MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH

Konrad-Zuse-Strasse 1 · D-18057 Rostock · Germany · Tel +49 (0) 3 81 20 81 - 0 · Fax +49 (0) 3 81 20 81 - 202 · www.demogr.mpg.de

MPIDR Working Paper WP 2023-029 I May 2023
https://doi.org/10.4054/MPIDR-WP-2023-029

A global perspective on the social structure of science

Aliakbar Akbaritabar I akbaritabar@demogr.mpg.de
Andres F. Castro Torres I castro@demogr.mpg.de
Vincent Larivière

This working paper has been approved for release by: Emilio Zagheni (sekzagheni@demogr.mpg.de),
Head of the Laboratories of Migration and Mobility and Population Dynamics and Sustainable Well-Being.

A global perspective on the social structure of science

Aliakbar Akbaritabar[a1], Andres F. Castro Torres[a,b], Vincent Larivière[c]

[a]Max Planck Institute for Demographic Research, Rostock, Germany, [b]Center for Demographic Studies, Autonomous University of Barcelona, Barcelona, Spain, [c]Université de Montréal, Montréal, Canada

[1]Corresponding author, Email: akbaritabar@demogr.mpg.de

Abstract

We reconstruct the career-long productivity, impact, (inter)national collaboration, and (inter)national mobility trajectory of 8.2 million scientists worldwide. We study the interrelationships among four well-established bibliometric claims about academics' productivity, collaboration, mobility, and visibility. Scrutinizing these claims is only possible with a global perspective simultaneously considering influential bibliometric variables alongside collaboration among scientists. We use Multiple Correspondence Analysis with a combination of 12 widely-used bibliometric variables. We further analyze the networks of collaboration among these authors in the form of a bipartite co-authorship network and detect densely collaborating communities using Constant Potts Model. We found that the claims of literature on increased productivity, collaboration, and mobility are principally driven by a small fraction of influential scientists (top 10%). We find a hierarchically clustered structure with a small top class, and large middle and bottom classes. Investigating the composition of communities of collaboration networks in terms of these top-to-bottom classes and the academic age distribution shows that those at the top succeed by collaborating with a varying group of authors from other classes and age groups. Nevertheless, they are benefiting disproportionately to a much higher degree from this collaboration and its outcome in form of impact and citations.

Keywords: bibliometric data; research productivity; scientific collaboration; scientific mobility; scholarly impact and citations

Introduction

Science is a social enterprise with hierarchy systems among its agents (*1–4*). Factors underpinning social hierarchies in science include differences within and between countries in institutional capacities and resources available for research (*5*) and socioeconomic inequalities among scholars such as gender (*6*, *7*), racial/ethnic (*2*), migration status (*8*, *9*), and social class differences in opportunities to access higher education and do research (*10–12*) —the overrepresentation of specific demographics in privileged positions within scientific systems are indicators of hierarchies (*13–15*). Differences in scholars' strategies in the search for prestige may also play a role in hierarchies in science (*16*). Because hierarchies can be seen as unwarranted and produce injustices, their durability depends, among other things, on taken-for-granted ideas about the necessity and benefits of hierarchical order. In the broader sphere of social and economic affairs, the belief that a market-oriented organization of the economy without state intervention is optimal legitimizes the existence of socioeconomic inequalities within and between societies (*17*, *18*), which in turn contributes to sustaining social hierarchies among nations and individuals (*19*). In all likelihood, as a sub-sphere of social relations, science works analogously. One complicating factor is that scientific research also is an inherently competitive endeavor, in which individual-based reputational incentives can undermine the motivation to collaborate (*20–22*).

Hierarchies in science rely on beliefs regarding the relevance of meritocracy for academic success and the inherent value of truth for science. Several metrics, such as the number of publications and citations from mainstream bibliometric databases, help fuel these beliefs. While these ideas and metrics are increasingly challenged by scholars from different perspectives (*23*, *24*), we need a global assessment of hierarchies in science and their strength and embeddedness in networks of scientific collaboration. This work contributes a quantitative and global assessment of hierarchies across fields of science based on a multivariate analysis of large-scale bibliometric information from 1996 to 2021. Because measuring hierarchies is only a first step in understanding their functioning, we make publicly available a dataset with country-level measures of scientific hierarchies for future research on the causes and consequences of the bibliometric stratification of scientific systems.

Existing inequalities in science and science hierarchies

Aggregated trends in scholars' collaboration, geographical mobility, productivity, and citations suggest that academia is growing in absolute numbers and expanding geographically. There are more coauthored papers in recent years compared to earlier decades (*25–27*), and more scholars experienced geographical mobility today than in the past (*8*, *9*, *28*). Likewise, studies have

shown that scholarly publications have increased and that digitization has made searching and citing easier. Greater productivity and increased citation capacities enhanced academic works' visibility and potential impact (*29*, *30*). Some of these analyses have pointed out that these rising trends are accompanied by an increased concentration of academic-success indicators in relatively few scholars or increased collaboration and rate of productivity per individual has not increased (*31*). From this perspective, the growth of academia and its geographical expansion require a critical examination of their consequences for inequality and the potential emergence of global hierarchies.

According to Scopus data, 33% of scholars have contributed to only one research paper throughout their careers and median number of authors in 28+ million publications in Scopus is 2 and 75% quintile is 4 authors, suggesting that a few highly productive researchers may drive rising trends in scholars' productivity (*32*, *33*). Likewise, according to Scopus data, approximately 27.2% of the publications have only one author, and more than 75% are authored by scholars from a single country. Likewise, most authors have been affiliated with a single country throughout their careers (87.5% or a single sub-national region, 73.5%) and have not experienced geographical mobility (*8*, *9*, *34*) and 36.8% of authors have been actively publishing over only one year. Bibliometric research has also shown that academic citations display a skewed distribution where only a tiny share of publications, journals, and authors receive disproportionately high citations which has increased recently (*35*). These studies suggest that academic-success indicators are concentrated on a few countries, institutions, and authors. We know less about the *interrelatedness* of these trends.

We argue that measuring hierarchies in science requires a multidimensional approach. This is because there are positive correlations, feedback effects, and synergistic connections among measures of academic success. More collaborations could lead to more citations, which in turn may translate into greater productivity and more opportunities for geographical mobility; greater mobility may expand scholars' networks, enhancing their potential pool of collaborators. The absence or lack of success in any of these realms may negatively affect performance in the others. Social hierarchies in science will likely emerge from the confluence of successful (and unsuccessful) academic paths in these interrelated realms: productivity, collaboration, geographical mobility, and citations.

Author level variables and career-long measurement

We rely on 12 well-established academic performance indicators for all authors with at least one publication in the Scopus database. Our analytical sample includes 8.2 million authors and 28+ articles and reviews. We excluded 41,278 authors because their publications have missing information in the field of science. The list below provides each bibliometric indicator's name and assignment among our main four categories, productivity, collaboration, mobility, and

visibility. These indicators are computed at the author level and comprise all individual publications indexed by Scopus since 1996 covering authors' career from one up to 25 years.

