IMT

SCHOOL
FOR ADVANCED
STUDIES
LUCCA

#04
2018

ISSN 2279-6894
**IMT LUCCA EIC WORKING
PAPER SERIES 04
October 2018**

RA **Economics and institutional change**

Brain-Circulation Network: The Global Mobility of the Life Scientists

Luca Verginer
Massimo Riccaboni

Research Area
Economics and institutional change

Brain-Circulation Network: The Global Mobility of the Life Scientists

Luca Verginer
IMT School for Advanced Studies Lucca

Massimo Riccaboni
IMT School for Advanced Studies Lucca;
Department of Managerial Economics, Strategy and Innovation, K.U. Leuven

ISSN 2279-6894
IMT LUCCA EIC WORKING PAPER SERIES #04/2018

Piazza San Ponziano 6, 55100 Lucca

Brain-Circulation Network: The Global Mobility of the Life Scientists

Luca Verginer[a,*], Massimo Riccaboni[a,b]

[a]*IMT Lucca School for Advanced Studies Lucca*
Piazza S. Francesco, 19, 55100 Lucca, Italy
[b]*KU Leuven, Department of Managerial Economics, Strategy and Innovation (MSI)*
Naamsestraat 69, 3000 Leuven, Belgium

Abstract

Global mobility and migration of scientists is an important modern phenomenon with economic and political implications. As scientists become ever more footloose it is important to identify general patterns and regularities at a global scale. At the same time cities, and especially global cities, have become important loci of economic and scientific activity. Limiting research to international migration, would disregard the importance of local innovation systems. The analysis of the mobility and brain circulation patterns at global scale remains challenging, due to difficulties in obtaining individual level mobility data. In this work we propose a methodology to trace intercity and international mobility through bibliographic records. We reconstruct the intercity and international mobility network of 3.7 Million Life Scientists moving between 9,745 cities. We present several features of the extracted network, offer evidence that the international innovation system is marked by national borders and linguistic similarity and show that international mobility largely contributes to the scientific output of national research systems. Moreover we find evidence to suggest that global cities attract highly productive scientist early in their careers.

Keywords: Network Analysis; Scientist Mobility; Brain Circulation; Global Cities; National Innovation Systems
JEL: F22, F66, J61, L65, O18, O15, O30, R12.

1. Introduction

Scientists are known to be highly mobile intellectuals, especially in the early phase of their careers. This has been true in the past (Cardwell, 1972; Mokyr, 2016; Serafinelli and Tabellini, 2017), but the size of the phenomenon has drastically increased in a globalized market for advanced human capital (Culotta, 2017; Geuna, 2015; OECD, 2017). Modern economies require a highly skilled labor force to maintain their competitive advantage and grow (Chambers et al., 1998; Solimano, 2008; Ozden and Rapoport, 2018; Zucker and Darby, 2007). Which makes it important to understand what determines this mobility. The authoritative manual on the "Global Mobility of Research Scientists" (Geuna, 2015, Ch.5, p.24) gives an overview of the current state of the research on the mobility of scientists and notes that research "on the mobility of researcher scientists is scarce because of a lack of reliable data to trace scientists along their careers". We contribute to this literature by constructing and analyzing a large scale and global scientist mobility dataset of 3.7 Million scientist working in 189 Countries and 9,745 cities.

Previous research on the mobility of scientists has used, among other approaches, large-scale surveys (Franzoni et al., 2012, 2014, 2018), and more recently massive bibliographic databases (Bohannon

[*] Corresponding author
Email addresses: luca.verginer@imtlucca.it (Luca Verginer), massimo.riccaboni@kuleuven.be (Massimo Riccaboni)

and Doran, 2017; Deville et al., 2014; Graf and Kalthaus, 2018). There are other sources of mobility information (e.g. Job search portals, social media), however papers offer the most direct and high frequency signal of scientific activity. We take advantage of the fact that scientists, especially in some disciplines, publish regularly in their career, and a lack of publications arguably signals its end. Inspired by bibliographic approaches we use Medline, a large publications repository primarily covering research in the life sciences.

This work focus the analysis on the level of the most important locations with activity in the life sciences (about 10 thousand populated places). We think that cities and especially global cities are an appropriate level to analyses mobility patterns and their role within the global economy in general (Taylor and Derudder, 2015; Sassen, 2016) and sciences in particular (Catini et al., 2015). We will also discuss implications at national level to complement the discussion on the more granular city level.

In this work we set out to characterize the geographic determinants of mobility, identify which cities lie at its center and show these "global cities" attract the most prolific scientists.

The rest of the paper is structured as follows. In Section 2 we show which data we use for our analysis and how the mobility network has been extracted. Then we characterize the basic properties of the global mobility network of the life-scientists describing topological and geographic features in Section 3. In Section 4 we present an analysis of "productivity" gains at scientist, city and country level, as a direct result of the observed mobility. We present the findings on the tendency of central cities to attract prolific scientists early in their career in Section 5. Finally in Section 6 we summarize the findings, offer an outlook for possible ways to extend the present analysis and discuss how this dataset might be used for different applications.

2. Reconstructing the Mobility Network

We reconstruct the mobility paths of life scientists through their publication history. Tracking scientist through their paper trail, an "activity based" approach, is best suited for the purpose of estimating mobility patterns and brain circulation phenomena. The use of papers as direct signal of production and location alleviates problems one might encounter in surveys and scraping of job listing services (e.g. LinkedIn). Most importantly, however publication are the actual output of interest when studying scientific output.

To reconstruct the mobility paths and estimate productivity we need to merge several sources of information. First we need a publication repository with a sufficient number of papers (Medline), proper disambiguation of the authors (Author-ity), assignment of these authors to locations (MapAffil) and a proxy for the quality of scientific output and by extension the authors themselves (SCImago).

In this section we introduce the four datasets, explain how they have been merged, how the mobility networks have been extracted and how we proxy author scientific production.

2.1. Data

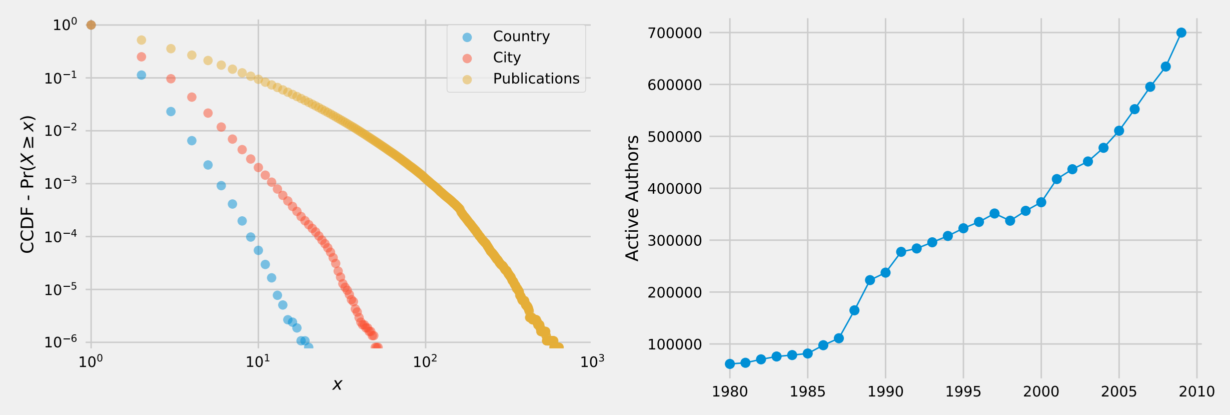
For the analysis we use four datasets, Medline, Author-ity, MapAffil, and SCImago.

Medline provides open access to more than 26 million records of scientific publications, with most of the corpus covering research in the life sciences. The data goes as far back as 1867 (earliest publication in the dataset) and is updated continuously. However we will focus on papers in the period between 1990 to 2009. We restrict our analysis to this period to have a good coverage and make use of existing high quality disambiguations of scientists (Author-ity) and affiliations (MapAffil), which are restricted to this time interval. MapAffil and Author-ity have been developed and published by Torvik (2015); Torvik and Smalheiser (2009).

MapAffil lists for a large portion of Medline papers the disambiguated city corresponding to the affiliation of each author as listed on the paper (ca. 37,396,671 author-locations). Author-ity developed by Torvik and Smalheiser (2009) contains the disambiguate names of 61,658,514 appearances of names on Medline papers (author-name instances). These author-name instances have been mapped to 9,300,182 disambiguated authors. MapAffil, also developed by Torvik (2015), is a disambiguation of affiliations listed on Medline papers. This dataset allows us to map the affiliation string to the city this affiliation is located in.

By merging Medline with Author-ity we obtain the necessary data to uniquely identify an author across publications. This information has been used in the past to reconstruct the global collaboration networks.[1]. The ability to reconstruct mobility comes from merging the previous two datasets with MapAffil. Without this last step, affiliations would not be disambiguate and we would have hundreds of different versions of "Boston University" in our dataset. Fortunately, MapAffil can accurately[2] map these various strings to a city.

By adding location information to the publication records we obtained for each author-publication pair a date and location. An example of which is available in Table 9 in the Appendix. From MapAffil we obtain as location the center of a city (low resolution), however these are mixed with locations at a higher resolution, which identify a suburb or part of a city. For example for "London, UK" we have the location (lat=51.5, lon=-0.126) but also 118 districts or city parts (i.e. "Bethnal Green, London, UK", "Goodmayes, Ilford, Redbridge, London, UK"). These have been reduced to the lowest common resolution So "Bethnal Green, London, UK" and "Goodmayes, Ilford, Redbridge, London, UK" would be mapped to "London, UK" at position (lat=51.5, lon=-0.126). And similarly the Boston neighborhoods "Jamaica Plain, Boston, MA, USA" and "Roslindale, Boston, MA, USA" are mapped to the lower resolution city center "Boston, MA, USA" (lat=42.359, lon=-71.057). By applying this method we obtain 9,745 urban areas.



(a) (b)

Figure 1: (a) Counter Cumulative Distribution of Countries, Cities and Publications per author. Each data point shows the probability to observe at least x unique Countries, Cities and Publications for a given author (i.e. $\Pr(X \geq x)$). (b) Number of unique active authors identified in Author-ity.

To have an appreciation for the number of unique cities, publications and countries any given author has been to or published in we show in Figure 1 their distribution. We see that all three distributions are highly skewed distributions (hence plotted in log-log) with a sharp decline for all values beyond 1. We see that only 10% of authors have at least two countries or 3 cities on their CV, or published at least 8 papers. Similarly only 1% of authors have worked in at least 3 different countries or 5 cities, or published at least 38 papers.

We analyze the affiliation path of 3,740,187 individuals, for which geo-location data is available in the period 1990 up to 2009. The coverage over time of these authors is available in Figure 1.

To estimate the quality of the researchers - required for Brain Circulation considerations - we augment the publication history with journal impact scores and research field classifications provided by SCImago. SCImago provides access to yearly "impact factor" scores for a large portion of journals indexed in Medline. We use this dataset to proxy the productivity of a scientists by the impact factor of the journal they publish in. SCImago calculates impact factors for journals starting from 1999 and backfills them. For this reason we do not use data for the brain circulation part of the analysis (Section 5) which reaches back several years

[1] Examples of co-authorship networks being used for research can be found in Newman (2001); Girvan and Newman (2002); Wagner and Leydesdorff (2005); Jackson and Rogers (2007)

[2] Torvik (2015) give a thorough explanation of their quality checks and provides estimates of the accuracy and precision.

before 1999 to reduce problems with deviations from the "true" citations per document in the journal. By considering only the period from 1999 to 2009 we have still 2,456,345 Scientists in our dataset, however only for 1,363,280 do we have complete coverage in SCImago. A detailed discussion on how the productivity indicators are constructed is available in Section 4. In addition to impact metrics we also use SCImago's journal classification to assign papers to thematic areas.

