Received: 16 January 2024 | Revised: 21 February 2024 | Accepted: 27 February 2024

DOI: 10.1111/hequ.12512

RESEARCH ARTICLE

Higher Education Quarterly WILEY

# Using large-scale bibliometric data in higher education research: Methodological implications from three studies

Marek Kwiek[1,2] | Hugo Horta[3] | Justin J. W. Powell[4]

[1]Institute for Advanced Studies (IAS), University of Poznan, Poznan, Poland

[2]German Centre for Higher Education Research and Science Studies (DZHW), Berlin, Germany

[3]Faculty of Education, University of Hong Kong, Hong Kong, China

[4]Faculty of Humanities, Education and Social Sciences, University of Luxembourg, Esch-sur-Alzette, Luxembourg

**Correspondence**
Marek Kwiek, University of Poznan, Institute for Advanced Studies (IAS), ul. Wieniawskiego 1, 61-712 Poznan, Poland.
Email: marek.kwiek@amu.edu.pl

**Funding information**
Ministerstwo Edukacji i Nauki

## Abstract

All fields of knowledge are challenged to adopt newer, more sophisticated methodologies to cope with growing complexity. Phenomena under study require further multidisciplinary and mixed methods collaborations to achieve expertise able to improve research strategies and practices. Furthermore, traditional methodological approaches face limits to their analytical reach. Here, we demonstrate opportunities from adopting newer, more sophisticated methodologies in the field of higher education (HE) research by comparing three case studies. We argue that such methods and data innovate the mapping and understanding of global HE. These studies uncover novel field characterizations, enabled via analysis of tens of thousands of HE authors and articles over several decades to assess how journal publication, topics, and levels of analysis (individual, organizational, and system) have evolved. Our results imply that to better understand the future of HE worldwide and to address growing challenges, newer methodological directions and data sources will be key to facilitate more comprehensive examinations of the globalizing field. However, our analysis also highlights the technical and learning challenges in implementing these methodologies; thus, we argue for the need to promote more sophisticated methodological training of current and future generations of HE researchers as well as strengthened collaborations across disciplinary, methodological, and cultural boundaries.

  

## 1 | INTRODUCTION

With new large-scale data sets, tools, and methodologies, we can study huge amounts of data related to higher education (HE), science, and the academic profession to discover patterns that would otherwise remain imperceptible (as it happens in other fields in the social sciences; Hesse et al., 2015). Leveraging comprehensive bibliometric data sets and conducting surveys enables us to trace academic careers, collaborations, and research productivity from cross-sectional or longitudinal perspectives and across countries, regions, cities, and organizations; to analyse academics as individuals and within research teams; ascriptive characteristics such as age and gender; and junior and senior researchers active in all academic disciplines—with different levels of granularity (up to thousands of research topics). This current, unsurpassed level of detail in HE research (individuals with their demographic and professional characteristics; specialized disciplines and research topics) combined with its scale (e.g., whole world, regions, or aggregates such as OECD or EU member countries) is remarkable. Providing unique opportunities, such new data-driven research directions are being applied from global, multi-level, and comparative perspectives (Fu et al., 2022; Horta et al., 2022; Kwiek, 2020). Synthesizing research topics and uncovering gaps supports a more sustainable and responsive research agenda for a globalizing multidisciplinary and stratified field whose research contributions, especially journal articles, continue to grow expansively.

Increasingly, it is possible to analyse higher education, research production, and the academic profession—or specific international, national, or disciplinary academic communities, including the global community of HE researchers—with increasing scope and, simultaneously, ever more precision.[1] These new opportunities to research academia and academics also imply certain limitations. To research active academic communities using large-scale bibliometric data—via measurements of research output or collaboration—practically means researching only publishing academics. Those who do not publish or who contribute in other formats, or those engaged in higher education management, are largely excluded from analysis, skewing the reality of a diverse community that includes many engaged practitioners. Consequently, large-scale studies of productivity, mobility, and collaboration refer to "publishers" of scholarship (and often to highly productive scholars, Kwiek, 2016) rather than to the full and highly diverse HE community. Similarly, the dominant English language scholarship published in leading journals cannot completely represent the world's HE systems, organizations, and individual experiences (Horta, 2018; Kwiek, 2019; Powell et al., 2017), even if the global participation in the HE research community has become much more inclusive in this century, as has scientific publishing generally (Baker & Powell, 2024).

However, large-scale bibliometric data drive the research agenda on global science and scientists in what is termed science of science, research on research, and quantitative science studies, as well as in scientometrics, economics of science, and beyond (e.g., Clauset et al., 2017; Fortunato et al., 2018; Gui et al., 2019; Ioannidis, 2018; Sugimoto & Larivière, 2018; Wang & Barabási, 2021). The crucial supply-side factor behind the increasing use of large-scale data in researching the HE field is the increasing availability of digital data on scholarly inputs and outputs. The demand-side factor is a more general pressure in HE research and policymaking—coming both from within the academia and from outside, from grant-making bodies, policymakers, and the media—to evaluate and understand beyond the single case study and small sample size studies by using larger data sets (with increasingly vast numbers of observations) that facilitate the drawing of valid and reliable conclusions.

Thus, here we explore, exemplify, and reflect on new opportunities that have emerged in the field of HE. We provide an overview of approaches to more fully utilize increasingly available large-scale bibliometric data. We compare three different case studies, each conducted independently by co-authors of this paper, with diverse data, tools, and methodologies.

## 2 | COMPARING THREE CASE STUDIES

We compare detailed case studies of research into the global HE research community and its contemporary collaboration and publication patterns that use newer global data sets and innovative techniques and methodologies.

We review methodological choices made and technical skills needed, discuss trade-offs, potential pitfalls, and limitations. The cases discussed are the research conducted on the evolving international research collaboration network of HE researchers (CS1; Fu et al., 2022); homophily among HE researchers, with a perspective based on co-authorships (CS2; Horta et al., 2022); and the stratification of the HE research community members based on their publishing intensity and country authorship affiliations (CS3; Kwiek, 2021).

Our retrospective, collaborative reflections required rethinking our diverse methodological approaches anew, applying a cross-case comparative perspective (see Table 1). Since all the research was completed, we reconstruct and compare our rationales, methodological choices, and their wider implications. The authors of each case study had to cope with new methodological challenges, without prior knowledge of what does and does not work in large-scale research based on bibliometric and survey data and why. Our selection of data sets and procedures also demanded continuous assessment, as these choices affected results. Ultimately, a new, more comprehensive layer has been added to more prevalent national, organizational, and especially, individual-level studies of the HE research community and its publishing and collaboration patterns, emphasizing complementarities between small-n studies and the large-n research profiled here.

The case studies use global bibliometric and survey data; bibliometric data come from both Clarivate Analytics' Web of Science (CS1) and Elsevier's Scopus (CS2, CS3) data sets. Importantly, samples are global (or networks of cross-national collaborations) rather than national. Observations include several layers: articles, journals, individual collaboration matrices, and individual academics. The limits are set high: more than 400,000 collaboration entries, 21,000 publications, and 27,000 individual scholars in the field. CS1 and CS3 use longitudinal data from about two decades, and CS2 uses longitudinal data analysed cross-sectionally. Beyond various combinations of technical expertise, our general finding is that fundamental data collection, integration, and analysis skills were needed, with some advanced visualization skills. There are limitations related to the one-time data collection approach used, yet no other approach was feasible. Any future (even small) revisions would likely require updates of all data for all years; thus, all needed bibliometric data should be available simultaneously.