1. The average number of coauthors per paper, *Avg. collaborations* (**collaboration/internationalization**)
2. The number of internationally coauthored publications, *Num. intl. pubs* (**collaboration/internationalization**)
3. The number of nationally coauthored publications, *Num. national pubs.* (**collaboration/internationalization**)
4. The number of coauthored papers, *Num. coauthored pubs.* (**collaboration/internationalization**)
5. The number of international changes in academic affiliation, *Num. intl. moves* (**mobility**)
6. The number of national changes in academic affiliation, *Num nat. moves* (**mobility**)
7. The number of affiliated organizations, *Num. organizations* (**mobility**)
8. The average number of citations per paper, *Avg. citations* (**impact/visibility**)
9. The total number of citations, *Total citations* (**impact/visibility**)
10. The fractional count of publications, *Fractional pubs.* (**productivity**)
11. The number of publications, *Total publications* (**productivity**)
12. The number of first-author publications, *First author publications* (**productivity**)

To favor comparability among scholars, we standardize these indicators by their academic age, measured as the years since the authors' first publication. We refer to this latter measure as "age." Average indicators (i.e., 1 and 8) do not require standardization as they are already expressed in relative terms. Age-standardized and average indicators were categorized into the maximum possible number of categories ensuring relative frequencies of at least 2% in all categories. To account for differences across disciplines in publication practices, we categorized variables separately for each of the six fields of science: Agricultural Sciences, Natural Sciences, Humanities, Medical and Health Sciences, Engineering and Technology, and Social Sciences.

This approach to variable coding is beneficial in the context of highly-skewed variables with heavy tails, as it allows us to: (i) include extreme values in the analysis, (ii) capture potential non-linear relationships among variables, (iii) preserve the distributional characteristics of each indicator, and (iv) avoids potential biases in correlational analyses due to outlier observations. The categories range from three for the number of international changes in academic affiliation in Agricultural Sciences to ten for the total number of citations in the Natural Sciences and Medical and Health Sciences. There are fewer categories in the number of international changes in academic affiliation because only 5% of scholars in Agricultural Sciences experienced international mobility.

A multidimensional measure of social stratification within scientific communities

We run a Multiple Correspondence Analysis (*36*) on the twelve categorized indicators for each macro field of science and in a separate analysis for aggregate of all fields that yielded similar results. We use the first three factorial coordinates of these six MCAs to cluster scholars into groups with similar academic performance profiles. We enhance comparability by conducting the cluster analysis independently for each academic-age group: One-year-old, two to five, six to nine, 10 to 14, 15 to 20, and 21 to 25. Hence, we conducted 36 hierarchical clusterings based on the Ward method followed by a cluster consolidation via the K-means algorithms. Neighboring solutions with five, six, seven, and eight clusters were assessed using the ratio of between to total variance. These assessments led us to focus on a six-cluster solution (see Supplementary Information). We term these clustering *bibliometric classes* and we use positional words to label them: *bottom*, *low*, *mid-low*, *mid-high*, *high*, and *top*. The marginal distribution of scholars across *bibliometric classes* measures the social stratification of science in each field. The differences between bibliometric classes in academic performance indicators capture the extent of hierarchies. We visualize these differences using factorial axes where distance implies differences and proximity implies similarity.

Authors' disambiguation algorithms were used (*37*) to assign papers to authors and to identify groups of authors who publish together in the global network of co-authorship. We group authors into scientific communities according to their degrees of proximity in collaboration networks. Scholars that coauthor papers are maximally close, whereas authors without any coauthor in common are maximal distal. We identify scientific communities using 18 different criteria for grouping authors based on their authorship proximity. Next, we examine the authors' distribution across bibliometric classes within these scientific communities. For this analysis, we pooled all academic-age groups and compared the distribution of authors within each scientific community according to their academic age and bibliometric class. A side-by-side comparison of the bibliometric classes and academic-age distributions within scientific communities and entropy measures for these two distributions allows for assessing the nature and strength of stratification across scientific communities.

Results

We represent scientific hierarchies and *bibliometric classes* using the first two MCA axes. We interpret these axes according to the variables' percentage contribution to their variance, as displayed in Figure 1. A vertical line is drawn at the mean percentage contribution, i.e., 8.3%. Markers to the right of this vertical line indicate variables with above-average contributions to the axes' variance.

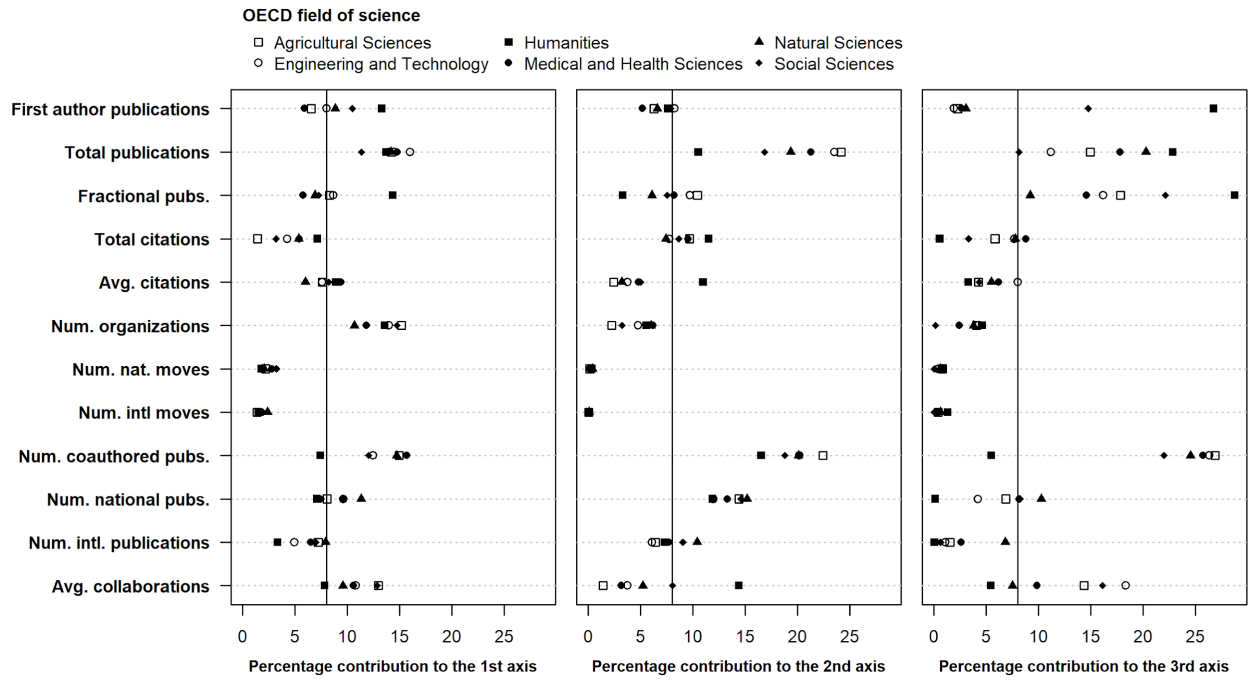

Fig 1. Variables' percentage contribution to the first three factorial axes by field of science and average contribution (vertical line).