2.2. Methodology

With the extracted publication, we can reconstruct the path for a given author over time as observed by the affiliations on the papers she publishes. In other words we have a path for author i over several years indicating where she passed through. It might and actually does happen, that an author has multiple publications in the same year as well as multiple locations[3]. Here we define what a move is and how we extract it from the empirically observed publication sequences. To determine a move, and just as importantly a non-move, we define mobility by determining the location of an author within a given time window before a year of interest (t) (i.e. the move year) and assess where she is located in the window after.

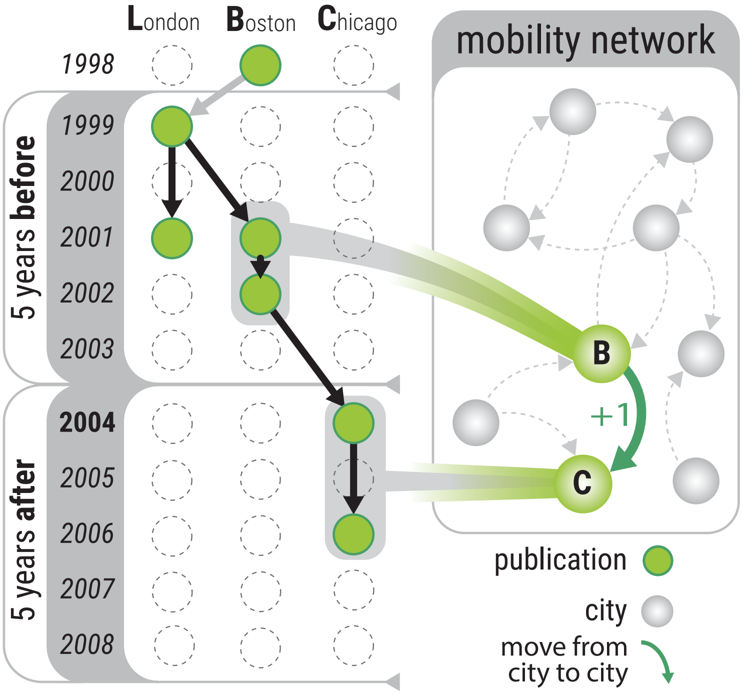


Figure 2: Creating the mobility network from Medline publications. The scientific publications by a single author are illustrated as a sequence of green circles from top to bottom. Each publication has a time (in rows) and location (in columns) associated with it. We take a buffer time (i.e. 5 years) before and after a candidate move from Boston (B) to Chicago (C) in 2004. In this example, we identify Boston as the source, since it is the longest sequence within the window and closest to the end of the move year. Similarly the destination is Chicago since it is the only observed city in the second window. Each move is tracked in a similar way and added to the mobility network by incrementing the edge weight accordingly.

Mores specifically to determine the source and destination of a move, for a given time interval we chose a candidate move-year (t) and a number of buffer years (b) around it (see Figure 2). To transform a publication path into a single edge representing a move we proceed e as follows. We chose a "move year" t of interest. The move year represents the year around which the decision to move happened. Next we choose a number of years around t defining two windows: **before** $[t-b, t)$ and **after** $[t, t+b)$. Given these two windows we

[3] an example of which is available in Table 9 of the Appendix

proceed to determine in which location any given author was before and after. If the locations differ then the author moved, otherwise she stayed.

To determine a unique starting position in window $[t-b, t)$ we choose the longest uninterrupted sequence of locations closest to t. Take for example the observed publication sequence as illustrated in Figure 2. Here we have the publication history $\{B_{1998}, L_{1999}, L_{2001}, B_{2001}, B_{2002}, C_{2004}, C_{2006}\}$, move year 2004 and a buffer of 5 years before and after. The Uppercase letter indicates the city and index the year. To determine the starting location we take all publications in the interval $[1999, 2004)$ and chose the locations with the longest sequence closest to 2004. In this example we observe 3 publications in B, but only 2 of these are within the $[1999, 2004)$ window, so we discard B_{1998}. On the other hand we observe 2 publications in L and one simultaneously with B. According to the aforementioned rule, we chose B as source since it closest to 2004 even though both L and B have 2 observations. As the destination of the move we chose C since in this case it is the only observed location in the window $[2004, 2009)$.

We chose this method, since it discards ambiguous affiliations in publication sequences with spurious affiliations (e.g. multiple affiliations in the same year but either of these appear only once).

This definition allows us to carry out several robustness checks in generating the network. For example we can increase the number of publications required in a given locations before and after to reduce the chance that a move was only temporary (e.g. visiting or double affiliations). Similarly we can restrict the size of the windows, thus requiring that authors have fewer holes in their publication history, however doing so will drop any scientist not publishing at least once in the two periods.

3. Descriptive Analysis of Mobility Network

In this section we offer an overview of several statistic describing the geographic mobility patterns of scientist at international and intercity level as well as an estimation of the centrality of cities within this network. We want to show that mobility does not only have a national component, but that analyzing it at the city level can give important insights into the position of countries within the international innovation system. First we show, that the most central cities in the international mobility network are US cities, with some minor exceptions. This observations, is confirmed by analyzing inflow and outflow patterns. In fact we find that these super-connected cities source their scientists from a wide range of cities and countries but their outflow is restricted to a smaller set of cities, suggesting that scientists passing through them remain in the core of the network. An analysis of the community structure of the mobility patterns suggests, not only that mobility is significantly influenced by national borders but that shared language can facilitated mobility.

Where we do only provide statistics for one network we refer to the mobility network for the move year 2004 with 5 years of buffer around it. In practice this means that the earliest publications we consider are from 1999. The starting city is determined in the period $[1999, 2004)$ and the destination is determined in the period $[2004, 2009)$. The analysis has been carried out also for 2003 and 2002 with window sizes ranging from 3 to 6, yielding similar results. We use this network because it is the most recent network for which we can be confident to have a good coverage of disambiguated authors and accurate SCImago scores.

3.1. City Centralities

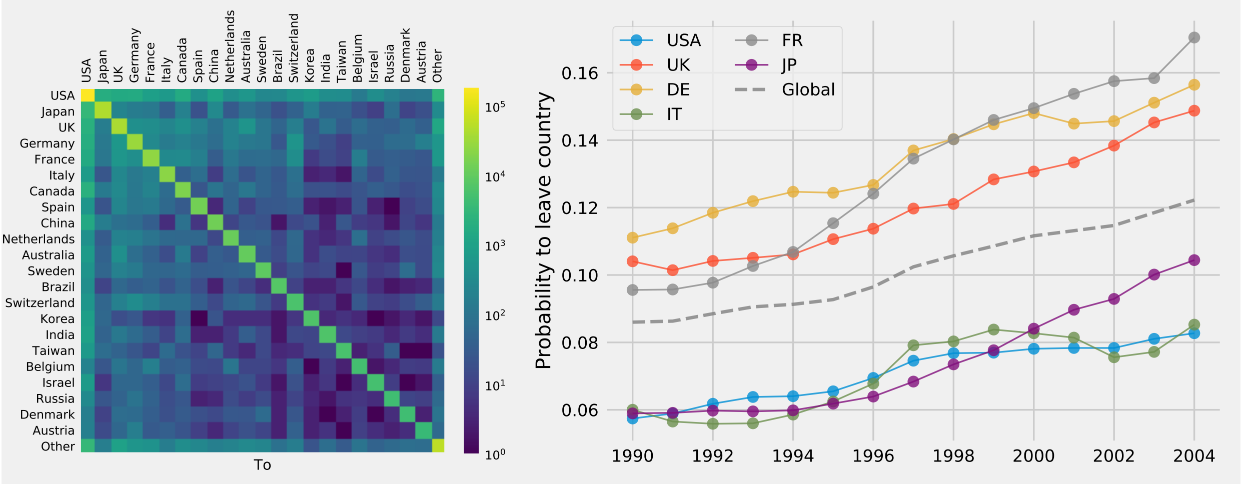
Which cities are at the center of the exchange of life scientists and how do different countries fare in this comparison? To answer this question we look at the 2004 Mobility network. Specifically we compute several standard network centrality measures to rank the position of cities. We determine which cities are part of highly connected “clubs” (k-core), would be the most likely location to find a scientist moving freely on the network (i.e. PageRank) and how many cities this city has access to (i.e. in/out-degree)

The k-core is defined as the set of nodes left after removing iteratively all nodes with degree less than k, until the graph is either empty or no more removal is possible. So for example in the case of an undirected, unweighted graph the 4-core contains all nodes which are connected to at least 4 other nodes which in turn are connected to 4 other nodes with the same property. The procedure filters out nodes which contribute to the degree of other nodes but do not themselves have many connections. This means that at a relatively low k most nuisance nodes (nodes which have few partners overall) are removed. PageRank is a commonly used centrality metric for directed weighted graphs. It estimates how likely a random walker traversing the network is to be found in a given node (Page et al., 1998). In the case of a mobility network, the measure can be understood as the stationary probability of a scientist to be found in any given city if she were to move following the strength and direction of the observed moves, with an occasional probability to be "teleported" to a random city.

In Table 1 we report the top 30 Cities as ranked by PageRank centrality along with k-core and degree rankings. The ranking reveals that US cities dominate the mobility network in the life sciences.

Among these top 30 cities only 9 are not US American and only 2 of these are from continental Europe: "Paris, France" and "Berlin, Germany". This ranking does not give a complete picture of the mobility network, but it suggests that cities are an important component. A more detailed analysis of the in and outflows (see Appendix, D-core decomposition) highlights the asymmetry in the global intercity exchange. We find that central cities in the US source their scientists from a wide verity of cities but they feed a smaller subset of cities .

3.2. National Border Effects



(a) (b)

Figure 3: (a) The Country to country mobility flows for the mobility network of 2004 with 5 year of buffer. On the main diagonal we find the *number* of all scientists who did not leave the country (i.e. the national scientist population). The rows are the source and the columns are the destination, with the color indicating the number. The countries are sorted according the size of their scientist population in the period 1999 to 2004. (b) Probability to leave country for selected countries and global mean (1990 to 2004). Note: the "country" is the country from which the move originates, not necessarily the nationality of the author.

Co-authorship networks have been found by Hoekman et al. (2010); Chessa et al. (2013) to be influenced by national borders resulting in collaborations being more likely within than across countries. In line with these findings we test the hypothesis, that countries have a stronger within mobility than across.

