Fitzgerald *et al. Appl Netw Sci* (2021) 6:38
https://doi.org/10.1007/s41109-021-00371-w

Applied Network Science

RESEARCH Open Access

Is academia becoming more localised? The growth of regional knowledge networks within international research collaboration

John Fitzgerald[1*], Sanna Ojanperä[1,2,4] and Neave O'Clery[1,3,4]

*Correspondence:
fitzgeraldj@maths.ox.ac.uk
[1] Mathematical Institute, Oxford University, Andrew Wiles Building, Radcliffe Observatory Quarter, Oxford OX2 6GG, UK
Full list of author information is available at the end of the article

Abstract

It is well-established that the process of learning and capability building is core to economic development and structural transformation. Since knowledge is 'sticky', a key component of this process is learning-by-doing, which can be achieved via a variety of mechanisms including international research collaboration. Uncovering significant inter-country research ties using Scopus co-authorship data, we show that within-region collaboration has increased over the past five decades relative to international collaboration. Further supporting this insight, we find that while communities present in the global collaboration network before 2000 were often based on historical geopolitical or colonial lines, in more recent years they increasingly align with a simple partition of countries by regions. These findings are unexpected in light of a presumed continual increase in globalisation, and have significant implications for the design of programmes aimed at promoting international research collaboration and knowledge diffusion.

Keywords: Research collaboration, Knowledge diffusion, Economic development, Geopolitics, Country network, Community detection, Data visualisation

Introduction

Following advances in transportation and communication technology over the past centuries, we're witnessing a rise in global interactions both in terms of cross-border trade and investment as well as flows of people and information. Termed globalisation, this complex process involves interaction and integration of people, businesses, and governments and while it primarily concerns economic aspects, social and cultural dimensions are similarly salient. In parallel to the process of globalisation, ongoing global economic restructuring has resulted in a transition towards a knowledge-based economy, where a 'greater reliance on intellectual capabilities than on physical or natural resources' (Powell and Snellman 2004) has meant a rise in production and services that are based on knowledge-intensive activities. The importance of specialised skills along with knowledge and information as forms of non-physical capital has grown, since economic growth increasingly derives from intangible intellectual property including copyrights, patents, trademarks, and trade secrets that work to make more effective use of inputs

and available resources. Knowledge diffusion and innovation underpin competition and fuel economic growth at almost all stages of development, as well as play a critical role in enabling responses to complex economic, environmental, and social challenges. Domestic and global research communities are central players in creating and diffusing knowledge and contributing to the development of new products and processes. These communities comprise of research and development activities, research laboratories, universities, and other educational institutions that, together with partners in the private sector and government, form innovation ecosystems (Jackson 2011).

The role of innovation, science and technology as drivers of economic growth and as vital enablers of sustainability is highlighted in the recent UNESCO (2015) Science Report, which showcases the trajectories of a large number of countries 'incorporating science, technology, and innovation in their national development agendas, in order to be less reliant on raw materials and move towards knowledge economies.' The desirability of fostering local skills and capacities for economic development is similarly echoed by recent work in economic complexity and economic geography (Hidalgo et al. 2007; Frenken and Boschma 2007; Hidalgo and Hausmann 2009; Hausmann and Hidalgo 2011; O'Clery et al. 2019a), analysing the growth of cities and regions. This literature finds the availability of diverse knowledge capacities or complex skills and capabilities as central to the development trajectories of regions, countries, and cities. They conceptualise knowledge and capabilities as geographically 'sticky', since tacit knowledge and abilities are a result of a workforce with skills learned on-the-job, and are thus not easily transportable. Research collaboration, in particular with academics from other regions likely in possession of novel or complementary skills and capabilities, could allow countries to upgrade their academic capacity and respond to unique societal and economic challenges more readily.

As countries and regions find themselves at various stages of the transformation towards—and readiness to join—the global knowledge economy (Ojanperä et al. 2017, 2019), the creation of scientific knowledge is more important than ever. Public and private sector funding is directed towards developing domestic research capabilities, and countries are putting policies in place to attract scientific talent from abroad (UNESCO 2015). The OECD Development strategy, implemented in partnership with the United Nations and the World Bank, as well as the OECD policy frameworks for Tertiary Education, Innovation, Development, and Gender Equity, all call for the promotion of regional and international research networks in order to further the dual pursuit of research communities everywhere, summarised by the Programme on Innovation (2012) as: 'knowledge generation per se and their specific role in attaining national development priorities'.

Reflecting this trend, the number of researchers and publications has been growing, with a 20 percent increase between 2007 and 2014 (UNESCO 2015). The extent of scientific collaboration has increased in parallel, both overall (Wagner-Döbler 2001; Meyer and Bhattacharya 2004), and internationally, between researchers based in different countries (Narin et al. 1991; Wagner and Leydesdorff 2005a; Wuchty et al. 2007; Jones et al. 2008; Gazni et al. 2012). Various factors have been suggested as underpinning the growing propensity to collaborate, including advancements in technologies facilitating remote collaboration (Ding et al. 2010), policy initiatives and

funding schemes to encourage international collaboration (Frenken et al. 2009; Ubfal and Maffioli 2011), specialisation requiring collaboration with researchers who may not be available within the local talent pool, cultivation of research impact and credibility (Kumar 2015), and avoidance of duplicating research efforts (Katz and Martin 1997).

Indeed, the internationalisation of research collaborations has received increasing attention over the past few decades. Collaboratively authored research has higher impact than research published by a sole author, both in terms of number of publications (Katz and Martin 1997; Lee and Bozeman 2005; Wuchty et al. 2007) and citations (Sooryamoorthy 2017; Gazni and Didegah 2011), while research published by international author teams tends to attract more citations than research authored by national teams (Narin et al. 1991; Katz and Martin 1997; Frenken et al. 2005). Furthermore, Jones et al. (2008) show that multi-university collaborations produce the highest impact papers when top-tier universities are included, and are increasingly stratified by in-group university rank.

An emergent body of literature on research collaboration networks—reviewed below—has primarily investigated ties between individuals or institutions, often focusing on particular disciplinary communities or bounded by a regional or sub-national context. Few studies, however, have looked in detail at changing patterns of international collaboration focusing on bilateral ties at the country level, and including all major disciplines. Instead studies tend to focus on particular disciplines, such as medicine or the life sciences. We look at research collaboration across all major disciplines, as it reflects the broad creation and diffusion of knowledge, which contributes to the development of new products and processes, or innovation across economic, social, and political domains.

The existing body of research on international research collaboration networks has deployed a variety of network methods, including network visualisation, local network measures focusing on the importance of nodes, models explaining network growth, and regression methods. In the present study, we apply a range of sophisticated methods deriving from network science and mathematical modelling, including historical profile clustering, calculation of the entropy of collaborations, community detection, and mutual information comparisons, which allow us to uncover patterns that have previously remained opaque.

Further, where studies have analysed a time period rather than investigated a snapshot, the time window tends to not span more than a decade or two. We address this research gap through analysing a dataset of international collaboratively authored scientific publications covering a range of disciplines published between 1970 and 2018. In doing so, we assess the extent to which countries learn from each other through 'borrowing' capabilities and specialisms from colleagues in other countries or regions, and thus induce knowledge flow. In the analysis to follow, we exploit a variety of network and mathematical modelling tools to analyse the temporal evolution of the global collaboration network to reveal what we term 'knowledge basins' [a concept related to 'skill basins' as proposed by O'Clery et al. (2019b)]. These are groups of countries which tend to collaborate frequently internally, but less frequently with other groups, thus forming localised (and potentially isolated) clusters of research output. These clusters evolve over time, aligning

with colonial and historical geopolitical alliances pre-2000, but coalescing more along geographical or regional lines since 2000.

The remainder of this paper is organised as follows. Section 2 will survey the relevant research on co-authorship networks. Our choice of data will be elaborated upon in Sect. 3, while Sect. 4 will introduce some preliminary analysis of the data. In Sects. 5 and 6 we will present our main research methodology and results with some discussion. Finally, Sect. 7 summarises our contribution, discusses the implications of our findings, and proposes avenues for future work.

Literature review

The literature on research networks has its roots in scientometrics, a sub-field of bibliometrics measuring and analysing scientific literature, but it additionally draws from related disciplines of information systems, information science, and science of science. While the creation of the Science Citation Index in 1964 and related studies (Burton and Kebler 1960; Garfield and Sher 1963; Kessler et al. 1962; Osgood and Xhignesse 1963; Price 1963; Tukey 1962) were seminal in establishing the field, the pioneering article by Price (1965) was the first one to investigate networks of scientific papers, and found that the network under study was scale-free with the in-degree (citations within an article) and out-degree (citations to an article) having power-law distributions. Since these early studies' focus on citation networks, the literature has branched out to comprise research on varied themes such as co-citation networks (documents are connected if they appear together in a reference list), co-word networks (words are connected if they appear together within a document), research collaboration (in particular through co-authorship of documents or collaborative grants), researcher mobility, and institutional boundaries. While these studies investigate varied topics, some themes that have received substantial research attention include identifying research fronts, evaluating the impact of individual authors in comparison to collaborations, and the relative influence of disciplines and journals.

Knowledge flows and co-authorship networks

This paper contributes to the literature on research collaboration—and specifically co-authorship—networks. In many cases, these are thought to be a proxy for knowledge flows, which are inherently challenging to define and measure. By knowledge we mean the creation and retention of knowledge by individuals or organisations, and by knowledge flows we mean the exchange or diffusion of ideas by individuals or organizations (Jaffe and Trajtenberg 1998). Such 'pure' knowledge and knowledge flows tend to be disembodied, and are non-rivalrous in the sense that one's consumption of knowledge does not prevent another from consuming the same knowledge. While these kinds of knowledge are difficult to measure chiefly due to their disembodied nature, some have suggested that the flows of certain knowledge-intensive products such as citations to patents could work as 'windows' to knowledge flows (Jaffe and Trajtenberg 1998). In a similar vein, internationally co-authored publications, which are considered a reliable proxy for research collaboration (Melin and Persson 1996; Glänzel and Schubert 2005; Heinze and Kuhlmann 2008), may be considered as 'windows' into knowledge flows between researchers located in different countries.