The variables that contribute the most to the first factorial axis are the total publications, number of organizations, number of coauthored publications, average colalborations, and first-author publications. Field differences are evident in the contribution of these variables to the first axis. For instance, compare the low contribution of "Num. coauthored pubs." and "Avg. collaborations" versus the above-average contribution of "First author publications" for the Humanities (square), i.e., a traditionally non-collaborative field. The reverse is valid for the Social Sciences (diamond), i.e., a more significant contribution of coauthored papers to the first axis compared to first-author publications. We labeled the first MCA axis as "Academic age, number of organizations, and individual productivity." A large coordinate in this axis represents older academic age, a relatively high number of organizations, and an above-average number of publications, as first-author in collaborations.

The variables that contribute the most to the second factorial axis are total, fractional (for some fields), and coauthored publications. In addition, the total number of citations and the number of national publications also contribute significantly to the second axis. We labeled the second axis as: "Total productivity, visibility, and collaborations."

The variables that contribute the most to the third factorial axis are first author publications, total publications, fractional publications, number of coauthored publications, and average collaborations. There is a large variety among fields of science in contribution of these variables.

Hence, according to the MCA results, the organization of scholars according to their bibliometric indicators revolves around two main dimensions: “Academic age, number of organizations, and individual productivity” on the one side, and “Total productivity, visibility, and collaborations,” on the other. Scholars’ productivity is distinctly comprised in both dimensions. In the first dimension, productivity goes along with age and first-author publication. In the second dimension, productivity is less dependent on age and is associated with collaborations and citations. Interestingly, none of the mobility measures contributes significantly to the first three MCA axes that could stem from the very small share of mobile authors (about 8% in international and 12% in national moves).

Fig. 2 displays the authors' distribution by science fields according to two synthetic measures of academic performance and the *bibliometric classes* detected via cluster analysis. Authors with identical bibliometric measures are grouped and represented as circles to reduce overplotting. Circles’ size is proportional to the number of authors with identical bibliometric profiles. Although we conduct the analysis for all ages and find similar results across those (gray background circles), we highlight the bibliometric stratification of those between 10 and 15 years of academic age. The top group comprises the most successful authors based on combining our 12 bibliometric measures. The bottom-left includes those at the bottom of academic achievement indicators’ distributions. Existing differences in academic practices (e.g., publication, collaboration, mobility, and citation) across fields of science require us to let axes’ scales be free and prevent scaled comparisons across them.

The clustering of authors according to their academic achievement is a direct measure of existing hierarchies in these fields of science. Despite disciplinary differences in size and scientific practices, the commonalities in the stratification of authors are notable. In all six fields of science, the top of the hierarchy comprises a minimum of 6% in Humanities to a maximum of 19% in Natural Sciences. The bottom class ranges from a minimum of 22% in Natural Sciences to a maximum of 32% in Engineering and Technology. On the contrary, the middle- and bottom classes unanimously position towards the bottom left quadrant, meaning they are always worse off. This structure replicates among other academic-age groups with more than one year of career.

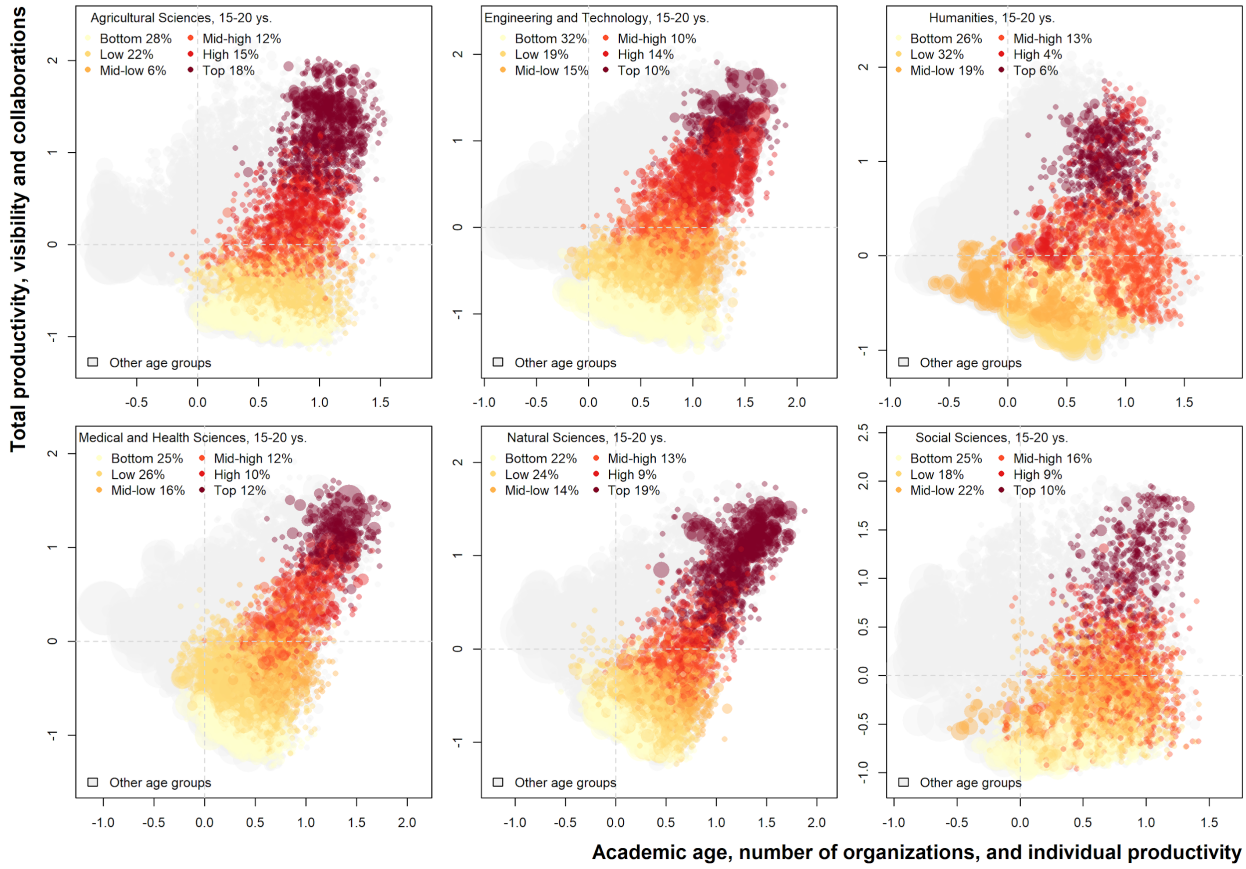

Fig. 2. Social structure of science and hierarchy of six identified clusters from top to bottom in six macro fields of science. Multiple Correspondence Analysis (MCA) results using the 12 most widely used bibliometric variables allowed identifying six classes of scientists from Bottom, Low, Middle low, Middle up, High, to Top. In all six fields of science and five-year career groups from a minimum of 1 to a maximum of 25 years of publication career indexed in Scopus, we see the same hierarchical structure appearing. A minority of the top class is identified which consists of less or about 10% (in most fields) of the most successful scientists indicated with dark red colors in the figure.