Figures 3 (a) shows the pattern of cross country mobility in 2004. Clearly most scientists do not leave their country (as indicated by the main diagonal). Note also that certain countries have few exchanges with all other countries, as indicated by having only few off diagonal elements brighter than the rest. This means that while the network is dense (i.e. all major countries have at least one exchange) there are preferences. Note also that the probability to leave the country has increased steadily year by year as can be seen in Figures 3 (b). The global probability to observe a move, i.e. that any given scientist moves abroad if we look at 5 years before and after, has never dipped since 1990. The listed countries fall into two categories, below the global mean and above. With the US, Japan and Italy clearly falling short of the global average, indicating a strong within mobility. Moves originating from the US tend to be mostly within the US. This

	Ranking				
City	k-core	PageRank	in-degree	out-degree	degree
Boston, MA, USA	1	1	1	2	2
London, UK	1	2	2	1	1
New York, NY, USA	1	3	6	4	5
Bethesda, MD, USA	1	4	3	5	4
Paris, France	1	5	5	3	3
Baltimore, MD, USA	1	6	4	7	7
Philadelphia, PA, USA	1	7	7	6	6
Chicago, IL, USA	1	8	9	8	8
San Francisco, CA, USA	1	9	13	18	14
Houston, TX, USA	1	10	8	9	9
San Diego, CA, USA	1	11	11	10	10
Tokyo, Japan	1	12	28	11	16
Atlanta, GA, USA	1	13	10	14	11
Seattle, WA, USA	1	14	12	12	12
Cambridge, MA, USA	1	15	15	15	15
Durham, NC, USA	1	16	18	21	19
Beijing, China	1	17	25	23	22
Toronto, ON, Canada	1	18	16	17	18
Los Angeles, CA, USA	1	19	20	33	27
Ann Arbor, MI, USA	1	20	19	20	20
Cambridge, Cambridgeshire, UK	1	21	16	16	17
Montreal, QC, Canada	1	22	23	28	25
Los Angeles, CA, USA	1	23	25	39	35
Stanford, CA, USA	1	24	22	26	23
Pittsburgh, PA, USA	1	25	23	28	25
New Haven, CT, USA	1	26	28	25	27
Berlin, Germany	1	27	31	31	31
Saint Louis, MO, USA	1	28	21	30	24
Seoul, Korea	1	39	59	70	62
Washington, DC, USA	1	30	35	24	30

Table 1: Ranking of top 30 Cities by centralities sorted by k-core and PageRank for the 2004 mobility network. Members of the EU (except UK) are **bold**

.

number has gone from 5% in 1990 to 8.1% in 2004, however compared to France (16.8%) and the global average (12%) it is low. Note however, that scientist based in the US do not leave the country as often as most other countries, but there is a substantial domestic exchange.

The international mobility patterns seen in Figure 3 suggest that international mobility varies by country and that there is more mobility within than across. The notion of "more within" and "less across" is made precise by the measure of *modularity* (Newman and Girvan, 2004). At a high level, modularity is a quality score of how well a given partitioning of nodes (i.e. set of cities) separates nodes which are well connected with each other but have few ties to members of other partitions. More specifically modularity measures the ratio of links falling within a given partition minus the ratio of links we would expect from a random network. A random network in this context is a network, which has the same degree sequence as the observed network, but rewired without regards for any underlying structure (see Newman and Girvan (2004) for more detail). Thus this null model represents a mobility network where scientists move without regard for geographic proximity or national borders. We estimate the communities by maximizing the modularity of the partition following the Louvain algorithm (Blondel et al., 2008) implemented by Traag (2017).

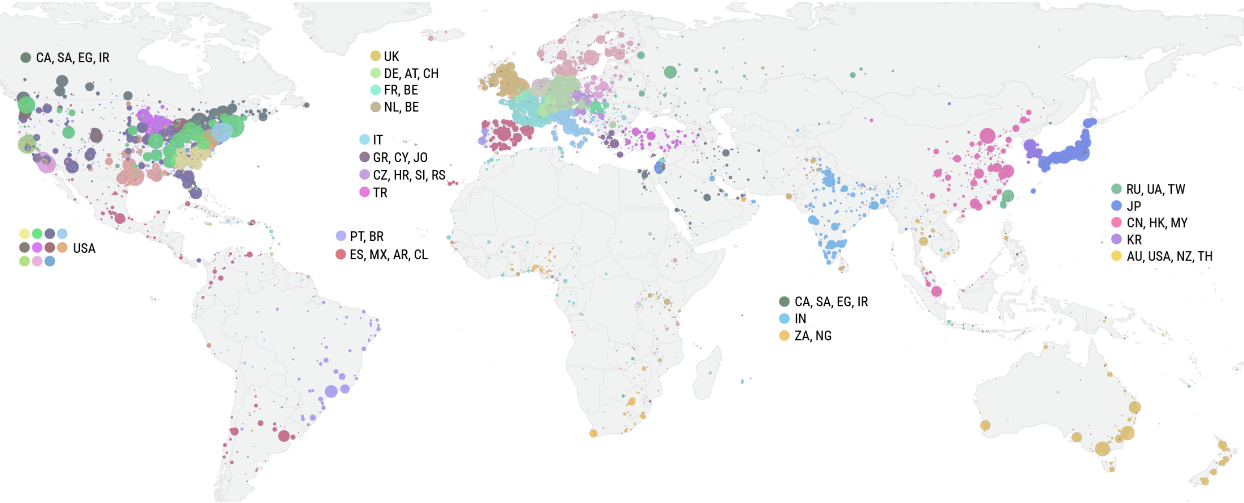


Figure 4: Community structure implied by the 2004 mobility network. Each node is a city and its color indicates to which community it belongs. The size is proportional to the sum of incoming and leaving authors.

If the null hypothesis that scientists move without regard for national borders were true, we should find that the community structure we obtain by maximizing the modularity does not coincide with any geographic or political boundaries.

However we do find that geography and national borders capture the community structure of the mobility network well (see Figure 4). A breakdown of countries as they fall within the various communities in 2004 is available in the Appendix (Table 10). The communities of the intercity mobility network in continental Europe, is clearly conditioned by national borders. For example, we find that the community to which Italy belongs is composed of 75% Italian cities, 6% US cities and several other minor percentages, the same goes for several other countries, which are the absolute majority within their community. However the picture changes when looking at North America. Here we also observe a national component in the form of Canada and Mexico being identified as separate communities, but within the US the identified communities are less spatially segregated than in the rest of the world.

Beyond the pure border effect the community structure reveals some additional patterns. We see that countries sharing a language are more likely to fall within the same community. For example three majority German speaking countries, Germany, Austria and Switzerland are identified as belonging to the same mobility community. Even more strikingly are Spain and Portugal. The two countries share a border but not a language. And we see that they are part of different communities. However as Table 10 (see Appendix) shows, Portugal and Brazil have a more significant exchange among themselves than Portugal has with Spain even though one is across the ocean and the other a next door neighbor. Similarly, Spain and Mexico are placed in the same community, both countries share a colonial history and language, as do Portugal and Brazil.

We should note that community detection through modularity maximization may fail to separate communities which are "too small" due to the method's "resolution limit" (Fortunato and Barthelemy, 2007). Ground truth communities, which are not of comparable size to the identified communities may be lumped together with larger communities or split up. In practice this could mean that we have lumped "small" communities together which probably should be kept separate, for example Greece, Cyprus and Jordan are placed in the same community. While Greece and Cyprus share a language the inclusion of Jordan in this community is most likely due to the fact that Jordan has had an exchange with the other two but was "erroneously" placed in the same community.

4. Mapping Brain Circulation

The concept of "Brain Drain", most prominent when discussing the mobility of scientists has been described by Geuna (2015, Ch.1, p.5) as an "unidirectional migration of skilled workers from less developed to more developed countries or regions". However as Agrawal et al. (2011); Saxenian (2005) argue, connections between migrant scientists and their home country persist and might facilitate knowledge flows in the opposite direction. Thus it is more appropriate to talk about brain circulation.

We present a high level overview of the flow of "talent/brains" at global scale taking various levels of aggregation into account. Specifically we want to look at the benefit scientists have from moving along certain paths/dyads and the gain in productivity a city has due to turnover (see Section 4.1). Similarly in Section 4.2 we describe at country level how international flows affects the scientific output.

To determine the productivity for any given scientist we use the journal "impact" factor data from SCImago. Specifically we use the "citations per document in the 2 years before the publication year" of the journal as the measure of quality of scientific output. To avoid inflating the output, we apply a fractional count, whereby any author receives for any paper coauthored with n authors and factor x the fraction x/n. We define several indicators, whose definition and description are summarized in Tables 2, 3 and 5.

To measure productivity we define two basic measures of scientific output, P_i^θ and r_i^θ, where i is the author and θ identifies in which window (before or after) her publications are aggregated. With $\theta = 0$ indicating the period before the move year t and $\theta = 1$ the period after. Specifically for every author i we obtain her publication list in the windows $\theta = 0 = [t-b, t)$ and $\theta = 1 = [t, t+b)$ (see Figure 2 window before and after). For each publication authored by i we then obtain the impact of the journal it is published in and divide it by the number of authors on that publication (i.e. fractional count). This yields for each author i a productivity before P_i^0 and after P_i^1. Additionally to take into account that authors might only start their career within the window we normalize this measure, such that it can be interpreted as the impact weighted annual productivity r_i^0. For example an author with $P_i^0 = 90$ who has started publishing in 1995 when considering the move year 1998 and a 5 year buffer would have a $r_i^0 = 90/\min(5, 1998-1995) = 30$ and similarly if the same author had published her first paper in 1990, r_i^0 would be $90/\min(5, 1998-1990) = 18$. Similarly for r_i^1 we divide by the buffer size b since she was by definition active from the beginning of that period (i.e. $90/5 = 18$).

4.1. Intercity mobility Gains

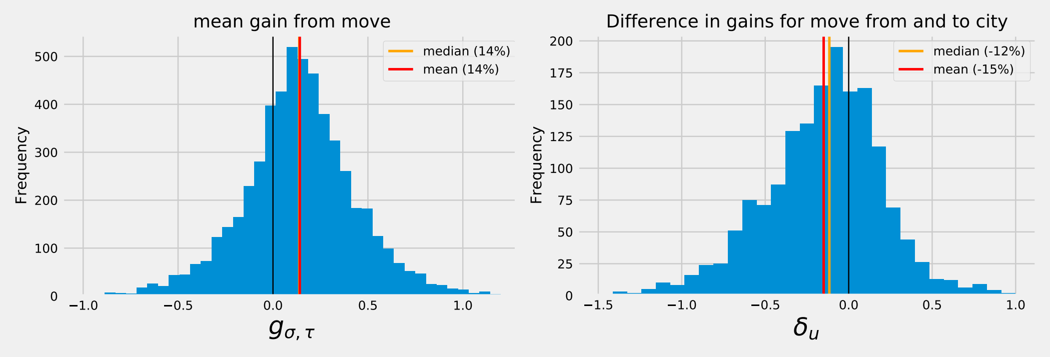
To understand the role and the importance of the cities in the international mobility and brain circulation network we define and compute several indicators of "productivity" gains. We want to quantify which routes/dyads confer the highest productivity gains on the scientist and if cities are able to replace the leaving scientists with incoming scholars.

To quantify and identify the gain a scientists can gain from moving from a given city σ to an other city τ we measure her impact weighted annual productivity before ($\psi^0_{\sigma,\tau}$) and after the move ($\psi^1_{\sigma,\tau}$) and compute the gain (i.e. log of ratio) and obtain $g_{\sigma,\tau}$. This measure represents the average gain scientists moving from σ to τ have experienced. Since a move might be due to productivity considerations and the global scientific

	Definition	Description
θ, t, b		t is the move year, b the number of buffer years around it, $\theta = 0$ is the period before $[t-b, t)$ and $\theta = 1$ after $[t, t+b)$.
$\mathcal{P}_i^\theta$		Set of papers produced by i in period θ
P_i^θ	$\sum_{p \in \mathcal{P}_i^\theta} w(p)$	Impact weighted fractional count of papers for author i in period θ. $w(\cdot)$ returns the impact factor of the journal the paper was published in that year, divided by the number of authors on the paper
r_i^0	$P_i^0 / \min\{\text{age}_i, b\}$	Annual productivity rate before the move
r_i^1	P_i^1 / b	Annual productivity rate after the move
$\mathcal{S}_{\sigma,\tau}$		Set of authors moving from source city σ to target city τ
$n_{\sigma,\tau}$	$\lvert \mathcal{S}_{\sigma,\tau} \rvert$	Number of scientist moving from σ to τ
$\rho_{\sigma,\tau}^\theta$	$\sum_{i \in \mathcal{S}_{\sigma,\tau}} r_i^\theta / n_{\sigma,\tau}$	Mean productivity rate in period θ for scientists moving from σ to τ
$\Psi_{\sigma,\tau}^\theta$	$\sum_{i \in \mathcal{S}_{\sigma,\tau}} P_i^\theta$	otal output for scientists moving from σ to τ in θ
$\psi_{\sigma,\tau}^\theta$	$\Psi_{\sigma,\tau}^\theta / n_{\sigma,\tau}$	Average output for scientists moving from σ to τ in θ

Table 2: Variables used in Brain Circulation calculations

output grows year by year we expect the global mean of $g_{\sigma,\tau}$ to be positive. And in fact we find that on average every move from any city to any other yields a gain of 14% (see Figure 5).