Co-authorship networks are some of the largest publicly available social networks and while they have received somewhat less research attention than citation networks, they enable a close examination of key aspects of what Newman (2004) terms as 'the structure of both academic knowledge and academic society'. The existing literature on co-authorship networks can roughly be divided into three streams based on the methodological approaches utilised, namely, bibliometric methods, survey-based methods, and network analysis. The studies applying a network analysis methodology form a somewhat more recent research area, and as our study falls within this stream, we will focus our discussion on the literature using related methodologies.

This literature investigates networks that vary in size from small groups e.g. related to a research institution (Fagan et al. 2018) to massive graphs e.g. depicting international patent citation networks (de Rassenfosse and Seliger 2020). The research field has gained notable interest after three seminal articles from Newman (2001a, 2001b, 2001c), which studied the micro and macro characteristics of seven large scientific co-authorship networks, and an article by Barabási et al. (2002) which examined the evolution and dynamics of these networks. Among further studies which looked at researcher collaboration networks, many focused on detecting popular or well positioned individuals (Fatt et al. 2010; Racherla and Hu 2010; Ye et al. 2012; Santos and Santos 2016). Newman (2001b) noted that scientific networks are highly clustered, with many triangles, while Goh et al. (2003) found that authors with a high betweenness centrality avoid collaboration with other authors who are similarly well-positioned, and rather seek less connected individuals.

Focusing on classifying the network structure, Newman (2001c) demonstrated that co-authorship networks could be characterised by the 'small world' property i.e., each author is not more than five or six steps away from each other within the network. Goh et al. (2002) found that the node degree distribution is scale free, indicating that while most authors have few collaborations, there are some that have numerous collaborations. Finding a similar pattern, Newman (2004) noted that biological scientists have significantly more coauthors than those publishing in mathematics or physics. Various studies have looked at the existence and size of the 'giant component', which seems to vary significantly across disciplines. Newman (2001c) found it comprises over 90 percent of authors in biomedical research, while Yan and Ding (2009) found it comprises just 20 percent of authors in library and information sciences. Hou et al. (2008) studied the network of authors within scientometrics and found that the two largest research clusters work on the same topic, but utilise different methodological approaches. Comparing network communities to the socioeconomic characteristics of the scholars, Rodriguez and Pepe (2008) found that communities best align with individuals working in the same department or institution suggesting that co-authorship is primarily driven by departmental and institutional affiliation.

International research collaboration

Studies adopting an international comparison include both regionally and globally focused approaches. Investigating the growth of international collaboration, Wagner and Leydesdorff (2005b) argue that the principle of preferential attachment—where those with more collaborations keep attracting proportionally more new

collaborations—explains the phenomenon. In support of this hypothesis, Ribeiro et al. (2018) identify a scale free node degree distribution for a global collaboration network comprising various scientific disciplines. Some authors argue that the core leading group consisting of the United States and Western nations has widened to include a much larger number of countries during the 1990's and 2000's (Leydesdorff et al. 2013). Other studies focusing on international research collaborations find that geographical distance and national borders continue to hinder cross-border collaboration (Frenken et al. 2009; Doria Arrieta et al. 2017). Looking at the patterns of medical research in Latin America and the Caribbean, Chinchilla-Rodríguez et al. (2012) find that the most productive countries collaborate mainly internally or with neighbouring countries, while small or developing countries tend to collaborate more distantly. Other studies suggest that the globalisation of science does not seem to have evolved uniformly across all countries and regions, as historical, sociotechnological, and geographical factors continue to play a key role (Geuna 2015; Scherngell 2013). This existing body of research adopts either a temporal snapshot into global research collaboration or covers a time window spanning up to two decades.

Data sources

Previous research has made use of bibliographic databases, academic search engines (ASEs), and services that offer a combination of these two functions. Bibliographic databases are comprehensive and reliable collections of information on academic outputs which allow users to efficiently query for information. ASEs on the other hand use computer algorithms to search the internet and recognize items which correspond to a query. They are less structured and subject to inconsistencies yet tend to be significantly larger in scope.

While it is challenging to measure the reach of these datasets, a recent article by Gusenbauer (2019) attempted to measure their respective sizes. The two largest scholarly bibliographic databases include Scopus (72 m records) and Web of Science (67 m records). The ASEs offer some significantly larger datasets, and comprise, among others, Google's academic index Google Scholar (387 m records), WorldWideScience (323 m records), AMiner (232 m records), Microsoft Academic (171 m records), Bielefeld Academic Search Engine (BASE) (118 m records), Q-Sensei Scholar (55 m records), and Semantic Scholar (40 m records). Aggregate services include ProQuest (280 m records) and EbscoHost (132 m records). While these sources of data have gained popularity within the field (Harzing and Alakangas 2016), each has their advantages and limitations depending on the geographic, disciplinary, or temporal scale of interest.

The Scopus database

Our dataset contains all co-authorship relations between authors of documents published between 1970 and 2018 which are indexed in Scopus. We chose Scopus as our data source because it has a high level of accuracy as is characteristic for bibliographic databases (Gusenbauer 2019; Gusenbauer and Haddaway 2019). It also has wide geographic, disciplinary, and temporal coverage including 24,600 active titles and 5000 publishers of scientific journals, books, and conference proceedings across the fields of science, technology, medicine, social sciences, and arts and humanities (Elsevier 2020).

Since we sought as comprehensive a dataset as possible, we decided not to consider the academic search engines because, while they are able to access the largest number of records, the query functions for them seem to be unreliable for detailed bibliometric data such as author affiliation (Mingers and Meyer 2017; Gusenbauer 2019). Similarly, while the aggregate services ProQuest and EbscoHost and the bibliographic database Web of Science provide more accurate results, it was not apparent whether our institutional access to these services would cover all constituent databases [a well-known shortcoming of these services (Gusenbauer 2019)].

While there are obvious advantages to using the Scopus dataset, there are nonetheless several known limitations including weaker coverage for the social sciences and humanities, and non-English publications (Aksnes and Sivertsen 2019). However, some have argued (Bennett 2013) that English has come to dominate academia as a 'lingua franca' leading to erosion of scholarly discourses in other languages and possibly introducing preferences for certain kinds of knowledge (Trahar et al. 2019). While quantifying this trend isn't possible within the scope of this analysis, it is likely to introduce a shift in original contributions from other languages to English over time and thus might increase the representativeness of our data. Furthermore, while Scopus does not include all possible academic outputs, the categories indexed are arguably some of the most salient kinds of academic outputs, and we would not expect that other omitted categories of outputs would introduce a specific geographic bias into our findings.

Since we are interested in collaborative relationships on a country level and across scientific disciplines, we first produce a dataset including all publications with authors in multiple countries (including papers with authors affiliated to multiple institutions in different countries), and aggregate this data to form yearly counts of co-authorship relations between countries based on the geographical location of each author's institution.[1] Specifically, if a paper or book is affiliated with institutions from more than two countries, e.g., Norway, UK, and India, three co-authorship relations will be included in this dataset: Norway–UK, Norway–India, and UK–India (which could be regionally aggregated to one within Europe co-authorship relationship and two between Europe and Asia co-authorship relationships). Subsequently, we further aggregate the data into ten time periods: 1970–1974, 1975–1979, 1980–1984, 1985–1989, 1990–1994, 1995–1999, 2000–2004, 2005–2009, 2010–2014, and 2015–2018. The final time period does not include 2019 as the Scopus database for this year is as of yet incomplete.[2]

Trends in the global production of knowledge

The production of academic publications is highly unequally distributed geographically. Figure 1a shows that the highest volume of publications is currently authored in the United States, United Kingdom, Germany, China, and India, while the lowest numbers can be found within Africa and Latin America. Looking back over the past five decades, Fig. 1b reveals that Asia is catching up with Europe and the Americas, while the growth of academic publishing is much slower for Africa and Oceania. We contrast this with the growth of co-authored publications and find that growth was much faster in Europe

[1] The resulting dataset may be found in supplementary file 'Annual international collaboration counts (Additional file 7)

[2] However as we only apply methods within each time period, this missing year does not prevent us from considering this final period. The aggregated dataset may be found in the supplementary file 'International publication counts aggregated to five year periods, unthresholded (Additional file 6).

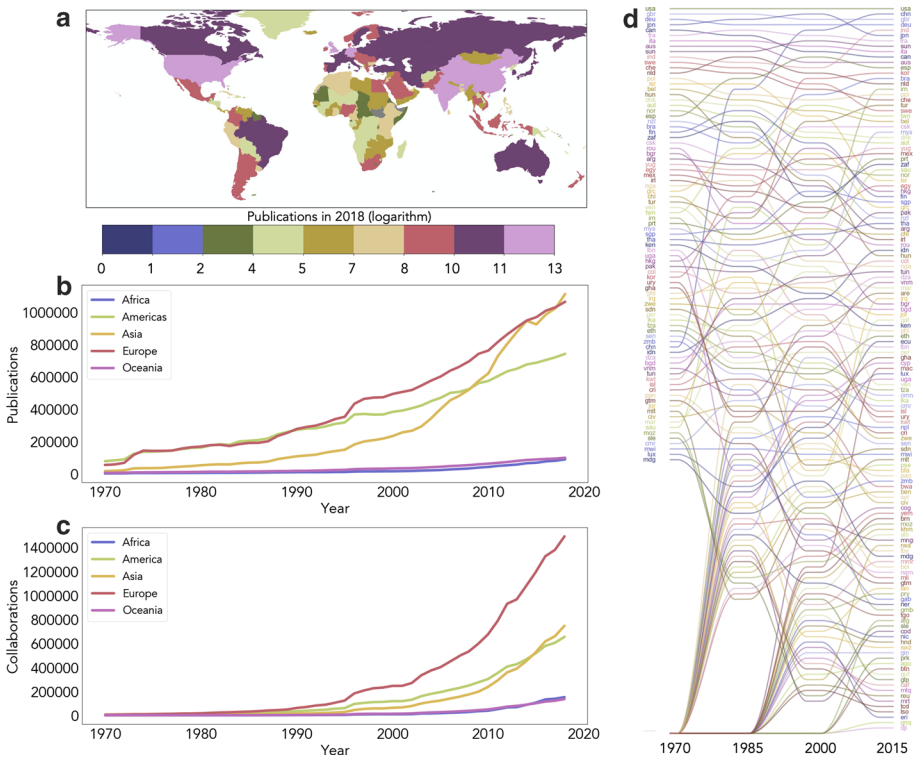

Fig. 1 **a** Maps the total number of academic publications in 2018 (log). **b**, **c** The temporal evolution of the total number of publications, and internationally coauthored publications, across geographical regions between 1970 and 2018 (Total collaboration counts may be found in supplementary file 'Total collaboration count by country, year' (Additioanl file 2), while aggregated publication counts for subfigure 1d are in supplementary file 'Total publications aggregated to five year periods (Additional file 4)). Finally, **d** displays the rank of countries in terms of the number of academic publications in 1970, 1985, 2000, and 2015. We observe some emerging economies such as China, Korea, Iran, and Malaysia rising significantly in rank towards the end of the time period

than other continents, in particular after the turn of the century. While Asia is catching up to the Americas, international collaborations are growing much slower than its share of overall publications.

