Research Policy 50 (2021) 104127

Contents lists available at ScienceDirect

Research Policy

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

Talent goes to global cities: The world network of scientists' mobility

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

[a] Chair of Systems Design ETH Zurich Weinbergstrasse 58, Zurich 8092, Switzerland
[b] IMT School for Advanced Studies Lucca, Piazza S. Francesco 19, Lucca 55100, Italy

ARTICLE INFO

JEL classification:
F22
F66
J61
O18
O15
O30
R12

Keywords:
Scientist mobility
Brain circulation
Global cities
Network effects
Network analysis

ABSTRACT

Global cities boast higher rates of innovation as measured through patent and scientific production. However, the source of the location advantage of innovation hubs is still debated in the literature, with arguments ranging from localized knowledge spillovers to network effects. Thanks to an extensive data set of individual scientist career paths, we shed new light on the role of scientist location choices in determining the superior innovative performance of global cities. We analyze the career paths of around two million researchers over a decade across more than two thousand cities around the globe. First, we show that scientists active in global cities are more productive in terms of citation weighted publications. We then show that this superior performance is in part driven by highly prolific scientists moving and remaining preferentially in global cities, i.e., central cities in the international scientist mobility network. The overall picture that emerges is that global cities are better positioned to attract and retain prolific scientists than more peripheral cities.

1. Introduction

Modern economies require highly skilled labor to sustain growth and keep their competitive advantage (Chambers et al., 1998; Ozden and Rapoport, 2018; Solimano, 2008; Verginer and Riccaboni, 2020; Zucker and Darby, 2007). High skill labour is especially important to cities that are home to a large portion of the world's population.[1] According to Bettencourt et al. (2007) and Schlapfer et al. (2014) global cities have higher rates of innovation in terms of patents and scientific production and, more recently, Belderbos et al. (2020) have found that half of all inventions have their origin in global cities. At the same time, scientists and inventors are highly mobile individuals, especially star inventors (Zacchia, 2018) and talented scientists in the early phase of their careers (Azoulay et al., 2017). High rates of mobility are not a new phenomenon (Cardwell, 1972; Mokyr, 2016; Serafinelli and Tabellini, 2017), but its size has increased in a globalized market for advanced human capital (Culotta, 2017; Geuna, 2015; OECD, 2017).

The confluence of these two trends: the superior innovation performance of global cities and sustained mobility of scientists begs the questions if and to what extent the latter boosts the former. Many factors might reasonably enhance knowledge production in large urban areas ranging from localized knowledge spillovers (Boschma, 2005; Cantwell and Piscitello, 2005; Feldman, 1999; Jaffe et al., 1993) to network effects (Agrawal et al., 2006; Almeida and Kogut, 1999; Alnuaimi et al., 2012; Breschi and Lenzi, 2016; Breschi and Lissoni, 2009). In this work, we concentrate on the effect of scientists' mobility on the performance of cities in terms of citation-weighted scientific production. We pinpoint one of the main drivers of the superior scientific performance of global cities which take advantage of their central position in the scientist mobility network. Specifically, we argue that scientist location choices can contribute to increasing returns to the centrality of cities in the mobility network: more prolific scientists gravitate towards global cities, which in turn generate a disproportionate share of the most impactful scientific production.

One of the main reason for the limited research on the causes and consequences of scientists' mobility is data availability. The main challenge is to trace individual movements of scientists in space and time on a global scale. In this work, we rely on scientific publications, the most direct and high-frequency signal of scientific activity, to quantify scientific output and to track scientists' mobility. Specifically,

* Corresponding author.
E-mail address: lverginer@ethz.ch (L. Verginer).

[1] 68% of the world population is projected to live in urban areas by 2050 (UN, 2018).

https://doi.org/10.1016/j.respol.2020.104127
Received 23 July 2019; Received in revised form 1 September 2020; Accepted 6 September 2020
Available online 21 September 2020

we extract the affiliations listed on published papers by 3.7 million disambiguated authors to reconstruct the global scientist mobility network across 189 countries and 7159 cities. We also leverage citation data, along with several other controls, to estimate the impact of mobility on scientific production. We observe an increase in the mobility of scientists over the past decade on a global scale, with significant differences across scientific fields (e.g., physicists are more mobile than physicians), countries (see also Verginer and Riccaboni (2020) on this) and career phases (similarly to Azoulay et al. (2017)). By using a Heckman selection model of the location choices of scientists, we find that global cities benefit from their central position in the international mobility network by attracting and retaining more prolific scientists early on in their careers.