Common challenges were data set choice and journal selection in the field: neither Scopus nor WoS uses "higher education" as an independent category in their journal classification, thus requiring a journal sampling strategy and the requisite methodological skills. Each case study implemented its own methodology of journal selection, with implications for results based on divergent journal coverage. Data preprocessing was an issue across cases, and complete access to raw data (or their aggregations in CS3) would be useful. Any changes performed in the databases retrospectively (for all previous years) could affect the results; however, a one-time approach was assessed as the only one feasible.

This manifests the power and implications of such methodological choices. The breadth of coverage of journals, languages, types of publications, and country contributions varied. Crucial steps in CS1 and CS3, relying on different data sets, were structurally similar: selecting the data set, with different limitations, specifically for social sciences; selecting journals in the field of HE research (based on topic identification, keyword analysis, bibliographic coupling, and citation patterns vs. based on journal percentile ranks); selecting publication type; selecting time period for longitudinal analyses; as well as selecting relevant network analysis measures versus defining core and elite journals; and defining full-timers in HE research based on lifetime productivity. We contrast the methodological choices made and some unavoidable limitations, including not covering national journals, not including publications other than articles (e.g., conference proceedings and books), and not providing a longer analysis than two decades. In CS3, including different lists of elite and core journals would manifest different stratification patterns of the HE research community.

These cases analyse an array of dimensions related to articles, journals, and individual scholars. For scholars, both ascribed and acquired attributes were used, originating from bibliometric data sets (CS1, CS3) and from bibliometric and survey data (CS2). Comparing across cases led us to consider dimensions not used in these analyses that, with additional data sets or through different proxies, could be integrated in future studies. Specifically, for CS1 and CS3, untapped potential lies in gender and biological age (available for some countries or available through a proxy of academic age); academic disciplines (available through a study of millions of cited references);

**TABLE 1** Methodological overview of three cases of large-scale bibliometric studies in HE research.

| | Case study 1 (Fu et al., 2022) | Case study 2 (Horta et al., 2022) | Case study 3 (Kwiek, 2021) |
|---|---|---|---|
| Number of observations (n) and their level | *n* = 9037 publications (type: article)<br>*n* = 13 journals | *n* = 913 researchers<br>*n* = collaboration matrix with 416,328 entries (co-authorships or absence of co-authorships) | *n* = 26,888 individual authors<br>*n* = 21,442 publications (type: article)<br>*n* = 41 journals |
| Data type | Bibliometric | Survey, bibliometric | Bibliometric |
| Data type: cross-sectional, longitudinal | Longitudinal (1998–2018); 13 journals in higher education (selected via bibliographic content analysis from 243 "education & educational research" journals) | Longitudinal but analysed in a cross-sectional fashion (pertaining to the publications of higher education researchers throughout their careers) | Longitudinal (1996–2018) and cross-sectional (1996–2018 combined) data on individuals (authors) and their publications in 6 "elite" and 41 "core" journals in higher education |
| Survey data source | – | Survey implemented by two of the authors; data not available for external parties due to ethical research agreements | – |
| Bibliometric data source | Web of Science | Scopus | Scopus |
| Skills required | Data collection, integration, and analysis; Social Network Analysis (gephi); ggplot for visualizations (R package) | Coding (crawler tool; collaboration matrix formation and storage—beyond capability of Excel to deal with its size); Data integration and analysis; Regression analysis in Stata (Firth regression to cope with rare event bias estimation) | Data collection, integration, and analysis; ggplot for visualizations (R package) |
| Challenges: data skills | Coding skill requirements in Python; Data management skills; Quantitative analytical skills; bibliometric data will require updates for future analyses | Coding skill requirements in Python; Data management skills; Quantitative analytical skills; Survey and bibliometric data will require updates for future analyse | Coding skill requirements in Python; Data management skills; Quantitative analytical skills; bibliometric data will require updates for future analyses |

TABLE 1 (Continued)

| | Case study 1 (Fu et al., 2022) | Case study 2 (Horta et al., 2022) | Case study 3 (Kwiek, 2021) |
|---|---|---|---|
| Challenges: data set | Defining "higher education" journals: WoS does not classify journals using a "higher education" category, thus the content analysis toolkit (CATAR) was used to identify journals with higher education content (based on topic identification, keyword analysis, bibliographic coupling, and citation patterns) | Matrix exceeds Excel capabilities. Each time a new indicator from the survey was needed that was not initially used, it demanded three days for a high-performance computer to integrate that indicator into the data set linked to the matrix | Defining "higher education" journals: Scopus ASJC (All Science Journal Classification) categories do not show "higher education" |
| Challenges: data preprocessing | Without complete access to the WoS raw data, all articles from selected HE journals needed to be downloaded (done on 15 May 2019 using a customized crawler program) | Needs complete access to Scopus; Need to develop or use a data crawler tool to obtain the information in a desired analytical format | Any changes in Scopus data for any prior year require recalculating all results |
| Methodological choices | 1. Deciding on global data set: Scopus, Web of Science (WoS), other: WoS selected<br>2. Defining relevant "higher education" journals using agglomerative hierarchical clustering and multidimensional scaling based on bibliographical coupling similarity<br>3. Articles were selected (no other publication formats analysed)<br>4. Defining time period for longitudinal analysis: two decades (1998–2018) because prior to that org. affiliation data had many missing values<br>5. Selection of relevant network analysis measures (further: two-mode network analysis) | Deciding on Scopus as the most useful and reliable source for bibliometric data due to better coverage of higher education journals | 1. Deciding on global data set: Scopus, Web of Science (WoS), other: Scopus selected<br>2. Defining "elite" versus "core" journals: Scopus CiteScore data used<br>3. Defining "full-timers" and "part-timers" in higher education: minimum 5 co-authored articles in 41 core journals for "full-timers", minimum 1 co-authored article for part-timers |

(Continues)