Fig. 3 shows the distribution of authors according to bibliometric classes (Panel A) and academic age groups (Panel B) across 19,970 scientific communities with at least 20 authors (99% of authors and 42.7% of communities). These communities are identified from the collaboration networks measured through co-authorship of publications. In panels A and B, scientific communities are represented by horizontal lines sorted from largest (on the top) to smallest and the deciles of the community-size distribution are indicated in the vertical axis. According to these panels, bibliometric-based stratification is similar to stratification based on age, suggesting that collaboration networks comprise authors of all ages and from all bibliometric classes. This similarity of bibliometric-class and academic age compositions is confirmed by Panel C, which displays the empirical density of the community-level entropy of authors' distribution by bibliometric classes and age groups. We display results for three community detection scenarios out of 18 that were assessed, to maintain the figure's clarity. The fact that all density curves are strongly skewed towards high entropy values (max entropy = 1) confirms our visual assessment of Panels A and B and suggests our results are robust to different community detection scenarios.

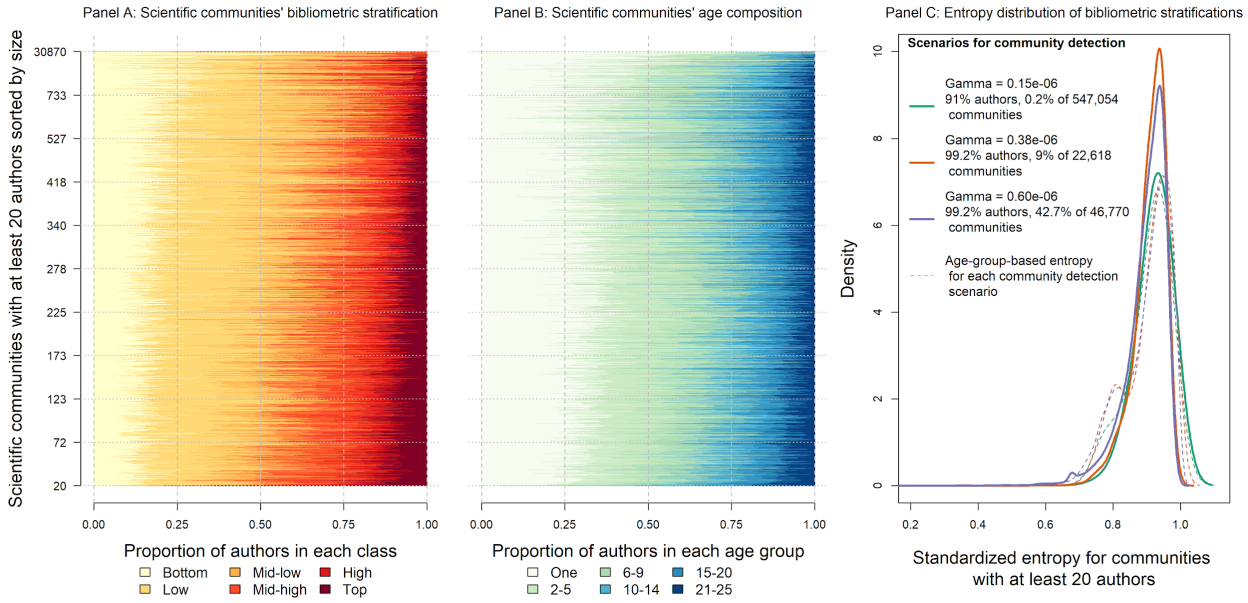

Fig. 3. Composition of communities of collaboration in terms of top to bottom classes (left) and age groups (middle) and entropy of stratifications (right). To investigate the trends shown in Fig. 2 further and control the collaboration structure among the classes, we turned to co-authorship networks of the studied 28 million publications. Networks of collaboration in terms of co-authoring scientific publications among 8.2 million authors worldwide allowed us to identify communities of collaboration. We used the Constant Potts Model (CPM) and its extension for bipartite networks with a varying range of 18 thresholds for the resolution parameter to detect communities. In all these detected communities (only 3 shown in panel C in the figure to preserve clarity), we investigated the class and age composition of members. Independent from the threshold used, all these communities have a heterogeneous composition of classes and age groups and analysis of entropies of this stratification indicates an inter-class and inter-age collaboration structure among the most and least prolific, collaborative/internationalized, and mobile scientists.

Discussion

This paper provided a quantitative assessment of the global hierarchies in science using bibliometric data across fields of science and within research communities. Our results show that a stratified hierarchical system exists, and it is as strong as stratification by academic age. We evaluated collaboration ties among classes and whether specific age groups dominate it. We provide the aggregated data to enable future research on the causes and consequences of this hierarchy.

Science is transmitted from senior scholars to the younger generations through a mentorship relationship that affects mentees' future success (*38–40*). Such mentoring and supervisor-supervisee relationships inherently have an age structure as junior scholars are trained by senior ones. Hence, we expect a share of observed scientific collaborations to be among junior and senior scholars. Nevertheless, our results show that the proportion of scholars who exit the system (i.e., one year olds who do not publish any longer) amount to 25% or more of the members of identified communities which cannot be solely representing the age structure of academia and could be driven by the performance measures described and the hierarchical structure inherent in them that drives a high proportion to exit the system prematurely. The probability of having higher impact and citations in the science system is disproportionately distributed and highly stratified (*35*).

Note that these bibliometric measures and indicators are widely used in national research assessment exercises (*41*, *42*) to determine who should be hired and promoted and whose research should be funded (*23*). Here, based on our analysis which was possible by adopting a global, multivariate, and multi-method framework to debunk the widely-spread myths about increased productivity, collaboration, internationalization, mobility, and impact among scientists, we call for a further elaborated investigation of these trends. We propose considering academic age, career cohorts and composition of a multitude of bibliometric variables instead of solely relying on one-indicator explanations which might be appealing to attract policy-makers' attention, but might be detrimental to our understanding of the science system, its social structure, and its inherent hierarchies and intersectional inequalities (*2*).

Materials and methods

We use all 28+ million article and review publications indexed in Elsevier's Scopus since 1996. We limit these publications to all of those written by the authors having identification numbers in Scopus and declared as "disambiguated" by Elsevier which has a 98.3% precision and a 90.6% recall (*37*). We further disambiguate the academic affiliation of authors in this set of publications using the Research Organization Registry's (ROR) Application Programming Interface (API) and geocode organizations' addresses to subnational units (see details in *43*). This reduces our

coverage of publications down from 36 million to 28.5 publications by 8.2 million disambiguated authors.

We use 12 well-established and widely-used bibliometric measures for productivity, (inter)national collaboration, (inter)national mobility, impact and citations (described in detail in Supplementary Information) to construct Multiple Correspondence Analysis (MCA) models with 3 axes. We further analyze the MCA results using cluster analysis to identify six clusters from top to middle and bottom.