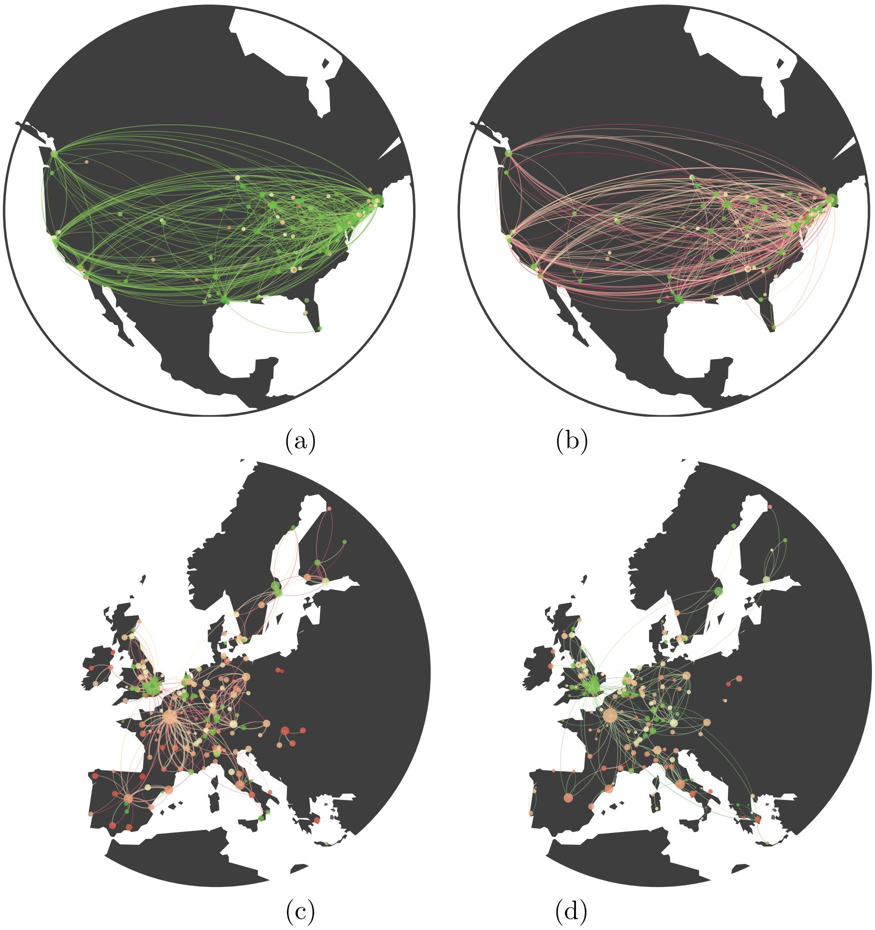


(a) (b)

Figure 5: (a) Distribution of average productivity gains ($g_{\sigma,\tau}$) for observed city to city flows (at least 10 moves, frequencies are dyads). (b) Distribution of gains for incomers net of gains for leavers (δ_u, frequencies are cities with at least 10 in and 10 out moves). Similarly for δ_u we show the distribution of cities falling within the specified bin.

By plotting $g_{\sigma,\tau}$ on a map (see Figure 7 and 6) and coloring the links according to its distance from the median, we see that most of the moves are green. Since there are as many red edges as green ones on the map, this implies that shorter moves, too small to be seen on the map, are below the median (i.e. red). We also notice that moves from the east to the west (edge direction is clockwise), especially the US are green, while moves from west to east are red (i.e. gains below the median).

Additionally we can look if there is an imbalance in the two possible direction the flow could takes place, i.e. $\xi_{\sigma,\tau} = ||g_{\sigma,\tau} - g_{\tau,\sigma}||$. Note that we can only compute this value for actually observed dyads. If the gain in any direction would be the same then $\xi_{\sigma,\tau}$ would be 0, however we find that this is not the case as it has a mean of 14%. This fact points to an imbalance in the direction of travel. We would expect that the direction with the higher gain to be chosen more often. However we do not find that the strength of the flow ($n_{\sigma,\tau}$) is correlated with the mean gain ($g_{\sigma,\tau}$). This is confirmation of the our visual intuition of red vs green edges on circulation map (Figure 7). This is an indication that there is not only a supply side (i.e.


(a)

(b)

(c)

(d)

Figure 6: Brain circulation network zoom USA and EU. Here we show only city to city connection within the USA and EU respectively. Each arc represents an observed movement of at least 15 people. Locations with neither in nor outflow or a scientist population of less then 50 are hidden. (a) and (b) show flows where $g_{\sigma,\tau}$ is above the median and (b, d) below. The colors and direction of flows are the same as in Figure 7.

	Definition	Description
$g_{\sigma,\tau}$	$\log(\rho^1_{\sigma,\tau}/\rho^0_{\sigma,\tau})$	Average productivity gain for scientists moving from σ to τ
$\xi_{\sigma,\tau}$	$\|\|g_{\sigma,\tau} - g_{\tau,\sigma}\|\|$	Absolute Difference in gains for scientists moving from σ to τ and vice versa.
δ_u	$g_{\star,u} - g_{u,\star}$	Difference between the productivity gain from moving to u and leaving it
Γ_u^{move}	$\log(\Psi^1_{\star,u}/\Psi^0_{u,\star})$	Increase in output for incoming relative to leavers
Γ_u^{stay}	$\log(\Psi^1_{u,u}/\Psi^0_{u,u})$	Increase in output for stationary scientists

Table 3: City level Brain Circulation indicators

scientists choose were to go) but also a demand side to scientist intercity mobility. Cities, in the form of universities and research centers, are discerning who they hire or reject.

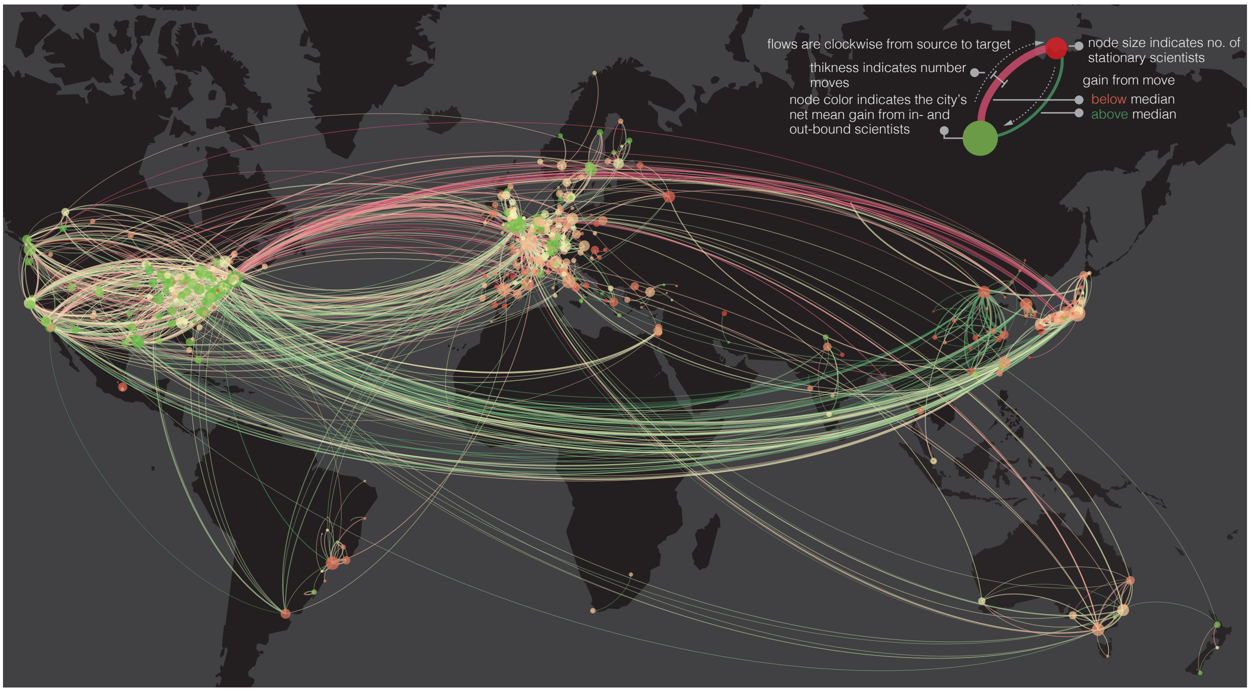

flows are clockwise from source to target
thikness indicates number moves
node color indicates the city's net mean gain from in- and out-bound scientists
node size indicates no. of stationary scientists
gain from move
below median
above median

Figure 7: Brain circulation network for the year 2004. The brain circulation map shows the intercity movement for cities (nodes) which have at least 15 authors incoming or 15 leaving and a stationary scientists population of at least 50. The thickness of the edges is proportional to the natural log of number of people moving between two cities. Locations with neither in nor outflow are dropped. Link colors show the average net gain ($g_{\sigma,\tau}$). Red colors indicates moves which are below the median of the shown links (14%) yellow close and green above. Node colors indicates the difference in gain for incoming and leaving scientist δ_u (see Section 4.1).

To determine if on average incoming scientists gain more than those they replace, an indicator that working in this city confers on the incoming scientist a large boost we look at δ_u (shown as node colors in Figure 7) and shown for a subset of global cities in Table 4. The global mean of this value is -25%. In other words, on average the gain from moving to any given city is less than leaving it. For example a move to "Boston, MA, USA" from anywhere confers on the scientists a 28% gain but anyone leaving for any other city gains only about 1%, which gives us $\delta_{\text{Boston}} = 0.27$. This value is depicted in Figure 7 and Figures 6. Here we see that the US contains several cities, which have a positive δ_u, while Europe has mostly negative δ_u.

These measures (i.e. g and δ) are interesting to the scientist making the decision to relocate. However cities have other priorities, i.e. increase scientific output. To quantify if cities benefit from the international exchange we look at two indicators Γ_u^{stay} and Γ_u^{move}. Γ_u^{stay} gives us for city u the increase in total output for stationary scientists. In other words, it measures the percentage increase in total scientific output for scientists who do not move. And Γ_u^{move} gives us the growth in total output coming from new scientist in period 1 relative to the output of the scientist who did leave in period 0. If $\Gamma_u^{\text{move}} > \Gamma_u^{\text{stay}}$ for a city u then the mobile scientist where able to produce enough scientific output to cover their predecessors and contributed positively to the total output growth of the city. From the histograms in Figure 8 we see that on average this is true ($\Gamma_u^{\text{move}} = 29\%$ and $\Gamma_u^{\text{stay}} = 23\%$). In fact we see in Table 4 $\Delta\Gamma_u$, the difference between growth from mobile scientists and the growth due to stationary scientists. The listed cities are the most central cities in the mobility network as identified in Section 1. We see again that US cities are able to manage the turnover better than central European cities, such as Paris, and Berlin. However within the US there are differences, with "Bethesda, MD, USA" for example being able to replace their scientific output with new scientists better than "Boston, MA, USA". This does not necessarily mean that they loose out, since these cities have a prolific stationary scientists populations, however is highlights several cities able to manage the turnover better than others.