Figure 1d displays the rank of countries in terms of the number of academic publications in 1970, 1985, 2000, and 2015. We observe that while countries in Europe and Asia as well as the United States and Canada are topping the list, some emerging economies such as China, Korea, Iran, and Malaysia significantly increased in rank towards the end of the time period.[3]

It is clear that national publication and co-authorship rates have been subject to significant change over the past decades. Before proceeding to disentangle co-authorship patterns over time, we desire a simple method to systematically uncover which countries are emerging as research leaders in terms of publication growth (relative to size). To do so, we follow the method described in Gargiulo et al. (2016). First, we calculate

[3] In the interest of readability, subfigure (d) omits any countries with less than 50 publications in one of the time periods, fewer than 100,000 inhabitants, and the group of small island developing nations (SIDS) except Singapore.

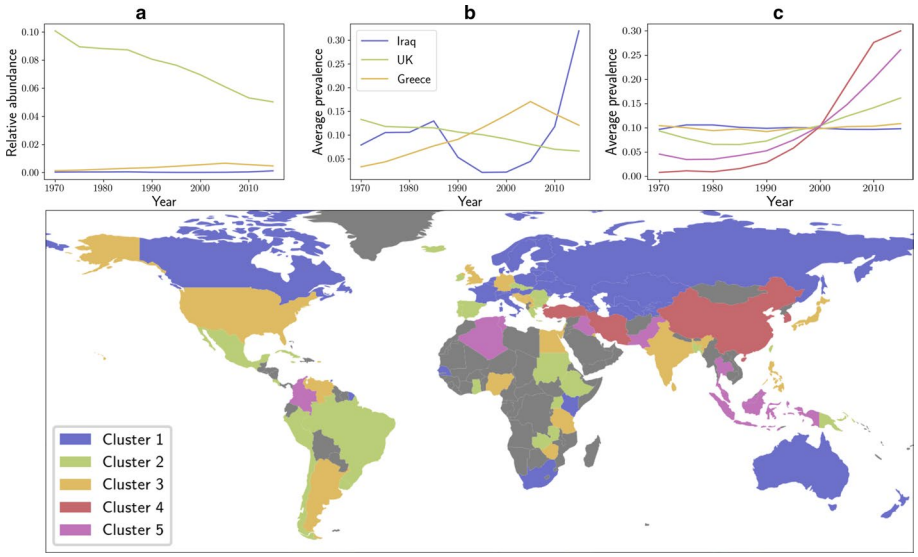

Fig. 2 **a** Displays relative abundance profiles for Iraq, the UK and Greece. While the decline in relative share of global publications from the UK is clear, it is difficult to compare with other countries due to differing overall levels of production. Thus in **b** we display the average prevalence for the same three countries (in same colours as **a**)—it is clear that while the relative share of the UK has slowly declined, for Greece it was increasing until 2010 (the subsequent decline may be due to the imposition of austerity). For Iraq we see a period of decline (during conflict) between 1990–2010, from which it is now recovering. In **c** we show the average prevalence of five groups of countries obtained from clustering their historical profiles. Two clusters display rapid growth in recent periods (red and purple), while two others display high variability (yellow and green). The final group (blue) features countries with stable or moderately declining profiles. Finally, in **d** we map these groups (in same colours as **c**), with the majority of the Global North belonging to stable or declining clusters while the Global South remains dynamic

the *relative abundance* of publications of each country within each time-step. That is, at each time-step, we compute the global share of publication activity of country i:

$$r_1(i,t) = \frac{n_i^{(t)}}{\sum_j n_j^{(t)}}, \tag{1}$$

where $n_i^{(t)}$ denotes the total number of publications produced by country i in time period t. However, as shown in Fig. 2a for the countries Iraq, the UK, and Greece, it is a poor measure to compare the historical profiles of countries with dramatically different levels of production. To overcome this, we normalise each country's relative abundance profile by its total production across the full time period to obtain a measure of *average prevalence*:

$$r_2(i,t) = \frac{r_1(i,t)}{\sum_{t'} r_1(i,t')}. \tag{2}$$

To ensure fair comparison, here we require each country to have produced more than 100 publications in each and every time period.[4] Figure 2b displays this metric for the same three countries, and now the relative trajectories of each country is clear: the UK has slowly declined in relative publication share, while Greece proportionally increased until 2010 (the subsequent decline may be due to the imposition of austerity). Iraq steeply declined in relative publication share from 1990 (possibly due to conflict after the invasion of Kuwait) and has only recovered more recently.

We then use these profiles to cluster countries with similar historical trajectories. We first calculate the Kolmogorov–Smirnov distance between each pair of country profiles [the supremum difference between the cumulative distribution of each profile (Smirnov and Smirnov 1939)], then use this distance matrix as the input for an agglomerative clustering algorithm. This algorithm works by first setting the maximum number of clusters to six (by looking at the corresponding clustering dendrogram), then finding the minimum threshold r such that the distance between any two points within each cluster is less than r, and there are at most six clusters—see e.g., Müllner (2011). These clustered profiles are displayed in Fig. 2c, where each line corresponds to the average prevalence value of a cluster—note that as such Clusters 1 and 3 seem comparable, but the variance of profiles within Cluster 3 is much greater.

In Fig. 2d we display a map of the world coloured by these clusters. Five profiles are typical: the blue cluster (Cluster 1) corresponds to countries with reasonably stable profiles over the period investigated, such as Norway and much of the Global North. The green and yellow clusters (Clusters 2 and 3) include countries with periods of relative growth and decline, such as the UK and Greece. Finally, the red and purple clusters (Clusters 4 and 5) correspond to countries that have greatly increased their publication share in recent years such as Iraq. Amongst these, every region of the Global South has countries which have considerably improved their trajectory in recent times, from Colombia in South America to China in Asia.

The international research landscape is clearly undergoing continued structural change with new leaders emerging from all corners of the globe. Here we ask, how has this shift in the geographic spread and dynamics of knowledge production shaped a reconfiguration of cross border collaboration ties?

The dynamics of international versus regional collaboration diversity

We wish to quantify how countries have changed their patterns of collaboration over time, focusing particularly on neighbouring and distant ties. One way to do this is to measure a countries' diversity of links to collaboration partners both within their own region and with countries in other regions.

In order to do this, we first calculate the Shannon entropy (see e.g., Evans et al. (2011); Kumar et al. (1986)) of the distribution of collaboration partners for each country. This provides us with a measure of the spread of collaborations for each country: values closer to one correspond to countries collaborating evenly with many countries around

[4] This thresholded dataset is available in supplementary file 'International collaboration counts aggregated to five year periods' (Additional file 5). Note Python code to reproduce this and other methods within this paper are provided in supplementary file 'Example code notebook' (Additional file 1).

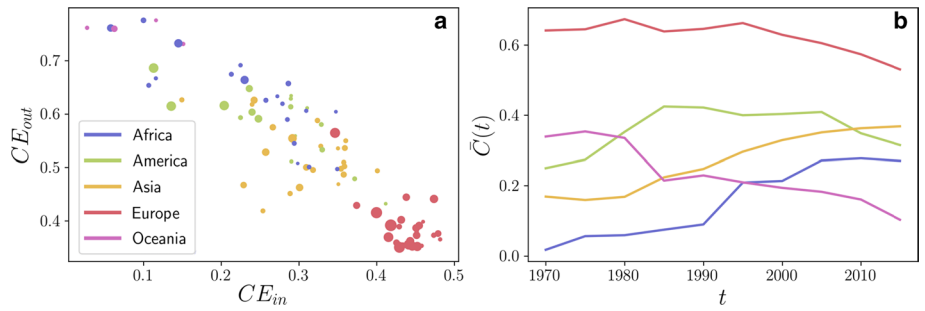

Fig. 3 **a** Displays the collaboration entropy measure within and outside of the region of each country in the final time period, with points scaled by the total number of collaborations. We observe that European countries tend to have diverse collaborations within their region relative to those further afield, while for many countries in Africa or the Americas the reverse is true. **b** Plots the average proportion of within-region entropy (as a share of total entropy) for each continent. We observe that Africa and Asia have greatly increased their focus on diverse within-region collaboration since 1970 (Data may be found in supplementary file 'Collaboration entropy data by country, year (Additional file 3))

the world, and low values correspond to a narrow focus on collaboration with few countries. To be specific, we define the *collaboration entropy* for country i as

$$CE(i,t) = -\frac{1}{\log(N-1)} \sum_{j \neq i} p_{ij}^{(t)} \log p_{ij}^{(t)}, \tag{3}$$

where N is the total number of countries in our dataset in the time period,[5] and

$$p_{ij}^{(t)} = \frac{n_{ij}^{(t)}}{\sum_{j \neq i} n_{ij}^{(t)}}, \tag{4}$$

where $n_{ij}^{(t)}$ is the number of collaborations of academics from country i with those in country j in time period t.

We are interested in investigating whether countries are collaborating more diversely within their region compared to outside their own region (continent). Hence, we decompose CE as follows:

$$CE_{in}(i,t) = -\frac{1}{\log(N-1)} \sum_{j \in J_u} p_{ij}^{(t)} \log p_{ij}^{(t)} \tag{5}$$

$$CE_{out}(i,t) = -\frac{1}{\log(N-1)} \sum_{j \in J_o} p_{ij}^{(t)} \log p_{ij}^{(t)}, \tag{6}$$

where J_u is the set of countries in the region of country i, and J_o is the set of countries outside the region of country i. Diversity increasing within regions when compared to diversity between regions suggests stronger regional clustering, and impacts a variety of network measures analysed later. In particular, if the total strength of internal

[5] Note: again only countries which produced more than 100 total publications in all time periods are included here so as to ensure comparability across time.

collaborations relative to external collaborations also increases (as verified in "Appendix 2" and shown in Fig. 6), it implies the formation of localised regional collaboration networks, or knowledge basins within which knowledge circulates more easily.