In parallel, we construct global bipartite networks of co-authorship among the 8.2 million authors, identify its largest connected (giant) component and detect communities of densely collaborating scientists. We use the Constant Potts Model (CPM) (*44*) and its extension to bipartite networks (*43*, *45*, *46*) with a varying range of 18 resolution parameters to cross-check the identified communities (see Supplementary Information).

Acknowledgements

We thank Cassidy R. Sugimoto for helpful comments on an earlier version of this manuscript.

References

1. W. Shrum, J. Genuth, W. B. Carlson, I. Chompalov, W. E. Bijker, *Structures of Scientific Collaboration* (MIT Press, 2007).
2. D. Kozlowski, V. Larivière, C. R. Sugimoto, T. Monroe-White, Intersectional inequalities in science. *PNAS*. **119** (2022), doi:10.1073/pnas.2113067119.
3. I. Chompalov, J. Genuth, W. Shrum, The organization of scientific collaborations. *Research policy*. **31**, 749–767 (2002).
4. W. Shrum, I. Chompalov, J. Genuth, Trust, Conflict and Performance in Scientific Collaborations. *Soc Stud Sci*. **31**, 681–730 (2001).
5. A. F. Castro Torres, D. Alburez-Gutierrez, North and South: Naming practices and the hidden dimension of global disparities in knowledge production. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2119373119 (2022).
6. A. Akbaritabar, F. Squazzoni, Gender Patterns of Publication in Top Sociological Journals. *Science, Technology, & Human Values* (2020), doi:10.1177/0162243920941588.
7. V. Larivière, C. Ni, Y. Gingras, B. Cronin, C. R. Sugimoto, Bibliometrics: Global gender disparities in science. *Nature*. **504**, 211–213 (2013).
8. X. Zhao, A. Akbaritabar, R. Kashyap, E. Zagheni, A gender perspective on the global migration of scholars. *Proceedings of the National Academy of Sciences*. **120**, e2214664120 (2023).
9. E. Sanliturk, E. Zagheni, M. J. Dańko, T. Theile, A. Akbaritabar, Global patterns of migration of scholars with economic development. *Proceedings of the National Academy of Sciences*. **120**, e2217937120 (2023).
10. V. Burris, The Academic Caste System: Prestige Hierarchies in PhD Exchange Networks. *Am Sociol Rev*. **69**, 239–264 (2004).
11. A. Clauset, S. Arbesman, D. B. Larremore, Systematic inequality and hierarchy in faculty hiring networks. *Sci. Adv.* **1**, e1400005 (2015).
12. P. Bourdieu, J. C. Passeron, *The Inheritors: French Students and Their Relation to Culture* (University of Chicago Press, 1979).
13. J. Alper, The Pipeline Is Leaking Women All the Way Along. *Science*. **260**, 409–411 (1993).
14. B. Hofstra, D. A. McFarland, S. Smith, D. Jurgens, Diversifying the Professoriate. *Socius*. **8**, 23780231221085120 (2022).
15. G. Marini, V. Meschitti, The trench warfare of gender discrimination: evidence from academic promotions to full professor in Italy. *Scientometrics*. **115**, 989–1006 (2018).
16. E. Leahey, C. L. Cain, Straight from the source: Accounting for scientific success. *Soc Stud Sci*. **43**, 927–951 (2013).
17. T. Pikkety, *Capital et idéologie* (Seuil, Paris, ed. 1st, 2019).
18. M. Mazzucato, *The entrepreneurial state: debunking public vs. private sector myths* (Penguin Books, Erscheinungsort nicht ermittelbar, 2018).
19. G. Therborn, *The killing fields of inequality* (Polity, Cambridge, 2013).
20. D. Penman, S. Goldson, Competition to collaboration: changing the dynamics of science. *Journal of the Royal Society of New Zealand*. **45**, 118–121 (2015).
21. R. Müller, Collaborating in Life Science Research Groups: The Question of Authorship. *High Educ Policy*. **25**, 289–311 (2012).
22. P. van den Besselaar, S. Hemlin, I. van der Weijden, Collaboration and Competition in

Research. *High Educ Policy*. **25**, 263–266 (2012).
23. C. R. Sugimoto, V. Larivière, *Measuring Research: What Everyone Needs to Know* (Oxford University Press, 2018).
24. J. Wilsdon, L. Allen, E. Belfiore, P. Campbell, S. Curry, S. Hill, R. Jones, R. Kain, S. Kerridge, M. Thelwall, J. Tinkler, I. Viney, P. Wouters, J. Hill, B. Johnson, The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management (2015), doi:10.13140/RG.2.1.4929.1363.
25. G. Abramo, C. A. D'Angelo, F. Di Costa, Research collaboration and productivity: is there correlation? *High Educ*. **57**, 155–171 (2009).
26. J. Melkers, A. Kiopa, The Social Capital of Global Ties in Science: The Added Value of International Collaboration: The Social Capital of Global Ties in Science. *Review of Policy Research*. **27**, 389–414 (2010).
27. S. Wuchty, B. F. Jones, B. Uzzi, The Increasing Dominance of Teams in Production of Knowledge. *Science*. **316**, 1036–1039 (2007).
28. C. R. Sugimoto, N. Robinson-Garcia, D. S. Murray, A. Yegros-Yegros, R. Costas, V. Larivière, Scientists have most impact when they're free to move. *Nature*. **550**, 29–31 (2017).
29. R. Sinatra, D. Wang, P. Deville, C. Song, A.-L. Barabasi, Quantifying the evolution of individual scientific impact. *Science*. **354**, aaf5239–aaf5239 (2016).
30. L. Liu, Y. Wang, R. Sinatra, C. L. Giles, C. Song, D. Wang, Hot streaks in artistic, cultural, and scientific careers. *Nature*. **559**, 396–399 (2018).
31. D. Fanelli, V. Larivière, Researchers' Individual Publication Rate Has Not Increased in a Century. *PLOS ONE*. **11**, e0149504 (2016).
32. J. P. A. Ioannidis, R. Klavans, K. W. Boyack, Thousands of scientists publish a paper every five days. *Nature*. **561**, 167–169 (2018).
33. M. F. Fox, I. Nikivincze, Being highly prolific in academic science: characteristics of individuals and their departments. *High Educ*. **81**, 1237–1255 (2021).
34. A. Akbaritabar, T. Theile, E. Zagheni, Global flows and rates of international migration of scholars. *MPIDR Working Paper*. **018** (2023), doi:https://dx.doi.org/10.4054/MPIDR-WP-2023-018.
35. M. W. Nielsen, J. P. Andersen, Global citation inequality is on the rise. *PNAS*. **118** (2021), doi:10.1073/pnas.2012208118.
36. B. Le Roux, H. Rouanet, *Geometric Data Analysis: from correspondence analysis to structured data analysis* (Dordrecht, 2004).
37. J. Baas, M. Schotten, A. Plume, G. Côté, R. Karimi, Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*. **1**, 377–386 (2020).
38. J. F. Liénard, T. Achakulvisut, D. E. Acuna, S. V. David, Intellectual synthesis in mentorship determines success in academic careers. *Nature Communications*. **9**, 4840 (2018).
39. Q. Ke, L. Liang, Y. Ding, S. V. David, D. E. Acuna, A dataset of mentorship in bioscience with semantic and demographic estimations. *Sci Data*. **9**, 467 (2022).
40. Y. Ma, S. Mukherjee, B. Uzzi, Mentorship and protégé success in STEM fields. *PNAS* (2020), doi:10.1073/pnas.1915516117.
41. A. Akbaritabar, G. Bravo, F. Squazzoni, The impact of a national research assessment on the publications of sociologists in Italy. *Science and Public Policy* (2021),

