better performance among all mentioned works. However, they included only the authors with an h-index higher than four and junior researchers whose predicting their scientific impact is more challenging have been excluded from their study.

We tackle these issues by investigating novel author- and paper-specific features for the prediction task and verifying their contribution to the h-index prediction for researchers with varying scientific experiences.

2.2 Influential factors on scientific impact

In the following, we categorize the features affecting the scientific impacts into three groups: demographic, paper/venue, and coauthor-based factors, and report the previous related studies.

2.2.1 Demographic factors

Academic mobility In contemporary science, collaboration plays a significant role, and international academic mobility affects the collaboration networks, which furthers knowledge transmission among countries and scholars. Therefore, many studies have focused on investigating its impact on science and scientists. Our recent study [17] revealed the positive impact of international mobility on the number of publications and received citations. However, mobile researchers do not necessarily perform better than those without mobility experience. Singh [18] found that differences in research outputs between returnee Ph.D. holders and those trained in their home country are field-specific and depend on their seniority. Netz et al. [19] reviewed the studies that investigated the effect of mobility on some scientific outcomes and found that most studies suggest a positive effect on mobility. But they reported some studies that demonstrated a negative effect on productivity and citation impact and proposed a positive impact of mobility only under specific circumstances. Liu et al. [20] found that international collaboration before mobility has an essential role in high performance after mobility. The reputation of institutions is another influential factor they discovered in their study.

Gender Gender differences in science and scientific impact have been the subject of many studies in various fields. A new study on the Breast Surgery Fellowship Faculty [21] found no noticeable gender difference between assistant professors but a higher h-index for men professors than women. [22] studied the gender gap in social sciences and found the difference in all career phases, especially in full professor positions. In contrast, the study's results by Lopez et al. [23] demonstrated a higher h-index for men among academic ophthalmologists. Still, controlling the range of publications, they found the same or more impact for women in the later career phases. The results of the study by Kelly et al. [24] indicated that although the h-index of men is higher than women for ecologists

and evolutionary biologists, there is no gender difference in the h-index once we control for publication rate. However, other studies [25, 26] examined the relationship between received citations and funding available from Web of Science data and found a weak correlation between them.

Income level In many countries, governments are the primary source of financial support for scientific progress. Gantman [27] demonstrated the positive effect of economic development on scientific productivity in all scientific fields. Confraria et al. [28] displayed a U-shape relationship between Gross Domestic Product (GDP) per capita and received citations and found the citation impact correlates positively with the nation's wealth after a certain GDP per capita level. However, their results showed that international collaboration is crucial for higher citation impact among all countries.

2.2.2 Paper and venue factors

Scientific field The average scholars' h-index of researchers differs among fields because productivity and the rate of citing vary from one to another [29, 30]. Iglesias and Pecharrom [31] showed the varying ranges of the h-index across fields and suggested a multiplicative correction to the h-index based on the scientific field to compare the scientists' research impact from different areas.

Journal quality Reputable journals increase the visibility of papers and the probability of receiving citations. Petersen and Penner [32] found that publishing in high-quality journals decreases the average time interval between the author's future publications in those journals and has a cumulative citation advantage for the author.

Open access Free access to publications in online form increases the probability of reading and citing papers. Various studies investigated the Open Access Citation Advantage (OACA), and most found a positive effect on received citations [33–36]. Langham-Putrow et al. [37] did a systematic review of the OACA and reported that among 143 studies, 47.8% confirmed OACA, 37% found no OACA, and 24% found OACA for a subset of their sample. Also, the result of our recent study [38] showed substantially higher citations for preprint papers, making publications freely available. Momeni et al. [39] examined the association of open access publishing with received citations and found a higher percentage of highly cited papers published in the open-access model than those in the closed-access model.

2.2.3 Coauthor factors

The number of the paper's citations received reveals the scientific impact of all authors, and hence it can vary according to their collaboration pattern. Hsu and Huang [40] found a positive correlation between the number of coauthors and received citations. Also, the result of the study by Puuska et al. [41] showed fewer citation scores for single-authored publications. Sarigöl et al. [42] tried to predict highly cited papers via the centrality of their authors in the co-authorship network and found a positive correlation between highly cited publications and highly centralized authors.

Other studies [41, 43] examined the citation impact of international coauthors and demonstrated a positive relation between international collaboration and received citations.

2.3 Prediction approaches

Many studies employed machine learning regression and classification approaches to predict the scientific impact of publications and researchers [6, 7, 9–11, 13]. The most common methods in these studies were regression models such as Support Vector Regression (SVR), Gradient Boosted Regression Trees (GBRT) or Gradient Boosting (GB), Gradient-Boosting Decision Tree (GBDT), Extreme Gradient Boosting (XGBoost), Random Forest (RF), K-nearest Neighbour (KNN), and Neural Networks (NN). Nie et al. [14] introduced a classification method to detect the academic rising stars (who have a lower research profile at the beginning of their career compared to other authors in the same career stage but may become prominent contributors in the future) and found better performance for KNN algorithm for small datasets, but a relatively stable result for GBDT, GB, RF, and RF with the change of dataset size. Ruan et al. [11] examined the performance of different regression algorithms and reported the best performance for Backpropagation neural network. Wu et al. [7] examined SVR, RF, GBRT, and XGBoost regression models for h-index prediction and obtained the best performance for XGBoost. The performance of methods for predicting the h-index in different ranges depends on applied features. By using prior impact-based features and regression models, previous studies [5–7] presented models with $R^2 > 0.90$ for the first predicting year and decreased in the next predicting years. However, none of these studies investigated the extent of the contribution of different features in the prediction task. Our study examines the contribution of features to the h-index prediction via feature selection/ranking approaches to understanding the influential factors better.

3 Data and methods

3.1 Describing the dataset

We used the in-house Scopus database maintained by the German Competence Centre for Bibliometrics (Scopus-KB), 2020 version, as the central resource of analyses and employed Scopus author Id to identify authors. We defined the career age of authors by the years between the first and last publication time. We took authors who started publishing after 1994 and used their publications until 2008 to calculate the features' value. We detected the gender status of authors by a combined name and image-based approach introduced by Karimi et al. [44], which results in a binary variable. We acknowledge that a person's gender can not be split into male and female, and if we consider the social dimensions, we have more gender identities.

To remove "not active authors" from the analyzed data, we included just those authors who had at least five years of career age, an h-index higher than zero and matched the threshold of one publication per three years in their career age. Excluding authors without gender status results in a final list of 1,824,203 authors. Table 1 presents some information about the distribution of analysed papers among main research domains (categorized by the All Science Journal Classification (ASJC) System in Scopus), the distribution of authors among gender, and career stages.

We applied the prediction model to three datasets containing the authors regarding their career development:

- Junior: researchers with a career age of fewer than five years (the first publication between 2005 and 2008)
- Mid-level: researchers with a career age between 5 and 9 years (the first publication between 2000 and 2004)

Table 1 The number of analyzed papers across scientific fields and gender and career stage distribution of authors

	Number	Percentage
Papers	40,352,318	
Health Sciences	10,608,222	26.3 %
Life Sciences	8,831,499	21.9 %
Physical Sciences	17,089,343	42.3 %
Social Sciences & Humanities	3,272,508	8.1%
Multidisciplinary	550,746	1.4%
Authors	1,824,203	
Gender:		
Female	543,517	30%
Male	1,280,686	70%
Career stage:		
Junior	265,368	15%
mid-level	533,768	29%
senior	1,025,067	56%

Table 2 Features used to train the machine learning models to predict the h-index

Feature group	Feature name	Description	Studies
Demographic	*CareerAge*	Years since first publication	[5]
	Gender	Zero for females and one for males	
	MobilityScore	Number of changing the affiliation at the country level	
	IncomeCurrentCountry	GDP Per Capita of current affiliation country	
Prior Impact	*CurrentHindex*	Current h-index	[5]; [6]; [7]
	PaperPerYear	Number of total papers divided by career age	[5]; [6]; [7]
	CitationPerPaper	Number of total citations among all papers until 2008 divided by the number of all papers	[5]; [6]; [7]
Paper/Venue	*PrimaryAuthorRatio*	Number of papers being as primary author divided by the number of all papers	
	OpenAccessRatio	Number of open access papers divided by the number of all papers	
	MainField	The scientific field with the highest amount of publications	
	HighRankPapersRatio	Number of publications in high-quality journals divided by the number of all papers	[5]
	DisciplineMobility	Number of unique disciplines authors has published paper divided by the number of all papers	
	KeywordPopularity	Number of publications with at least one popular keyword divided by the number of all papers	
	EnglishPapersRatio	Number of English papers divided by the number of all papers	
Coauthor	*MaxCoauthorHindex*	Maximum h-index of coauthors among all papers	[15]
	CoauthorPerPaper	Number of unique coauthors among all publications divided by the number of all papers	[7]
	InternationalCoauthorRatio	Number of papers with international collaboration divided by the number of all papers	

- Senior: researchers with a career age of over ten years (the first publication between 1995 and 1999).

3.2 Feature engineering

Table 2 shows variables used to estimate the future h-index of researchers. In this table, we mentioned the previous studies that employed any of the features for the prediction task. In the following, we explain how we calculated the features:

- *Gender:* It has a value equal to one for males and zero for females.
- *MobilityScore:* This feature indicates the frequency of movement between countries by tracking the authors' affiliations over their publications. More details about calculating this feature are available in our previous study [17].
- *IncomeCurrentCountry:* This feature indicates the countries' income level based on the GDP per capita of the affiliation country in the last publication. We used the World Bank information[1] to measure it.
- *PrimaryAuthorRatio:* We defined the primary author as the first or corresponding author. We computed the value of this feature by dividing the number of publications in which the researcher is the primary author to all publications.
- *OpenAccessRatio:* We extracted the article's access status from the Unpaywall dataset (a service that provides full-text articles from open access resources[2]). An open-access article can be any form of gold, green, or bronze. We declare that we could match from 8,953,939 investigated papers only 5,476,852 (61%) with Unpaywall's articles. To calculate the proportion of open access papers, we considered the number of detected as open access to the number of whole articles of the author.
- *MainField:* We identified the field of authors from the field of the journals in which they publish, and in Scopus are classified under four broad subject clusters.[3] The field with the most publications will be the main field of the author.
- *HighRankPapersRatio:* We used the journal ranking based on their quality to evaluate the rank of papers. To assess the quality of journals, we calculated the h-index of journals from 1995 to 2015. Because of different citation patterns among disciplines, journals' h-index can have varying ranges for different disciplines, which should be normalized. We applied the percentile rank approach inspired by Bornmann and Lutz [45] and computed the h-index's rank among all journals inside its discipline. We used Scopus's classification system to find the journals' disciplines. In this system, journals are classified into 27 subject categories.[4] In this percentile rank approach, each journal within a category ranks 0 (lowest h-index) to 100 (highest h-index). Journals with the same h-index have the same rank. If the journal belongs to more than one category, we used the weighted Percentile Ranking wPR) [46]. Based on this approach, wPR will be calculated using the formula:

$$wPR = \frac{PR_{sc1} * n_{sc1} + PR_{sc2} * n_{sc2} + \cdots + PR_{sci} * n_{sci}}{n_s c1 + n_s c2 + \cdots + n_{sci}}. \quad (1)$$

Whereby *sci* is the *i*th subject category that the journal belongs to and n_{sci} is the number of journals in this subject category, and PR_{sci} is PR of the journal in it. Journals with a wPR higher than 50% are assumed to be high quality. Finally, we counted the proportion of the author's publications in high-quality journals among all their publications for the variable *HighRankPapersRatio*.