We plot CE_{in} versus CE_{out} for the final time period for all countries in Fig. 3a, where the size of the points is scaled by the total number of collaborations. We observe that European countries (shown in red) seem to collaborate more diversely with each other than with the rest of the world, while for many African countries (shown in blue) the reverse seems to be the case.

In order to assess the dynamics of inter- and intra-region collaboration diversity over time, we compute the proportion of within-region entropy (as a share of total entropy) for each country:

$$C(i,t) = \frac{CE_{in}(i,t)}{CE_{in}(i,t) + CE_{out}(i,t)}. \tag{7}$$

We plot the mean value—across countries in a region—of this quantity over time in Fig. 3b. We may observe that within-region diversity was high but has been slowly declining in Europe since 1995, suggesting the region is broadening its focus to some extent. On the other hand, within-region collaboration diversity increased significantly for the Americas and Asia from the 1980s, and for Africa after 1990. However, there appears to be a general small decline in within-region diversity (relative to out-of-region collaboration diversity) in the final two time periods for the Americas, suggesting a recent opening up of their collaboration networks.

The evolving structure of research clusters in the global collaboration network

The change in research focus, from international to regional collaboration, observed in the previous section provokes a more general investigation of how knowledge flows (as proxied by academic collaborations) may have changed over time. In particular, we ask whether these trends have translated into an overall consolidation of regional ties, creating isolated clusters or pools of knowledge production.

To uncover the complex structure of these flows, we construct a network where the nodes are countries, and the edges correspond to the number of collaborations between countries i and j at time t, $n_{ij}^{(t)}$, such that the network at this time has adjacency matrix, $A^{(t)}$, with the corresponding i, jth entries.

Prior to further analysis, to immediately visualise significant partnerships, we follow a similar procedure to that proposed by Neffke and Henning (2013) for estimating skill-overlap between industry pairs based on inter-industry job transitions. The logic behind doing so is similar to that for revealed comparative advantage (RCA, see e.g. Balassa (1965)), in that measures calculated on the network formed by raw counts are typically dominated by those locations with the highest overall production (i.e. USA, China and similar). Instead, we normalise the observed counts by the capacity of each country, measured by total collaborations, using a configuration model-like approach, apply a transformation to help account for the spread of subsequent results, and finally apply a thresholding step. The details may be found in "Appendix 3".

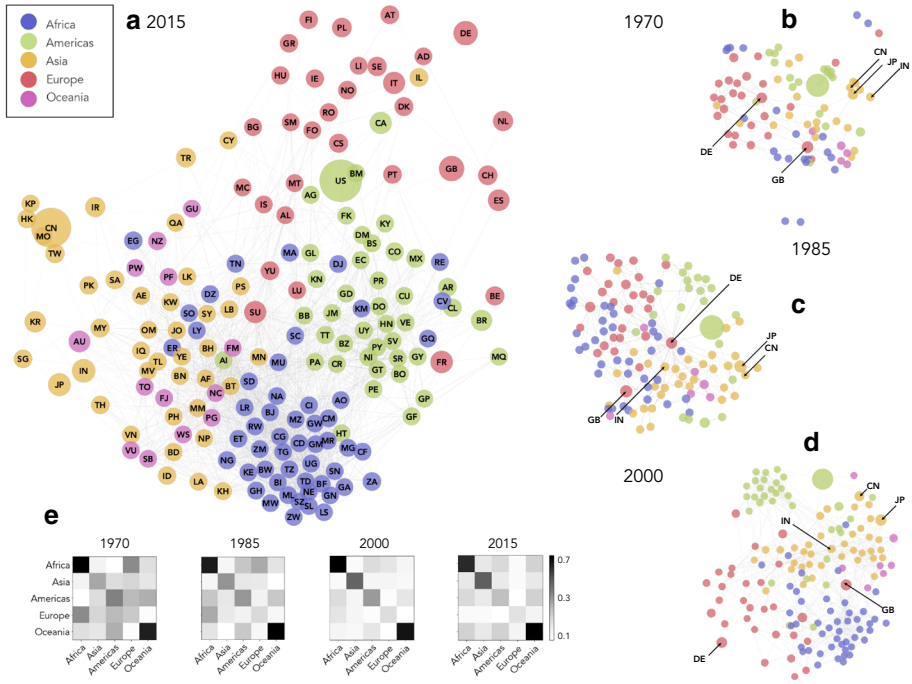

Fig. 4 **a** The global collaboration network constructed for data between 2015–2018. The algorithm ForceAtlas is used to layout the network, with edges above a 0.5 edge-weight threshold shown. Node sizes are scaled according to the total number of publications during this period, labeled with the ISO2 code corresponding to each country, and coloured according to geographical regions. **b**–**d** show analogous networks constructed for periods 1970–1974, 1985–1989, and 2000–2004. Finally, **e** displays mean edge weights between the five continents in the form of adjacency matrices. We observe a concentration of edge weights within regions, as seen by the shaded diagonal entries, particularly from 2000. Colours correspond to the same continent in each plot

In Fig. 4, we display this transformed network, with edges with strengths given by Eq. (15), for the 5-year periods commencing in 1970, 1985, 2000, and for 2015–2018, where countries which belong to the same continent have the same colour, and the size of each country is proportional to their total number of publications within that time period. The spring algorithm ForceAtlas in Gephi is used to layout each network, and edges above a 0.5 threshold are shown.

We observe that countries tend to cluster together geographically in latter time periods. This can be seen with respect to the United Kingdom and Germany: in earlier time periods they occupy fairly central 'positions', but in the latter time periods locate more closely to other European countries. On the other hand, while we note that the rise of publication volume in China and India is visible particularly over the past two decades, the positions of these countries along with Japan remain relatively close to their regional groups. In Fig. 4e, we display the mean edge weights between the five continents in the form of aggregated adjacency matrices. We observe the emergence of a defined diagonal from the year 2000, while the off-diagonals grow paler. This indicates that intra-regional collaborations have strengthened, while the inter-continental collaborations appear to decline. Once again, we observe that in the most recent time period this trend may be beginning to change, with more intercontinental partnerships emerging.

In order to explore the increasing 'regionalisation' of research collaboration, we wish to extract information from the networks about groups of countries engaged in intense research collaboration across time. Exploring such groupings is a key focus of network science, known as *community detection*. Loosely speaking, this corresponds to a partition of nodes into communities for which within-community links are significantly stronger that between-community links. It is often found to be the case that these naturally arise in the real world, e.g. in social, neurological, or indeed academic networks as under consideration here (Newman and Girvan 2004). Here, such communities reveal groupings of countries which engage in significant research collaboration—and analysis of their evolution over time enables us to extract a quantitative description of the changing global research landscape.

While a variety of methods exist (see e.g. Javed et al. (2018) for an overview), the approach we take is that of optimisation of linearised stability (Delvenne et al. 2010; Lambiotte et al. 2008, 2011). Given a partition X, this method involves computing a sum of the deviations of the network edges within each community from a weighted *configuration null model* (where edges are shuffled randomly but node strengths are preserved). Mathematically,

$$Q_{\text{conf}}(X) = \frac{1}{2m} \sum_{i,j} \left\{ A_{ij} - \gamma \frac{k_i k_j}{2m} \right\} \delta(x_i, x_j), \tag{8}$$

where

$$k_i = \sum_j A_{ij}, \quad 2m = \sum_i k_i, \tag{9}$$

are respectively the strength of node i and total edge weight of the network, x_i is the community of node i (thus $\delta(x_i, x_j) = 1$ if i and j are in same community and is zero else), and γ is a so called 'resolution' parameter. This final parameter controls the contribution of the null model to the sum, and so affects which partition will be optimal—larger values favour recovering smaller communities, and *vice versa*. Under the configuration null model, the expected strength of link between i and j is $k_i k_j / 2m$—i.e., the total strength of node j times the probability of connecting to node i. In particular, using this null model, if $\gamma = 1$ linearised stability is identical to the conventional Newman-Girvan modularity (Newman and Girvan 2004). This linearised form is also effectively identical to another method previously introduced in Reichardt and Bornholdt (2006) for modularity at different network scales. This tuning parameter is highly useful, as it allows us to avoid to some extent the resolution limit that typical modularity has been shown to face (Fortunato and Barthelemy 2007), in that it is possible to fail to detect non-trivial small communities.

The principal idea behind stability is that if we follow walkers around the network, which jump between nodes with a probability proportional to the edge weight, then over time sets of nodes where walkers spend a prolonged period suggest denser connections within such a set than to outside, i.e. they form form a community. The period of time for which we track such walkers naturally leads to the resolution parameter γ. More details on this are provided in "Appendix 4".