The rest of this work is structured as follows. Section 2 surveys the extant literature on scientist mobility and introduces the specific hypotheses we will explore. Then in Section 3 we describe the dataset we use as well as the methodology we developed to trace the mobility of scientists based on bibliometric data. Section 4 illustrates the results of the empirical analysis, and finally in Section 5 we discuss our contribution to the literature and the policy implications of our findings.

2. Theoretical background and research hypotheses

Cities are central to national economic activity and are part of a global network of interconnected urban areas. The celebrated works by Jacobs (1969, 1984) illustrate how cities function as cauldrons of culture, creativity and economic activity (Sassen, 2016; Taylor and Derudder, 2015). Similarly, Florida (2005) argues that cities play a central role in 21st-century creative capitalism and thus deserve more attention. Cities collaborate and compete in various fields for resources and human capital (Belderbos et al., 2020). The widely cited work by Bathelt et al. (2004) outlines that local activities are essential for innovation ("local buzz") but are moderated by global interactions ("global pipelines", i.e., long-distance connections). A result that has also been shown empirically in Scholl et al. (2018). The observation that cities are hubs of scientific production embedded in a global network, or ecosystem, of cities, is the central theme of this work.

Glaeser (1999) and Bettencourt (2013) argue that the creation and concentration of know-how in cities increase their attractiveness for highly skilled and creative individuals. Some evidence in support of the fundamental idea that mobility plays a crucial role in science and innovation comes from Breschi et al. (2017); Fink et al. (2017); Franzoni et al. (2012, 2014) and Franzoni et al. (2018), among others. In their seminal work, Jaffe et al. (1993) show that patent citations are up to six times more likely to be between patents in the same urban area than would be expected from a control set of patents. Breschi and Lissoni (2009) show that the proximity effect on patent citations gets smaller but still significant when controlling for the social network of inventors. Both social and geographical proximity increase the probability of knowledge flows (Agrawal et al., 2008). In particular, Almeida and Kogut (1999) report that inventor mobility plays a fundamental role in patent citations and knowledge flows. This finding is corroborated by Agrawal et al. (2006) who found that patent citations come disproportionately from inventors' prior locations. Even though most of the research done so far has used patent citations as a proxy of knowledge flows, we argue that the same effects should apply to the citations to scientific papers (Pan et al., 2012). Therefore, scientists located in global cities are likely to receive more citations. First, spatial proximity will increase the likelihood to be cited. Second, centrality in the scientist mobility network of global cities magnifies the social proximity effect increasing the visibility of scientific production at a distance.

For these reasons, we maintain that scientists in global cities attract more citations, and propose our first hypothesis.

Hypothesis 1 Scientists in global cities attract more citations than peers in peripheral cities (**H1**).

Many possible mechanisms could reasonably explain the leading role of global cities, i.e., central cities in the network of knowledge flows. For instance, spatial and social proximity of researchers makes it more likely that their research is cited, access to more and possibly better research institutions, the presence of multinational companies, dense local networks and research infrastructures contribute to scientific success. On top of this, we propose that a significant part of the productivity gap may be driven by scientist mobility, especially the mobility of most prolific scientists. Precisely, we maintain that global cities, as measured by their centrality in the world network of scientist mobility, preferentially attract and retain more prolific scientists, which in turn contribute to the superior scientific performance of global cities.

Only recently, thanks to the availability of geo-referenced data about patents (Li et al., 2014; Morrison et al., 2017) and scientific publications (Catini et al., 2015; Torvik and Smalheiser, 2009) with disambiguated individual inventors and scientists, it has been possible to investigate the impact of spatial mobility on individual productivity. Recent empirical works have revealed the importance of scientific productivity as a positive predictor of mobility (Azoulay et al., 2017; Ganguli, 2015).