[1] https://www.weforum.org/agenda/2020/08/world-bank-2020-classifications-low-high-income-countries/.

[2] https://unpaywall.org/.

[3] https://service.elsevier.com/app/answers/detail/a_id/14882/supporthub/scopus/~/what-are-the-most-frequent-subject-area-categories-and-classifications-used-in/.

[4] https://service.elsevier.com/app/answers/detail/a_id/14882/supporthub/scopus/~/what-are-the-most-frequent-subject-area-categories-and-classifications-used-in/.

Table 3 Descriptive statistics of features. This table shows the mean standard deviation for numerical features and distribution of authors based on their gender, mobility status and main field

Feature name	Mean	Standard deviation	Distribution
CareerAge	9.35	3.69	
Gender	0.70	0.46	70% male, 30% female
MobilityScore	0.50	1.08	27% mobile, 73% non-mobile
IncomeCurrentCountry	35,052.63	14,024.40	
CurrentHindex	6.13	6.17	
PaperPerYear	2.00	2.39	
CitationPerPaper	11.47	22.18	
PrimaryAuthorRatio	0.36	0.29	
OpenAccessRatio	0.19	0.23	
MainField			H: 29%, L:23%, P:37%, S:6%, M:4% *
HighRankPapersRatio	0.01	0.06	
DisciplineMobility	0.47	0.45	
KeywordPopularity	0.53	0.28	
EnglishPapersRatio	0.92	0.20	
MaxCoauthorHindex	15.51	14.86	
CoauthorPerPaper	3.74	30.39	
InternationalCoauthorRatio	0.21	0.25	

*H: Health Sciences, L: Life Sciences, P: Physical Sciences, M:Multiple Fields.

- *DisciplineMobility:* This feature indicates the number of unique fields the author has published during the entire academic age divided by the number of whole papers.
- *KeywordPopularity:* This feature indicates the proportion of papers with popular keywords. First, we ranked keywords based on the frequency of occurrence in papers from the same discipline (27 subject categories) and publication year to measure the keyword popularity for a paper. Next, we gave a value of 1 to the paper with a ranking above 0.5; otherwise, 0. Finally, we summed up these values over all papers and divided them by the number of all papers.
- *EnglishPapersRatio:* This feature measures the ratio of papers written in English.
- *CoauthorPerPaper:* This feature displays the number of unique coauthors, which is normalized by dividing by the number of all papers.
- *CoauthorMaxHindex:* To assess the effect of the scientific impact of coauthors, we used the maximum h-index among all coauthors as an alternative measure of the Godfather Effect [15].
- *InternationalCoauthorRatio:* This feature specifies the number of international collaborators for all papers. To calculate it, first, we counted the number of papers with at least one coauthor having a different country in the affiliation than the author and then divided it by the number of all papers.

We provided descriptive statistics for investigated features in Table 3 to describe the data.

3.3 Applied methods for the prediction task

We tackled the h-index prediction as a regression problem comparable to previous studies [5–7, 11, 16]. We explored the performance of four different machine learning methods, namely SVR, RF, GB, and XGBoost. Among these, XGBoost emerged as the top-performing method, consistent with the findings reported by [7]. Consequently, we utilized the XGBoost approach for our h-index prediction task. XGBoost is a scalable end-to-end tree boosting system introduced by Chen and Guestrin [47]. It efficiently implements

Table 4 Different feature combinations to predict the h-index

Feature group	Feature name	Feature combination								
		1	2	3	4	5	6	7	8	9
Demographic	*CareerAge*		✓		✓			✓		✓
	Gender		✓		✓			✓		✓
	MobilityScore		✓		✓			✓		✓
	IncomeCurrentCountry		✓		✓			✓		✓
Prior impact	*CurrentHindex*	✓	✓	✓	✓	✓				
	CitationPerPaper	✓	✓	✓	✓	✓				
Paper/venue	*PrimaryAuthorRatio*	✓	✓	✓	✓		✓	✓	✓	✓
	OpenAccessRatio	✓	✓	✓	✓		✓	✓	✓	✓
	MainField	✓	✓	✓	✓		✓	✓	✓	✓
	HighRankPapersRatio	✓	✓	✓	✓		✓	✓	✓	✓
	DisciplineMobility	✓	✓	✓	✓		✓	✓	✓	✓
	EnglishPapersRatio	✓	✓	✓	✓		✓	✓	✓	✓
	KeywordPopularity	✓	✓	✓	✓		✓	✓	✓	✓
Coauthor	*MaxCoauthorHindex*			✓	✓				✓	✓
	CoauthorPerPaper			✓	✓				✓	✓
	InternationalCoauthorRatio			✓	✓				✓	✓

Gradient Boosting in terms of speed and is appropriate for solving problems using minimal resources. We need to have the data in numerical form to apply this method. We utilized one hot encoder to convert the categorical values to integers. In this encoding method, each value of the categorical variable will be converted to a feature with a binary value, where 1 represents the data value and 0 is used for all other values. So, for *MainField* with five values, we have five features, and the feature with a value equal to 1 indicates the *MainField*. To evaluate the model, we utilized the Mean Absolute Percentage Error (MAPE) to measure the error as a percentage, which is appropriate to compare the performance of a model for the different datasets, as used by some previous studies [6–8]. Because MAPE is affected by outliers [48], we also utilized symmetric Mean Absolute Percentage Error (sMAPE), which is scaled to percentage too and is more resistant to outliers [47]. In addition, we used Root Mean Square Error (RMSE) to evaluate the performance of models, as in prior works [5, 8, 9]. We used the 5-fold cross-validation procedure to evaluate the models.

We defined different feature combinations based on the attributes of the author, paper, venue, and coauthors to see which feature categories are better for short/long-term prediction. Table 4 shows the different feature combinations utilized to train the model.

Prior studies regarded varying time frames to estimate the future h-index [5, 7, 49] and examined several years from one to five-year and [49] for five-year and ten-year time frames. The prediction performance declined as the prediction time frame increased in all studies. We considered the h-index as our target from one to ten years in the future (h-index from 2009 to 2018). It enables us to measure the extent of predicting performance in the future.

To examine the importance of each feature in the prediction task, we employed a feature selection technique, *Recursive Feature Elimination* (RFE), which removes recursively features and builds a model based on the remaining features [50, 51].

4 Results

In this section, we present the results of our analysis, focusing on the relationship between various features and the future h-index of researchers. Before delving into the specific

findings, we address the potential multicollinearity problem in Sect. 4.1 by examining the dependencies between features. We analyze the Pearson correlation between independent variables and visualize the results using a heatmap. Next, we explore the correlation between the introduced features and the future h-index in 2009, 2014, and 2018. This analysis allows us to examine the statistical association between variables, providing insights into the strength and direction of these relationships. However, it's important to note that the correlations captured by the correlation analysis primarily represent linear associations between features and the h-index.

To capture the non-linear relationship between the h-index and the investigated features, we apply ML prediction models in Sect. 4.2. First, in Sect. 4.2.1, we identify the most important factors for predicting the h-index using the feature selection method, RFE. This step helps us narrow down the key variables. Then, in Sect. 4.2.2, we examine the effectiveness of these models for researchers with different career ages, focusing on the temporal dimension.

4.1 Correlation analysis

Before investigating the relationship between various features and future h-index, we examine the dependencies between features to avoid the potential multicollinearity problem. Figure 1 presents the Pearson correlation between independent variables. We see a strong correlation between *PaperPerYear* and *CurrentHindex*; therefore, to avoid multicollinearity in regression and classification models, we exclude *PaperPerYear* from the data for prediction tasks.

To examine the affecting factors on the h-index, we first provide the correlation between features introduced in Table 2 and future h-index. Table 5 presents the Pearson correlation coefficient between the features (except for *MainField*, a categorical variable) and h-index in 2009, 2014, and 2018. The highest correlation coefficient for two prior impact-based features (*CurrentHindex*, *PaperPerYear*) displays the strong association of this kind of feature with the future h-index. The higher correlation coefficient between the future h-index and the number of papers (*PaperPerYear*) than the number of citations (*CitationPerPaper*) reveals that productivity has a more significant impact than received citations on the h-index. Among non-prior impact-based features, *MaxCoauthorHindex* has the highest correlation with the h-index and suggests the strong relation of coauthors' reputation with the future h-index. The negative value for *DisciplineMobility* suggests that authors who publish in several scientific fields have a lower h-index than those who publish in a specific field.

Most of the correlations between the influential factors and the h-index demonstrate consistent patterns across different time frames, indicating similar effects in both the short and long term. While correlation analysis offers informative perspectives about the strength and direction of these relationships, it primarily captures linear associations between variables. However, we will employ machine learning algorithms in the next section to uncover non-linear associations and delve deeper into the temporal dimension of the relationship for researchers in different career stages. This approach allows us to examine the complex interactions and temporal dynamics between the factors and the h-index, specifically analyzing how they vary across different career stages. It provides a more comprehensive understanding of their relationship and enables us to make accurate predictions beyond what correlation analysis alone can reveal.

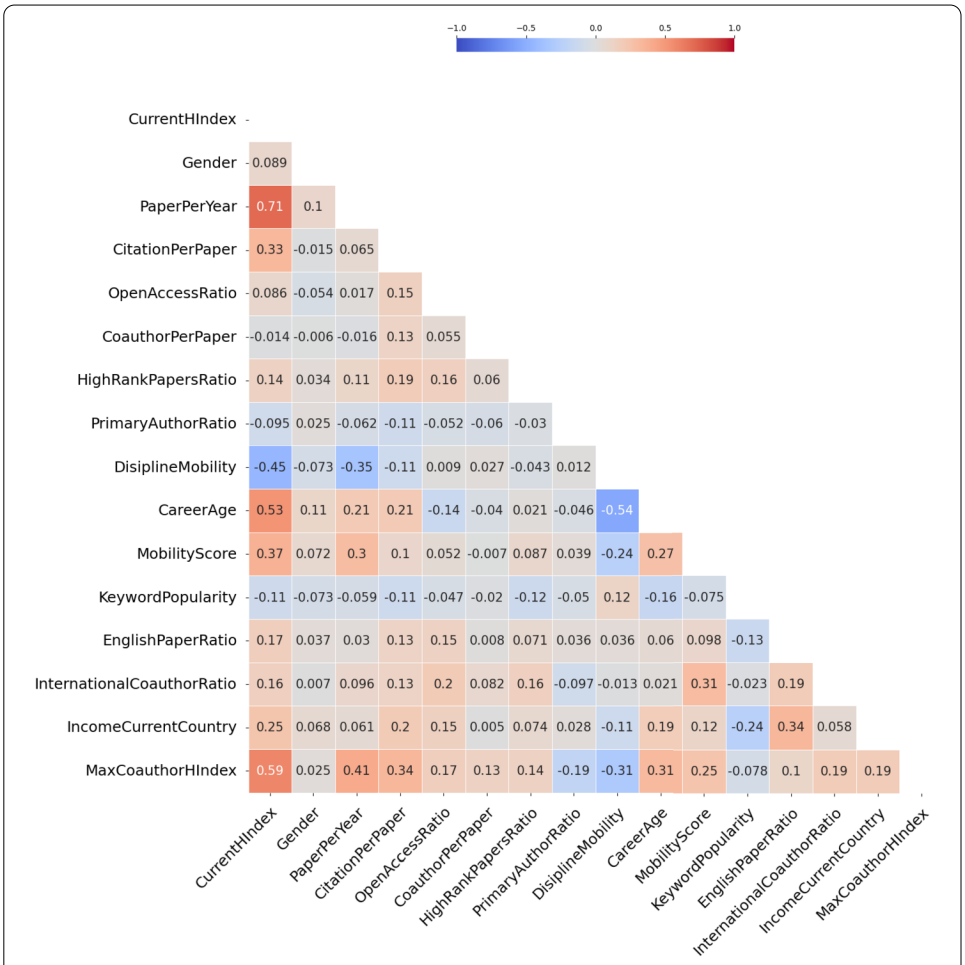

Figure 1 Pearson correlation heatmap of independent variables. We observe a particularly strong positive correlation between 'PaperPerYear' and 'CurrentHindex' in this heatmap