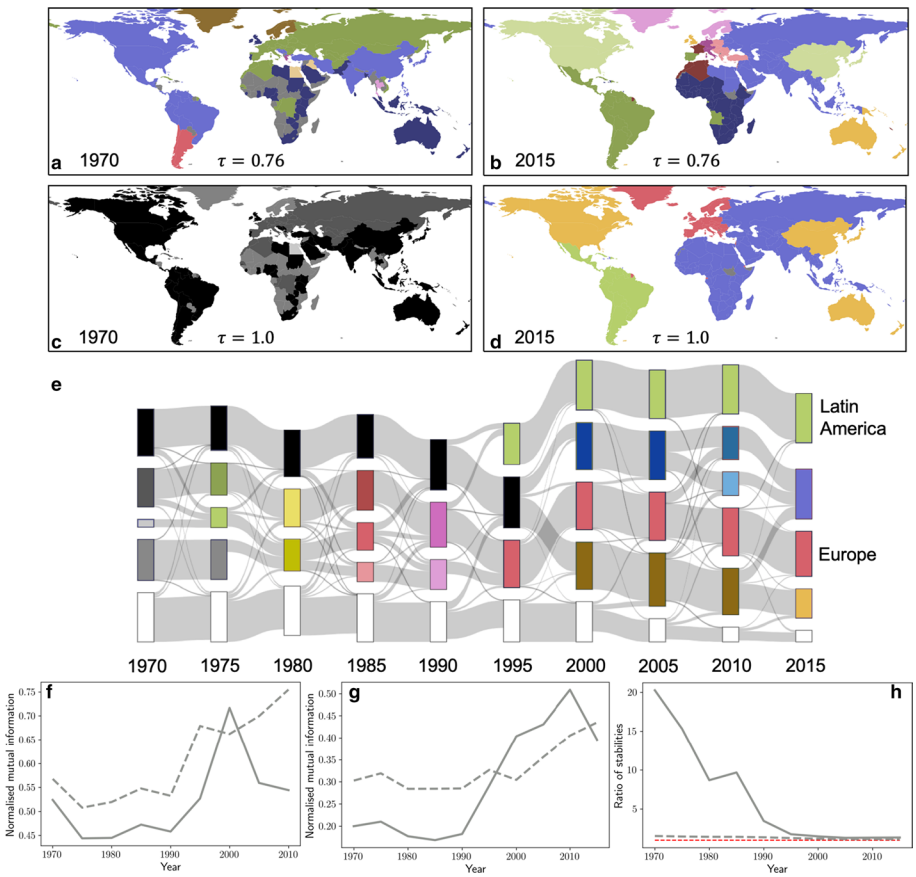

Fig. 5 In **a**, **b**, we display the communities found in the global collaboration network constructed from data between 1970–1974 and 2015–2018 respectively, using $\tau = 1/\gamma = 0.76$. We can observe for instance that in 2015, Europe in particular appears highly (sub-)regional, and much of Oceania becomes its own separate community at this finer scale. **c**, **d** Show communities in these same periods, coloured as in the Sankey plot in **f** below, but found instead with $\tau = 1.0$. We observe that while in 1970 these communities were globally distributed (in some cases along colonial lines), in 2015 they appear to overall be more regionally focused. This is supported by **e**, where communities are connected between adjacent periods if their Jaccard (similarity) index is greater than 0.6. **f** Compares the community structure of adjacent time periods using the NMI, showing generally increasing stability of the community structure over time. **g** Compares each partition to a partition where nodes are split into continents, highlighting the increasing similarity of the communities to continents over time. Finally, **h** displays the ratio of the stability of each partition to the continental partition, with the dashed red line at 1 thus corresponding to equal stability, further supporting the latter insight. For each of **f**, **g**, **h**, results for communities found with $\tau = 1.0$ are shown solid, and dashed for $\tau = 0.76$

In order to find a node partition X which maximises this function, a typical approach is to use a greedy algorithm by Blondel et al. (2008). This works by initially placing each node in its own community, then iteratively merging nodes with those adjacent to themselves if an increase in linearised stability is achieved. This process is stochastic in the sense that it may produce a slightly different optimum partition depending on the order in which nodes are 'visited'. It is efficient as only local information (nearest neighbours) to the node is necessary at each step. Recently there has been a further improvement with a similar logic, known as the Leiden algorithm (Traag et al. 2019): this appears to result in higher linearised stability with lower computational cost, and so will be used here. Through studying the variation of information (a

metric for comparing partitions) as described in "Appendix 4", we find that two resolution times $\tau = 1/\gamma$ of interest are $\tau = 1.0$ (i.e. actually conventional modularity) and $\tau = 0.76$, which provides a finer-grained view of the network.

We display the best partitions $X^{(t)}$ found from applying this optimisation process, with $\tau = 1.0$, to the network constructed for each time period in Fig. 5e. Following a similar approach to that of Pietiläinen and Diot (2012) and Fagan et al. (2018), 'flows' between two communities A and B are scaled according to the Jaccard index $J(A,B) = |A \cap B|/|A \cup B|$. We first assign each community a colour arbitrarily, then compare adjacent time periods and retain the previous colour if $J(A,B) > 0.6$. The white community corresponds to countries outside of the time period under consideration. This figure contains a wealth of information, for instance evidencing that collaboration patterns often changed more regularly in earlier, more turbulent decades, before beginning to settle from 1995 onwards. It may be seen for example that Europe consolidates as a block at this scale from 1995 onwards, shortly after the formation of the European Economic Area (EEA). We observe that in the final time period, there are four communities which roughly correspond to the regions of Europe and Latin America, North America with China, Australia and nearby countries, and the rest of the world. The community of North America *et al.* may be an artefact of the USA and China being the two major global producers, and suggests that an alternative null model could be more suitable depending on the goal of analysis—we explore the deviation from the null model further in "Appendix 5", but leave the development of such an alternative to future work.

In order to further investigate the rate of change of the modular structure over time, and the observed 'regionalisation' of research collaboration ties, we wish to quantify the similarity between each partition and its preceding partition, and between each partition and the 'continental partition' (where countries are assigned to a community based on continent). While the Jaccard index is good measure for comparing pairs of communities, to compare partitions we instead calculate the normalised mutual information (also known as the symmetric uncertainty (Witten and Frank 2002)). This is defined by

$$NMI(X,Y) = 2\frac{\sum_{i,j} r_{ij} \log\left(r_{ij}/p_i q_j\right)}{\sum_k p_k \log p_k + \sum_\ell q_\ell \log q_\ell}, \tag{10}$$

for two partitions X and Y, where n is the number of nodes, and $p_i = |X_i|/n$ (the share of nodes in community i of X), $q_i = |Y_i|/n$ (the share of nodes in community i of Y), and $r_{ij} = |X_i \cap Y_j|/n$ (the share of nodes in both community i of X, and community j of Y).

We compare the partitions obtained in adjacent time-steps through calculating the normalised mutual information: i.e. $NMI(X^{(t)}, X^{(t+1)})$, where $X^{(t)}$ is the partition obtained for time period t. In Fig. 5f, we display the values of this function over time at two different scales, with $\tau = 1.0$ shown solid, and the finer scale $\tau = 0.76$ shown dashed. We observe that after an initial period of change, recent years have seen relatively stable global research communities form at the finer scale, while at the more aggregate scale there is still some change (primarily due to splits in the large, 'rest of the world', community shown in purple). Next, we construct a new partition, C, which divides the world into five continents (communities): each country is assigned to their continent, i.e.

Africa, America, Asia, Europe, and Oceania. In order to see how similar each partition is to this continental partition, we calculate $NMI(X^{(t)}, C)$ for all t. Figure 5g confirms what we had suspected from previous figures in that there has been a clear trend towards regionalisation of research ties at both scales, particularly between 1990–2010.

As a final check, we compare the stability of each detected partition to the stability of the continental partition. Since stability is a measure of partition quality, we would expect the stability of the continental to approach that of the detected partition in latter time periods. It is important to understand the difference in quality between these partitions, particularly as there is inherent randomness to the optimisation algorithm used, and it only guarantees convergence to a local optima. In other words, the 'optimal' partition we find could in fact be only marginally better than the continental partition in early decades, even if the partitions themselves were very different as measured by NMI. We cannot compare raw values of stability across time, as it varies with respect to network size/density etc.—as such, we compute the ratio

$$Q_{rat}(X^{(t)}, C) = \frac{Q_{conf}(X^{(t)})}{Q_{conf}(C)}. \tag{11}$$

The ratio of the stability scores tells us how well the geographic (or continental) partition 'performs' as a set of communities compared to those detected by our community detection algorithm. Figure 5h confirms that, as expected, this ratio declines over time. More specifically, we observe that the continental partition was of significantly lower quality in earlier time periods, particularly for the scale with $\tau = 1.0$, suggesting this was not a good 'description' of the network structure at that time. In later periods, the ratio approaches 1 (shown dashed red) at both scales, suggesting that the continental partition is increasingly a good fit for the network structure.

Discussion

The creation and diffusion of knowledge between nations is crucial for the advancement of skills and capabilities, critical drivers of economic development. Patterns of knowledge diffusion via research collaboration on a global level, however, remain poorly understood. We address this research gap through analysing a worldwide dataset of international scientific publications spanning all major disciplines over five decades. We find that collaboration ties appear to have become more localised since 2000, with researchers prioritising regional co-authorship relative to more distant ties. We corroborate this insight via an analysis of the evolving modular structure of the global collaboration network, finding a recent stabilisation of research clusters along increasingly regional lines.

These findings were unexpected given the generally accepted wisdom on the onward march of globalisation, and thus have a number of significant implications. On one hand, this could be a positive signal: research expertise is growing in many previously underequipped nations and regions, and hence scholars no longer have to look as far afield as they once did. Regional research efforts may be driven by resident researchers focusing their efforts on addressing particular economic, social, and political concerns within the region. Specific research programmes have been introduced to strengthen scientific

collaboration within regions such as the Horizon 2020 (soon to give way to Horizon Europe) initiative in Europe. Further, regional research and development programmes increasingly make use of the 'smart specialisation' model in research, whereby countries with well-defined domains of specialisation (e.g., in research and innovation) are seen as more likely to produce research excellency in specific areas. These countries are then chosen as sites for related regional research programmes and institutes, with the aim of anchoring and nurturing these localised sites of expertise. This model was originally developed by the European Union in order to address a transatlantic gap in R&D but has since been adopted by many regions and countries (Gómez Prieto et al. 2019)—a trend that our research findings would seem to support and perhaps a driving force behind some of the patterns we have identified. On the other hand, such a retrenchment may be worrying, given what we know about the importance of capability building and knowledge diffusion through 'on-the-job' learned experience, leading to an uneven distribution of capabilities across regions. Indeed, the role of donors in strengthening local research capacity through international collaborations in many lower and middle-income economies has been deemed crucial, as these countries tend to lack research capacity and face problems translating research into impact. In this respect, it seems more important than ever that large research funders such as those within the EU and the US support international collaboration on a scale that far outstrips current levels. While funders, such as the US Agency for International Development's (USAID) Partnerships for Enhanced Engagement in Research or the UK's Newton Fund already have dedicated mechanisms to support North-South research partnerships, this could mean expansion of National Science Foundation (NSF) programmes to allow non-US research leads, or a U-turn in the recent decrease in funding allocated to the much-feted Fulbright programme (which supported two of the authors of this paper to spend time in the US). One bright spot is the recent growth of development-oriented research funding in the UK, the Global Challenges Research Fund (which supported this work), that not only supports equitable UK-developing nation collaboration but mandates it. It is only with large scale investment in such programmes that international research collaboration will continue to play a vital role in global capability building.