TABLE 1 (Continued)

| | Case study 1 (Fu et al., 2022) | Case study 2 (Horta et al., 2022) | Case study 3 (Kwiek, 2021) |
|---|---|---|---|
| Implications of methodological choices | 1. The data set does not include journals indexed only in Scopus<br>2. Other relevant higher education journals were not analysed<br>3. Other pub. formats (e.g., conference proceedings) were not analysed<br>4. Longitudinal analysis limited to two decades, so early development of field not analysed<br>5. Made relative country contributions to various key topics and levels in HE research visible | 1. Data from the survey essentially representative of active higher education researchers<br>2. Some data from the survey reports the respondent's perspectives and is not impervious to respondent's bias<br>3. Only journals from higher education in Scopus chosen, meaning that publications essentially mean papers, omitting book chapters, books, and other publication forms | 1. The data set does not include journals indexed only in WoS (and not in Scopus)<br>2. the list of 6 "elite journals" could be different (shorter, longer) even when Scopus is used<br>3. Higher requirements (more articles) shorten the list of full-timers, lower requirements (fewer articles) lengthen the list. The list of part-timers is stable |
| Major analytic dimensions | Individual authors; time (1998–2018); country affiliation; journal type (publishing HE research); publications (type: article); topical clusters (topic diversity measured using Shannon diversity index), language of publications | Ascribed characteristics: gender and age; geographical attributes: location of author affiliations at institutional, city and national levels; Career attributes of the authors, including research funding, time allocations, etc (all self-reported), h-index<br>Acquired attributes: Personality and research agenda preferences (self-reported); Co-authorships | Individual authors (Scopus Author ID); time (1996–2018); country affiliation; clusters of country affiliations; journal type (elite vs. core); total individual productivity (full-timers vs. part-timers); publications (type: article) |
| Other dimensions (not examined in the case study) | Gender, biological age, academic discipline, collaboration patterns (across gender, age, disciplines); type of organization (and affiliation), characteristics of HE system | Certain career (educational and career mobility), reputation of the university where one works; Broader region (e.g., Europe); and intensity of collaboration | Gender, biological age, academic discipline, collaboration patterns (across gender, age, disciplines, countries, regions) |
| Methods used | Shannon diversity index | Firth regression | Herfindahl–Hirschman Index (HHI) |

collaboration patterns across gender, age, disciplines, countries, and world regions; and reputation or ranking positions of the employing university, nationally or globally.

Specific methods applied included social network analysis (two-mode), topic identification, bibliographic coupling, and Shannon diversity index in CS1; Firth regression to cope with rare event bias estimation in CS2; and Herfindahl–Hirschman Index (HHI) in CS3. Regression analyses were performed in Stata, high-performance computers were used, and coding skills in Python were needed.

## 2.1 | Case study 1. Examining the evolving international collaboration network of higher education researchers

Here, we examine the contribution of two increasingly popular methods in bibliometrics and scientometrics, namely automated content analysis and topic modelling based on analyses of scientific articles (Kozlowski et al., 2021) and social network analysis applied to the case of HE research (Dusdal et al., 2021). Longitudinally, we investigate the rise of international research collaborations (IRCs) by joining this bibliographic topic modelling and relational analyses of co-authorship patterns based on authors' organizational affiliations. We chart the evolution, across two decades (1998–2018), of the global network of higher education researchers via their international research collaborations (see analyses presented in Fu et al., 2022). The natural and social sciences—and, more gradually, educational research—have become (much) more collaborative (e.g., Aman & Botte, 2017), especially over recent decades. Indeed, contemporary science reflects expanding and diverse forms of collaboration. Yet, IRCs remain challenging to conceptualize and carry out (Dusdal & Powell, 2021), even as this cross-border teamwork is crucial for advances in HE research, in particular for international and comparative work (Kosmützky & Krücken, 2014) and to answer questions of global isomorphism (Zapp et al., 2021). In many fields, scientists generally have abandoned scientific nationalism, but education researchers continue to emphasize the uniqueness of their own national and organizational context(s). Higher education and science advance in an age of "global mega-science," in which transnational, sometimes truly global, research based on international and intercultural teamwork facilitates continued pure exponential growth in scientific knowledge (Baker & Powell, 2024; Powell et al., 2017). Thus, we ask whether and to what extent these general patterns hold for the multidisciplinary field of HE.

Increasingly, HE research has become theoretically and methodologically sophisticated (e.g., Huisman & Tight, 2021) and utilizes large-scale data, including bibliometric data sets, especially Scopus and Web of Science (WoS). Analyses based on large samples of journal articles in the leading, largely Anglophone, journals show HE research becoming much more collaborative in the 21st century, yet with persistent regional and country differences as well as persistent stratification (e.g., Akbaritabar & Barbato, 2021; Avdeev, 2021; Fu et al., 2022; Vlegels & Huisman, 2021).

Pertinent research questions include firstly, how did the network (country) affiliations of international research collaborations (IRCs) in HE research evolve over two decades (1998–2018)? To chart the dynamics of growth and differentiation of levels of analysis of research, we analysed co-authorships in a sample of 9067 articles in thirteen selected HE journals. Secondly, which topical preferences did these international teams develop and at which level(s) of analysis?

Methodologically, automated bibliometric topic identification of articles from 1998 until 2018 enabled us to identify thematic convergence/divergence over time and shifts in level(s) of analysis of the collaborative research conducted. To answer these questions, a two-mode social network analysis was needed (Latapy et al., 2008). We combined article data (theme and level of analysis) and author affiliation data to analyse IRCs (see Newman, 2001). The project database includes 9067 articles from Clarivate Analytics' WoS. Since "higher education" does not have its own category in the WoS classification, we selected 243 journals in "Education & Educational Research" of the 2018 Journal Citation Reports and, using the Content Analysis Toolkit for Academic Research (CATAR),[2]

bibliographical coupling similarity provided clusters of journals by analysing the degree of overlap among references. Thirteen journals were selected, paper clustering proceeded based on bibliographic similarity, and articles with similar citation patterns were categorized into groups, later merged into larger categories to create a topic tree of global HE research.

Text mining techniques to extract key terms from paper title and abstract and to visualize clusters are a methodological innovation to show the development of huge volumes of scholarship across space and time. Two-mode network analysis finds correspondence between two sets of units (articles, authors). Measuring degree centrality can uncover the number of co-authors' international relationships, showing the stratification dynamics of the core-periphery network structure of IRCs in HE research globally.

This century has witnessed pure exponential rises in the production of HE research and in IRCs: the number of contributing countries rose from 36 to 85 countries within just twenty years, reflecting global trends across all disciplines (Baker & Powell, 2024; Fu et al., 2022). This emphasizes more inclusion of diverse perspectives and analysis of more diverse contexts worldwide. The proportion of IRCs of all HE publications in the sample of leading HE journals grew six-fold, from 3% to 18%, over this period. While quantitatively significant, the rise of co-authorship largely strengthened existing collaborations, less than enhancing diversity. We found not much variation in collaboration propensity across the global network, with co-authoring mainly focused on the same countries, esp. the Anglophone countries (various continents) as well as the enhanced centrality of non-Anglophone Europe and China.

In terms of research themes and level(s) of analysis, we found stability in the distribution of levels of analysis, over the past decade, as co-authored papers were mainly focused on the individual level (70%), with organizational (20%), and system level (10%) foci found far less often. In terms of topics, the thematic diversity of co-authored papers increased, especially after 2003.

We now turn to reflect on the importance of these briefly presented methods (and representative findings) and the potential of large-scale data with a focus on IRCs for HE's global research community and on what these methods and findings may contribute. The field's boundaries have extended enormously, becoming more inclusive spatially as many more countries host researchers active in publishing in the leading HE journals—an important condition for responsiveness to current worldwide developments and challenges. However, relatively uniform patterns of collaboration indicate that issues of language diversity (and the costs of Anglophone dominance in leading journals) need to be addressed. These journals are limited in the amount of research they can publish and are not fully utilizing global knowledge stores in all languages. The persistent stratification found in numerous dimensions questions the sustainability of relying heavily on leading journals for global analyses, as these may discount HE issues in non-Anglophone countries and the long-neglected Southern hemisphere.