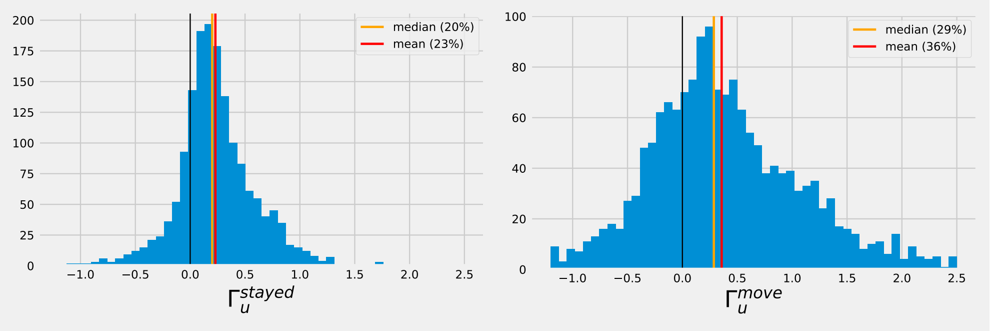


Figure 8: Distribution of gain indicators for the global network. The distributions shows statistics for routes which had at least 10 moves along them.

4.2. National Gains

Brain circulation is a major concern at country level and we can estimate the contribution to the growth of the national innovation systems from international mobility, but also domestic mobility. This allows us to compare knowledge output (change in total impact weighted output) across countries and identify which countries were the primary direct beneficiaries of international mobility.

The output produced within a country can be accounted for in the following way. Knowledge produced by authors staying in their city (S) moving domestically (D), *coming in from abroad* (I) and *leaving the country* (L). The total output for a given time period within a country before A^0 and after A^1 are given by $A^0 \equiv S^0 + D^0 + L^0$ and $A^1 \equiv S^1 + D^1 + I^1$ respectively. Note that in A^0 the output contains the production of those individuals who will leave the country L^0 in the second period and A^1 the production of those that will come in the second period I^1. Based on this breakdown we can define indicators identifying the growth due to the three types of scientists. Overall growth ν_A for the country, ν_S growth due to stationary scientists, ν_D growth due to nationally mobile scientists and most relevant for the brain circulation discussion ν_I, the gain due to international turnover. Additionally to have a indication of the generational turnover we also report the mean age of incoming ($\overline{\text{age}}_I$) and leaving scientists ($\overline{\text{age}}_L$). These indicators are defined in more detail in Table 5. The results for the largest countries in the dataset for the interval before ($\theta = 0 = [1999, 2004)$) and ($\theta = 1 = [2004, 2009)$) are reported in Table 6.

From Table 6 we see that the USA, with international turnover, has increased its scientific output by 14% overall (ν_A). Among all the three types of scientists, the growth due to new arrivals (I) is largest, 61%. This also compared to growth due to stationary and domestically mobile scientists (9% and 17%, respectively).

City	Γ_u^{move}	Γ_u^{stay}	$\Delta\Gamma_u$	$g_{\star,u}$	$g_{u,\star}$	δ_u
Tokyo, Japan	-0.17	0.13	-0.31	0.04	0.16	-0.12
Saint Louis, MO, USA	-0.16	0.06	-0.22	0.20	-0.04	0.24
Paris, France	-0.08	0.11	-0.19	0.04	0.14	-0.10
Berlin, Germany	0.13	0.24	-0.10	0.12	0.20	-0.08
London, UK	0.04	0.12	-0.08	0.21	0.13	0.08
Washington, DC, USA	0.08	0.16	-0.07	0.13	0.08	0.05
New York, NY, USA	0.00	0.04	-0.04	0.20	-0.02	0.22
New Haven, CT, USA	-0.01	0.01	-0.02	0.28	-0.01	0.29
Chicago, IL, USA	0.08	0.09	-0.01	0.21	0.04	0.17
Stanford, CA, USA	0.01	-0.02	0.03	0.18	0.07	0.11
Seoul, Korea	0.62	0.58	0.04	0.23	0.51	-0.28
Boston, MA, USA	0.17	0.13	0.04	0.23	0.01	0.22
Cambridge, MA, USA	0.08	0.01	0.07	0.09	0.16	-0.06
Montreal, QC, Canada	0.19	0.11	0.08	0.11	0.08	0.03
Westwood, Los Angeles, CA, USA	0.16	0.07	0.09	0.23	0.11	0.12
Philadelphia, PA, USA	0.17	0.07	0.09	0.20	0.01	0.18
San Francisco, CA, USA	0.11	0.01	0.10	0.05	0.01	0.03
Toronto, ON, Canada	0.20	0.10	0.10	0.16	0.07	0.09
Cambridge, Cambridgeshire, UK	0.20	0.10	0.11	0.17	0.03	0.15
Seattle, WA, USA	0.22	0.09	0.13	0.15	0.10	0.05
Baltimore, MD, USA	0.23	0.08	0.15	0.25	0.06	0.19
San Diego, CA, USA	0.08	-0.08	0.16	0.27	-0.13	0.40
Ann Arbor, MI, USA	0.29	0.12	0.16	0.23	0.07	0.16
Houston, TX, USA	0.25	0.07	0.18	0.24	-0.02	0.26
Los Angeles, CA, USA	0.25	0.07	0.18	0.09	0.13	-0.05
Durham, NC, USA	0.34	0.06	0.28	0.16	-0.01	0.17
Bethesda, MD, USA	0.39	0.07	0.33	0.23	-0.04	0.27
Pittsburgh, PA, USA	0.49	0.13	0.35	0.30	0.17	0.13
Beijing, China	1.32	0.87	0.45	0.36	0.75	-0.39
Atlanta, GA, USA	0.64	0.11	0.52	0.22	0.19	0.02

Table 4: The Γ_u values (growth due to turnover) for the 30 most central cities as listed in Section 3.1. The indicator of scientists gains from a move there u ($g_{\star,u}$), u $g_{u,\star}$, δ_u are also listed. The cities are ordered by Γ_u

	Definition	**Description**
S^θ	$\sum_{i\in\mathcal{S}_{u,u}} P_i^\theta$	Output of stationary scientist in domestic city d
D^θ	$\sum_{i\in\mathcal{S}_{u,d}} P_i^\theta$	Output of scientist moving from domestic city u to domestic city d
I^θ	$\sum_{i\in\mathcal{S}_{f,d}} P_i^\theta$	Output of scientist coming from foreign city f to a domestic city d.
L^θ	$\sum_{i\in\mathcal{S}_{d,f}} P_i^\theta$	Output of scientist leaving the country for a foreign city f.
A^0	$S^0 + D^0 + L^0$	Total Output in the country before the move year
A^1	$S^1 + D^1 + I^1$	Total Output in the country after the move year
ν_A	$(A^1 - A^0)/A^0$	National output growth of output
ν_S	$(S^1 - S^0)/S^0$	Output growth of from stationary scientists
ν_D	$(D^1 - D^0)/D^0$	Output growth from domestically mobile scientists
ν_I	$(I^1 - L^0)/L^0$	Output growth from international exchange
$\overline{\text{age}}_I$		Average age (years from first publication) for incoming
$\overline{\text{age}}_L$		Average age of leaving scientists
$\Delta\overline{\text{age}}$	$\overline{\text{age}}_L - \overline{\text{age}}_I$	Age difference between Leaving and Incoming scientists

Table 5: Country level Brain Circulation indicators

The stationary ($S = 0.71$) and domestically mobile scientists ($D = 0.23$) represents the largest portion of the population, however international exchange has had a net benefit on the output growth. However not all countries have a higher than average growth from international mobility (i.e. $\nu_I > \nu_A$). This suggests that not all countries have the same direct gain from international exchange. Moreover looking at the age differential between incoming and leaving scientists we see that the average scientists moving to the US (6.7) are younger than the ones they replace (7.9). This means that the US has been able to rejuvenate their scientific labor force, while simultaneously increasing their scientific output.

Clear beneficiaries of international exchange beyond the US, are Australia, Canada, Spain and Switzerland with $\nu_I > \nu_D > \nu_D$ and with a substantial contribution (i.e. more than 10% of output share). Argentina for example, has experienced only 5% output growth, the second lowest in the list and has lost 48% of output due to international exchange. Japan is also striking, the scientist leaving are young (6.7) compared to the scientists moving to Japan (8.6). This is accompanied by a negative loss from international exchange -21%. All other countries in this comparison either loose out or the effect is ambiguous. What is clear, is that international exchange as measured by direct scientific output does not benefit everyone in the same way.

	ν_A	ν_S	ν_D	ν_I	$\overline{\text{age}}_I$	$\overline{\text{age}}_L$	$\Delta\overline{\text{age}}$
Argentina	5%	23% (0.71)	79% (0.02)	-48% (0.27)	8.98	8.13	0.85
Australia	32%	28% (0.80)	24% (0.07)	56% (0.13)	7.34	6.93	0.40
Austria	18%	22% (0.77)	-7% (0.04)	9% (0.18)	7.35	7.34	0.01
Belgium	23%	28% (0.79)	15% (0.07)	0% (0.14)	8.23	8.21	0.02
Brazil	46%	53% (0.78)	57% (0.08)	1% (0.14)	7.66	6.88	0.77
Canada	20%	16% (0.73)	19% (0.11)	36% (0.16)	7.14	7.08	0.06
China	41%	117% (0.62)	212% (0.15)	58% (0.23)	4.52	3.36	1.16
Denmark	18%	18% (0.81)	18% (0.06)	17% (0.13)	8.00	7.93	0.07
Finland	3%	8% (0.77)	2% (0.10)	-21% (0.13)	9.06	8.92	0.14
France	10%	14% (0.76)	12% (0.09)	-10% (0.16)	8.14	7.47	0.66
Germany	16%	18% (0.66)	16% (0.19)	8% (0.15)	7.29	7.31	-0.02
India	42%	65% (0.66)	73% (0.10)	-33% (0.24)	7.91	5.64	2.27
Israel	14%	22% (0.74)	27% (0.09)	-23% (0.17)	9.67	7.12	2.54
Italy	32%	32% (0.82)	31% (0.10)	36% (0.08)	8.49	8.00	0.49
Japan	9%	12% (0.66)	13% (0.24)	-21% (0.10)	8.57	6.71	1.85
Korea	71%	66% (0.63)	87% (0.19)	76% (0.18)	5.55	4.70	0.85
Netherlands	26%	28% (0.74)	18% (0.13)	19% (0.12)	7.57	7.86	-0.29
Russia	13%	27% (0.73)	41% (0.01)	-27% (0.26)	8.03	7.37	0.66
Spain	35%	35% (0.81)	21% (0.07)	44% (0.11)	7.68	6.45	1.23
Sweden	10%	18% (0.74)	13% (0.08)	-25% (0.18)	8.47	8.07	0.40
Switzerland	14%	8% (0.67)	2% (0.08)	33% (0.24)	7.53	7.94	-0.42
Taiwan	37%	37% (0.74)	51% (0.15)	17% (0.11)	6.90	7.94	-1.04
UK	16%	14% (0.74)	25% (0.13)	17% (0.13)	7.16	7.27	-0.12
USA	14%	9% (0.71)	17% (0.23)	61% (0.06)	6.69	7.89	-1.20

Table 6: National scientific output growth figures for selected countries (at least 3,000 stationary scientists in 1999–2004). See Table 5 for definitions. In parentheses the proportion of the total output in the first period (A^0) by category.