While previous work comparing data sources on academic publishing highlight the comparative strength of the Scopus dataset, we are aware that there are limitations to this data given our interest in comparison between countries. The database's coverage is thought to be weaker for the social sciences and humanities, and for literatures in other languages than English (Aksnes and Sivertsen 2019). Furthermore, we cannot ascertain that the indexing of work from publishers located in countries where academia is less well resourced is as complete as for countries with more established academes. Additionally, our dataset includes journal articles, books, and conference proceedings, but no other types of academic outputs. Ideally one would complement this analysis with additional material from academic search engines such as Google Scholar, which contains up to four times as many documents as Scopus. However, due to well-known issues such as document duplication, false citations, and unstable search indexing, this would require a major investment in data cleaning and processing.

There are a few clear avenues for future work. Firstly, much remains to be understood about the nature and evolution of global collaboration networks, and the roles

of individual actors. For example, it would be interesting to investigate whether we can identify global hegemons, countries playing the role of gatekeeper between lesser regional partners and the rest of the world. Similarly, our work suggests that colonial links and geopolitical alliances shaped, for a time, regional basins of knowledge. Has this transition from historical blocs to regional clusters positively (or negatively) impacted the research capacity of less developed nations? In other words, who have been the winners and losers from this shift in network structure?

Finally, there is ample scope and reason to further investigate the structure of global research ties and knowledge diffusion beyond inter-country links. First and foremost, research quality and disciplinary focus is often highly institution—rather than country—specific. Furthermore, funding programmes often target specific fields and institutions. For these reasons amongst others, it would be fruitful to dis-aggregate the global collaboration network by institution and field. Perhaps collaborations in certain disciplines are heavily demarcated along regional lines, while others flourish under international collaboration. Perhaps top-tier institutions maintain international links, while second-tier institutions focus more on regional ties. Additionally, there are a large number of possible metrics one might compare to co-authorship ties, including researcher mobility patterns. I.e, is the recent regional retrenchment in collaboration patterns also observed in researcher mobility patterns?

Appendix 1: Data processing

As our data covers over five decades, our analysis spans such geopolitical changes as the reunification of East Germany and West Germany in 1990, the collapse of the Soviet Union in 1991, the breakup of Yugoslavia from 1991 to 1992 and the dissolution of Czechoslovakia to the Czech Republic and Slovakia in 1993 as well as smaller transitions such as post-colonial transitions during the 1970s and early 1980s, the independence of Bangladesh from Pakistan in 1971, Palestinian declaration of independence in 1988, the independence of Namibia from South Africa in 1990, unification of North and South Yemen in 1990, and the independence of Eritrea from Ethiopia in 1993, East-Timor from Indonesia in 1999, and South Sudan from Sudan in 2011. Since we are interested in observing international collaboration across the network of countries over time, some of our methods require the network to remain relatively consistent over time and in order to achieve this, we adjust for the larger geopolitical transitions by keeping the Soviet Union, Yugoslavia, Germany, and Czechoslovakia as single nodes throughout the analysis. We consider this operationalisation justified by the fact that beyond fulfilling our methodological requirements, the relationship of these larger regions to the rest of the global academia follows rather constant trends (beyond initial disruptions following the political changes), which gives us confidence that the academic institutions continue working in a relatively similar fashion before and after the changes.

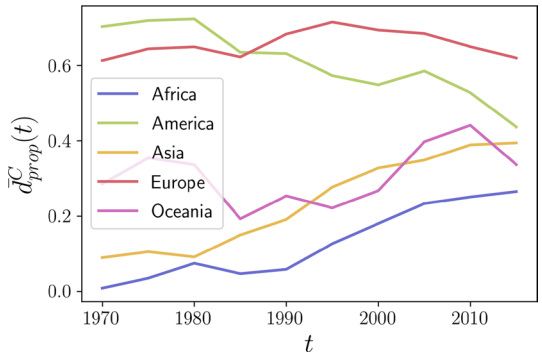

Fig. 6 In this figure, we display the average1 proportion of within-region strength (as a share of total strength) for each continent. We observe that Africa and Asia have greatly increased their focus on within-region collaboration since 1970, while countries in America have on average broadened their collaboration profiles

Appendix 2: Within region strength

We may define the total collaboration strength within (resp. outside) the region for each country by

$$d^C_{in}(i,t) = \sum_{j \in J_u} n^{(t)}_{i,j}, \qquad d^C_{out}(i,t) = \sum_{j \in J_o} n^{(t)}_{i,j}, \tag{12}$$

then as previously performed with entropy define the proportion of total strength within the region by

$$d^C_{prop}(i,t) = \frac{d^C_{in}(i,t)}{d^C_{in}(i,t) + d^C_{out}(i,t)}. \tag{13}$$

Now taking the average of this measure over each continent, we display results in Fig. 6. We see a similar picture to those for our entropy measure, where Africa and Asia in particular have greatly increased their focus on within-region collaboration.

Appendix 3: Transformation of collaboration counts for visualisation in Fig. 4

For better highlighting significant partnerships when visualising the international collaboration network in Fig. 4, we apply a transformation to the raw counts of collaborations. Specifically, the *collaboration significance* may be defined as

$$s^{(t)}_{i,j} = \frac{n^{(t)}_{i,j} / \sum_{k \neq i} n^{(t)}_{i,k}}{\sum_{\ell \neq j} n^{(t)}_{\ell,j} / \sum_{m,n} n^{(t)}_{m,n}}. \tag{14}$$

This corresponds to the ratio of the actual number of collaborations between country pairs to those expected under a configuration model [see e.g. Newman and Girvan

(2004)]—values larger than one correspond to more collaborations occurring than expected at random. As this measure is highly skewed, we re-scale it so that values lie between −1 and 1:

$$\hat{p}_{i,j}^{(t)} = \frac{s_{i,j}^{(t)} - 1}{s_{i,j}^{(t)} + 1}. \tag{15}$$

We then finally set $\hat{p}_{i,j}^{(t)} = 0$ if $\hat{p}_{i,j}^{(t)} < 0$, i.e., those pairs for which fewer collaborations occurred than would be expected, and visualise the resulting network in Fig. 4.

Appendix 4: Using stability in community detection

The key step is the relation of modularity to the stability of communities under Laplacian dynamics Lambiotte et al. (2008), as defined by the formula

$$\dot{p}_i = \sum_j L_{ij} p_j. \tag{16}$$

Here the appropriate matrix L to relate to modularity depends on the type of network under consideration. As our network is undirected, we may use the normalised Laplacian matrix

$$L_{ij} = \frac{A_{ij}}{k_j} - \delta_{ij}, \tag{17}$$

and the stability of the partition is then defined to be Lambiotte et al. (2008)

$$R(\tau) = \sum_{i,j} \left[(e^{\tau L})_{ij} p_j^* - p_i^* p_j^* \right] \delta(g_i, g_j), \tag{18}$$

where $p_i^* = k_i/2m$ is the stationary solution to (16). Expanding the exponential matrix to first order in τ,

$$(e^{\tau L})_{ij} = \delta_{ij} + \tau L_{ij} + \mathcal{O}(\tau^2), \tag{19}$$

thus naturally leads to the inclusion of a *resolution parameter* $\gamma = 1/\tau$ in the modularity equation, as ignoring the constant term and dividing by τ we then have

$$R_{lin}(\tau) = \frac{1}{2m} \sum_{i,j} \left\{ A_{ij} - \gamma \frac{k_i k_j}{2m} \right\} \delta(g_i, g_j), \tag{20}$$

an answer to the resolution problem previously mentioned.

In matrix form, Eq. (20) may be written as Delvenne et al. (2013)

$$R_{lin}(\tau) = \mathrm{Tr}\left[H^\top \left(\frac{A}{2m} - \gamma \pi^\top \pi \right) H \right], \tag{21}$$

where H is the $N \times c$ matrix with $H_{ij} = 1$ if country i is in community j and zero else, and Tr corresponds to the trace operator. For a given Markov time τ (or equivalently resolution parameter γ), we seek the partition that maximises this function. As this

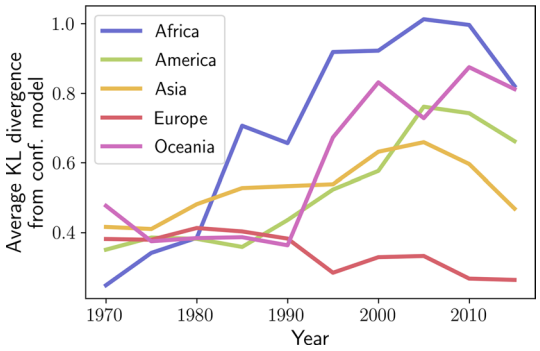

Fig. 7 Average KL divergence (as defined in Eq. (23)) for each continent between the observed collaboration counts and those expected by the configuration model. Over time, the configuration model has become a better description for Europe, while other continents have further differed

optimisation problem is NP-hard, as stated in the main text, we use a greedy method from Traag et al. (2019).

Suitable resolution parameter ranges are typically discovered through calculation of the variation of information. For two partitions X and Y of a set A into disjoint subsets, $X = \{X_1, X_2, \ldots, X_k\}$ and $Y = \{Y_1, Y_2, \ldots, Y_l\}$, this measure is defined as follows. Let $n = \sum_i |X_i| = \sum_j |Y_j| = |A|$, $p_i = |X_i|/n$, $q_j = |Y_j|/n$, $r_{ij} = |X_i \cap Y_j|/n$. Then the normalised variation of information between the two partitions is:

$$\mathrm{VI}(X, Y) = -\frac{1}{\log N} \sum_{i,j} r_{ij} \left[\log(r_{ij}/p_i) + \log(r_{ij}/q_j)\right]. \tag{22}$$

As there is some stochasticity in the output of the optimisation process, we run the method many times for each resolution, and collect the resulting partitions. If the average variation of information between each pair of such partitions is small, then this suggests that this parameter choice provides somewhat more robust communities.

Appendix 5: How good a null model is the configuration model?