Table 5 Pearson correlation coefficient between the features and h-index in the future for three different years. *CurrentHindex*, *PaperPerYear*, and *CitationPerPaper* are prior impact-based features, and the rest are non-prior impact-based features

Feature	H-index		
	2009	2014	2018
CareerAge	0.48	0.38	0.32
Gender	0.09	0.08	0.07
MobilityScore	0.44	0.43	0.41
IncomeCurrentCountry	0.23	0.21	0.19
CurrentHindex	0.99	0.95	0.87
PaperPerYear	0.73	0.75	0.73
CitationPerPaper	0.31	0.26	0.23
PrimaryAuthorRatio	−0.09	−0.08	0.-0.06
OpenAccessRatio	0.10	0.14	0.15
EnglishPapersRatio	0.17	0.16	0.15
KeywordPopularity	−0.09	−0.07	−0.05
HighRankPapersRatio	0.14	0.15	0.15
DisciplineMobility	−0.45	−0.42	−0.39
MaxCoauthorHindex	0.58	0.58	0.55
CoauthorPerPaper	−0.01	0.02	0.04
InternationalCoauthorRatio	0.17	0.19	0.19

4.2 Prediction analysis

In this section, we present the prediction results of our study, highlighting the influence of different features on predicting the h-index. Firstly, in Sect. 4.2.1, we evaluate the importance of these features using the Recursive Feature Elimination (RFE) method. Then, in Sect. 4.2.2, we examine the effectiveness and stability of various feature combinations in predicting the h-index. We analyze the predictive performance across different time frames and for researchers at different career stages, providing valuable insights into the temporal dynamics and the impact of features on the h-index prediction task.

4.2.1 Feature impact

We evaluate the importance of features in the prediction task by ranking them via the RFE method. Table 6 demonstrates the feature ranking for selecting the predictors in the model. For *MainField*, we used one hot encoder, which converts each unique category value to a feature (five features for five fields). The features highlighted in blue are the top five features in the selection process. We observe that paper-specific features are most relevant among all career stages. Also, coauthor-specific features are among the most important features to predict the h-index for the researchers in junior and mid-level career stages. It suggests that the coauthor's characteristics have more influence on the h-index for these researchers than seniors.

4.2.2 Career stage and temporal dimension of model performance

Before we show the result of the analyses, we make some comparisons between the performance of our model and previous works. Wu et al. [7] have already compared their performance with other studies [5, 6, 49] and presented the best performance among all

Table 6 Ranking of features for selection in predicting the h-index with the RFE method. The five most relevant features (with a ranking between 1 and 5) are highlighted in blue. It demonstrates variations in feature importance across career stages and prediction years. 'CurrentHindex' consistently ranks as the top feature, indicating its significant influence. Additionally, the most influential features vary by career stage, highlighting the complexity of research impact factors

Career stage	**Junior**			**Mid-level**			**Senior**		
Prediction year	**2009**	**2014**	**2018**	**2009**	**2014**	**2018**	**2009**	**2014**	**2018**
Feature:	Rank	Rank	Rank	Rank	Rank	Rank	Rank	Rank	Rank
CareerAge	7	5	4	3	2	3	3	4	2
Gender	20	18	16	19	18	16	19	19	19
MobilityScore	18	14	12	12	8	4	18	18	16
IncomeCurrentCountry	14	16	17	13	14	13	9	9	9
CurrentHindex	1	1	1	1	1	1	1	1	1
CitationPerPaper	11	15	15	6	6	7	7	5	6
PrimaryAuthorRatio	6	10	9	10	9	9	16	11	10
OpenAccessRatio	3	8	7	4	7	8	6	2	5
EnglishPapersRatio	4	11	13	9	16	17	15	17	18
KeywordPopularity	10	13	14	11	13	15	11	14	13
MainField									
Health Sciences	12	3	5	17	19	18	5	15	17
LifeSciences	15	17	18	15	17	19	8	6	3
multiple fields	19	20	20	20	20	20	20	20	20
Physical Sciences	13	2	2	16	4	5	13	7	8
Social Sciences	16	19	19	14	15	14	4	3	4
HighRankPapersRatio	9	9	11	5	12	12	12	16	15
DisciplineMobility	2	12	10	2	11	11	2	12	14
MaxCoauthorHindex	5	6	3	8	5	6	10	10	11
CoauthorPerPaper	17	4	6	18	10	10	17	8	7
InternationalCoauthorRatio	8	7	8	7	3	2	14	13	12

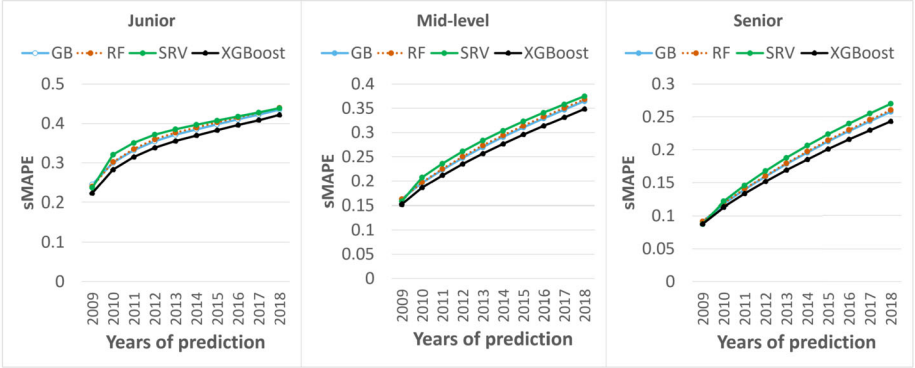

Figure 2 Comparison of predictive performance using sMAPE metric among four machine learning algorithms (SVR, RF, GB, and XGBoost) for researchers' h-index prediction at different career stages from 2009 to 2018. The analysis utilized feature combination 1 as the predictor

these studies. They excluded the authors with an h-index of less than four from the investigated data. They achieved the minimum MAPE of 0.063 for the first prediction year by employing more prior impact-based features. We could reach the minimum MAPE of 0.068 by applying this condition to investigated authors. Instead, two-thirds of the authors will be discarded in our analyses. Because of losing too much data, particularly from young scholars, we didn't apply this condition and implemented our models with all authors, despite reducing the performance. To evaluate the predictive performance, we conducted a comparison among four machine learning algorithms: SVR, RF, GB, and XGBoost, using feature combination 1, which includes all features. The results are illustrated in Fig. 2, demonstrating that XGBoost outperforms the other methods across all career stages. As a result, we proceed with this method for further analyses.

Table 7 showcases the performance metrics, including RMSE, MAPE, and sMAPE, for all three groups of researchers (junior, middle-level, and senior) across the years 2009, 2014, and 2018. It provides a detailed overview of the model's performance, enabling a direct comparison of the metrics for each group and year. Lower values of these metrics indicate better predictive performance. We observe a decline in performance for all groups of researchers across all metrics from the near future (2009) to the far future (2018). While the models for seniors generally demonstrate better performance compared to the other groups, the decline in performance is more pronounced for researchers in later career stages. Specifically, in terms of RMSE for junior researchers, the range varies from 0.6 (combination 4, considering all features) in 2009 to 5.46 (combination 1, considering only prior-impact features) in 2018. For seniors, the range is from 0.74 (combination 1) in 2009 to 6.93 (combination 8) in 2018. We observe a greater decline in performance for seniors in the far future compared to juniors. When considering MAPE and sMAPE, which provide performance in percentage, we can better compare the model's performance across career stages. Although these metrics show better performance for researchers in later career stages, the performance is more stable for juniors. For instance, combination 4 exhibits the best performance for juniors, with sMAPE ranging from 0.22 to 0.42, while for seniors, it ranges from 0.09 to 0.24. Furthermore, despite combinations containing prior-impact features exhibiting better performance in the near future (2009) for all researcher groups, we observe that for juniors, combinations without prior-impact features approach

Table 7 Comparison of XGBoost regression model performance to predict the feature h-index in one, five, and ten years (2009, 2014, and 2018) implemented on three datasets (junior, middle, and senior researchers). RMSE, MAPE, and sMAPE are the metrics to assess performance

Feature combination	Metric	Junior			Middle-level			Senior		
		2009	2014	2018	2009	2014	2018	2009	2014	2018
1	RMSE	0.62	3.01	5.15	0.68	2.85	4.94	0.75	3	5.09
	MAPE	0.24	0.52	0.62	0.16	0.33	0.45	0.09	0.2	0.28
	sMAPE	0.23	0.39	0.45	0.16	0.29	0.36	0.09	0.19	0.25
2	RMSE	0.61	2.91	4.99	0.67	2.78	4.81	0.75	2.94	4.97
	MAPE	0.24	0.49	0.59	0.16	0.32	0.43	0.09	0.2	0.28
	sMAPE	0.23	0.38	0.43	0.15	0.28	0.35	0.09	0.19	0.25
3	RMSE	0.61	2.85	4.91	0.68	2.75	4.77	0.75	2.9	4.9
	MAPE	0.24	0.5	0.6	0.16	0.33	0.44	0.09	0.2	0.28
	sMAPE	0.23	0.38	0.44	0.15	0.28	0.36	0.09	0.19	0.25
4	RMSE	0.6	2.78	4.81	0.67	2.68	4.67	0.74	2.85	4.8
	MAPE	0.24	0.48	0.57	0.16	0.32	0.43	0.09	0.2	0.27
	sMAPE	0.22	0.37	0.42	0.15	0.28	0.35	0.09	0.19	0.24
5	RMSE	0.67	3.23	5.46	0.72	3.05	5.23	0.78	3.24	5.49
	MAPE	0.28	0.57	0.68	0.17	0.37	0.49	0.09	0.23	0.31
	sMAPE	0.27	0.42	0.47	0.17	0.31	0.39	0.1	0.21	0.28
6	RMSE	1	3.27	5.43	1.87	3.56	5.5	4.04	5.75	7.52
	MAPE	0.37	0.56	0.65	0.41	0.44	0.53	0.41	0.4	0.44
	sMAPE	0.31	0.41	0.46	0.32	0.35	0.4	0.32	0.32	0.34
7	RMSE	0.97	3.19	5.3	1.8	3.48	5.38	3.79	5.47	7.24
	MAPE	0.36	0.54	0.62	0.39	0.43	0.51	0.38	0.38	0.42
	sMAPE	0.31	0.4	0.44	0.31	0.34	0.39	0.31	0.31	0.33
8	RMSE	0.96	2.97	5.02	1.75	3.33	5.23	3.64	5.23	6.93
	MAPE	0.35	0.53	0.62	0.38	0.41	0.5	0.35	0.36	0.4
	sMAPE	0.3	0.4	0.44	0.31	0.33	0.38	0.29	0.29	0.32
9	RMSE	0.94	2.92	4.93	1.69	3.28	5.15	3.47	5.05	6.74
	MAPE	0.34	0.51	0.6	0.36	0.41	0.49	0.34	0.34	0.39
	sMAPE	0.3	0.39	0.43	0.3	0.33	0.38	0.28	0.29	0.32

the performance of models with prior-impact features in the long term (2018). In some cases, these combinations even outperform models with prior-impact features. This finding suggests that non-prior impact-based features are more reliable predictors for the future h-index of junior researchers, compared to seniors. In summary, seniors generally exhibit better performance, but juniors demonstrate more stable performance and the potential for improved long-term predictions using non-prior impact-based features.