5. Preference for Global cities: Regression Analysis

The topological, geographic and impact gain analysis in the previous sections suggest that there is a spatial component to the mobility patterns of scientist, that certain cities are more central within this network and that not all moves offer the same gain for a mobile scientist. We test the hypothesis that more central cities (k-core, page rank or degree), not only attract a lot of scientist but attract more productive scientist. If it is indeed true that more productive scientists move preferentially to more central cities, we expect productivity to be positively correlated with the centrality of the destination. That is, after we control for various factors and account for selection bias in our data (i.e. not all scientists move), we should find that scientific output before the move is indicative of a move to a more central city.

We estimate a Heckman two stage regression to account for the fact that the majority of scientists do not move and as such we would not observe a change in the centrality of their relocation choice. The focal variable of this analysis are "Productivity^0" and "Centrality^1". The variable "Productivity^0" measures how prolific a scientists was before she moved. Specifically this is the log of r^0, which is described in detail in Section 4.1. All other controls used in the regression are listed in Table 7.

Variable	**Description**
Centrality^0	The centrality of the source city (i.e. PageRank)
Centrality^1	The centrality of the destination city
Moved	1 if the author moves to a different city, 0 otherwise.
Productivity^0	The log of the annual productivity rate r^0(see Section 4.1)
Pr(Move other Fields)	The proportion of authors moving away from the source city which do not publish in the same field as the focal researcher.
$\mathbb{I}$(Year)	The move year (i.e. 2000, 2002, 2004)
$\mathbb{I}$(Age Group)	Age is measured as the difference in years from first publication to the move year. The age-groups are split such that the cohorts are of comparable size.
Intermove	The years between the last observation in the first period and the first in the second.
$\mathbb{I}$(Country)^0	The country in which the author was working in period 0
log(km dist)	The log of the distance from source to target city in kilometers
$\mathbb{I}$(Field)^0	The SCImago thematic area the authors publishes most in period 0

Table 7: Regression Variables

Note that not all countries and fields are present in sufficient number or are relevant for the analysis. For this reason we drop an author from the dataset if one of the following applies: (1) the author is a member of a country which has less than 500 scientists or (2) the author publishes predominantly in fields for which there are less than 500 papers in the period. These are mostly fields which are not considered life-sciences but are in Medline (e.g. Economics).

As the measure of "Centrality^0" and "Centrality^1" we use PageRank since it is proportional to the stable distribution of a random walker on the observed mobility network. The PageRank of a city can be interpreted as the null model where the relocation choice is simply done at random without regards for productivity, distance or other features we assume are important, but following the empirically observed flows between cities. As a robustness check we also estimate the model for k-core and degree centrality, which yield qualitatively similar results (see Appendix Table 11 and Table 12). The dataset is constructed by combining three mobility network (2000, 2002 and 2004) all with a buffer of 5 years.

To estimate the Heckman model and correct for self selection of scientists into the population of mobile scientists, we use as an exclusion restriction the probability to leave the city for all scientist not belonging to the focal field (Pr(Move other Fields)). For example, for authors predominantly publishing in "Embryology" the probability to move is computed as the fraction of scientists leaving the city in the same period, but who do not publish on "Embryology". The rational to use this variable as an exclusion restriction is that if

	Pr(move)		PageRank destination	
PageRank source	-5.262	(-0.23)	0.0166	(0.76)
Productivity^0	0.0209***	(4.28)	0.000146***	(4.16)
Pr(Move other Fields)	1.564***	(6.10)		
log(km dist)			0.000108*	(2.26)
2002	0.0289***	(6.05)	0.0000352	(0.84)
2004	0.0331***	(4.42)	0.0000375	(0.70)
2002 × Productivity^0	0.00269	(0.86)	0.0000236*	(2.35)
2004 × Productivity^0	-0.00455	(-1.35)	0.0000100	(0.68)
inter-move	0.109***	(64.38)	-0.0000751***	(-4.25)
Constant	-1.754***	(-4.05)	0.00289***	(3.77)
Year Effects	Yes		Yes	
Origin country effects	Yes		Yes	
Age effects	Yes		Yes	
Field effects	Yes		Yes	
Observations	1,363,280		433,023	
tanh(ρ)	-0.12 (-4.84)	log(σ)	-5.73 (-35.80)	
Log pseudo-likelihood	1,000,412			

Table 8: Regression results for mobility and relocation choice. Results of the Heckman two stage regression for the PageRank. The standard errors have been clustered at source city for the first stage (Pr(move)) and on the destination city for the second stage (i.e. the centrality of the destination)

we observe a lot of mobility originating from a city, it stands to reason that it increases the propensity of the focal author to move as well. By excluding the focal field we reduce the likelihood that the focal author is influenced by competition or imitation of peers working in the same field.

$$\begin{aligned}\text{Moved}_i =& \gamma_0 + \gamma_1 \text{Centrality}_i^0 + \gamma_2 \text{Productivity}_i^0 + \\ & \gamma_3 \Pr(\text{Move other Fields})_i + \gamma_4 \text{Intermove}_i + \\ & \boldsymbol{\gamma_a} \mathbb{I}(\text{Age Group})_i + \boldsymbol{\gamma}_y \mathbb{I}(\text{Year})_i + \boldsymbol{\gamma}_f \mathbb{I}(\text{Field})_i^0 + \boldsymbol{\gamma}_c \mathbb{I}(\text{Country})_i^0 + \\ & \boldsymbol{\gamma_{pa}} \mathbb{I}(\text{Year})_i \times \text{Productivity}_i^0 + \\ & v_i\end{aligned} \tag{1}$$

$$\begin{aligned}\text{Centrality}_{1i} =& \beta_0 + \beta_1 \text{Centrality}_i^0 + \beta_2 \text{Productivity}_i^0 + \\ & \beta_3 \text{Intermove}_i + \\ & \boldsymbol{\beta_a} \mathbb{I}(\text{Age Group})_i + \boldsymbol{\beta}_y \mathbb{I}(\text{Year})_i + \boldsymbol{\beta}_f \mathbb{I}(\text{Field})_{0i} + \boldsymbol{\beta}_c \mathbb{I}(\text{Country})_i^0 + \\ & \boldsymbol{\beta_{pa}} \mathbb{I}(\text{Year})_i \times \text{Productivity}_i^0 + \\ & \log(\text{km dist}) + u_i\end{aligned} \tag{2}$$

In the first stage (1) we estimate the probability that a given author decides to relocate And in the second stage (2) the PageRank of the destination is estimated conditional on observing a move.

5.1. Regression Results

The results of the regression are shown in Table 8. We find that in the second stage the propensity to move to a more central city is positively correlated with "Productivity^0". This confirms our hypothesis that controlling for various factors, more prolific scientists (before) tend to move to more central locations. However this effect changes with age, with young scientists having a substantially higher propensity to move to a more central city than more senior scientists (see Figure 9).

With regard to the first stage, note that Pr(Move other Fields) is positive. This means that for any given location the probability to observe a move is positively correlated with the probability to move of other scientists, not working in the same field. So in fact, we do find that the exclusion restrictions has the desired sign. The probability to observe a move (see column Pr(move)) does not depend on the PageRank of the source city. However we do find that Productivity^0 controlling for various factors has a significantly positive effect on the probability to move. This effect holds across centrality measures (see Appendix Tables 11 and 12).

From Section 3.2, we have observed an increased tendency to move abroad over the years. This is also confirmed by the increasing propensities to move (i.e. 2000, 2002, 2004). However the influence of "Productivity^0" has remained constant across snapshots. Additionally we find that having larger holes in the publication history (i.e. long Intermove) is strongly indicative of a observing a move. This is to be expected since the longer we do not have a signal of presence the chances of finding a scientist again in the same location decrease. Additionally a larger hole in the publication history means that the starting and destination locations are weak signals of actual presence in the and could have been spurious.

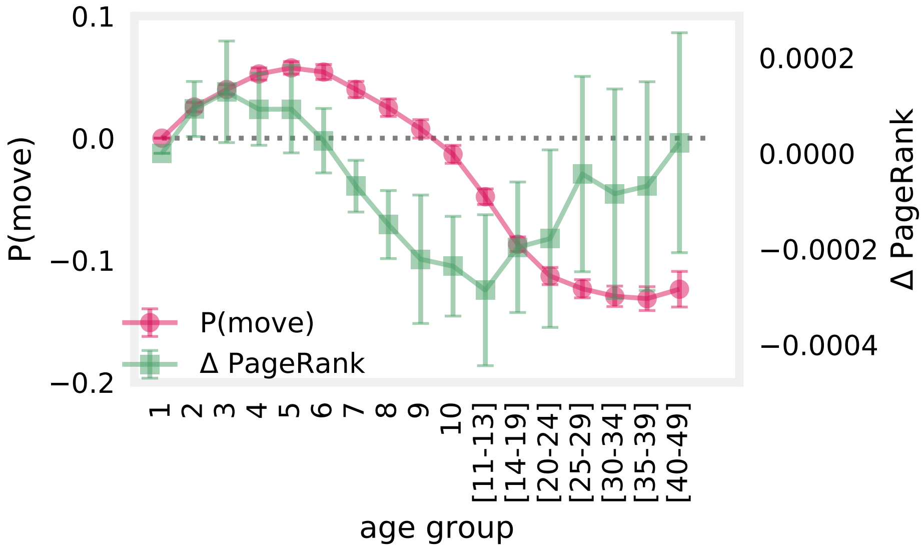


Figure 9: Probability and centrality gain by age This figure shows the marginal effects in probability to move compared to an author with age=1 (a year after the first publication in MEDLINE) and the marginal PageRank increase in destination due to age. An author with a 9 years career has the same probability to be observed moving as an author at the beginning of her career, however the PageRank of the chosen destination will be on average lower (error bars indicate 95% confidence interval).

We can also observe a cyclical pattern in the probability to move by looking at the marginal effects by age-group (see Figure 9). In the years after the first publication the marginal probability to move increases but declines after about 5 years. At country level we find also differences in the propensity to move (see Figure 10). The US is identified as having the most mobile scientists population, followed closely by UK, Switzerland and Germany. Note that while mobility from the US is low, as we have seen in Section 3.2, however the overall mobility is high, which implies that most of the mobility is domestic, this is also confirmed by the analysis on country gains. We also find that the probability to move by field of research "$\mathbb{I}$(Field)" varies greatly (see Figure 14), with "Physics and Astronomy" being considerably more mobile than fields such as Dermatology.