As suggested in the main text, the grouping of major producers—specifically the USA and China—together in a single community in recent years, despite not necessarily having similar partners other than each other, may imply that the configuration null model used is not the optimal choice of null model for uncovering significant partnerships. To further investigate this, we may study how closely the observed distribution of collaborations for each country lies to that predicted by the configuration model. One way of assessing the proximity of two such probability distributions is the Kullback–Liebler (KL) divergence (see e.g. Cover and Thomas (1991)). For two discrete probability distributions P and Q that have the same support, χ say, to find the information gain from using P (which can be the real observed data) in place of our model, Q, we calculate

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \chi} P(x) \log \left(\frac{P(x)}{Q(x)} \right). \tag{23}$$

In our situation, for country i in year t, we compare the empirical distribution of collaborations with all other countries, i.e. $P_i^{(t)}(j) = p_{ij}^{(t)} = n_{ij}^{(t)} / \sum_k n_{ik}^{(t)}$, to that predicted by the configuration null model, where the expected number of links between countries i and j is $k_i^{(t)} k_j^{(t)} / 2m^{(t)}$ (followed by analogous normalisation for each country to form a probability distribution). To ensure the support of the empirical distribution and the configuration model match, we perform additive smoothing (see e.g. Schütze et al. (2008)), i.e. we add one to the count of collaborations between each pair of countries prior to normalising.

In Fig. 7 we display the average of the resulting KL divergence for each continent across time. We observe that while in 1970 the configuration model was a comparably good choice for all continents, there has since been a large deviation. Europe has on average increasingly collaborated as the model would predict, suggesting that preferential attachment is a major mechanism driving international collaborations at the aggregate level, while other continents—particularly Africa—have collaborated in more and more 'surprising' patterns relative to the model. This decreasingly suitable description of some regions true collaboration implies that an alternative could be used to further highlight groups of closely partnered countries, though in doing so note we would lose the dynamical interpretation of communities, and the associated stability function. We leave the development of a suitable alternative model to future work.

Abbreviations

ASE: Academic search engine; IHERD: Organization for economic co-operation and development project on innovation, higher education and research and for development; NSF: National science foundation; OECD: Organization for economic co-operation and development; RD: Research and development; UNESCO: United Nations educational, scientific and cultural organization.

Supplementary Information

The online version supplementary material available at https://doi.org/10.1007/s41109-021-00371-w.

Additional file 1. Example code notebook.

Additional file 2. Total collaboration count by country, year.

Additional file 3. Collaboration entropy data by country, year.

Additional file 4. Total publications aggregated to five year periods.

Additional file 5. International collaboration counts aggregated to five year periods.

Additional file 6. International collaboration counts aggregated to five year periods, unthresholded.

Additional file 7. Annual international collaboration counts.

Acknowledgements

This work uses Scopus data provided by Elsevier through the International Center for the Study of Research (ICSR) Lab.

Authors' contributions

NOC, JF and SO designed the study and the methodology, JF and SO processed and analysed the results.

Funding

This publication is based on work partially supported by the EPSRC Centre For Doctoral Training in Industrially Focused Mathematical Modelling (EP/L015803/1) in collaboration with Elsevier (John Fitzgerald). Sanna Ojanperä is currently funded by The Alan Turing Institute doctoral grant TU/C/000020. This article was completed with support from the PEAK Urban programme, funded by UKRI's Global Challenge Research Fund, Grant Ref. ES/P011055/1 (Neave O'Clery).

Availability of data and materials
The data that support the findings of this study are available from Elsevier (Scopus) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Elsevier.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Mathematical Institute, Oxford University, Andrew Wiles Building, Radcliffe Observatory Quarter, Oxford OX2 6GG, UK. [2]Oxford Internet Institute, University of Oxford, Oxford, ASD, Oxford, UK. [3]The Bartlett Centre for Advanced Spatial Analysis, University College London, London, UK. [4]The Alan Turing Institute, London, UK.

Received: 3 March 2020 Accepted: 18 March 2021
Published online: 31 May 2021

References

Aksnes DW, Sivertsen G (Feb 2019) A criteria-based assessment of the coverage of scopus and web of science. J Data Inf Sci 4(1):1–21 https://doi.org/10.2478/jdis-2019-0001. https://content.sciendo.com/configurable/contentpage/journals$002fjdis$002f4$002f1$002farticle-p1.xml
Aqib JM, Shahzad YM, Latif S, Qadir J, Baig A (2018) Community detection in networks: a multidisciplinary review. J Netw Comput Appl 108:87–111
Balassa B (1965) Trade liberalisation and "revealed" comparative advantage. Manch Sch 33(2):99–123
Barabási AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. Physica A 311(3–4):590–614. https://doi.org/10.1016/S0378-4371(02)00736-7
Bennett K (2013) English as a Lingua Franca in Academia. Interpreter Transl Trainer 7(2):169–193. https://doi.org/10.1080/13556509.2013.10798850
Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008
Burton RE, Kebler RW (1960) The half-life of some scientific and technical literatures. Am Doc 11(1):18–22
Chinchilla-Rodríguez Z, Benavent-Pérez M, de Moya-Anegón F, Miguel S (2012) International collaboration in Medical Research in Latin America and the Caribbean (2003–2007). J Am Soc Inf Sci Technol 63(11):2223–2238. https://doi.org/10.1002/asi.22669
Cover TM, Thomas JA (1991) Information theory and statistics. Elem Inf Theory 1(1):279–335
de Price DJS (1963) Little science, big science. Number 1962 in George B. Pegram lecture series. Columbia University Press, New York
de Rassenfosse G, Seliger F (Feb 2020) Sources of knowledge flow between developed and developing nations. Sci Public Policy 47(1):16–30. ISSN 0302-3427. https://doi.org/10.1093/scipol/scz042. https://academic.oup.com/spp/article/47/1/16/5580327
Delvenne JC, Yaliraki SN, Barahona M (2010) Stability of graph communities across time scales. Proc Natl Acad Sci 107(29):12755–12760
Delvenne J-C, Schaub MT, Yaliraki SN, Barahona M (2013) The stability of a graph partition: a dynamics-based framework for community detection. In: Dynamics on and of complex networks, vol 2, pp 221–242. Springer
Ding WW, Levin SG, Stephan PE, Winkler AE (2010) The impact of information technology on academic scientists' productivity and collaboration patterns. Manag Sci 56(9):1439–1461. https://doi.org/10.1287/mnsc.1100.1195
Doria Arrieta OA, Pammolli F, Petersen AM (2017) Quantifying the negative impact of brain drain on the integration of European science. Sci Adv 3(4):e1602232. https://doi.org/10.1126/sciadv.1602232
Elsevier (2020) Scopus | the largest database of peer-reviewed literature | Elsevier. https://www.elsevier.com/en-gb/solutions/scopus
Evans TS, Lambiotte R, Panzarasa P (2011) Community structure and patterns of scientific collaboration in business and management. Scientometrics 89(1):381–396
Fagan J, Eddens KS, Dolly J, Vanderford NL, Weiss H, Levens JS (2018) Assessing research collaboration through co-authorship network analysis. J Res Admin 49(1):76–99
Fatt C, Ujum E, Ratnavelu K (2010) The structure of collaboration in the Journal of Finance. Scientometrics 85(3):849–860. https://doi.org/10.1007/s11192-010-0254-0
Fortunato S, Barthelemy M (2007) Resolution limit in community detection. Proc Natl Acad Sci 104(1):36–41
Frenken K, Boschma RA (2007) A theoretical framework for evolutionary economic geography: industrial dynamics and urban growth as a branching process. J Econ Geogr 7:635–649
Frenken K, Hoekman J, Kok S, Ponds R, van Oort F, van Vliet J (2009) Death of distance in science? A gravity approach to research collaboration. In: Pyka A, Scharnhorst A (eds) Innovation networks: new approaches in modelling and analyzing, understanding complex systems. Springer, Berlin, pp 43–57. https://doi.org/10.1007/978-3-540-92267-4_3
Frenken K, Hölzl W, de Vor F (2005) The citation impact of research collaborations: the case of European biotechnology and applied microbiology (1988–2002). J Eng Technol Manag 22(1–2):9–30. https://doi.org/10.1016/j.jengtecman.2004.11.002
Garfield E, Sher IH (1963) New factors in the evaluation of scientific literature through citation indexing. Am Doc 14(3):195–201. https://doi.org/10.1002/asi.5090140304