doi:10.1093/scipol/scab013.
42. T. Zacharewicz, B. Lepori, E. Reale, K. Jonkers, Performance-based research funding in EU Member States—a comparative assessment. *Science and Public Policy*. **46**, 105–115 (2019).
43. A. Akbaritabar, A quantitative view of the structure of institutional scientific collaborations using the example of Berlin. *Quantitative Science Studies*. **2**, 753–777 (2021).
44. J. Reichardt, S. Bornholdt, Detecting fuzzy community structures in complex networks with a Potts model. *Physical Review Letters*. **93**, 218701 (2004).
45. V. A. Traag, P. Van Dooren, Y. Nesterov, Narrow scope for resolution-limit-free community detection. *Phys. Rev. E*. **84**, 016114 (2011).
46. A. Akbaritabar, G. Barbato, An internationalised Europe and regionally focused Americas: A network analysis of higher education studies. *European Journal of Education*. **56**, 219–234 (2021).
47. A. J. Lotka, The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*. **16**, 317–323 (1926).
48. J. R. Cole, H. Zuckerman, The productivity puzzle. *Advances in Motivation and Achievement. Women in Science. JAI Press, Greenwich, CT* (1984).
49. J. R. Cole, H. Zuckerman, Marriage, Motherhood and Research Performance in Science. *Scientific American*. **256**, 119–125 (1987).
50. A. Akbaritabar, N. Casnici, F. Squazzoni, The Conundrum of Research Productivity: a Study on Sociologists in Italy. *Scientometrics*. **114**, 859–882 (2018).
51. M. F. Fox, Publication Productivity among Scientists: A Critical Review. *Social Studies of Science*. **13**, 285–305 (1983).
52. M. Gauffriau, P. O. Larsen, I. Maye, A. Roulin-Perriard, M. von Ins, Comparisons of results of publication counting using different methods. *Scientometrics*. **77**, 147–176 (2008).
53. S. F. Way, A. C. Morgan, A. Clauset, D. B. Larremore, The misleading narrative of the canonical faculty productivity trajectory. *Proc Natl Acad Sci USA*. **114**, E9216–E9223 (2017).
54. L. B. Ellwein, M. Khachab, R. H. Waldman, Assessing research productivity: evaluating journal publication across academic departments. *Academic Medicine*. **64**, 319–25 (1989).
55. G. Abramo, C. A. D'Angelo, National-scale research performance assessment at the individual level. *Scientometrics*. **86**, 347–364 (2011).
56. E. Leahey, From Sole Investigator to Team Scientist: Trends in the Practice and Study of Research Collaboration. *Annu. Rev. Sociol.* **42**, 81–100 (2016).
57. S. Milojević, Principles of scientific research team formation and evolution. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 3984–3989 (2014).
58. J. Wang, D. Hicks, Scientific teams: Self-assembly, fluidness, and interdependence. *Journal of Informetrics*. **9**, 197–207 (2015).
59. H. F. Moed, G. Halevi, A bibliometric approach to tracking international scientific migration. *Scientometrics*. **101**, 1987–2001 (2014).
60. H. F. Moed, M. Aisati, A. Plume, Studying scientific migration in Scopus. *Scientometrics*. **94**, 929–942 (2013).
61. G. Laudel, Studying the brain drain: Can bibliometric methods help? *Scientometrics*. **57**, 215–237 (2003).
62. R. Kashyap, R. G. Rinderknecht, A. Akbaritabar, D. Alburez-Gutierrez, S. Gil-Clavel, A. Grow, J. Kim, D. R. Leasure, S. Lohmann, D. V. Negraia, D. Perrotta, F. Rampazzo, C.-J.

Tsai, M. D. Verhagen, E. Zagheni, X. Zhao, Digital and Computational Demography (2022), , doi:10.31235/osf.io/7bvpt.
63. A. Miranda-González, S. Aref, T. Theile, E. Zagheni, Scholarly migration within Mexico: analyzing internal migration among researchers using Scopus longitudinal bibliometric data. *EPJ Data Sci.* **9**, 34 (2020).
64. P. Block, C. Stadtfeld, G. Robins, A statistical model for the analysis of mobility tables as weighted networks with an application to faculty hiring networks. *Social Networks*. **68**, 264–278 (2022).
65. M. Kato, A. Ando, National ties of international scientific collaboration and researcher mobility found in Nature and Science. *Scientometrics*. **110**, 673–694 (2017).
66. Z. Chinchilla-Rodríguez, L. Miao, D. Murray, N. Robinson-García, R. Costas, C. R. Sugimoto, A Global Comparison of Scientific Mobility and Collaboration According to National Scientific Capacities. *Front. Res. Metr. Anal.* **3** (2018), doi:10.3389/frma.2018.00017.
67. H. D. Boekhout, V. A. Traag, F. W. Takes, Investigating scientific mobility in co-authorship networks using multilayer temporal motifs. *Network Science*. **9**, 354–386 (2021).
68. G. Vaccario, L. Verginer, F. Schweitzer, The mobility network of scientists: analyzing temporal correlations in scientific careers. *Appl Netw Sci*. **5**, 1–14 (2020).
69. L. Waltman, A review of the literature on citation impact indicators. *Journal of Informetrics*. **10**, 365–391 (2016).
70. L. Egghe, The Hirsch index and related impact measures. *Ann. Rev. Info. Sci. Tech.* **44**, 65–114 (2010).
71. R. Burrows, Living with the H-Index? Metric Assemblages in the Contemporary Academy. *The Sociological Review*. **60**, 355–372 (2012).
72. J. E. Hirsch, An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*. **102**, 16569–16572 (2005).
73. J. Wang, Citation time window choice for research impact evaluation. *Scientometrics*. **94**, 851–872 (2013).
74. P. Donner, Effect of publication month on citation impact. *Journal of Informetrics*. **12**, 330–343 (2018).
75. OECD, Revised Field of Science and Technology (FOS) classification in the Frascati Manual (Classification, Field of science and technology classification, FOS, Frascati, Methodology, Research and development) (2007), (available at https://www.oecd.org/science/inno/38235147.pdf).
76. L. Lebart, A. Morineau, M. Piron, *Statistique Exploratoire Multidimensionnelle* (Dunod, Paris, ed. 2, 1997).
77. C. E. Pardo, P. C. Del Campo, Combinación de métodos factoriales y de análisis de conglomerados en R: El paquete FactoClass. *Revista Colombiana de Estadistica*. **30**, 231–245 (2007).
78. L. Kaufman, P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis* (Wiley, New York, 1990).
79. M. E. J. Newman, Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E*. **64**, 016132 (2001).
80. M. E. J. Newman, Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E*. **64**, 016131 (2001).
81. B. Keegan, D. Gergle, N. Contractor, Do Editors or Articles Drive Collaboration?