There are numerous challenges facing globally networked science, including disciplinary and methodological divisions of labour that hinder synthesis and comprehensive understanding. The methods presented to utilize large-scale data sets promise reduced selectivity and counter methodological nationalism, especially when IRCs also support comparative and transnational perspectives. Synthetic overviews of the topical specialization and the dynamics of research production facilitate responses by individual teams to help fill in the considerable gaps in our knowledge base. Longitudinal analyses promise better understanding of factors in developing HE research and its globally networked community; alas, too few studies explicitly track changes over time (but see Fu et al., 2022).

Spatially, multi-level approaches and mixed methods are crucial to extend the possibilities for analysis and explanation of this dynamic global field—and to facilitate the (re)direction of researcher resources and scientific attention to those issues that get drowned out by research agendas popular among wealthy Anglophone countries. For a more responsive body of HE research, we need more emphasis on the meso-level (organizations) and the macro-level of systems, cross-country comparisons, and (trans)regional research, especially as scholars, policymakers, and other stakeholders try to make sense of existing research—and select which avenues of research promise important discoveries.

We next turn to the second case study, on homophily in the field of HE research.

## 2.2 | Case study 2. Homophily in higher education research: A perspective based on co-authorships

For a field to attain legitimacy and recognition, two issues are important. First is the existence of a community that shares interest in a set of phenomena that overlaps; second is sharing an identity. The confluence of interests and identity is likely to lead to the formation of a group of scholars that will produce and share information and knowledge, and eventually collaborate. Research collaborations have been of relevance to sociologists of science since the first studies on how science is made and on how scientists operate. However, most studies that have been done on collaborations have focused on Science, Technology, Engineering, Mathematics and Medicine (STEMM) fields; a scarcity of studies focus on research collaborations in the social sciences and humanities. The reasonings for this vary. STEMM fields tend to be the ones in which research collaborations are most prevalent, a demand that originates in blue sky projects that have a high degree of complexity, and therefore need to combine several fields of expertise and access to instrumentation. Collaboration in these fields also comes naturally from the research environment, which is the laboratory, where research projects have a leader, but tend to be participated in by different elements of the same laboratory, including seasoned researchers, postdocs, and students (thus making the laboratory a natural research-teaching setting; see Clark, 1995). The diverse team members that participate in the project tend to all be listed as co-authors. The grand objectives (i.e., research challenges) that STEMM communities tend to agree upon, such as addressing climate change or colonizing Mars, require a complex array of expertise and specializations to come together and promote multidisciplinary and transdisciplinary approaches—global in terms of scope and effort. All of these aspects promote a sense of community purpose around those challenges to be tackled, thus leading to a convergence of research objectives to be focused on, and inexorably leading to the formation of research collaborations (Robinson, 2019). Since there is an accumulated history of these practices, STEMM scientists are socialized to do such research from the beginning of their research training and to be proficient in the two universal languages of science: English and mathematics. This greatly facilitates their potential for worldwide and cross-disciplinary communication and engagement. It also enables the dissemination of research focused on journals that have an international readership.

The research work in the social sciences and humanities manifests different habitus and research collaborations tend to be sparser because not all the issues under study are of global relevance (Aksnes & Sivertsen, 2023). Most social scientists focus on national and local phenomena, that may or not be of broader interest, and therefore be presented much more often in national languages, which are closely tied to national culture and society. Their research is also largely bound by localized or nationalized cultural, political, economic, and social dynamics that mostly reduce the potential group of collaborators to fewer scientists in their own country, region or city. Social scientists and those from the humanities do not often work in laboratories, but rather by themselves at their desks, and their research interests are more defined individually than collectively. Only relatively recently have social scientists started to publish more in collaboration with others and in journals, while researchers from the humanities still rely heavily on books as their main means of disseminating research findings; they continue to prefer single authorships (Kwiek, 2020).

Studies on research collaborations usually take two perspectives: studies of research collaboration determinants and of geographies of collaboration. The first type of studies tries to understand who publishes the most and what conditions lead to greater frequency of collaborations. These factors can be related to positioning in the field, organization, or educational and professional pathways, or to institutional and organizational work conditions that affect time allocation, but also reflect career incentives and other factors. Regressions tend to be the analytical method. The second type of studies tends to assume a more visual and descriptive perspective, and often combines bibliometric and social network analysis to discern the centrality of research collaborations: which countries, cities, and organizations collaborate the most, and who collaborates with whom, usually at the macro (country) and meso (organization) levels. However, what has been sparsely studied so far is what drives people to choose with whom to collaborate in fields of the social sciences, especially HE studies.

This is an important phenomenon, because it goes beyond understanding what drives one to collaborate (that we know relates to knowledge demands and extrinsic motivations, related to career progression, for example), and may encompass a set of variables that relate more to research preferences and cognition (that is mostly understudied when it comes to collaborations). This is particularly important in the field of HE, where research collaborations are rising but still limited, particularly when it comes to collaboration between countries, and where the core of researchers in the field is still somewhat limited (Santos & Horta, 2018). It is also important to move towards more sophisticated ways of using data and conducting analyses so that new insights can be drawn from larger data sets that enable combinations of information that may be distributed among two or more sources. A recent study on preferences of collaborators in the field of higher education was conducted by Horta et al. (2022) brought novel insights concerning collaborative patterns of HE researchers, and since it combined information of two data sources, it became a prime case to analyse in the context of this paper. Next, we describe the procedures to develop such a study and report its findings.

One way to identify the research collaborations that an academic (or organization) is involved in is by sending a questionnaire and asking respondents to identify the research collaborations in which they have been involved. Yet this straightforward asking comes with many problems: (1) Respondents will want to know which collaborations the leaders of the study are interested in, and for simplicity's sake, the requirement will likely rest in co-authorships. Successful research collaborations introduce selection bias; (2) If respondents have numerous (co-authored) publications, writing them all down will take time, and it is unlikely that many academics would have sufficient time to fill in all this information and related items, resulting in low response rates; (3) Information ordering and quality is a further challenge, as some respondents start with the latest publications, while others would start with the oldest. Furthermore, respondents may not include all types of publications, limiting information to articles only, and constraining information quantity and quality; and (4) Self-reports, especially those requiring a substantial amount of effort may inadvertently provide erroneous information, thus increasing errors in subsequent analyses. Checking the information provided with other sources may be difficult for the surveyor because of inordinate efforts and, in some cases, the information may not be available elsewhere or only exist until a past temporal period that may be outdated or erroneous (many academics' webpages have outdated publication information).