To further illustrate the performance trends over time, Fig. 3 focuses on the sMAPE metric and covers the years from 2009 to 2018. It offers a visual representation of the prediction efficiency of different feature combinations for researchers at different career stages throughout the entire time span. In this figure, the lower sMAPE for combinations including prior impact-based features indicates the higher performance for these combinations, but losing the performance with the passing years for these combinations is more than other combinations.

To compare the prediction efficiency between different career stages, we implemented the prediction model for authors from three career stages and presented the performance (sMAPE) in Fig. 3(a). We observe a better performance for the combinations containing prior impact-based features for all researchers' groups in the near future. Still, they lose more performance than combinations without prior impact-based features in the distant future. Interestingly, the performance of non-prior impact-based models (e.g., combina-

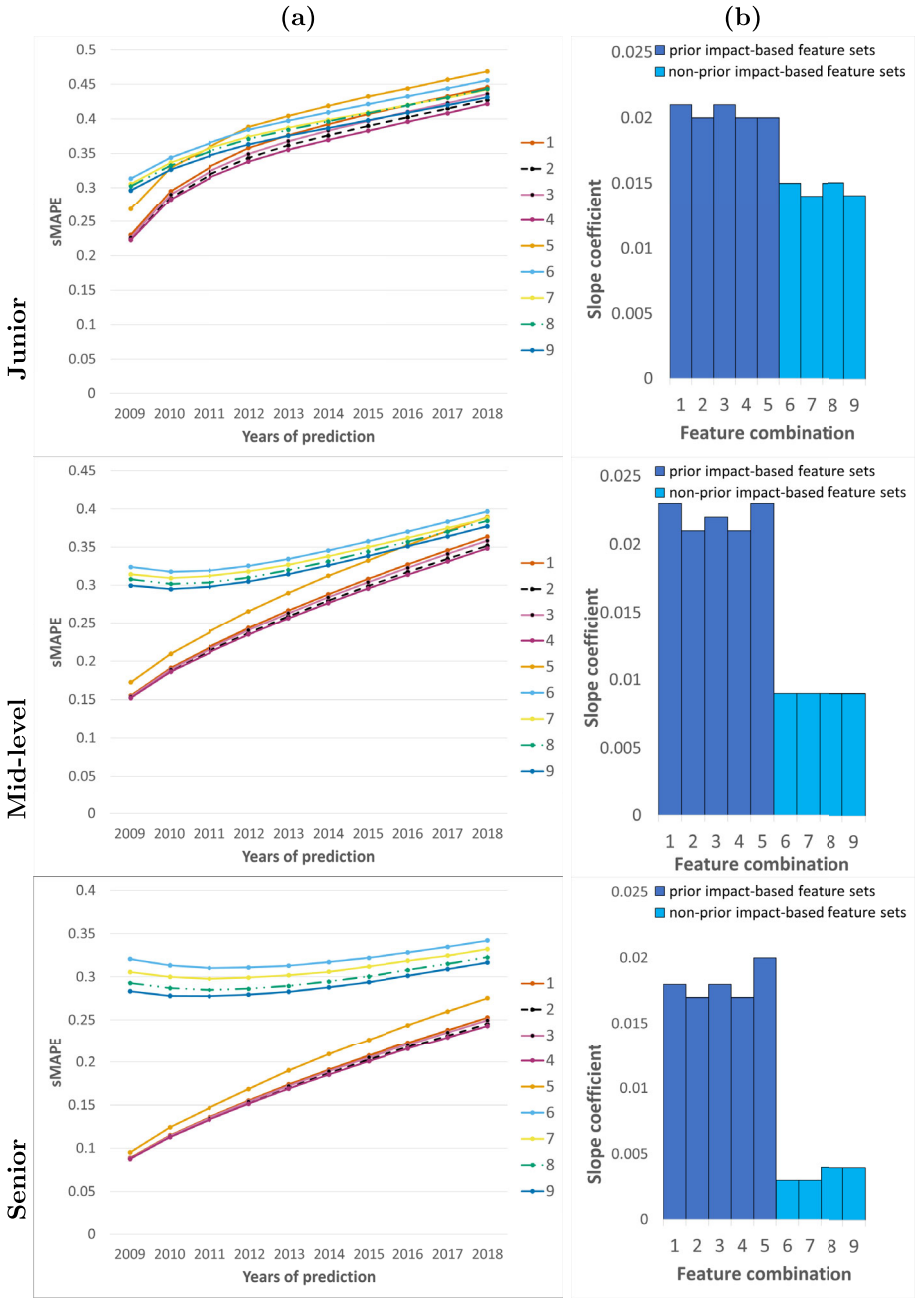

Figure 3 Comparison of predictive performance (a) and slope coefficients (b) over ten years for different feature combinations trained with the XGBoost regression method among researchers of varying experience levels (junior, mid-level, and senior). (a) illustrates the performance of predicting models using the sMAPE metric. (b) displays the corresponding slope coefficients, indicating the performance change over time. The dark/light blue columns in (b) represent feature combinations, including/excluding prior impact-based features

tions 8 and 9) for junior researchers, which is worse than prior impact-based models (e.g., combinations 1 and 5) in the earlier years, dominates them in the long term. We see a similar result for researchers at the mid-level (better performance for combinations 8 and 9

than combination 5). This suggests that non-prior impact-based features are more reliable in predicting the future h-index of younger researchers over distant periods.

To quantify the extent of performance degradation for the two groups of combinations (prior and non-prior impact-based features), we calculated the slope coefficient for model performances reported in Fig. 3(a). The slope coefficient (m) was computed using the least squares method [52] with the following equation:

$$m = \frac{\sum(x - \bar{x}y - \bar{y})}{\sum(x - \bar{x})^2}, \tag{2}$$

where x represents the years from 2009 to 2018, y represents the sMAPE in the corresponding year and $\bar{x}$ and $\bar{y}$ are their respective averages over the ten-year period.

The presented slope coefficient in Fig. 3(b) reveals insights into the stability of the models' performance. A lower slope coefficient signifies greater stability, indicating that the model's performance changes more slowly and consistently over the ten-year period. Conversely, a higher slope coefficient indicates that the model's performance fluctuates more significantly.

In general, we observed a higher slope coefficient (indicating more significant performance loss over time) for feature combinations with prior impact-based features (in dark blue) compared to other feature combinations for researchers at any career stage. The lower value for combinations containing non-prior impact-based features (in light blue) indicates that they are more stable predictors in the long term, although at a modest performance level.

5 Limitations

In this study, we considered just journal papers and not conference papers, and it causes bias issues, especially for disciplines in which authors publish their studies mainly as conference proceedings papers. Another limitation is the problem concerning data reliability and validity in calculating the features. For example, to obtain the proportion of open-access publications, we identified the access form of articles in 2019 on Unpaywall. Many journals have changed their business model to open-access or closed-access. We can not be sure about the accessibility of papers at the time of publishing and two years time windows that we considered to calculate the number of received citations. Also, we measured the mobility feature similar to our previous paper [17], and the mentioned limitations in that paper exist for this feature too.

6 Main findings and discussion

In this study, we comprehensively investigated the impact of different feature categories on predicting the h-index for researchers at various career stages. By employing a machine learning approach and extensive feature analysis, our main objective was to understand the factors influencing researchers' future scholarly impact and how these factors differ based on their career stage.

The contributions of this research are threefold, as outlined in the introduction. Firstly, we explored the impact of various features on predicting researchers' h-index across different career stages by employing the feature selection technique, RFE, and implementing predictive models for various feature combinations. This analysis gave us valuable insights into the predictive power of different attributes and their varying effectiveness at different

career phases. Our analysis of Table 7 and Fig. 3(a) revealed that models with prior impact-based features demonstrated better performance than those without these features. This finding suggests that prior impact-based features are more reliable predictors of future scholarly impact, particularly for researchers in later career stages, both in the short and long term. Conversely, the smaller performance gap between models with prior impact-based feature combinations and models without such features for junior researchers in the short term, and the superiority of models with non-prior impact-based features over models with prior impact-based features in the long term (as shown in Table 7), indicates that non-prior impact-based features play a more prominent role, particularly in long-term predictions, for younger researchers. This implies that these non-prior impact-based features could be valuable for identifying rising stars with strong potential for future scientific impact.

Secondly, our investigation delved into the temporal dimension of feature performance, encompassing both prior impact-based and non-prior impact-based features. We made notable observations by examining different feature combinations and their predictive power over time. Prior impact-based features exhibited the highest predictive accuracy in the short term, but their performance significantly declined in the long term compared to other features. This finding underscores the importance of considering non-prior impact-based features for enhancing long-term predictions.

Lastly, we introduced novel author (e.g., demographic characteristics) and paper/venue-specific features to estimate the author's h-index and assessed their impact on prediction tasks through feature selection analysis. The results revealed interesting insights into the individual contributions of these features to researchers' scientific impact. Among the introduced features, gender showed the weakest predictive power, suggesting that gender has almost no impact on the scientific impact, which is desirable. However, *OpenAccessRatio* emerged as one of the top five powerful predictors for junior and mid-level seniors in the short term and held a similar position for seniors in the long term. In contrast, *DisciplineMobility* ranked as the second top predictor for researchers from any career stage in the short term but exhibited weaker predictive power in the long term. The higher ranking of *MaxCoauthorHindex* in predicting the h-index for researchers in earlier career stages, both in the short and long term, highlighted the significance of co-authors and their reputation in forecasting future h-index values. Additionally, *InternationalCoauthorRatio* was among the top five predictors for mid-level researchers in the long term, while the *MainField* also held a place among the top five predictors, indicating a strong association of the h-index with specific research fields. Notably, *SocialSciences* featured as one of the top predictors for senior researchers, while *PhysicalSciences* played a similar role for junior and mid-level researchers in the long term, suggesting that predicting the h-index of seniors and certain disciplines in the long term is more feasible. On the other hand, *MobilityScore* demonstrated no significant impact on the h-index for any of the three groups of researchers, except for mid-level researchers in the long term, where it ranked fourth. Finally, other newly introduced features, such as *KeywordPopularity* and *PrimaryAuthorRatio*, had minimal impact due to their low ranking in the feature selection process.

Additionally, the results of the correlation analysis were consistent with the feature selection findings. A positive moderate correlation coefficient was observed between the authors' international mobility and their future h-index. However, given the low proportion of mobile researchers (about 27%), this author's feature proved less effective in predicting

the h-index when accounting for other factors. Conversely, we found a very weak correlation between gender and the h-index, with gender displaying the lowest importance in predicting the h-index among all features. The results also underscored the importance of focusing on the study's field to achieve a better scientific impact. Paper/venue-specific features were shown to have more impact on the future h-index than the author's demographic and co-authorship characteristics.