Multilevel Statistical Network Analysis of Wikipedia Coauthorship, 10 (2012).
82. V. Traag, *Algorithms and Dynamical Models for Communities and Reputation in Social Networks* (Springer, 2014).
83. V. A. Traag, L. Waltman, N. J. van Eck, From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep*. **9**, 5233 (2019).
84. V. A. Traag, G. Krings, P. Van Dooren, Significant Scales in Community Structure. *Sci Rep*. **3**, 2930 (2013).

Supplementary Information for: A global perspective on the social structure of science

Aliakbar Akbaritabar[a1], Andres F. Castro Torres[a,b], Vincent Larivière[c]

[a]Max Planck Institute for Demographic Research, Rostock, Germany, [b]Center for Demographic Studies, Autonomous University of Barcelona, Barcelona, Spain, [c]Université de Montréal, Montréal, Canada

[1]Corresponding author, Email: akbaritabar@demogr.mpg.de

Here we provide more detail on the literature, and bibliometric measures widely used for productivity, collaboration/internationalization, mobility and impact. In addition, we provide more detail on our methods and robustness analyses carried out to control the validity and replicability of our results.

Background

The literature has proposed various measures for productivity, collaboration/internationalization, and mobility. There are detailed discussions on the pros and cons of each measure.

Productivity. The number of publications is the most used measure of research productivity (*41*, *47–50*). Some have emphasized the need to place authors at the core of analysis and consider individual characteristics, personality, early age behavior etc. (See a review of productivity measures in *51*). Others have compared different counting methods (e.g., whole counting versus fractional counting which allocates a fraction of productivity to contributing authors or the mean number of papers that one has contributed to them (*52*). Some propose a career-based productivity measurement to see the ups and downs of one's productivity (*53*). Some measures for the productivity of authors consider the impact factor of journals and author name position in the by-lines of the paper e.g., in form of $P = \sum_{i=1}^{N} W_i * Z_i$ (*54*). Other researchers have proposed a microeconomic function to calculate fractional scientific strength (FSS) for research productivity. Authors discuss how it is possible to use this formula to calculate aggregate research productivity in the institution, academic disciplinary sector or even national levels formulated as $FSS_R = \frac{1}{t} \sum_{i=1}^{N} \frac{C_i}{\bar{C}} f_i$ (*55*). Nevertheless, most of these measures correlate highly with the more basic measures of productivity such as N of publications.

Collaboration/Internationalization. We consider collaboration and internationalization as two sides of the same coin. Research which is not collaborative (e.g., solo-author publications (*56*)) cannot by definition be international. Hence, it is beneficial to consider internationalization in combination with the notion of team science (*27*). While the list of authors of scientific papers are growing (*31*) this does not necessarily mean the productivity of individual authors are increased. Some research has differentiated between the notions of "article team" and "project team" (*57*) which could affect one's productivity during their career due to being embedded in a context of highly prolific collaborators and teams or having former collaboration ties with more prolific others (*58*).

Mobility. Here we briefly review the most innovative approaches in inferring mobility from a change in academic affiliation addresses (*8*, *9*, *34*, *59–63*) and highlight their advantages and shortcomings. Some works model mobility similar to a social selection process versus social influence. The authors (*64*) proposed an extension to the usual exponential random graph models (ERGMs), but they only considered mobility. While they include the immobile actors, they do not consider scientific collaboration's effect on facilitating mobility (*65*, *66*). Another work (*67*) draws on considering affiliation as an author's attribute and checking if a large proportion of their current collaborators are in the same institute or elsewhere. The authors take a triangle as the unit of analysis and count many typologies of graphlets. They "infer" mobility, without defining who has moved, using the collaboration e.g., if authors had a collaboration inside the same institution, and at a later time, their collaboration involves two institutions in the same country, that indicates that one author has moved nationally inside the same country, or if the collaboration has become an international one, that indicates one author has moved abroad. Their method of inferring a move using collaboration does not consider actual moves and instead infers it. It can be prone to neglecting an individual's specific trajectory (which is dependent on their attributes, e.g., gender, discipline, country of origin etc.) and mixing it with their collaborators' moves and trajectory. Other work (*68*) has taken "cities", "institutions" and "countries" as the nodes in a network and the authors move between these nodes. Their idea of *aggregating individual authors into larger entities* could neglect the author's trajectories and attributes (e.g., former/future collaborations) in shaping mobility patterns. In addition, they do not consider immobility.

Impact and citations. Many measures have been proposed for the scientific impact and visibility (*24*, *69*). This abundance of metrics stems from funding agencies and policymakers who wish to have a "one-number" measure that says it all about the scientist, their productivity and their impact. One such measure was the famous H-Index (*70–72*) which has received many criticisms and proposed extensions and variations over time. There have been discussions on the effect of selected time window post-publication on these impact measures and the time needed for publications to mature their count of citations (*73*) or even the publication month and how they could affect impact measurement (*74*). We decided to choose the two simplest measures of impact which consider a) the total number of citations received for all publications as an aggregated sum and b) divide that sum by the count of publications in form of an average.

Our twelve selected variables. Based on the reviewed literature and measures proposed for the four main variables, i.e., productivity, collaboration/internationalization, mobility, and impact, we selected the following twelve for our analysis.

- **Productivity**:
    - Total count of publications
    - Total count of publications as the first author
    - Fractional count of publications which considers the number of contributing authors
- **Collaboration/internationalization**:
    - Number of coauthored publications
    - Number of nationally coauthored publications
    - Number of internationally coauthored publications
    - The average number of coauthors
- **Mobility**:
    - Number of affiliated organizations throughout publication career
    - Number of national moves based on changes in academic affiliation addresses
    - Number of international moves based on changes in academic affiliation addresses
- **Impact and citations**:
    - Aggregated total count of citations
    - Average citations throughout publication career

Further detail on methods

For academic age, we take the first publication year as the start of publication career and the latest indexed publication date in Scopus as the end of publication career. This allows us to cover a maximum of 25 years of publication career due to our licence limits which starts from 1996. We divide this into six groups of one-year-olds, 2-5, 6-9, 10-14, 15-20, and 21-25 years of career.

It is a difficult task to assign one macro field of science and discipline to an academic's publications throughout their career. We use the six macro OECD fields of science (*75*) and a mapping of them to Scopus ASJC categories assigned to all publications of one scientist. We calculate the proportion of publications in each macro field. We then take the field with the highest proportion of publications as the field where the said scientist publishes the most. We are aware of the limitations of such an approach and further normalize the field assignment based on the count of total publications in a specific discipline.

Multiple Correspondence Analysis (MCA) and cluster analysis

We select MCA and clustering analysis because they are designed to capture structural aspects of dataframes (*36*, *76–78*). In our case, we use them for identifying the main axes of differentiation of authors' bibliometric performance as well as their grouping into the bottom, middle and top-performance authors. Stratified analysis by OECD macro field of science and authors' academic age account for disciplinary differences in bibliometric records and warrant comparability, respectively.