A way to tackle this is to rely on relatively high-quality bibliometric information that is provided by publication indexing services, such as Scopus (there are issues with it as well; mentioned later). Scopus provides an author ID for each author that has publications in indexed journals (with a good coverage across scientific fields), some conference proceedings, and books. Pretty much like the case of using surveys mentioned above, Scopus publications only refer to successful research endeavours, and this is certainly a limitation when considering dimensions of research collaborations because these are only the most visible result based on formal co-authorships of successful research projects (Laudel, 2002). However, this information is retrieved from the publication sources and their reliability is much higher than self-reported data. Furthermore, the data that these indexing services provide encompass a range of publication information that academics do not have easily at hand: impact factors, h-indexes, citations, self-citations, (co-)author's affiliations, immediacy indexes, just to mention a few. In short, the publication data for each author is already compiled and much easier to work with, without needing to ask respondents and reducing their work compiling data, probably erroneously. Then, questionnaires can focus on asking other questions that only the respondent knows and cannot be found anywhere else, such as opinions about particular research practices or preferences. Such questionnaires deliver data on personality and personal cognitive, work, and relational preferences and orientations.

By using the Scopus author ID, one can first collect publication data for all the potential respondents, their affiliation, collaboration networks, and other information as well as identify the respondent e-mail address. This information can be compiled through scrapping tools that authors can learn to use online or can be compiled by colleagues in computer science. A questionnaire then will need to be built around a particular set of previously established research objectives and questions. In the case of the homophily study, the research question was to

understand what determines homophily in co-authorship in HE research. Therefore, the research question was mostly oriented towards knowing more about academics' research agendas (Horta & Santos, 2016), their educational and professional backgrounds, personality, attitudes, and preferences concerning research practices, and socio-demographic information. This information was needed to better understand co-authorship preferences because the analytical plan was to understand what linked authorship to specific characteristics. The literature on homophily highlights two main type of traits: ascribed traits, which are the characteristics that individuals possess, such as age and gender (this indicator has been recently disputed as ascribed, but the homophily literature still considers it to be so), and acquired traits, which result from educational and professional pathways, and life experience, plus concrete achievements, such as published publications (Santos et al., 2024). Among ascribed traits, the academic's personality and research agendas were included because they are constructed; the result of life paths (Santos et al., 2024). From the regional studies and scientometrics literature, geographical proximity indicators were added to the analysis: same organization, city, and country (e.g., Ponds et al., 2007).

Even without asking for collaboration and publication information, the questionnaire still required about 40 minutes to be fully completed. An online survey was conducted from May to November 2015, involving several waves obtaining a response rate of 24%, which is good considering the lower response rates to surveys in general, and online questionnaires in particular. The responses provided in the survey were then linked to their corresponding Scopus ID, therefore joining the two data sources (note that in the informed consent form, this was stated so that the respondents were aware that they were not providing fully anonymized responses). After survey data were collected, organized, and cleaned, bibliometric data was collected for the working sample of survey respondents: co-authorships were identified, through the creation of a matrix of co-authorships between the respondents (whereas a co-authorship corresponds to 1, and a lack of co-authorship to 0). For the 913 higher education researchers in the sample, a set of more than 416,000 entries was entered, leading to a fairly large co-authorship matrix. The compilation of this matrix again requires some technical expertise, and a computer science collaborator may provide essential help. Servers with high computational data capacity also helped to compile the matrix (this process took around a week).

The pairs of authors formed the unit of analysis. For each pair (co-authorship or not), ascribed, acquired, and geographical characteristics were computer-generated. For binary variables, such as gender by ones (in case of match) or zeros, and for continuous variables in terms of deltas, that is, differences between values. For example, if the age of one academic was fifty and the other academic was forty, the value to be inserted would be 10 (i.e., the difference). The data had a clear dependent variable, which was binary. One for co-authorship and zero for non-co-authorship. Therefore, a logistic regression was used for the analysis. As expected for a social science field, the co-authorship density of the matrix was low, since HE researchers often do not collaborate, and when they collaborate, usually only with one or two co-authors. This led to use of a specific regression treatment (Firth regression) to account for rate event bias. The findings (see Horta et al., 2022) bring novel contributions to the field, namely the importance of geographical proximity above anything else for co-authorship of HE researchers, and also the prominent explanatory power of acquired attributes over ascribed attributes that basically suggest that similar ways of working, ambition, and other traits explain co-authorships rather than age or gender. We next turn to the third case study that examines persistent stratification of the field.

## 2.3 | Case study 3. The stratification of the higher education research field

The HE research field has its own journals and (inter)national research communities. While the lists of international journals in the field have been compiled and discussed for decades (Bray & Major, 2011; Hutchinson & Lovell, 2004; Silverman, 1987), access to the data allowing to examine its research (or publishing) communities in detail has been limited. The focus of research on HE research has traditionally been on journals and publications rather than researchers. The global community of HE researchers, however, consists of individual academics

scattered across the world, from its core contributing countries to its distant peripheries, only recently entering the global scholarly conversation continuing since the 1970s. We had expected that globally, as there were dozens of thousands of publications in the field over the past few decades, but examining the HE research community worldwide was long beyond the scope of technical possibilities. While survey-based international comparative academic profession studies have been widespread, with at least 500 papers published stemming from the CAP and EUROAC international comparative research projects (Carvalho, 2017; Fumasoli et al., 2015), research on HE researchers clearly lagged behind (Kwiek, 2019).

New opportunities to examine the field have emerged over the past few years as global databases indexing publications and citations (e.g., Scopus; WoS) have become widely available for research purposes. Finally, it is possible to take steps forward from studying publications to studying their authors and disciplinary-based communities (Kwiek, 2021). We expected a global core of scholars in the field; perhaps several thousands of them publishing in the past few decades; and also large numbers of peripheral scholars in the field: those who appeared once or twice in international journals, never coming back to the global circulation of ideas in HE research. We had had only guesses about the proportions of these core and episodic scholars by country and world region, and about their dynamics. The globalization of science has been in full swing and the processes involving social sciences in general involved also HE research (e.g., internationalization of research, new publishing patterns with more emphasis on journals and English, team work instead of solo work, global focus as complementary to national focus). To the many dimensions of stratification of the field, one more could be added at individual and country levels: the intensity of participation in the global scholarly conversation as measured by globally indexed publications. We speculated that the Anglo-Saxon countries, and especially the USA, dominate the field, but what about contributors from other world regions viewed from a longitudinal perspective, and through the lenses of both selected prestigious journals and all globally indexed journals in the field? These questions could not be probed before structured large-scale data sets were made available to students of the academic profession.

Consequently, our focus was on stratification in the global HE research community and the changing geography of country affiliations in two types of journal: elite and core, with different levels of selectivity and acceptance rates, different contributors and readership, as well as diverse thematic coverage. Our goal was to map the HE research community globally, both in terms of country affiliations and the intensity of involvement in the HE research enterprise. Different approaches could lead to different results, but the key insight was that the community could be examined comprehensively: globally, large-scale, by country, by journal prestige, and their temporal dynamics. To the best of our knowledge, no research previously examined authorship affiliation patterns in elite HE journals in detail over the period of two decades; and no research previously showed the powerful stratification of the global HE research community, divided between scholars heavily involved in research and publishing and scholars involved in the field just once: between global "full-timers" and global "one-timers".