As in any labour market, there is a demand-side effect, i.e., cities attract researchers, and a supply-side effect, i.e., scientists apply for positions. Among the supply side factors (Moretti and Wilson, 2017) an extensive literature has stressed the importance of the local scientific community in terms of proximity, knowledge sharing and localized spillovers (Azoulay et al., 2017). In this respect, global cities offer an attractive local scientific community and better access to knowledge networks. On the demand side, scientific productivity plays a crucial role in determining which scientist will successfully move and remain in global cities.

In other words, the scientific community of global cities encourages more prolific scientists to move there, thus inducing a virtuous cycle, whereby success begets success.

Moreover, prolific scientists who are already working in central cities are better positioned to exploit existing social ties within and between cities, and therefore they are more likely to remain. If this is the case, it would suggest that beyond attracting more productive scientist, global cities offer better career prospects. All in all, the combined effect of supply-side and demand-side factors make talented scientists more likely to gravitate toward global cities.

We formalize this idea with the following hypotheses.

Hypothesis 2.1 More prolific scientists, as measured by citation weighted scientific output, are more likely to move to global cities (**H2.1**).

Hypothesis 2.2 More prolific scientists are more likely to remain in global cities (**H2.2**).

Hypothesis H1, H2.1 and H2.2 will be addressed in Section 4. Before that, in the next section, we introduce the dataset we built up and the methodology we developed to trace scientist mobility and to reconstruct the international mobility network between cities.

3. Data and methodology

The main hurdle in the study of mobility patterns at individual and city levels has been data availability. The authoritative manual on the "Global Mobility of Research Scientists" gives an overview of the 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 with their careers" Geuna (2015, Ch.5, p.24).

Previous research on the mobility of scientists has used, among other approaches, large-scale surveys (Franzoni et al., 2012; 2014; 2018) and, more recently, large bibliographic datasets (Bohannon and Doran, 2017; Deville et al., 2014; Graf and Kalthaus, 2018; Vaccario et al., 2020).

For our analysis, we need data to reconstruct global scientist mobility

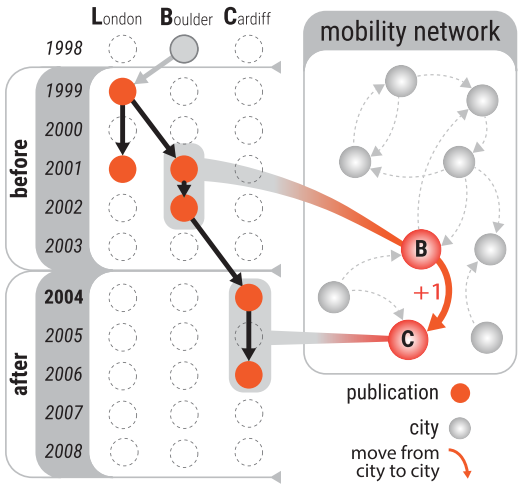

Fig. 1. Extracting the mobility network from Medline papers. The papers by an author i are shown as a sequence of red circles from top to bottom. Each publication has a date (in rows) and a city (in columns). We take a buffer time of 5 years before and after 2004. Here, we identify Boulder as the origin city, since it is where the longest sequence of publications has been produced in the buffer period and closest to 2004. The target city of the move is Cardiff since it is the only observed city in the buffer period after 2004. The move between Boulder and Cardiff is then added to the mobility network by increasing the corresponding edge weight by one unit. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

between cities. Moreover, we need data on the "impact" of publications authored by researchers in those cities. Therefore, we collect publications from Medline, disambiguate authors using Author-ity (Torvik and Smalheiser, 2009), assign locations using MapAffil (Torvik, 2015) and count citations by combining data from Microsoft Academic Graph, Pubmed Central and AMiner. We also include some country-level control variables from CEPII (Mayer and Zignago, 2011). All of these datasets are available for research for free, either publicly or upon request from the relative authors. A detailed description of the datasets and how they have been processed is available in Appendix A.1.

By merging these datasets, we can identify an author across publications uniquely. An example of a career trajectory for a specific author is provided in Table A1. A central problem in studying cities is finding a good definition of their boundaries (Rozenfeld et al., 2011; Zipf, 1949). In this study, we rely on the definition of a "location" provided by Google Maps. A detailed description of how we identify and define cities is available in Appendix A.3.