The performances of proposed models indicate that still more features that don't depend on the history of publications and citations are required to forecast the future h-index of young researchers. For example, [13, 15] focused on analyzing the co-authorship network to investigate the relationship between the structural role of authors in the network and the future h-index. Using such intensive network analysis in our study could improve the performance, particularly for junior researchers with lower impact history in their profiles. Additionally, the textual content of papers examined by [13] and topic authority by [49] could be combined with the introduced features in this study to enhance the predictive power of our models. By incorporating these additional features alongside the ones introduced in our research, we may offer a more comprehensive understanding of researchers' future scholarly impact and lead to more accurate predictions for early-career academics.

7 Conclusion

This study aims to reveal the factors associated with the future h-index of researchers based on bibliometric data, which allowed us to have various researchers groups from different countries and scientific fields for more comprehensive analyses. The results can be informative for researchers to understand how bibliometric characteristics of authors and papers can influence the future h-index and for policymakers to support them by focusing on the factors having positive relations with scientific success. We admit that the h-index, which is the most popular metric to assess the scholars, suffers from some limitations (e.g., field-dependent [53], incapable of comparing researchers in different career stages [24] and detect authors with extremely highly cited papers [54], can be manipulated by self-citations [55]). Our work is not about promoting the h-index, but acknowledging its deficiencies to better understand what factors influence it. Without understanding these factors, researchers cannot understand its biases. Hence we actually contribute to understanding the deficiencies. In addition, possible bias by missing data (e.g., including only authors with gender status) can affect the validity of models. In addition, margin error has not been indicated in this study, and the reliability level of these models is uncertain.

To predict the scientific impact, we employed artificial intelligence (AI) models, which are supposed to mimic human decision-making for assessment and don't necessarily lead to ethical and desirable results. One ethical issue is considering certain features that cause discriminatory effects or introduce bias against certain groups in the predicting model [56, 57], which we don't intend in this study. For example, investigating gender as a predictor in the prediction model was to study gender inequality in science for more attention in policy-making.

Acknowledgements
We acknowledge the support of the German Competence Center for Bibliometrics (grant: 01PQ17001) for maintaining the used dataset for the analyses.

Funding
Open Access funding enabled and organized by Projekt DEAL. This work is financially supported by BMBF project OASE, grant number 01PU17005A.

Abbreviations
GB, Gradient Boosting; GBDT, Gradient-Boosting Decision Tree; GBRT, Gradient Boosted Regression Trees; GDP, Gross Domestic Product; KNN, K-nearest neighbour; MAPE, Mean Absolute Percentage Error; NN, Neural Networks; OACA, Open Access Citation Advantage; RF, Random Forest; RFE, Recursive Feature Elimination; RMSE, Root Mean Square Error; sMAPE, symmetric Mean Absolute Percentage Error; SVR, Support Vector Regression; wPR, weighted Percentile Ranking; XGBoost, Extreme Gradient Boosting.

Availability of data and materials
We don't have permission to redistribute Spocus's raw data, but processed data used for the analyses are available and documented in the Git repository Git repository.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author contributions
SD supervised this study, and PM was the project leader that supported it financially. Material preparation, data collection, Methodology, analysis, Validation, and Visualization were performed by FM. FM wrote the first draft of the manuscript, and all authors commented on previous versions. All authors read and approved the final manuscript.

Authors' information
Fakhri Momeni is a research associate at GESIS – Leibniz Institute for the Social Sciences in Cologne and Ph.D. student in information science at Heinrich Heine University in Duesseldorf. Dr. Philipp Mayr is a team leader (Information & Data Retrieval) at GESIS in Cologne, department Knowledge Technologies for the Social Sciences (KTS). Prof. Dr. Stefan Dietze is Professor of Data & Knowledge Engineering at Heinrich Heine University Duesseldorf and Scientific Director of the Knowledge Technologies department for the Social Sciences at GESIS in Cologne.

Author details
[1]GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany.
[2]Heinrich-Heine-University, Universitätsstr. 1, 40225 Düsseldorf, Germany.

Received: 8 July 2022 Accepted: 23 September 2023 Published online: 06 October 2023

References
1. Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci 102(46):16569–16572
2. Egghe L et al (2006) An improvement of the h-index: the g-index. ISSI Newsl 2(1):8–9
3. Kaur J, Radicchi F, Menczer F (2013) Universality of scholarly impact metrics. J Informetr 7(4):924–932
4. Daud A, Abbasi R, Muhammad F (2013) Finding rising stars in social networks. In: International conference on database systems for advanced applications. Springer, Berlin, pp 13–24
5. Ayaz S, Masood N, Islam MA (2018) Predicting scientific impact based on h-index. Scientometrics 114(3):993–1010
6. Weihs L, Etzioni O (2017) Learning to predict citation-based impact measures. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, Los Alamitos, pp 1–10
7. Wu Z, Lin W, Liu P, Chen J, Mao L (2019) Predicting long-term scientific impact based on multi-field feature extraction. IEEE Access 7:51759–51770
8. Bai X, Zhang F, Lee I (2019) Predicting the citations of scholarly paper. J Informetr 13(1):407–418
9. Abrishami A, Aliakbary S (2019) Predicting citation counts based on deep neural network learning techniques. J Informetr 13(2):485–499
10. Jiang S, Koch B, Sun Y (2021) Hints: citation time series prediction for new publications via dynamic heterogeneous information network embedding. In: Proceedings of the web conference 2021, pp 3158–3167
11. Ruan X, Zhu Y, Li J, Cheng Y (2020) Predicting the citation counts of individual papers via a bp neural network. J Informetr 14(3):101039
12. Kossmeier M, Heinze G (2019) Predicting future citation counts of scientific manuscripts submitted for publication: a cohort study in transplantology. Transpl Int 32(1):6–15
13. Nikolentzos G, Panagopoulos G, Evdaimon I, Vazirgiannis M (2021) Can author collaboration reveal impact? The case of h-index pp 177–194
14. Nie Y, Zhu Y, Lin Q, Zhang S, Shi P, Niu Z (2019) Academic rising star prediction via scholar's evaluation model and machine learning techniques. Scientometrics 120(2):461–476
15. McCarty C, Jawitz JW, Hopkins A, Goldman A (2013) Predicting author h-index using characteristics of the co-author network. Scientometrics 96(2):467–483
16. Dong Y, Johnson RA, Chawla NV (2016) Can scientific impact be predicted? IEEE Trans Big Data 2(1):18–30
17. Momeni F, Karimi F, Mayr P, Peters I, Dietze S (2022) The many facets of academic mobility and its impact on scholars' career. J Informetr 16(2):101280
18. Singh V (2018) Comparing research productivity of returnee-phds in science, engineering, and the social sciences. Scientometrics 115(3):1241–1252
19. Netz N, Hampel S, Aman V (2020) What effects does international mobility have on scientists' careers? A systematic review. Res Eval 29(3):327–351
20. Liu J, Wang R, Xu S (2021) What academic mobility configurations contribute to high performance: an fsqca analysis of csc-funded visiting scholars. Scientometrics 126(2):1079–1100

21. Radford DM, Parangi S, Tu C, Silver JK (2022) h-index and academic rank by gender among breast surgery fellowship faculty. J Women's Health 31(1):110–116
22. Carter TE, Smith TE, Osteen PJ (2017) Gender comparisons of social work faculty using h-index scores. Scientometrics 111(3):1547–1557
23. Lopez SA, Svider PF, Misra P, Bhagat N, Langer PD, Eloy JA (2014) Gender differences in promotion and scholarly impact: an analysis of 1460 academic ophthalmologists. J Surg Educ 71(6):851–859
24. Kelly CD, Jennions MD (2006) The h index and career assessment by numbers. Trends Ecol Evol 21(4):167–170
25. Leydesdorff L, Bornmann L, Wagner CS (2019) The relative influences of government funding and international collaboration on citation impact. J Assoc Inf Sci Technol 70(2):198–201
26. Smirnova N, Mayr P (2023) A comprehensive analysis of acknowledgement texts in web of science: a case study on four scientific domains. Scientometrics 128(1):709–734
27. Gantman ER (2012) Economic, linguistic, and political factors in the scientific productivity of countries. Scientometrics 93(3):967–985
28. Confraria H, Godinho MM, Wang L (2017) Determinants of citation impact: a comparative analysis of the global south versus the global North. Res Policy 46(1):265–279
29. Malesios C, Psarakis S (2014) Comparison of the h-index for different fields of research using bootstrap methodology. Qual Quant 48(1):521–545
30. Lillquist E, Green S (2010) The discipline dependence of citation statistics. Scientometrics 84(3):749–762
31. Iglesias J, Pecharromán C (2007) Scaling the h-index for different scientific isi fields. Scientometrics 73(3):303–320
32. Petersen AM, Penner O (2014) Inequality and cumulative advantage in science careers: a case study of high-impact journals. EPJ Data Sci 3:1
33. Xie F, Ghozy S, Kallmes DF, Lehman JS (2022) Do open-access dermatology articles have higher citation counts than those with subscription-based access? PLoS ONE 17(12):0279265
34. Blair LD, Odell JD (2020) The open access policy citation advantage for a medical school
35. Ottaviani J (2016) The post-embargo open access citation advantage: it exists (probably), it's modest (usually), and the rich get richer (of course). PLoS ONE 11(8):0159614
36. Amjad T, Sabir M, Shamim A, Amjad M, Daud A (2022) Investigating the citation advantage of author-pays charges model in computer science research: a case study of Elsevier and Springer. Libr Hi Tech 40(3):685–703
37. Langham-Putrow A, Bakker C, Riegelman A (2021) Is the open access citation advantage real? A systematic review of the citation of open access and subscription-based articles. PLoS ONE 16(6):0253129
38. Fraser N, Momeni F, Mayr P, Peters I (2020) The relationship between biorxiv preprints, citations and altmetrics. Quant Sci Stud 1(2):618–638
39. Momeni F, Dietze S, Mayr P, Biesenbender K, Peters I (2023) Which factors are associated with Open Access publishing? A Springer Nature case study. Quant Sci Stud 4(2):353–371
40. Hsu J-W, Huang D-W (2011) Correlation between impact and collaboration. Scientometrics 86(2):317–324
41. Puuska H-M, Muhonen R, Leino Y (2014) International and domestic co-publishing and their citation impact in different disciplines. Scientometrics 98(2):823–839
42. Sarigöl E, Pfitzner R, Scholtes I, Garas A, Schweitzer F (2014) Predicting scientific success based on coauthorship networks. EPJ Data Sci 3:1
43. Ni P, An X (2018) Relationship between international collaboration papers and their citations from an economic perspective. Scientometrics 116(2):863–877
44. Karimi F, Wagner C, Lemmerich F, Jadidi M, Strohmaier M (2016) Inferring gender from names on the web: a comparative evaluation of gender detection methods. In: Proceedings of the 25th international conference companion on World Wide Web, pp 53–54
45. Bornmann L, Mutz R (2014) From p100 to p100': a new citation-rank approach. J Assoc Inf Sci Technol 65(9):1939–1943
46. Bornmann L, Williams R (2020) An evaluation of percentile measures of citation impact, and a proposal for making them better. Scientometrics 124(2):1457–1478
47. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd international conference on knowledge discovery and data mining, pp 785–794
48. Blasco BC, Moreno JJM, Pol AP, Abad AS (2013) Using the r-mape index as a resistant measure of forecast accuracy. Psicothema 25(4):500–506
49. Dong Y, Johnson RA, Chawla NV (2015) Will this paper increase your h-index? Scientific impact prediction. In: Proceedings of the eighth ACM international conference on web search and data mining, pp 149–158
50. Artur M (2021) Review the performance of the Bernoulli naïve Bayes classifier in intrusion detection systems using recursive feature elimination with cross-validated selection of the best number of features. Proc Comput Sci 190:564–570
51. Zhao L, Deng F, Zhang X, Yu N (2022) Rfe based feature selection improves performance of classifying multiple-causes deaths in colorectal cancer. In: 2022 7th International Conference on Intelligent Informatics and Biomedical Science (ICIIBMS), vol 7. IEEE, Los Alamitos, pp 188–194
52. Newbold P, Carlson WL, Thorne B (2013) Statistics for business and economics. Pearson Education, Upper Saddle River
53. Grech V, Rizk DE (2018) Increasing importance of research metrics: journal impact factor and h-index. Springer, Berlin
54. Egghe L (2006) Theory and practise of the g-index. Scientometrics 69(1):131–152
55. Bartneck C, Kokkelmans S (2011) Detecting h-index manipulation through self-citation analysis. Scientometrics 87(1):85–98
56. Asaro PM (2019) Ai ethics in predictive policing: from models of threat to an ethics of care. IEEE Technol Soc Mag 38(2):40–53
57. Zuiderveen Borgesius F et al (2018) Discrimination, artificial intelligence, and algorithmic decision-making. Línea. Council of Europe

Publisher's Note

Chapter 7

Conclusion

7.1 Summary and Main Results

This thesis is the culmination of several studies investigating the factors influencing scientific impact. In the following, we summarize the key findings of these studies.