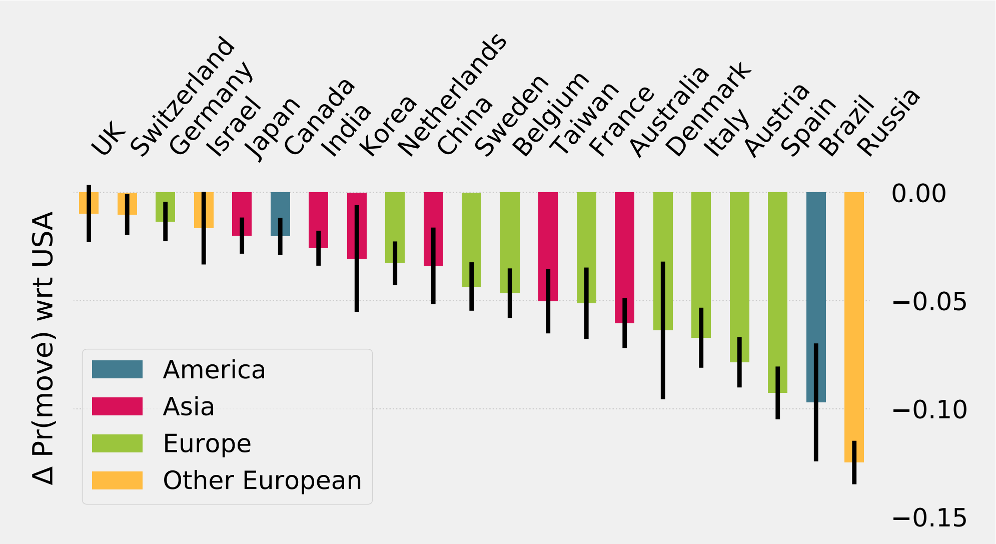


Figure 10: The figures shows the marginal effect on the probability to move compared to the base case USA and the 95% confidence interval (black bars). A negative value such as Taiwan (-5%) means that keeping everything else fixed, a scientists in Taiwan is 5% less likely to move than a colleague in the US. Only countries are shown here for which we have observed at least 3,000 scientists in the country in the period 2000–2004.

6. Discussion and Conclusions

Our results highlights several features of the mobility patterns of life scientists in the period 1999 to 2009. In general we find that not all cities are equal and borders do matter, however shared language can reduce barriers. We find that not all countries and cities benefit equally from intercity and international exchange. It is apparent that for the analyzed period European cities are not well represented within the global life sciences research system. Moreover gains in national scientific output, as highlighted by the output growth due to international turnover, do not provide a clear signal that international exchange is unequivocally beneficial to all participants. The results clearly point to the USA being a prime beneficiary, which according to the data is able to attract young and prolific scientists.

This study makes four contributions. First, it introduces a novel approach to extract mobility networks from bibliographic data and augments it with quality indicators. Second, it characterizes the international flows of life scientist highlighting the importance of national barriers. Third, it quantifies the gains from mobility to scientists, cities and countries. And finally it offers evidence that young prolific scientist move to global cities early in their career.

Our study has several strengths. We are able to reconstruct intercity mobility networks for specific timer intervals, making it potentially useful for event studies, although here we have focused primarily on the 2004 cross section. The dataset has an extensive coverage of life scientists spanning multiple countries, career stages and productivity levels (i.e. not only star scientists). However, it is not without limitations. First, we do not have detailed personal information on the scientists such as gender, birth date or citizenship (only the origin of mobility, which may not coincide). This information is available in smaller but more focused datasets such as the ones used by (Franzoni et al., 2012; Graf and Kalthaus, 2018). Second we are restricted to 2009 by Author-ity, making the findings less current than we would like, however a more recent high quality Medline author disambiguation could alleviate this problem. And third, this dataset covers primarily life scientists, omitting a large chunk of potentially relevant disciplines.

In this work we have limited ourself to a descriptive analysis of the mobility network, omitting causality claims. However the richness of the dataset makes it potentially useful for use in determining causal relocation factors. The global nature and good temporal coverage means that several natural experiments can be identified, which can help to isolate the determinants of mobility. An example of this, is the estimation of the impact of stem cell legislation in the US on stem cell scientist mobility (US states offer various degrees of support). Similarly, the effect of regional projects (e.g. opening a new research campus), aiming to improve scientific output or innovation, can be quantitatively analyzed. This dataset, in conjunction with natural language processing techniques and text mining, can also be used to follow the mobility and diffusion of new ideas and concepts in the life sciences. By estimating the relative importance of mobility and collaboration research policies optimizing diffusion could be devised.

In conclusion, this papers has described a method to extract mobility networks from bibliographic data, used the resulting mobility networks to characterize the mobility patterns and output gains of life scientists at city, intercity and national level as well as provided evidence that prolific scientists gravitate preferentially towards global cities early in their career.

Acknowledgments

We would like to thank Vetle Torvik and Neil Smalheiser for granting us access to Author-ity and MapAffil and Prof. Frank Schweitzer for helpful discussions and comments. We would also like to thank participants of the inet Workshop, where we presented an early draft of this paper. Discussions with participants at the GeoInno 2018 conference in Barcelona have also been very helpful.

References

Agrawal, A., Kapur, D., McHale, J., Oettl, A., 2011. Brain drain or brain bank? The impact of skilled emigration on poor-country innovation. Journal of Urban Economics 69, 43–55. doi:10.1016/j.jue.2010.06.003.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, P10008. doi:10.1088/1742-5468/2008/10/P10008.

Bohannon, J., Doran, K., 2017. Introducing ORCID. Science 356, 691–692. doi:10.1126/science.356.6339.691.

Cardwell, D.S.L., 1972. Turning points in western technology; a study of technology, science and history. New York, NY Science History Pub.

Catini, R., Karamshuk, D., Penner, O., Riccaboni, M., 2015. Identifying geographic clusters: A network analytic approach. Research Policy 44, 1749–1762. doi:10.1016/j.respol.2015.01.011.

Chambers, E., Foulon, M., Handfield-Jones, H., Hankin, S., Michael III, E., 1998. The war for talent. The McKinsey Quarterly 3, 44–57. doi:10.1080/03071840308446873.

Chessa, A., Morescalchi, A., Pammolli, F., Penner, O., Petersen, A.M., Riccaboni, M., 2013. Is Europe Evolving Toward an Integrated Research Area? Science 339, 650–651. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1227970, doi:10.1126/science.1227970.

Culotta, E., 2017. People on the move: The science of migrations. Science URL: http://www.sciencemag.org/news/2017/05/people-move-science-migrations, doi:10.1126/science.aan6884.

Deville, P., Wang, D., Sinatra, R., Song, C., Blondel, V.D., Barabási, A.L., 2014. Career on the move: Geography, stratification, and scientific impact. Scientific reports 4, 4770.

Fortunato, S., Barthelemy, M., 2007. Resolution limit in community detection. Proceedings of the National Academy of Sciences 104, 36–41. URL: www.pnas.orgcgidoihttp://www.pnas.org/cgi/doi/10.1073/pnas.0605965104, doi:10.1073/pnas.0605965104.

Franzoni, C., Scellato, G., Stephan, P., 2012. Foreign-born scientists: mobility patterns for 16 countries. Nature Biotechnology 30, 1250.

Franzoni, C., Scellato, G., Stephan, P., 2014. The mover's advantage: The superior performance of migrant scientists. Economics Letters 122, 89–93.

Franzoni, C., Scellato, G., Stephan, P., 2018. Context Factors and the Performance of Mobile Individuals in Research Teams. Journal of Management Studies 55, 27–59. doi:10.1111/joms.12279.

Geuna, A., 2015. Global mobility of research scientists: The economics of who goes where and why. Elsevier, Academic Press.

Giatsidis, C., Thilikos, D.M., Vazirgiannis, M., 2013. D-cores: measuring collaboration of directed graphs based on degeneracy. Knowledge and Information Systems 35, 311–343. doi:10.1007/s10115-012-0539-0.

Girvan, M., Newman, M.E.J., 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America 99, 7821–6. doi:10.1073/pnas.122653799.

Graf, H., Kalthaus, M., 2018. International research networks: Determinants of country embeddedness. Research Policy 47, 1198–1214. doi:10.1016/j.respol.2018.04.001.

Hoekman, J., Frenken, K., Tijssen, R.J., 2010. Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. Research Policy 39, 662–673. doi:10.1016/J.RESPOL.2010.01.012.

Jackson, M.O., Rogers, B.W., 2007. Meeting Strangers and Friends of Friends: How Random Are Social Networks? American Economic Review 97, 890–915. doi:10.1257/aer.97.3.890.

Mokyr, J., 2016. A culture of growth: the origins of the modern economy. Princeton University Press.

Newman, M.E.J., 2001. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences 98, 404–409. doi:10.1073/pnas.98.2.404.

Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Physical Review E 69, 026113. doi:10.1103/PhysRevE.69.026113.

OECD, 2017. OECD Science, Technology and Industry Scoreboard 2017. OECD Publishing, Paris.

Ozden, C., Rapoport, H., 2018. Cross-country perspectives on migration and development: Introduction. The Economic Journal .

Page, L., Brin, S., Motwani, R., Winograd, T., 1998. The PageRank Citation Ranking: Bringing Order to the Web. World Wide Web Internet And Web Information Systems 54, 1–17. doi:10.1.1.31.1768.

Sassen, S., 2016. The global city: Strategic site, new frontier, in: Managing Urban Futures. Routledge, pp. 89–104.

Saxenian, A., 2005. From Brain Drain to Brain Circulation: Transnational Communities and Regional Upgrading in India and China. Studies in Comparative International Development 40, 35–61. doi:10.1007/BF02686293.

Serafinelli, M., Tabellini, G., 2017. Creativity over time and space .

Solimano, A., 2008. The international mobility of talent: Types, causes, and development impact. volume 394. Oxford University Press Oxford.

Taylor, P.J., Derudder, B., 2015. World city network: a global urban analysis. Routledge.

Torvik, V.I., 2015. MapAffil: A Bibliographic Tool for Mapping Author Affiliation Strings to Cities and Their Geocodes Worldwide. D-Lib Magazine 21. doi:10.1045/november2015-torvik.

Torvik, V.I., Smalheiser, N.R., 2009. Author name disambiguation in MEDLINE. ACM Transactions on Knowledge Discovery from Data 3, 1–29. URL: http://portal.acm.org/citation.cfm?doid=1552303.1552304, doi:10.1145/1552303.1552304.

Traag, V., 2017. vtraag/louvain-igraph: 0.6.1. doi:10.5281/zenodo.1054103.

Wagner, C.S., Leydesdorff, L., 2005. Network structure, self-organization, and the growth of international collaboration in science. Research Policy 34, 1608–1618. doi:10.1016/J.RESPOL.2005.08.002.

Zucker, L.G., Darby, M.R., 2007. Star scientists, innovation and regional and national immigration. Technical Report. National Bureau of Economic Research.

Appendix

	Year	Affiliation City	PubMedID
1	2003	Stony Brook, NY, USA	12703729
2	2003	Stony Brook, NY, USA	12595470
3	2005	Kansas City, KS, USA	15936007
4	2005	Stony Brook, NY, USA	15791955
5	2005	Stony Brook, NY, USA	15944300
6	2005	Milwaukee, WI, USA	16299285
7	2007	Milwaukee, WI, USA	17311921
8	2007	Milwaukee, WI, USA	17490406
9	2008	Boston, MA, USA	18566416
10	2008	Stony Brook, NY, USA	18591234

Table 9: Example of career path of a specific author (Zhang Y.). For each record we have the year of publication, the city of the affiliation and the relative PubMed ID identifying the paper

Asymmetric mobility D-core

To further understand the diversity in the exchange between cities we look at the D-core decomposition (Giatsidis et al., 2013).

The D-core analysis (directed generalization of the k-core) allows us to analyze simultaneously the centrality and "coreness" of a city while taking the asymmetric nature of global mobility into account (i.e. the cities feeding scientists to a given location are not the same they receive scientists from). This algorithms instead of a list of k-Core, yields a matrix of (in-degree; out-degree)-cores, see Figure 11.