Numerous methodological choices have to be made in mapping the global HE research community: the global database needs to be selected (here: Scopus); the list of core journals (including elite journals) needs to be compiled, with input from either peers via surveys, previous analyses of journals in the field, or from more objective and comparable global citation-based indicators (here: Scopus CiteScore journal percentile ranks); the timeframe needs to be selected, with careful consideration of data availability (here: 1996–2018); publication types need to selected (here: research article only); the classes of HE researchers need to be defined, with different productivity thresholds (here: full-timers defined as the authors of at least 5 articles—lower thresholds would increase the numbers of full-timers, higher thresholds would decrease their numbers, with implications for their geographical distribution); academics in the final sample need to be unambiguously ascribed a single affiliation country. Finally, an assumption needs to be made that the difference between examining "Scopus Author IDs" and examining "real individuals" representing the HE research community is in practice negligible, although theoretically important, following current data on average precision (i.e., the ratio of publications correctly assigned to the author) of 98.1% and an average recall (i.e., the ratio of publications captured) of 94.4% in Scopus (Baas et al., 2020, p. 379). Interestingly, the list of six elite journals, based on sophisticated bibliometric measures of citation numbers and

citation-driven prestige in our global data set, was often identical to those used in previous studies based on peer review and general insights (e.g., Tight, 2014).

Our focus was thus on the authors and authorship patterns in global HE journals. To trace the patterns, 6334 articles published in six top journals in 1996–2018 were studied in the wider context of 21,442 articles from 41 core journals in the field (Kwiek, 2021). The six "generic" rather than "topic-specific" elite journals were selected for in-depth analysis: *Higher Education*, *Studies in Higher Education*, *Higher Education Research and Development*, the *Journal of Higher Education*, *Research in Higher Education*, and the *Review of Higher Education*. The whole counting (rather than fractional counting) method was used: each author of a multi-authored article was given one credit, and all country affiliations were studied. There were 11,688 country affiliations in our data set and the metadata for each article included the Scopus Author ID (unique identification number), the Scopus document ID, organizational and country affiliations, and all cited references. The final list included 26,888 unique authors treated as the global (Scopus-indexed) HE research community. Author profiles are generated using a combination of algorithms and manual curation and authorship assessment is sophisticated: "the end-to-end accuracy is measured continuously by several metrics. Moreover, regular spot checks are run on aspects of author profiles, such as canonical names or affiliations" (Baas et al., 2020, p. 379), making Scopus a trustworthy data source for research on academics and academic communities.

Viewed through a proxy of publications in 41 core journals over the past two decades, the global HE research community comprises no more than 27,000 individual academics. However, the scale of their participation in the field remains highly skewed. We studied all articles by 8226 academics published in the six elite journals and found a stunning stratification pattern: full-timers at one end and one-timers at the other end (on stratification in academic careers, see Kwiek, 2019). The number of full-timers was 274 (or 3.33% of all authors). These constitute the publishing core of the global HE research community. However, most of those who contributed to the six elite journals (6485 or 78.81%) published just one article in the twenty years studied: we termed them one-timers. About a thousand (997 or 12.2%) scholars authored or co-authored 2 articles, 305 (or 3.7%) authored or co-authored 3 articles and 165 (or 2.0%)—4 articles. The total number of academics associated with the 21,442 articles published in the 41 core journals is 26,888, of whom 878 (3.27%) were full-timers and 21,389 (79.55%) were one-timers. In other words, eight in ten academics in the HE field remain on the publishing periphery, having (co-)authored just one single article in elite or core journals. Somehow surprisingly, for the vast majority of researchers active in the field, HE is not their prime research interest and globally visible HE journals are not their prime publishing locus.

In view of the recent global expansion of HE research, we also examined the changing role of major Continental European, East Asian, and 66 "other" countries combined. We studied whether the increase in article numbers in the six elite journals is driven by newcomers to the field—or by the traditionally dominant US and other Anglo-Saxon countries. Changing authorship patterns were analysed in terms of changing percentages and numbers of author affiliations. In the new geography of elite higher education, relative newcomers are gaining at the expense of the traditionally dominant US. Indeed, the single biggest affiliation loser in terms of shares is the US, and the biggest affiliation winner is Continental Europe, where affiliations almost doubled. In terms of changing numbers of author affiliations over time, the data on elite journals reflect both changing national engagement in global HE research and changing international collaboration patterns in the field. Between 1996 and 2018, the bulk of global HE research published in elite journals was produced in Anglo-Saxon countries, including the USA (70.0%), Continental Europe (16.7%), and East Asia (5.1%). These are the major participants in the global research conversation, with gradually increasing participation from other world regions (8.2%, from 66 countries). The changing distribution of country affiliations shows the relative weakening of the field in the US and its relative strengthening in Continental Europe, East Asia, and elsewhere (see also Horta & Jung, 2014).

Our ongoing research on the global HE field changing over time has also been guided by other themes belonging to a wider research agenda: (1) the changing global gender distribution in the field (using gender-defining procedures for each individual); (2) the changing global age distribution in the field (using academic age, highly

correlated with biological age, see Kwiek & Roszka, 2022); (3) the changing global disciplinary distribution, inquiring whether scholars involved in publishing in the HE field come from education, sociology, political science, economics or other fields (using lists of lifetime publications and all cited references to define their dominating disciplines); (4) the changing global gender homophily patterns in collaboration in the field (all-male, all-female, and mixed-sex collaborations); and (5), the changing global seniority-related homophily patterns in collaboration in the field (all-older, all-younger, and mixed-age collaborations).

Detailed analyses of the global HE research community along the above lines of inquiry and beyond are not feasible without longitudinal, large-scale publication and citation data sets of the Scopus type (Kwiek & Roszka, 2024). Such comprehensive analyses are also fully replicable for other timeframes, journal collections, and academic fields. The major point is our change in focus: from publications (and their metadata) to individuals (and their characteristics) as units of analysis.

The publishing patterns briefly reported here show that global HE research community is highly stratified; few scholars publish intensively, and many publish just once, coming from other fields and other journals and then just disappearing again. The community is stratified between about 3% of heavy publishers and about 80% of incidental one-time authors, with the remaining 17% comprising part-timers with 2–4 articles in the HE journals in their output. Most authors in the field can thus be policy-oriented practitioners, administrators or academics focused on teaching rather than research. Consequently, sustained scholarly conversation in the field may be hindered by the omnipresence of these infrequent contributors.

## 3 | CONCLUSIONS

Key in the studies that we authored and conjointly reanalysed here is our global analytical approach. We applied sophisticated methodologies to study worldwide transformations of systems and scholars involved in the field of HE research in their entirety. We believe that the relatively small field of HE research deserves expansive and detailed global accounts made possible by using large-scale data. This enables us to develop the knowledge base and drive in-depth understanding of the field but also contribute to the analytical and methodological sustainability of the field, by keeping it *au pair* with analytical and methodological advances in other social science fields and beyond.[3] In our view, much of the HE research community may simply be unaware of the availabilities and potentials that new large-scale data (both commercial and open access) and new methodologies provide. The general perception in this diverse global community may be that immense resources and an extensive learning curve are needed to successfully utilize these tools, yet this does not seem to be true. This might be the case were HE research mostly an individual endeavour, but the field is becoming ever more collaborative, cross-disciplinary in methodological approaches and with increasing teamwork and the resulting co-authored publications. Indeed, there seem to be plentiful opportunities in HE research for statisticians and data scientists to collaborate with community members. The answer to the biggest hurdle in the adoption of the new data sets and methods reported in this paper is collaboration; becoming cross-disciplinary and multi-method experts ourselves is possible, but an arduous task. Clearly, the most efficient way to overcome new challenges in leveraging these large data sets is to collaborate with colleagues who have such skills.