In Chapter 4, we investigated the OA effects and influential factors on the publishing model. The first study's results in Section 4.2.1 analyzed the conversion of journals from a CA to an OA model and its impact on publication volumes and citation metrics. The findings showed that journals that flipped to OA generally experienced an increase in their impact factors and article- and journal-level citation metrics. However, these effects varied across scientific fields. It was also observed that the number of published articles increased after flipping, indicating a higher tendency among authors to submit to OA journals. The willingness to submit to OA journals with APCs can be influenced by funding availability, with authors in health and life sciences more willing to invest in OA publishing.

In Section 4.2.2, we presented the paper investigating the relationship between author-specific and structural factors, OA publishing, and OA citation advantage using a case study of articles published by Springer Nature in 2017 and 2018. The study examines the income level of authors' affiliation countries, the proportion of gold OA publications, and the impact of citations for different publishing models and income levels. The results show

that authors from lower-middle-income countries with eligibility for APC discounts have a lower proportion of gold OA publications than others, indicating that discounted APCs may still be too expensive for authors from these countries. However, authors from low-income countries have a higher proportion of gold OA publications due to receiving APC waivers. The OA model, whether gold OA or hybrid, has the highest percentage of highly cited papers for all countries, and higher income levels are associated with a higher proportion of highly cited articles. Correlation, regression, and machine learning analyses are conducted to identify factors related to OA publishing, and the most influential factors are found to be the country's income and prior experience with OA publishing. In summary, the study's results indicate that OA publishing contributes to inequality in scientific success. The observed "Matthew effect," characterized by the phenomenon of "the rich get richer, and the poor get poorer" [67, 111], demonstrates the unintended consequence of OA. This effect highlights the need to address the disparities and challenges associated with OA publishing to achieve a more equitable scholarly landscape.

The study introduced in Chapter 5 analysed worldwide scholars' international mobility to present the differences in mobility patterns based on individuals' factors and disparities in mobility programs. We found gender inequality in mobility across countries and scientific fields. Future work can investigate the reasons and motivations for lower/higher proportion of mobile researchers in different communities to solve parity issues in academic mobility. We examined the mobility outcomes and found a generally positive scientific impact for researchers. The next analysis can compare the mobility impact of researchers from different countries in more detail, considering their income level, language, and cultural background, to better understand the barriers to a positive outcome for researchers and countries.

Chapter 6 featured a study that examined the contribution of various features in predicting researchers' h-index for authors in different career stages. The models were implemented for

junior, mid-level, and senior researchers, and the accuracy of the models was compared. It was found that more experienced researchers had higher performance consistent with previous studies. Prior impact-based features were identified as reliable predictors of researchers' future impact, particularly in the later career stages. Non-prior impact-based features played a more prominent role in predicting the long-term impact of younger researchers. The study introduced novel author and paper/venue-specific features and analyzed their impact on h-index prediction. It was found that paper/venue-specific features had more impact on future h-index than demographic and co-authorship characteristics. The study also found a positive moderate correlation between authors' international mobility and their future h-index. However, with low proportions of mobile researchers, this feature had limited effectiveness in predicting the h-index. Gender was found to have a very weak correlation with the h-index and the lowest importance among all features. The study suggested that more features that don't depend on the history of publications and citations, such as the textual content of papers [81], topic authority [25], and author position in collaboration networks based on centrality measures, are required to forecast the h-index of young researchers.

7.2 Limitations and Future Work

7.2.1 Open Access Effect

Journals Flipping to OA

The study presented in Section 4.2.1 had a relatively small sample size, which may not represent all journals that have flipped to OA. Future research could utilize larger-scale aggregators of OA information, such as Unpaywall or CORE (COnnecting REpositories), to obtain a more comprehensive list of flipped journals. The accuracy of Unpaywall data in

classifying articles as OA or CA should also be addressed. Another limitation was the lack of data on submissions to flipped journals, which could better reflect author preferences for OA publishing. Future studies should consider this aspect to provide a more comprehensive understanding of the authors' behavior. The study did not exclude outliers at the article and journal level, which could have affected the measures. Future work should be more sensitive to outliers and consider their impact on the analysis.

The study did not consider the specific business models of flipped journals. These models, which include APC-driven models and support from societies or library presses, bring forth various economic challenges that can potentially impact editorial decisions and acceptance rates. The role of the publisher itself and its platform quality and visibility should also be considered. Long-term changes in institutional and funder support and increasing pressure to transition to OA will further shape the findings. Future research should investigate the complex interactions among these factors.

Future work should incorporate qualitative information from stakeholders such as publishers, funders, libraries, and societies to understand their attitudes, expectations, and motivations for journal flipping and OA publishing. Interviews and qualitative research can complement quantitative bibliometric analyses and provide a more comprehensive understanding of the implications of changing publication models.

Factors Associated with Open Access Publishing

One limitation of the study presented in Section 4.2.2 is that it only includes articles from one publisher, Springer Nature, which may limit the generalisability of the findings to articles published by other publishers. Additionally, the study did not account for journals that may have flipped from a CA model to OA or vice versa, potentially introducing errors in

the results. The accuracy of external data, such as Springer Nature's and Unpaywall's, also affects the precision of the results. The study faced challenges in identifying the gender of authors, particularly for Asian names, resulting in biases in the analyses.

For future work, the study suggests examining other publishers to understand how different APC policies among publishers impact OA publishing. Controlling for the articles' language in the analyses is also recommended to explore the role of languages in international OA publishing and citation advantages. Expanding the analysis to include publishers with non-English content can provide a more comprehensive understanding of the factors influencing OA publishing and citation impacts worldwide.

7.2.2 International Mobility Effect

Several limitations of the study discussed in Chapter 5 should be acknowledged. Using bibliometric data and Scopus author ID introduced potential issues, such as incomplete coverage of authors' article sets and potential errors in specifying mobility and career stages. The fixed periods assumed for early and mid-career stages may not be universally applicable and should be further investigated. The study also did not differentiate between academic researchers and those outside academia, and temporary mobilities like research visits were not considered, warranting future examination.

The collaboration networks were constructed separately for each discipline, potentially affecting the network positions of scholars with interdisciplinary backgrounds. The size of networks and communities to which researchers belong were not controlled, which could impact future career perspectives. Additionally, the study did not account for the authorship position in publications, and future research could explore the influence of mobility on authorship positions.

Future work should also consider destination countries and different types of mobility, such as immigrants and returnees, to provide a more comprehensive understanding of the advantages and disadvantages of mobility for both origin and destination countries. Investigating mobility within scientific fields and its relation to geographic mobility will contribute to understanding knowledge transfer between fields. Analyzing co-authorship networks among different types of mobile researchers can uncover collaboration patterns and their impact on other researchers.

The impact of the COVID-19 pandemic on scholars' careers and collaboration patterns is another area for future investigation. Comparing collaboration patterns before and after the pandemic, and examining the importance of physical mobility versus virtual collaboration during restricted mobility periods, can provide insights into the potential alternatives for scientists.

In conclusion, this study identified various dimensions of academic mobility and its effects on scholars' careers. While highlighting limitations, the findings present opportunities for future research to address the complexities and inequalities associated with academic mobility.

7.2.3 Predicting the Researchers' h-index

The study outlined in Chapter 6 acknowledged several limitations. Firstly, it only considered journal papers and not conference papers which may introduce bias, especially in disciplines where authors predominantly publish as conference proceedings papers. Secondly, data reliability and validity were mentioned as potential issues, particularly regarding the accessibility of papers at the time of publishing and the two-year time windows used to calculate the number of citations received. The mobility feature used in the study also inherited the limitations mentioned in a previous paper.

Future work was proposed to address these limitations and further improve h-index prediction. The study suggested using non-prior impact-based features to evaluate the rising stars with weak scientific impact in their profiles. It also recommended incorporating textual content analysis of papers and topic authority to enhance the performance of h-index prediction models. Additionally, exploring the relationship between the structural role of authors in the coauthorship network and the future h-index, particularly for junior researchers, was highlighted as an area for future research. The study also emphasized the need for more comprehensive analyses considering researchers from different countries and scientific fields to understand better the factors influencing the h-index. Finally, the study acknowledged the limitations and deficiencies of the h-index as a metric and suggested further research to understand its biases and deficiencies. Also, the findings of this study highlight the challenges faced to assess the feature impact of young researchers, especially women, who experience career interruptions due to parental leave and consequently have lower productivity and citation impact. Future work will focus on studying gender disparity in career interruptions due to parental leave through bibliometric and survey studies. Machine learning prediction models can be applied to identify factors that describe the scientific outcomes of researchers, independent of their current productivity, and to forecast the long-term impact of these researchers. Additionally, the development of an adjusted h-index that accounts for career interruptions can be explored as a means to assess researchers more fairly.

7.3 Closing

In this thesis, we investigated the factors influencing the future impact of researchers. One of the intentions behind our motivations for this research was to address the issue of ”inequality in science” and strive towards a system where researchers’ assessment is solely based on the

content of their research, enabling a fair evaluation of their scientific impact. Moreover, we aimed to promote equality by ensuring that all researchers have equal opportunities to pursue and excel in their scientific achievements, irrespective of their background. Through our investigation of the Open Access effect, we found that while OA publishing improves the visibility of publications and their citation rates, it contributes to inequalities in presenting research. Examining the mobility effect, we observed positive impacts for researchers while uncovering disparities in international mobility across different communities. These issues should be a concern for science policy, as supporting all researchers engaged in science can enhance international collaboration, productivity, and the visibility of their work. Our study also addressed the gender gap and highlighted the underrepresentation of women in mobility, potentially due to family responsibilities. Furthermore, our study on h-index prediction highlighted the challenges faced in assessing the impact of young researchers, particularly women, who experience career interruptions due to parental leave, resulting in lower productivity and citation impact.

Reflecting on our research journey, we acknowledge the ongoing need for continuous efforts to create a more equitable scientific landscape. We are committed to advocating for fairness, diversity, and equal opportunities in science. By raising awareness of these issues and promoting evidence-based policies, we believe we can contribute to transforming the scientific community into a more inclusive and balanced environment that fosters innovation and excellence.