Gargiulo F, Caen A, Lambiotte R, Carletti T (2016) The classical origin of modern mathematics. EPJ Data Sci 5(1):26
Gazni A, Didegah F (2011) Investigating different types of research collaboration and citation impact: a case study of Harvard University's publications. Scientometrics 87(2):251–265. https://doi.org/10.1007/s11192-011-0343-8
Gazni A, Sugimoto CR, Didegah F (2012) Mapping world scientific collaboration: authors, institutions, and countries. J Am Soc Inform Sci Technol 63(2):323–335. https://doi.org/10.1002/asi.21688
Geuna A (2015) Global mobility of research scientists: the economics of who goes where and why. Academic Press. Google-Books-ID: I7rjAwAAQBAJ
Glänzel W, Schubert A (2005) Analysing scientific networks through co-authorship. In: Moed HF, Glänzel W, Schmoch U (eds) Handbook of quantitative science and technology research: the use of publication and patent statistics in studies of S&T systems. Springer, Dordrecht, pp 257–276. https://doi.org/10.1007/1-4020-2755-9_12
Goh K-I, Oh E, Kahng B, Kim D (2003) Betweenness centrality correlation in social networks. Phys Rev E Stat Nonlinear Soft Matter Phys 67(1 Pt 2):017101. https://doi.org/10.1103/PhysRevE.67.017101
Goh K-I, Eulsik O, Jeong H, Kahng B, Kim D (2002) Classification of scale-free networks. Proc Natl Acad Sci USA 99(20):12583–12588. https://doi.org/10.1073/pnas.202301299. http://www.pnas.org/content/99/20/12583.abstract
Gusenbauer M (2019) Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics 118(1):177–214. https://doi.org/10.1007/s11192-018-2958-5
Gusenbauer M, Haddaway NR (2019) Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res Synth Methods. https://doi.org/10.1002/jrsm.1378
Gómez PJ, Demblans A, Palazuelos MM (2019) Smart specialisation in the world, an EU policy approach helping to discover innovation globally. Technical report, Publications Office of the European Union, Luxembourg. https://s3platform.jrc.ec.europa.eu/-/smart-specialisation-in-the-world-an-eu-policy-approach-helping-to-discover-innovation-globally?inheritRedirect=true
Harzing A-W, Alakangas S (2016) Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. Scientometrics 106(2):787–804. https://doi.org/10.1007/s11192-015-1798-9
Hausmann R, Hidalgo CA (2011) The network structure of economic output. J Econ Growth 16(4):309–342
Heinze T, Kuhlmann S (2008) Across institutional boundaries? Research collaboration in German public sector nanoscience. Res Policy 37(5):888–899. https://doi.org/10.1016/j.respol.2008.01.009
Hidalgo CA, Hausmann R (2009) The building blocks of economic complexity. Proc Natl Acad Sci 106(26):10570–10575
Hidalgo CA, Klinger B, Barabási A-L, Hausmann R (July 2007) The product space conditions the development of nations. Science 317(5837):482–487. ISSN 0036-8075, 1095–9203. https://doi.org/10.1126/science.1144581. https://science.sciencemag.org/content/317/5837/482. American Association for the Advancement of Science Section: Research Article
Hou H, Kretschmer H, Liu Z (2008) The structure of scientific collaboration networks in Scientometrics. Scientometrics 75(2):189–202. https://doi.org/10.1007/s11192-007-1771-3
Jackson D (2011) What is an innovation ecosystem. Technical report 1(2), National Science Foundation
Jaffe AB, Trajtenberg M (April 1998) International knowledge flows: evidence from patent citations. Working paper 6507, National Bureau of Economic Research. http://www.nber.org/papers/w6507
Jones BF, Wuchty S, Uzzi B (2008) Multi-university research teams: shifting impact, geography, and stratification in science. Science (New York) 322(5905):1259–1262. https://doi.org/10.1126/science.1158357
Kumar S (2015) Co-authorship networks: a review of the literature. Aslib J Inf Manag 67(1):55–73. https://doi.org/10.1108/AJIM-09-2014-0116
Kumar U, Kumar V, Kapur JN (1986) Normalized measures of entropy. Int J Gener Syst 12(1):55–69
Lambiotte R, Delvenne J-C, Barahona M (2008) Laplacian dynamics and multiscale modular structure in networks. Preprint arXiv:0812.1770
Lambiotte R, Sinatra R, Delvenne J-C, Evans TS, Barahona M, Latora V (2011) Flow graphs: interweaving dynamics and structure. Phys Rev E 84(1):017102
Lee S, Bozeman B (2005) The impact of research collaboration on scientific productivity. Soc Stud Sci 35(5):673–702. https://doi.org/10.1177/0306312705052359
Leydesdorff L, Wagner C, Woo PH, Adams J (Jan 2013) International collaboration in science: the global map and the network. arXiv:1301.0801 [cs]
Melin G, Persson O (1996) Studying research collaboration using co-authorships. Scientometrics 36(3):363–377. https://doi.org/10.1007/BF02129600
Meyer M, Bhattacharya S (2004) Commonalities and differences between scholarly and technical collaboration. Scientometrics 61(3):443–456. https://doi.org/10.1023/B:SCIE.0000045120.04489.80
Mingers J, Meyer M (2017) Normalizing Google Scholar data for use in research evaluation. Scientometrics 112(2):1111–1121. https://doi.org/10.1007/s11192-017-2415-x
Mirton KM, Heart FE (1962) National Science Foundation (US), Massachusetts Institute of Technology, Libraries, and Lincoln Laboratory. Analysis of bibliographic sources in the physical review (vol. 77, 1950 to vol. 112, 1958). Massachusetts Institute of Technology, Cambridge, Mass., OCLC: 339864
Müllner D (2011) Modern hierarchical, agglomerative clustering algorithms. Preprint arXiv:1109.2378
Narin F, Stevens K, Whitlow E (1991) Scientific co-operation in Europe and the citation of multinationally authored papers. Scientometrics 21(3):313–323. https://doi.org/10.1007/BF02093973
Neffke F, Henning M (2013) Skill relatedness and firm diversification. Strateg Manag J 34(3):297–316
Newman M (2001a) Scientific collaboration networks. I. Network construction and fundamental results. Phys Rev E Stat Nonlinear Soft Matter Phys 64(1 Pt 2):016131
Newman M (2001b) Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Phys Rev E 64(1):7. https://doi.org/10.1103/PhysRevE.64.016132
Newman M (2001c) The structure of scientific collaboration networks. Proc Natl Acad Sci 98(2):404–409. https://doi.org/10.1073/pnas.98.2.404. https://www.pnas.org/content/98/2/404

Newman M (2004) Coauthorship networks and patterns of scientific collaboration. Proc Natl Acad Sci 101(suppl 1):5200–5205
Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113
Ojanperä S, Graham M, Straumann R, De Sabbata S, Zook M (March 2017) Engagement in the knowledge economy: regional patterns of content creation with a focus on sub-Saharan Africa. Inf Technol Int Dev 13:33–51. ISSN 1544-7529. http://itidjournal.org/index.php/itid/article/view/1479
Osgood CE, Xhignesse LV (1963) Characteristics of bibliographical coverage in psychological journals published in 1950 and 1960. Institute of Communicatios Research, University of Illinois, Champaign
O'Clery N, Curiel RP, Lora E (2019a) Commuting times and the mobilisation of skills in emergent cities. Appl Netw Sci 4(1):118
O'Clery N, Flaherty E, Kinsella S (2019b) Modular structure in labour networks reveals skill basins. arXiv:1909.03379 [econ, q-fin]
Pietilänen A-K, Diot C (2012) Dissemination in opportunistic social networks: the role of temporal communities. In: Proceedings of the 13th ACM international symposium on mobile ad hoc networking and computing, pp 165–174
Powell WW, Snellman K (2004) The knowledge economy. Annu Rev Sociol 30(1):199–220. https://doi.org/10.1146/annurev.soc.29.010202.100037
Price DJ (1965) Networks of scientific papers. Science (New York) 149(3683):510–515. https://doi.org/10.1126/science.149.3683.510
Programme on Innovation (2012) Higher Education and Research and for Development (IHERD). Research universities, Networking the Knowledge Economy
Racherla P, Clark H (2010) A social network perspective of tourism research collaborations. Ann Tour Res 37(4):1012–1034. https://doi.org/10.1016/j.annals.2010.03.008
Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Phys Rev E 74(1):016110
Ribeiro LC, Rapini MS, Silva LA, Albuquerque EM (2018) Growth patterns of the network of international collaboration in science. Scientometrics 114(1):159–179. https://doi.org/10.1007/s11192-017-2573-x
Rodriguez MA, Pepe A (2008) On the relationship between the structural and socioacademic communities of a coauthorship network. J Inf 2(3):195–201. https://doi.org/10.1016/j.joi.2008.04.002
Sanna O, Mark G, Matthew Z (2019) The digital knowledge economy index: mapping content production: the journal of development studies: vol 55, no 12. J Dev Stud 55(12):2626–2643. https://doi.org/10.1080/00220388.2018.1554208
Santos JAC, Santos MC (2016) Co-authorship networks: collaborative research structures at the journal level. Tour Manag Stud 12(1):05–13.https://doi.org/10.18089/tms.2016.12101. http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S2182-84582016000100001&lng=en&tlng=en
Scherngell T (ed) (2013) The geography of networks and R&D collaborations. Adv Spatial Sci. https://doi.org/10.1007/978-3-319-02699-2. https://www.springer.com/gp/book/9783319026985
Schütze H, Manning CD, Raghavan P (2008) Introduction to information retrieval, vol 39. Cambridge University Press, Cambridge
Smirnov N, Smirnov NV (1939) On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bull Math Univ Moscou 2:3–14
Sooryamoorthy R (2017) Do types of collaboration change citation? A scientometric analysis of social science publications in South Africa. Scientometrics 111(1):379–400. https://doi.org/10.1007/s11192-017-2265-6
Sylvan Katz J, Martin BR (1997) What is research collaboration? Res Policy 26(1):1–18. https://doi.org/10.1016/S0048-7333(96)00917-1
Traag VA, Waltman L, van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9(1):1–12
Trahar S, Juntrasook A, Burford J, von Kotze A, Wildemeersch D (2019) Hovering on the periphery? Decolonising writing for academic journals. Compare J Comp Int Educ 49(1):149–167. https://doi.org/10.1080/03057925.2018.1545817
Tukey JW (1962) Keeping research in contact with the literature: citation indices and beyond. J Chem Doc 2(1):34–37. https://doi.org/10.1021/c160004a011
UNESCO (2015) UNESCO science report: towards 2030. Technical report, Imprimerie Centrale, Luxembourg. https://unesdoc.unesco.org/ark:/48223/pf0000235406
Ubfal D, Maffioli A (2011) The impact of funding on research collaboration: evidence from a developing country. Res Policy 40(9):1269–1279. https://doi.org/10.1016/j.respol.2011.05.023. http://www.sciencedirect.com/science/article/pii/S0048733311001041
Wagner CS, Leydesdorff L (2005) Mapping the network of global science: comparing international co-authorships from 1990 to 2000. Int J Technol Global 1(2):185–208
Wagner-Döbler R (2001) Continuity and discontinuity of collaboration behaviour since 1800—from a bibliometric point of view. Scientometrics 52(3):503–517. https://doi.org/10.1023/A:1014208219788
Wagner CS, Leydesdorff L (2005) Network structure, self-organization, and the growth of international collaboration in science. Res Policy 34(10):1608–1618. https://doi.org/10.1016/j.respol.2005.08.002. http://www.sciencedirect.com/science/article/pii/S0048733305001745
Witten IH, Frank E (2002) Data mining: practical machine learning tools and techniques with Java implementations. ACM SIGMOD Rec 31(1):76–77
Wuchty S, Jones BF, Uzzi B (May 2007) The increasing dominance of teams in production of knowledge. Science 316(5827):1036–1039 https://doi.org/10.1126/science.1136099. https://science.sciencemag.org/content/316/5827/1036
Yan E, Ding Y (2009) Applying centrality measures to impact analysis: a coauthorship network analysis. J Am Soc Inf Sci Technol 60(10):2107–2118. https://doi.org/10.1002/asi.21128. http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=44295774&site=ehost-live&authtype=ip,uid

Ye Q, Song H, Li T (2012) Cross-institutional collaboration networks in tourism and hospitality research. Tour Manag Perspect 2–3:55–64. https://doi.org/10.1016/j.tmp.2012.03.002

Publisher's Note