Across three case studies, we showed various modalities of working with large-scale bibliometric data in the HE field, allowing also for global replications for other fields, other disciplinary communities, and other timeframes. We provided insights into what is currently feasible and post factum reflections to compare three contemporary approaches. We provided a general technical and methodological guide to research the HE field from a global perspective. Others can further refine these methods, apply them in different contexts, and for different purposes. We describe new methods and new tools enabling expanded views of the HE research field and to analyse its members and communities globally: how extensive they are, how they collaborate within networks, what topics they choose, and how they change over time, among other important questions. The core of the field

is very small, while the periphery, with marginal publishing involvement (at least via internationally visible articles), is large. Maintained disparities in country and world region contributions question the short-term responsiveness and long-term sustainability of contemporary HE research. These findings emphasize persistent stratification along numerous dimensions, not only who collaborates with whom. This may be viewed as problematic if we believe HE research is responsible for ensuring effective knowledge transfer and participation of researchers from everywhere and with different skill sets.

Our research shows that encompassing discussions are needed to advance the scholarly conversation about the field and its communities. Trade-offs between what is ideal and what is manageable remain necessary. For example, while such data are gradually becoming more available, analysts need to be careful as new error types emerge, possibly leading to faulty results (consistent with the principle of "garbage in – garbage out" in data science). We need to accept some level of uncertainty just to move forward to construct global overviews of the field, with its collaborations, homophily, and stratifications, assuming that different types of analysis imply different levels of precision and illuminated dimensions. Simultaneously, there are risks in using new data sets and tools, with their own inherent limitations. Academic journal reviewers may be more likely to reject submissions applying new methods to new data. Reviewers may not fully understand the added-value that these new methods and data bring compared to existing studies or they may not have the expertise to judge their correct application (thus rejecting the paper, erring on the side of caution). However, even in conservative publishing, reward, and grant systems, there should always be sufficient space for the coexistence of established and newer methodological approaches, with the boundaries of acceptability changing with innovative methods and new data sources. We believe that the global HE research community, as it adapts to new and ongoing challenges, will be skilfully avoiding the field's stagnation as it complements established methods and traditional data with the new.

Finally, we emphasize that in this type of research, multidisciplinary and multi-method collaborations are often necessary. Complementary skills of statisticians and computer scientists and traditional HE researchers will be required. Studies about the HE research field have begun to discover the power of large-scale data. We believe new tools and methods can enlarge the field's horizons, useful in advancing the scholarly conversation on academia and academics within organizational, national, and global contexts. Therefore, we compared three case studies conducted with novel methodological approaches to leverage the largest bibliometric data sets, hitherto largely unexplored in studies on the global higher education research community.

## AUTHOR CONTRIBUTIONS

**Marek Kwiek:** Conceptualization; investigation; writing – original draft; methodology; writing – review and editing. **Hugo Horta:** Conceptualization; investigation; writing – original draft; methodology; writing – review and editing. **Justin Powell:** Conceptualization; investigation; writing – original draft; methodology; writing – review and editing.

## ACKNOWLEDGEMENTS

Marek Kwiek gratefully acknowledges the support provided by the NDS grant no. NdS/529032/2021/2021.

## CONFLICT OF INTEREST STATEMENT

No potential conflict of interest was reported by the authors.

## DATA AVAILABILITY STATEMENT

We used data from Scopus (owned by Elsevier) and Web of Science (owned by Clarivate), proprietary scientometric databases. For legal reasons, data from Scopus received through collaboration with the ICSR Lab and data from Web of Science cannot be made openly available.

## ORCID

*Marek Kwiek* https://orcid.org/0000-0001-7953-1063

*Hugo Horta* https://orcid.org/0000-0001-6814-1393

*Justin J. W. Powell* https://orcid.org/0000-0002-6567-6189


## ENDNOTES

[1] Although in this paper we focus on HE researchers and academics mostly, the potential of these new data and methodologies applies equally to the study of other HE stakeholders, such as students.

[2] To use the Content Analysis Toolkit for Academic Research (CATAR), see https://github.com/SamTseng/CATAR.

[3] This idea is also aligned with the need for those researching higher education to engage with more sophisticated methodologies to drive the field forward (as argued recently by Huisman, 2024).

## REFERENCES


Akbaritabar, A., & Barbato, G. (2021). An internationalised Europe and regionally focused Americas: A network analysis of higher education studies. *European Journal of Education*, *56*(2), 219–234. https://doi.org/10.1111/ejed.12446

Aksnes, D. W., & Sivertsen, G. (2023). Global trends in international research collaboration, 1980–2021. *Journal of Data and Information Science*, *8*(2), 26–42.

Aman, V., & Botte, A. (2017). A bibliometric view on the internationalization of European educational research. *European Educational Research Journal*, *16*(6), 843–868. https://doi.org/10.1177/1474904117729903

Avdeev, S. (2021). International collaboration in higher education research. *Scientometrics*, *126*, 5569–5588. https://doi.org/10.1007/s11192-021-04008-8

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, *1*(1), 377–386. https://doi.org/10.1162/qss_a_00019

Baker, D. P., & Powell, J. J. W. (2024). *Global Mega-Science: Universities, Research Collaborations, and Knowledge Production*. Stanford University Press.

Bray, N. J., & Major, C. H. (2011). Status of journals in the field of higher education. *Journal of Higher Education*, *82*(4), 479–503.

Carvalho, T. (2017). The study of the academic profession – contributions from and to the sociology of professions. In J. Huisman & M. Tight (Eds.), *Theory and Method in Higher Education Research* (pp. 59–76). Bingley.

Clark, B. R. (1995). *Places of inquiry: Research and advanced education in modern universities*. University of California Press.

Clauset, A., Larremore, D. B., & Sinatra, R. (2017). Data-driven predictions in the science of science. *Science*, *no. 355(6324)*, 477–480. https://doi.org/10.1126/science.aal4217

Dusdal, J., & Powell, J. J. W. (2021). Benefits, motivations, and challenges of international collaborative research: A sociology of science case study. *Science and Public Policy*, *48*(2), 235–245. https://doi.org/10.1093/scipol/scab010

Dusdal, J., Zapp, M., Marques, M., & Powell, J. J. W. (2021). Higher education organizations as strategic actors in networks: Institutional and relational perspectives meet Social Network Analysis. *Theory and Method in Higher Education Research*, *7*, 55–73. https://doi.org/10.1108/S2056-375220210000007004

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., & Vespignani, A. (2018). Science of science. *Science*, *359*(6379), eaao0185. https://doi.org/10.1126/science.aao0

Fu, Y. C., Marques, M., Tseng, Y.-H., Powell, J. J. W., & Baker, D. P. (2022). An evolving international research collaboration network: Spatial and thematic developments in co-authored higher education research, 1998–2018. *Scientometrics*, *127*, 1403–1429. https://doi.org/10.1007/s11192-021-04200-w

Fumasoli, T., Goastellec, G., & Kehm, B. M. (Eds.). (2015). *Academic Work and Careers in Europe*. Springer.