Bibliography

[1] Giovanni Abramo, Ciriaco Andrea D'Angelo, and Flavia Di Costa. The effect of academic mobility on research performance: The case of italy. *Quantitative Science Studies*, 3(2):345–362, 2022.

[2] Ali Abrishami and Sadegh Aliakbary. Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics*, 13(2):485–499, 2019.

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

[4] Dag W Aksnes, Kristoffer Rørstad, Fredrik N Piro, and Gunnar Sivertsen. Are mobile researchers more productive and cited than non-mobile researchers? a large-scale study of norwegian scientists. *Research Evaluation*, 22(4):215–223, 2013.

[5] Valeria Aman. Does the scopus author id suffice to track scientific international mobility? a case study based on leibniz laureates. *Scientometrics*, 117(2):705–720, 2018.

[6] Valeria Aman. Transfer of knowledge through international scientific mobility: introduction of a network-based bibliometric approach to study different knowledge types. *Quantitative Science Studies*, 1(2):565–581, 2020.

[7] Tehmina Amjad, Nafeesa Shahid, Ali Daud, and Asma Khatoon. Citation burst prediction in a bibliometric network. *Scientometrics*, 127(5):2773–2790, 2022.

[8] Éric Archambault, Didier Amyot, Philippe Deschamps, Aurore Nicol, Françoise Provencher, Lise Rebout, and Guillaume Roberge. Proportion of open access papers

published in peer-reviewed journals at the european and world levels—1996–2013. *Unknown*, 2014.

[9] Samreen Ayaz, Nayyer Masood, and Muhammad Arshad Islam. Predicting scientific impact based on h-index. *Scientometrics*, 114(3):993–1010, 2018.

[10] Tobias Backes. Effective unsupervised author disambiguation with relative frequencies. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*, pages 203–212, 2018.

[11] Xiaomei Bai, Fuli Zhang, and Ivan Lee. Predicting the citations of scholarly paper. *Journal of Informetrics*, 13(1):407–418, 2019.

[12] Siddikov Ilyosjon Bakhromovich. Development trends and transformation processes in academic mobility in higher education in uzbekistan and the world. *Bakhromovich, Siddikov Ilyosjon, and Maxamadaliev Lutfillo." Development of ecological culture in students in the process of education of history of uzbekistan.–2021*, 2021.

[13] Stefano H Baruffaldi, Marianna Marino, and Fabiana Visentin. Money to move: The effect on researchers of an international mobility grant. *Research Policy*, 49(8):104077, 2020.

[14] Stefano Horst Baruffaldi, Marianna Marino, and Fabiana Visentin. International mobility and research careers: Evidence from a mobility grant program. In *Academy of Management Proceedings*, page 12523. Academy of Management, 2017.

[15] Harald Bauder. International mobility and social capital in the academic field. *Minerva*, 58(3):367–387, 2020.

[16] Marine Bernard, Bastien Bernela, and Marie Ferru. Does the geographical mobility

of scientists shape their collaboration network? a panel approach of chemists' careers. *Papers in Regional Science*, 100(1):79–99, 2021.

[17] Hanjo Boekhout, Inge van der Weijden, and Ludo Waltman. Gender differences in scientific careers: A large-scale bibliometric analysis. *arXiv preprint arXiv:2106.12624*, 2021.

[18] Paolo Boldi, Sebastiano Vigna, and Matteo Zignani. Gender differences in collaboration patterns in computer science: An analysis of the DBLP bibliography. *Journal of Informetrics*, 13(1):321–338, 2019.

[19] S Fortunato CT Bergstrom K Borner, JA Evans D Helbing S Milojevic, AM Petersen F Radicchi R Sinatra, B Uzzi A Vespignani L Waltman, D Wang, and AL Barabasi. Science of science. *Science*, 359:6379, 2018.

[20] Jeroen Bosman and Bianca Kramer. Open access levels: a quantitative exploration using web of science and oadoi data. Technical report, PeerJ Preprints, 2018.

[21] Emiel Caron and Nees Jan van Eck. Large scale author name disambiguation using rule-based scoring and clustering. In *Proceedings of the 19th international conference on science and technology indicators*, pages 79–86. CWTS-Leiden University, Leiden, 2014.

[22] Yibo Chen, Zhiyi Jiang, Jianliang Gao, Hongliang Du, Liping Gao, and Zhao Li. A supervised and distributed framework for cold-start author disambiguation in large-scale publications. *Neural Computing and Applications*, pages 1–16, 2021.

[23] Sonia Conchi and Carolin Michels. Scientific mobility: An analysis of germany, austria, france and great britain. Technical report, Fraunhofer ISI Discussion Papers-Innovation Systems and Policy Analysis, 2014.

[24] Michelle L Dion, Jane Lawrence Sumner, and Sara McLaughlin Mitchell. Gendered citation patterns across political science and social science methodology fields. *Political analysis*, 26(3):312–327, 2018.

[25] Yuxiao Dong, Reid A Johnson, and Nitesh V Chawla. Will this paper increase your h-index? scientific impact prediction. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 149–158, 2015.

[26] Yuxiao Dong, Reid A Johnson, and Nitesh V Chawla. Can scientific impact be predicted? *IEEE Transactions on Big Data*, 2(1):18–30, 2016.

[27] Pablo Dorta-González and María Isabel Dorta-González. The influence of funding on the open access citation advantage. *Journal of Scientometric Research*, 12(1):68–78, 2023.

[28] Olof Ejermo, Claudio Fassio, and John Källström. Does mobility across universities raise scientific productivity? *Oxford Bulletin of Economics and Statistics*, 82(3):603–624, 2020.

[29] Jamal El-Ouahi and Vincent Larivière. On the lack of women researchers in the middle east and north africa. *Scientometrics*, pages 1–28, 2023.

[30] Jamal El-Ouahi, Nicolas Robinson-García, and Rodrigo Costas. Analyzing scientific mobility and collaboration in the middle east and north africa. *Quantitative Science Studies*, 2(3):1023–1047, 2021.

[31] Lina A Elfaki, Jessica GY Luc, Mara B Antonoff, David T Cooke, Rakesh C Arora, Nikki Stamp, Thomas K Varghese Jr, and Maral Ouzounian. Sex differences in authorship in cardiothoracic surgery during the early coronavirus disease 2019 pandemic. *JTCVS open*, 11:265–271, 2022.

[32] faceplusplus. face++. https://www.faceplusplus.com/, 2023. Accessed: 2023.

[33] Daniele Fanelli, Matteo Schleicher, Ferric C Fang, Arturo Casadevall, and Elisabeth M Bik. Do individual and institutional predictors of misconduct vary by country? results of a matched-control analysis of problematic image duplications. *PloS one*, 17(3): e0255334, 2022.

[34] Ana Fernández-Zubieta, Aldo Geuna, and Cornelia Lawson. Productivity pay-offs from academic mobility: should i stay or should i go? *Industrial and Corporate Change*, 25 (1):91–114, 2016.

[35] Márcia R Ferreira, Philippe Mongeon, and Rodrigo Costas. Large-scale comparison of authorship, citations, and tweets of web of science authors. *Journal of altmetrics*, 4 (1), 2021.

[36] Eitan Frachtenberg. Research artifacts and citations in computer systems papers. *PeerJ Computer Science*, 8:e887, 2022.

[37] Chiara Franzoni, Giuseppe Scellato, and Paula Stephan. Context factors and the performance of mobile individuals in research teams. *Journal of Management Studies*, 55(1):27–59, 2018.

[38] Nicholas Fraser, Fakhri Momeni, Philipp Mayr, and Isabella Peters. The relationship between biorxiv preprints, citations and altmetrics. *Quantitative Science Studies*, 1 (2):618–638, 2020.

[39] Yassine Gargouri, Chawki Hajjem, Vincent Larivière, Yves Gingras, Les Carr, Tim Brody, and Stevan Harnad. Self-selected or mandated, open access increases citation impact for higher quality research. *PloS one*, 5(10):e13636, 2010.

[40] GENDERIZE. Genderize.io. https://genderize.io/, 2023. Accessed: 2023.

[41] Panagiotis Giannos, Konstantinos Katsikas Triantafyllidis, Maria Paraskevaidi, Maria Kyrgiou, and Konstantinos S Kechagias. Female dynamics in authorship of scientific publications in the public library of science: A 10-year bibliometric analysis of biomedical research. *European Journal of Investigation in Health, Psychology and Education*, 13(2):228–237, 2023.

[42] Gali Halevi, Henk F Moed, and Judit Bar-Ilan. Researchers' mobility, productivity and impact: Case of top producing authors in seven disciplines. *Publishing Research Quarterly*, 32:22–37, 2016.

[43] Frank Havemann and Birger Larsen. Bibliometric indicators of young authors in astrophysics: Can later stars be predicted? *Scientometrics*, 102(2):1413–1434, 2015.

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

[45] Muhammad Ilyas, Ghulam Mustafa, Hafiz Mohkam Jamil, and Omid Dehzangi. An overview of co-authorship network analysis: State-of-the-art and challenges. *Information Processing & Management*, 57(2):102073, 2020.

[46] Song Jiang, Bernard Koch, and Yizhou Sun. Hints: Citation time series prediction for new publications via dynamic heterogeneous information network embedding. In *Proceedings of the Web Conference 2021*, pages 3158–3167, 2021.

[47] Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International conference companion on World Wide Web*, pages 53–54, 2016.

[48] Fariba Karimi, Philipp Mayr, and Fakhri Momeni. Analyzing the network structure

and gender differences among the members of the networked knowledge organization systems (nkos) community. *International Journal on Digital Libraries*, 20(3):231–239, 2019.

[49] Hirotaka Kawashima and Hiroyuki Tomizawa. Accuracy evaluation of scopus author id based on the largest funding database in japan. *Scientometrics*, 103(3):1061–1071, 2015.

[50] Molly M King, Carl T Bergstrom, Shelley J Correll, Jennifer Jacquet, and Jevin D West. Men set their own cites high: Gender and self-citation across fields and over time. *Socius*, 3:2378023117738903, 2017.

[51] Michael Kossmeier and Georg Heinze. Predicting future citation counts of scientific manuscripts submitted for publication: a cohort study in transplantology. *Transplant International*, 32(1):6–15, 2019.

[52] Thomas Krämer, Fakhri Momeni, and Philipp Mayr. Coverage of author identifiers in web of science and scopus. *arXiv preprint arXiv:1703.01319*, 2017.

[53] Matthias Kuppler. Predicting the future impact of computer science researchers: Is there a gender bias? *Scientometrics*, 127(11):6695–6732, 2022.

[54] Keiko Kurata, Tomoko Morioka, Keiko Yokoi, and Mamiko Matsubayashi. Remarkable growth of open access in the biomedical field: analysis of pubmed articles from 2006 to 2010. *PloS one*, 8(5):e60925, 2013.

[55] Allison Langham-Putrow, Caitlin Bakker, and Amy Riegelman. Is the open access citation advantage real? a systematic review of the citation of open access and subscription-based articles. *PloS one*, 16(6):e0253129, 2021.

[56] Vincent Larivière and Rodrigo Costas. How many is too many? on the relationship between research productivity and impact. *PloS one*, 11(9):e0162709, 2016.

[57] Vincent Larivière and Cassidy R Sugimoto. Do authors comply when funders enforce open access to research?, 2018.

[58] Regula Julia Leemann. Gender inequalities in transnational academic mobility and the ideal type of academic entrepreneur. *Discourse: Studies in the Cultural Politics of Education*, 31(5):605–625, 2010.