Following technical requirements for an adequate MCA, we recode all measures of authors' academic achievements into a number of categories that preserve as much as possible the distributional features of the numerical variables and yield categories with relative frequencies of a least 5% (*76*). This procedure was applied separately for each OECD macro field of science. Working with categorical variables allows us to preserve all observations, including extreme outliers with very high or very low measurements that are typically excluded from bibliometric analysis and variables with high skewness. Additionally, categorical data are well suited to capture nonlinear relations among variables. These procedures yielded six databases, one for each OECD macro field of science, with 12 categorical variables.

Next, we apply an MCA on each of these six databases. The MCA represents each author as a set of p factorial coordinates, where p equals the number of categories minus the number of variables (number of categories - 12 = p). Because variable categorization was conducted separately for each OECD macro field of science, the value of p varies as follows: Agricultural Sciences (68), Engineering and Technology (62), Humanities (52), Medical and Health Sciences (74), Natural Sciences (83), and Social Sciences (74). Importantly, factorial axes are ordered by the proportion of variance they comprised being the first the one with the largest share, and they are orthogonal. These two properties are important for the clustering analysis.

We use the top 25% factorial axes for computing author-level dissimilarity matrices by OECD macro fields of science and authors' academic age groups (One, 2 to 5, 6 to 9, 10 to 14, 15 to 20 21 to 25). The 36 dissimilarity matrices were the inputs for a two-step clustering analysis. We use the Ward method to perform hierarchical clustering. Next, a visual inspection of the dendrograms along with practical considerations on the number of clusters will permit further statistical analysis. We selected six to nine cluster solutions. Comparing the percentage of explained variance across these solutions, we decide to preserve a six-cluster solution for all OECD fields of science and academic age groups. The centers of these clusters and the dissimilarity matrices were used as inputs for the consolidation of the clustering solution via the K-means algorithm (*77*).

Hence, the K-means-consolidated solution of six clusters by academic age and OECD field of science measures the author's position within the space of our 12 measures of academic

achievements. Given the nested nature of higher and lower-order cluster solutions, we are confident that our main conclusions are not affected by the selection of six clusters.

Bipartite network modeling and community detection

We construct bipartite co-authorship networks using ties between publications as the first set of nodes and authors as the second set of nodes. Studies on co-authorship networks usually use a one-mode projection of these bipartite networks (*79*, *80*) whereby removing one of the two node types (i.e., modes), the network of the ties between members of the other type is constructed. The problem with this projection is twofold. First, some differing structures in a bipartite network yield the same one-mode structure causing information loss about the underlying structure. Second, the one-mode projection can present an artificially higher density and connectivity due to publications with a high number of authors which project to maximally connected cliques. We claim that by adopting methods and modeling strategies specifically developed for bipartite networks (*43*, *46*, *81*) we are able to treat the shortcomings.

To identify communities of co-authorship, we use bipartite community detection specifically the Constant Potts Model (CPM) (*82*). This is a specific version of the Potts model (*45*) which resolves the resolution limit problem in modularity that can obstruct the detection of small communities in very large networks. We use the implementation in the Leidenalg library in Python programming language. The recently proposed improvements in the algorithm ensures that even in very large networks, such as the one investigated here, identified communities do not have internal disconnections that can occur using modularity and the Louvain algorithm (*83*). Community detection emphasizes the importance of links *within* communities rather than those *between* them. CPM uses a resolution parameter *gamma* ("constant"), leading to communities such that the link density between the communities (external density) is lower than gamma and the link density within communities (internal density) is higher than gamma (*45*). We set the resolution (i.e., gamma) to 18 different values to test a varying number of scenarios and densities. These include gamma equal to 0.0001 (which finds 57,553 bipartite communities), 0.0006 (154,269), 0.001 (221,638), 0.006 (871,227), 0.01 (1,223,914), 0.06 (2,989,159), 1.5177403574950853e-07 (554,864), 6e-05 (47,026), 5.007588701787476e-05 (58,183), 3e-05 (54,749), 0.000015 (36,181), 0.0000075 (34,509), 0.00000375 (22,697), 0.000002 (28,118), 0.0000012 (39,684), 6e-7 (115,995), 1.2e-8 (47,501), 2.4e-10 (1).

Selected scenarios for bipartite community detection. After evaluating described 18 different values for the resolution parameter and investigating the community's composition based on age and bibliometric classes (see Fig. 3 in the main text), we selected 3 scenarios roughly with 22k, 47k and 550k detected communities as illustrative examples representing a wide range to visualize them in Fig 3 (in the main text). 1) Gamma is equal to 3.75e-06 which finds 22,697 communities. 2) Gamma is equal to 6e-05 which finds 47,026 communities and 3) Gamma is equal to the giant component's density, 1.5177403574950853e-07, which finds

554,864 communities. The theoretical reason for using the giant component's density as a resolution parameter is that literature (*84*) has shown this to be a turning point in the number of communities detected from too many to too few.

We set a random seed in our community detection hence using the same resolution parameter and the seed will replicate the membership in communities. Please note that in our current analysis, the choice of gamma or the number of communities detected is not important, the most important question for us was the "composition" of these detected communities and whether they include a homogeneous population, i.e., merely one bibliometric class or age group populating a specific set of communities with higher (or lower) collaboration density or if communities consisted of a heterogeneous population with multiple bibliometric classes and age groups. This was proven to be the case irrelevant of the choice of the resolution parameter used and the heterogeneous composition of the communities did not change. Hence, we selected only three scenarios as illustrative examples to visualize in the main text.

Additionally, we controlled if the identified communities were composed of a disproportionately larger share of specific countries or disciplines. For this analysis, and similar to the case of Fig 3 in the main text, we limit the communities to those having a minimum of 20 authors. That decreases the count of communities from 22,697 to 2,030, from 47,026 to 19,992, and from 554,864 to 909 for the three selected scenarios, with gamma equal to 3.75e-06, 6e-05, and 1.5177403574950853e-07, respectively. We found that the mean number of countries per community is 35.04, 13.75, and 8.6 and (median: 33, 12, and 2, and standard deviation: 26.23, 9.19, and 24.98) for the three selected scenarios, i.e., gamma equal to 3.75e-06, 6e-05, and 1.5177403574950853e-07, respectively. The majority of communities were dominated by one country. After excluding the communities with less than 20 authors, the median increased to 33, 12, and 2 in the three selected scenarios, respectively. This is in line with our speculation that science is still being produced nationally and internationalization is the exception than the rule (*43*, *46*). Some communities include authors from a diverse array of countries, but these are the minority. In terms of the macro fields of science and disciplines, we found that the mean number of disciplines per community is 4.94, 4.11, and 2.64 (median: 6, 4, and 2, and standard deviation: 1.51, 1.17 and 1.35) for the three selected scenarios, i.e., gamma equal to 3.75e-06, 6e-05, and 1.5177403574950853e-07, respectively. This indicates that the interdisciplinary mode of producing science is the exception and most of the identified communities are dominated by one macro field of science.