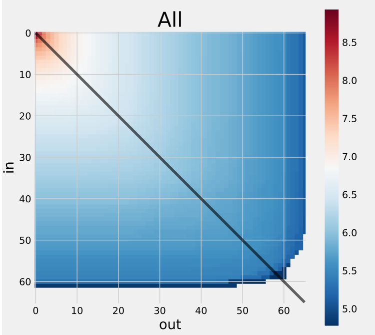
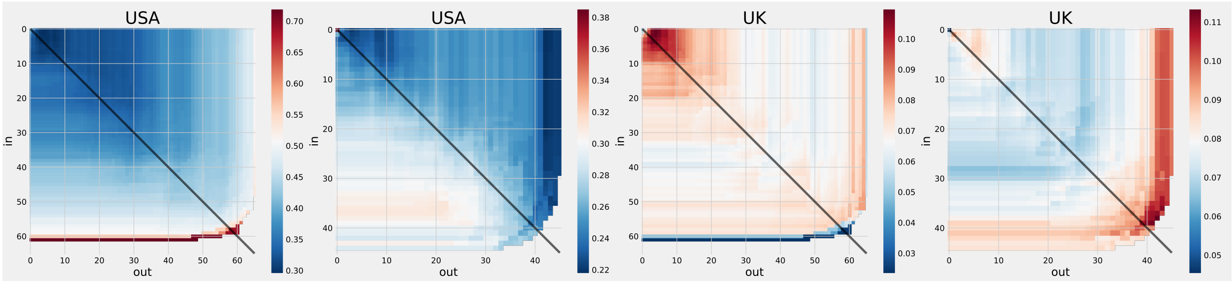


Figure 11: Every square at coordinate (k; l) corresponds to the (in; out)-core of the mobility network in 2004. The color indicates the number of cities in the D-core (in $\log_{10}$).

In the All matrix (see Figure 11) the squares have been colored according to the size of the core and the rest according to the proportion of cities therein contained belonging to that country (see. Figure 12). We observe in the "All" matrix, that the D-core decomposition is skewed toward having high out and low in-degrees, as can be seen by the slightly brighter blue on the right frontier than on the bottom portion (high in, low out). Comparing this result with Figure 12 for the USA, we see that US cities are represented more in the high in, low out portion of the plot, the opposite of what we would expect if all cities were distributed according to "All". This means that most US cities which are in the most central cores have more cities feeding into them than they are feeding. In other words central US cities on average source from a wide variety of cities but their scientists move to a less diverse set of cities.

Since the size of each country influences the number of cities they contain it could be the case that what we observe represents mainly the national configuration of cities. To explore this idea further we can look at how the D-core looks like if we remove all national connections. In other words we leave for each city only its international connections such that the D-core they are part of is only induced by being part of an international network (see right side of Figure 12). We observe that the USA is still marked by a strong presence in cores with stronger IN degree than OUT. This suggests that US cities sources from a wide variety of cities, but are the origin of moves to a more restricted set of cities.



(a) (b) (c) (d)

Figure 12: D-core profiles for USA (a-b) and UK (c-d). The coordinate $(k;l)$ corresponds to the $(k;l)$-core of the mobility network. The matrices show the proportion cities belonging to the US contained within a given $(k;l)$-core. The color scale has been adjusted such that the average proportion across all D-cores is white. Complete network (a, c); Only international moves (b, d). Note: the transparent cells on the border belong to empty $(k;l)$-cores.

National Border Effects

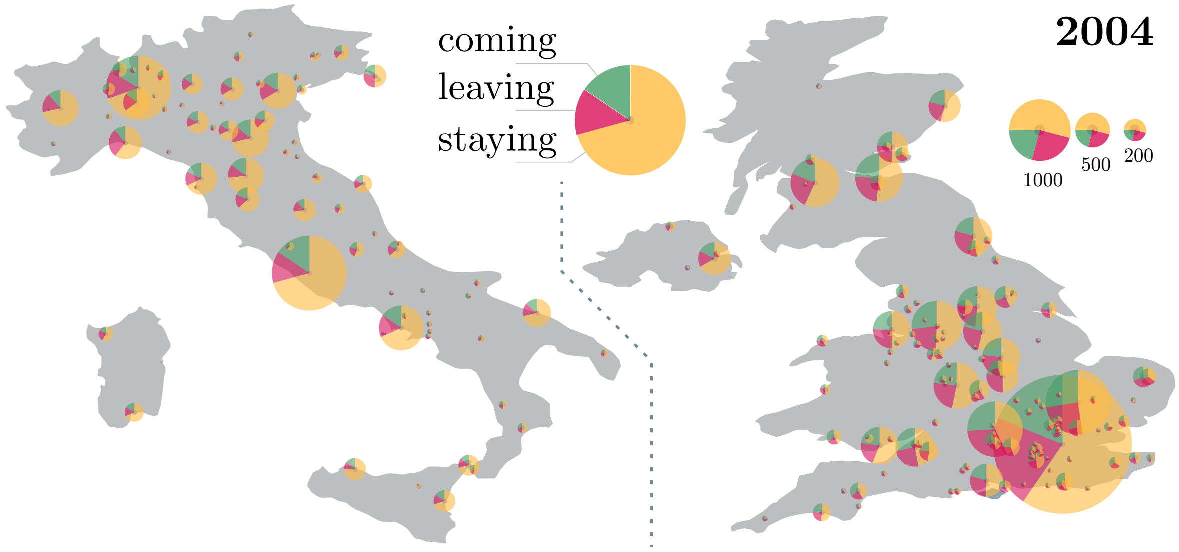


Figure 13: These two maps show the scientists population for the UK and Italy, which have at least 30 scientists stationary there in the period 1999 to 2008 (move year 2004). The pie chart indicates the proportion of scientist which are incoming, leaving and or staying. The size of the pie-char is proportional to the sum of all three types.

Table 10: Breakdown of communities by country. For each modularity class the number of cities belonging to a country are listed along with their proportion of cities in the class. For example community 15 is composed of 25% Swedish, 22% Finnish, 20% Norwegian and 16% Danish cities as well as 17% smaller cities. Note: only countries with at least 5% member cities are listed.

Community	Country, c	% of cities of country c in the community	No. Cities
1	USA	92%	392
2	USA	91%	161
3	Spain	27%	65
	Mexico	22%	53
	Argentina	11%	27
	Chile	6%	14
	USA	6%	15
4	USA	94%	355
5	USA	98%	197
6	USA	85%	197
7	France	57%	225
	Belgium	10%	40
8	Germany	69%	444
	Switzerland	14%	88
	Austria	7%	46
9	Russia	45%	53
	Taiwan	21%	25
	USA	20%	23
	Ukraine	9%	10
10	USA	95%	151
11	USA	91%	192
12	Australia	41%	81
	USA	12%	24
	Thailand	11%	22
	New Zealand	10%	20
13	Czech Republic	36%	39
	Croatia	16%	17
	Slovenia	12%	13
	Serbia	11%	12
	Slovakia	9%	10
	USA	6%	6
14	Netherlands	60%	110
	Belgium	14%	26
15	USA	51%	41
16	Sweden	25%	86
	Finland	22%	75
	Norway	20%	69

Community	Country, c	% of cities of country c in the community	No. Cities
	Denmark	16%	55
17	Korea	88%	43
	USA	10%	5
18	USA	89%	71
19	USA	98%	122
20	Japan	81%	91
21	Canada	54%	127
	Iran	11%	25
	Saudi Arabia	9%	22
	USA	8%	19
	Egypt	7%	17
22	Greece	80%	35
	Jordan	9%	4
	Cyprus	7%	3
23	Brazil	58%	55
	Portugal	21%	20
24	Israel	70%	39
	USA	21%	12
25	China	70%	171
	Hong Kong	7%	17
	Malaysia	7%	18
26	Poland	92%	57
27	UK	66%	315
28	Turkey	91%	53
29	Italy	75%	121
	USA	6%	10
30	USA	83%	84
31	Nigeria	47%	34
	South Africa	39%	28
32	India	73%	122
	USA	8%	14

Regression Analysis

	Pr(move)		k-core destination	
k-core source	-0.000362	(-0.45)	0.0128	(0.77)
productivity	0.0218***	(4.68)	4.313***	(6.57)
inter-move	0.109***	(64.92)	-2.583***	(-6.77)
2002	0.0335**	(3.12)	10.62***	(10.07)
2004	0.0452	(1.72)	29.04***	(16.70)
2002 × productivity	0.00283	(0.88)	0.943**	(3.26)
2004 × productivity	-0.00414	(-1.12)	1.543***	(3.59)
log(km distance)			5.301***	(5.43)
Pr(move other fields)	1.510***	(6.12)		
Constant	-1.711***	(-3.89)	118.5***	(6.79)
Origin country effects	Yes		Yes	
Age effects	Yes		Yes	
Field effects	Yes		Yes	
Observations	1,363,280		433,023	
tanh(ρ)	-0.14 (-4.60)	log(σ)	4.46 (39.80)	
Log pseudo-likelihood	-3,193,349			

t statistics in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 11: Heckman two stage regression for the centrality measure "k-core". The standard errors have been clustered at source city for the first stage (Pr(move)) and on the destination city for the second stage (i.e. the centrality of the destination)

	Pr(move)		norm. degree	
norm. degree source	-0.574	(-0.38)	0.0174	(0.92)
productivity	0.0217***	(4.35)	0.00249***	(6.53)
inter-move	0.109***	(64.91)	-0.00126***	(-6.05)
2002	0.0294***	(6.15)	0.000775*	(2.17)
2004	0.0352***	(4.12)	0.00310***	(5.81)
2002 × productivity	0.00261	(0.83)	0.000194	(1.21)
2004 × productivity	-0.00470	(-1.39)	0.0000164	(0.07)
log(km distance)			0.00241***	(4.30)
Pr(move other fields)	1.534***	(6.33)		
Constant	-1.733***	(-3.98)	0.0498***	(5.12)
Origin country effects	Yes		Yes	
Age effects	Yes		Yes	
Field effects	Yes		Yes	
Observations	1,363,280		433,023	
tanh(ρ)	-0.14 (-4.64)	log(σ)	-3.117 (-31.04)	
Log pseudo-likelihood	-75,548.17			

t statistics in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 12: Regression results for Mobility and relocation choice. This are the results of the Heckman two stage regression for the centrality measure "normalized degree". The standard errors have been clustered at source city for the first stage (Pr(move other fields)) and on the destination city for the second stage (i.e. the centrality of the destination)

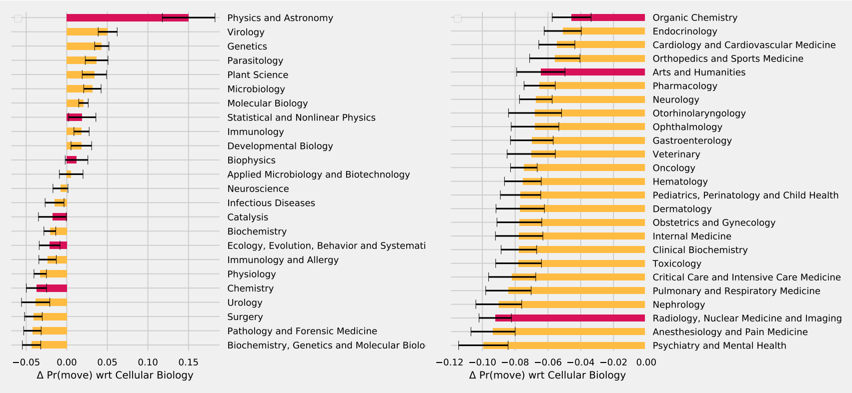


Figure 14: Marginal probability to move compared to Cellular Biology (i.e. the largest field). The 95% confidence interval is illustrated as black bars.

IMT
SCHOOL
FOR ADVANCED
STUDIES
LUCCA

Piazza San ponziano 6, 55100 Lucca, Italy.
www.imtlucca.it