[59] Marc J Lerchenmueller, Olav Sorenson, and Anupam B Jena. Gender differences in how scientists present the importance of their research: observational study. *bmj*, 367, 2019.

[60] Maggi WH Leung. Unsettling the yin-yang harmony: An analysis of gender inequalities in academic mobility among chinese scholars. *Asian and Pacific Migration Journal*, 23(2):155–182, 2014.

[61] Colby Lil Lewis. The open access citation advantage: Does it exist and what does it mean for libraries? *Information Technology and Libraries*, 37(3):50–65, 2018.

[62] Feng Li and Li Tang. When international mobility meets local connections: Evidence from china. *Science and Public Policy*, 46(4):518–529, 2019.

[63] Abdelghani Maddi and David Sapinho. Does open access really increase impact? a large-scale randomized analysis. In *26th International conference on science and thechnology indicators/ STI2022*, 2022.

[64] Daniel Maliniak, Ryan Powers, and Barbara F Walter. Gendered citation patterns in international relations journals. *International Organization*, 67(4):889–922, 2013.

[65] Mark J McCabe and Christopher M Snyder. Identifying the effect of open access on citations using a panel of science journals. *Economic inquiry*, 52(4):1284–1300, 2014.

[66] Erin C McKiernan, Philip E Bourne, C Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, Brian A Nosek, Karthik Ram, Courtney K Soderberg, et al. How open science helps researchers succeed. *elife*, 5, 2016.

[67] Robert K Merton. The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63, 1968.

[68] Helena Mihaljević and Lucía Santamaría. Disambiguation of author entities in ads using supervised learning and graph theory methods. *Scientometrics*, 126(5):3893–3917, 2021.

[69] Henk F Moed. The effect of "open access" on citation impact: An analysis of arxiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, 58(13):2047–2054, 2007.

[70] Philippe Moguérou et al. A double gender-family inequality phenomenon in the international mobility of young researchers. *online][Cit. 12. 7. 2007]. Dostupné z:< http://129.3*, 20, 2004.

[71] Fakhri Momeni and Philipp Mayr. Analyzing the research output presented at european networked knowledge organization systems workshops (2000-2015). In *NKOS@ TPDL*, pages 7–14, 2016.

[72] Fakhri Momeni and Philipp Mayr. Evaluating co-authorship networks in author name disambiguation for common names. In *International Conference on Theory and Practice of Digital Libraries*, pages 386–391. Springer, 2016.

[73] Fakhri Momeni and Philipp Mayr. Using co-authorship networks for author name disambiguation. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 261–262, 2016.

[74] Fakhri Momeni, Philipp Mayr, Nicholas Fraser, and Isabella Peters. What happens when a journal converts to open access? a bibliometric analysis. *Scientometrics*, 126 (12):9811–9827, 2021.

[75] Fakhri Momeni, Fariba Karimi, Philipp Mayr, Isabella Peters, and Stefan Dietze. The many facets of academic mobility and its impact on scholars' career. *Journal of Informetrics*, 16(2):101280, 2022.

[76] Fakhri Momeni, Stefan Dietze, Philipp Mayr, Kristin Biesenbender, and Isabella Peters. Which factors are associated with open access publishing? a springer nature case study. *Quantitative Science Studies*, pages 1–26, 2023.

[77] Fakhri Momeni, Philipp Mayr, and Stefan Dietze. Investigating the contribution of author-and publication-specific features to scholars'h-index prediction. *EPJ Data Science*, 12(1):45, 2023.

[78] M. E. J. Newman. Social network analysis of collaboration networks in science. *Advances in Complex Systems*, 4(1):89–92, 2001.

[79] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[80] Yubing Nie, Yifan Zhu, Qika Lin, Sifan Zhang, Pengfei Shi, and Zhendong Niu. Academic rising star prediction via scholar's evaluation model and machine learning techniques. *Scientometrics*, 120(2):461–476, 2019.

[81] Giannis Nikolentzos, George Panagopoulos, Iakovos Evdaimon, and Michalis Vazirgiannis. Can author collaboration reveal impact? the case of h-index. In *Predicting the Dynamics of Research Impact*, pages 177–194. Springer, 2021.

[82] Alan Olsen. International mobility of australian university students: 2005. *Journal of Studies in International Education*, 12(4):364–374, 2008.

[83] ORCID. Orcid id. https://info.orcid.org/what-is-orcid/, 2023. Accessed: 2023.

[84] Orion Penner, Raj K Pan, Alexander M Petersen, Kimmo Kaski, and Santo Fortunato. On the predictability of future impact in science. *Scientific reports*, 3(1):1–8, 2013.

[85] Alexander M Petersen. Multiscale impact of researcher mobility. *Journal of The Royal Society Interface*, 15(146):20180580, 2018.

[86] Fernando Pinto. The effect of university graduates' international mobility on labour outcomes in spain. *Studies in Higher Education*, 47(1):26–37, 2022.

[87] Heather Piwowar, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, and Stefanie Haustein. The state of oa: a large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, 6: e4375, 2018.

[88] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[89] Andreas Rehs. A supervised machine learning approach to author disambiguation in the web of science. *Journal of Informetrics*, 15(3):101166, 2021.

[90] researcherID. researcherid. https://webofscience.help.clarivate.com/Content/wos-researcher-id.htm, 2023. Accessed: 2023.

[91] Lauren A Rivera. Hiring as cultural matching: The case of elite professional service firms. *American sociological review*, 77(6):999–1022, 2012.

[92] Nicolás Robinson-Garcia, Cassidy R Sugimoto, Dakota Murray, Alfredo Yegros-Yegros, Vincent Larivière, and Rodrigo Costas. The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics*, 13(1):50–63, 2019.

[93] Nicolas Robinson-Garcia, Rodrigo Costas, and Thed N van Leeuwen. Open access uptake by universities worldwide. *PeerJ*, 8:e9410, 2020.

[94] Xuanmin Ruan, Yuanyang Zhu, Jiang Li, and Ying Cheng. Predicting the citation counts of individual papers via a bp neural network. *Journal of Informetrics*, 14(3): 101039, 2020.

[95] Reyaz Rufai, Sumeer Gul, and Tariq Ahmad Shah. Open access journals in library and information science: the story so far. *Trends in information management*, 7(2): 218–228, 2011.

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

[97] Evi Sachini, Nikolaos Karampekios, Pierpaolo Brutti, and Konstantinos Sioumalas-Christodoulou. Should i stay or should i go? using bibliometrics to identify the international mobility of highly educated greek manpower. *Scientometrics*, 125:641–663, 2020.

[98] Mert A. Sapmaz, Dr. Dileep Mani, Jayashree Sankaran, and Vikram Padaki. Understanding research collaboration through social network analysis. *2017 IEEE International Conference on Big Data (Big Data)*, pages 1180–1189, 2017.

[99] Marie Sautier. Move or perish? sticky mobilities in the swiss academic context. *Higher Education*, 82(4):799–822, 2021.

[100] Anna Severin, Matthias Egger, Martin Paul Eve, and Daniel Hürlimann. Discipline-specific open access publishing practices and barriers to change: an evidence-based review. *F1000Research*, 7, 2018.

[101] Marc-André Simard, Gita Ghiasi, Philippe Mongeon, and Vincent Larivière. Geographic differences in the uptake of open access. In *Proceedings of the 18th International Conference on Scientometrics and Informetrics (ISSI 2021)*, pages 1033–1038, 2021.

[102] Neil R Smalheiser, Vetle I Torvik, et al. Author name disambiguation. *Annual review of information science and technology*, 43(1):1, 2009.

[103] PLOS ONE Staff. Correction: The post-embargo open access citation advantage: It exists (probably), it's modest (usually), and the rich get richer (of course). *PloS one*, 11(10):e0165166, 2016.

[104] Alexander Subbotin and Samin Aref. Brain drain and brain gain in russia: Analyzing international migration of researchers by discipline using scopus bibliometric data 1996–2020. *Scientometrics*, 126(9):7875–7900, 2021.

[105] Cassidy R Sugimoto, Chaoqun Ni, Jevin D West, and Vincent Larivière. Gender homophily in scholarly communication: A large-scale analysis of scientific collaborations. *Information Processing & Management*, 53(1):50–61, 2017.

[106] Alexander Tekles and Lutz Bornmann. Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. *Quantitative Science Studies*, 1(4):1510–1528, 2020.

[107] Brian Uzzi, Satyam Mukherjee, Michael Stringer, Ben Jones, and Ryan Williams. Collaboration in computer science: A network science approach. *Proceedings of the National Academy of Sciences*, 110(Supplement 1):1837–1842, 2013.

[108] Maarten CAG Van der Sanden, Nees Jan Van Eck, and Peter Van den Besselaar. Co-authorship network analysis in the social sciences: The structure and development of interdisciplinary fields, 2000–2018. *Journal of Informetrics*, 15(1):101166, 2021.

[109] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. Gendered patterns of collaboration in wikipedia. In *Proceedings of the 9th International Conference on Web and Social Media*, pages 113–122. AAAI, 2015.

[110] Ludo Waltman, Nees Jan van Eck, and Jan van den Berg. Gender differences in research collaboration. *Scientometrics*, 118(2):507–526, 2019.

[111] Jian Wang. Unpacking the matthew effect in citations. *Journal of Informetrics*, 8(2): 329–339, 2014.

[112] Jue Wang, Rosalie Hooi, Andrew X Li, and Meng-hsuan Chou. Collaboration patterns of mobile academics: The impact of international mobility. *Science and Public Policy*, 46(3):450–462, 2019.

[113] Mingyang Wang, Zhenyu Wang, and Guangsheng Chen. Which can better predict the future success of articles? bibliometric indices or alternative metrics. *Scientometrics*, 119(3):1575–1595, 2019.

[114] Yinqiu Wang, Hui Luo, and Yunyan Shi. Complex network analysis for international talent mobility based on bibliometrics. *International Journal of Innovation Science*, 11(3):419–435, 2019.

[115] Luca Weihs and Oren Etzioni. Learning to predict citation-based impact measures. In *2017 ACM/IEEE joint conference on digital libraries (JCDL)*, pages 1–10. IEEE, 2017.

[116] Ziming Wu, Weiwei Lin, Pan Liu, Jingbang Chen, and Li Mao. Predicting long-term scientific impact based on multi-field feature extraction. *IEEE Access*, 7:51759–51770, 2019.

[117] Jingfeng Xia, Sarah B Gilchrist, Nathaniel XP Smith, Justin A Kingery, Jennifer R Radecki, Marcia L Wilhelm, Keith C Harrison, Michael L Ashby, and Alyson J Mahn. A review of open access self-archiving mandate policies. *portal: Libraries and the Academy*, 12(1):85–102, 2012.

[118] An Zeng, Zhesi Shen, Jianlin Zhou, Jinshan Wu, Ying Fan, Yougui Wang, and H Eugene Stanley. The science of science: From the perspective of complex systems. *Physics reports*, 714:1–73, 2017.

[119] Lin Zhang, Gunnar Sivertsen, Huiying Du, Ying Huang, and Wolfgang Glänzel. Gender differences in the aims and impacts of research. *Scientometrics*, 126:8861–8886, 2021.

[120] Zhenyue Zhao, Yi Bu, Lele Kang, Chao Min, Yiyang Bian, Li Tang, and Jiang Li. An investigation of the relationship between scientists' mobility to/from china and their research performance. *Journal of Informetrics*, 14(2):101037, 2020.