Gui, Q., Liu, C., & Du, D. (2019). Globalization of science and international scientific collaboration: A network perspective. *Geoforum*, *105*, 1–12. https://doi.org/10.1016/j.geoforum.2019.06.017

Hesse, B. W., Moser, R. P., & Riley, W. T. (2015). From big data to knowledge in the social sciences. *The Annals of the American Academy of Political and Social Science*, *659*(1), 16–32.

Horta, H. (2018). Higher-education researchers in Asia: The risks of insufficient contribution to international higher-education research. In J. Jung, H. Horta, & A. Yonezawa (Eds.), *Researching higher education in Asia. History development and future* (pp. 15–36). Springer. https://doi.org/10.1007/978-981-10-4989-7_2

Horta, H., Feng, S., & Santos, J. M. (2022). Homophily in higher education research: A perspective based on co-authorships. *Scientometrics*, *127*(1), 523–543. https://doi.org/10.1007/s11192-021-04227-z

Horta, H., & Jung, J. (2014). Higher education research in Asia: An archipelago, two continents or merely atomization? *Higher Education*, *68*(1), 117–134. https://doi.org/10.1007/s10734-013-9695-8

Horta, H., & Santos, J. M. (2016). An instrument to measure individuals' research agenda setting: The multi-dimensional research agendas inventory. *Scientometrics*, *108*(3), 1243–1265. https://doi.org/10.1007/s11192-016-2012-4

Huisman, J. (2024). The use of methods: Are higher education scholars lazy or insufficiently skilled? *Higher Education Research and Development*, *43*(1), 260–266.

Huisman, J., & Tight, M. (Eds.). (2021). *Theory and Method in Higher Education Research*, *7*. Bingley. https://doi.org/10.1108/S2056-375220210000007012

Hutchinson, S. R., & Lovell, C. R. (2004). A review of methodological characteristics of research published in key journals in higher education. *Research in Higher Education*, *45*(4), 383–403.

Ioannidis, J. P. A. (2018, March 13). Meta-research: Why research on research matters. *PLoS Biology*, *16*(3), 1–6. https://doi.org/10.1371/journal.pbio.2005468

Kosmützky, A., & Krücken, G. (2014). Growth or steady state? A bibliometric focus on international comparative higher education research. *Higher Education*, *67*(4), 457–472. https://doi.org/10.1007/s10734-013-9694-9

Kozlowski, D., Dusdal, J., Pang, J., & Zillian, A. (2021). Semantic and relational spaces in science of science. *Scientometrics*, *126*, 5881–5910. https://doi.org/10.1007/s11192-021-03984-1

Kwiek, M. (2016). The European research elite: A cross-national study of highly productive academics in 11 countries. *Higher Education*, *71*(3), 379–397. https://doi.org/10.1007/s10734-015-9910-x

Kwiek, M. (2019). *Changing European Academics. A Comparative Study of Social Stratification, Work Patterns and Research Productivity*. Routledge. https://doi.org/10.4324/9781351182041

Kwiek, M. (2020). What large-scale publication and citation data tell us about international research collaboration in Europe: Changing national patterns in global contexts. *Studies in Higher Education*, *45*, 1–21. https://doi.org/10.1080/03075079.2020.1749254

Kwiek, M. (2021). The prestige economy of higher education journals: A quantitative approach. *Higher Education*, *81*, 493–519. https://doi.org/10.1007/s10734-020-00553-y

Kwiek, M., & Roszka, W. (2022). Academic vs. biological age in research on academic careers: A large-scale study with implications for scientifically developing systems. *Scientometrics*, *127*, 3543–3575. https://doi.org/10.1007/s11192-022-04363-0

Kwiek, M., & Roszka, W. (2024). Once highly productive, forever highly productive? Full professors' research productivity from a longitudinal perspective. *Higher Education*, *87*, 519–549. https://doi.org/10.1007/s10734-023-01022-y

Latapy, M., Magnien, C., & Del Vecchio, N. (2008). Basic notions for the analysis of large two-mode networks. *Social Networks*, *30*(1), 31–48.

Laudel, G. (2002). What do we measure by co-authorships? *Research Evaluation*, *11*(1), 3–15. https://doi.org/10.3152/147154402781776961

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *PNAS*, *98*(2), 404–409.

Ponds, R., van Oort, F., & Frenken, K. (2007). The geographical and institutional proximity of research collaboration. *Papers in Regional Science*, *86*(3), 423–443.

Powell, J. J. W., Fernandez, F., Crist, J. T., Dusdal, J., Zhang, L., & Baker, D. P. (2017). Introduction: The worldwide triumph of the research university and globalizing science. In J. J. W. Powell, D. P. Baker, & F. Fernandez (Eds.), *The century of science: The global triumph of the research university* (pp. 1–36). Bingley. https://doi.org/10.1108/S1479-367920170000033003

Robinson, M. (2019). The CERN community; a mechanism for effective global collaboration? *Global Policy*, *10*(1), 41–51.

Santos, J. M., & Horta, H. (2018). The research agenda setting of higher education researchers. *Higher Education*, *76*(4), 649–668. https://doi.org/10.1007/s10734-018-0230-9

Santos, J. M., Horta, H., & Feng, S. (2024). Homophily and its effects on collaborations and repeated collaborations: A study across scientific fields. *Scientometrics*. https://doi.org/10.1007/s11192-024-04950-3

Silverman, R. J. (1987). How we know what we know: A study of higher education journals. *The Review of Higher Education*, *11*(1), 39–59.

Sugimoto, C. R., & Larivière, V. (2018). *Measuring research*. Oxford University Press.

Tight, M. (2014). Working in separate silos? What citation patterns reveal about higher education research internationally? *Higher Education*, *68*(3), 379–395.

Vlegels, J., & Huisman, J. (2021). The emergence of the higher education research field (1976–2018). *Higher Education*, *81*, 1079–1095. https://doi.org/10.1007/s10734-020-00600-8

Wang, D., & Barabási, A. (2021). *The science of science*. Cambridge University Press. https://doi.org/10.1017/9781108610834

Zapp, M., Marques, M., & Powell, J. J. W. (2021). Blurring the boundaries: Organizational actorhood and isomorphic change in global higher education. *Comparative Education*, *57*(4), 538–559. https://doi.org/10.1080/03050068.2021.1967591

**How to cite this article:** Kwiek, M., Horta, H., & Powell, J. J. W. (2024). Using large-scale bibliometric data in higher education research: Methodological implications from three studies. *Higher Education Quarterly*, *00*, e12512. https://doi.org/10.1111/hequ.12512