3.1. Tracing a move

We reconstruct the mobility of scientists based on the changes in their affiliations in published papers. We rely on the track record of papers by an individual i over time to identify the sequence of cities the author passed through. Note that, an author may have multiple publications in the same year possibly with different affiliations, an example of which is available in Table A1 in the Appendix. From these records, we propose a method to extract moves. We define a move as a change of the city where an author was located in a period before a given year (t) (i.e. the move year) and afterwards.

More precisely, to determine the source and target cities of a move, we choose a candidate move-year (t) and a buffer of b years around it (see Fig. 1). To obtain from a series of papers a single move, we apply the following procedure. We select a "move year" t and a given number of years *before* $[t-b, t)$ and *after* $[t, t+b)$ year t. We identify the location of author i in the two periods $[t-b, t)$ and $[t, t+b)$. When the locations differ, we count a move.

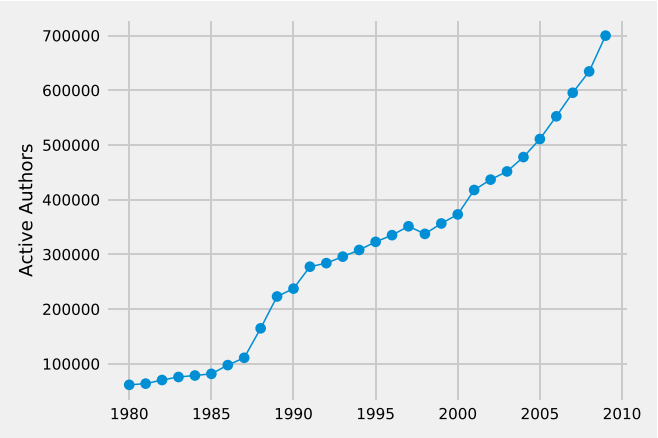

Fig. 2. Number of unique active authors identified in Author-ity per year.

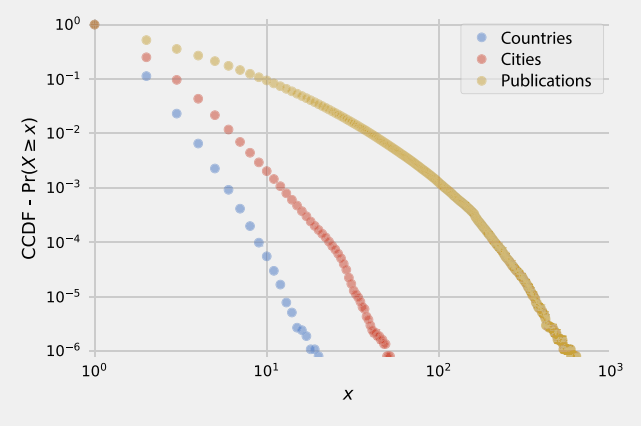

Fig. 3. Counter Cumulative Distribution of countries, cities and publications per author, double log scale. Each data point shows the probability to observe at least x unique countries, cities and publications for a given author (i.e. $\Pr(X \geq x)$).

To select a unique starting location in the period $[t-b, t)$ we take the longest uninterrupted sequence of locations closest to t. Consider, as an example, the publication sequence in Fig. 1. In this example, we have the publication sequence $\{B_{1998}, L_{1999}, L_{2001}, B_{2001}, B_{2002}, C_{2004}, C_{2006}\}$, move year $t = 2004$ and a buffer of $b = 5$ years. The capital letter indicates the city and the index the year of a publication. To determine the origin city, we look at all papers in the interval [1999, 2004) and choose the city with the longest sequence of publications closest to 2004. In this example, the author has published three papers in B, but only two are within the [1999, 2004) window. Similarly, there are two publications in L and one in the same year in B. According to our rule, we identify B as the origin of the move because it is closest to 2004. We choose B even though there are two papers in L and B.[2] Finally, the destination of the move is C since it is the only observed city in the time window [2004, 2009).

Iteratively applying this method to the career paths of all authors in the dataset yields a directed and weighted mobility network for a year t

[2] City L has been discarded as the origin city because it appears only in the year 1999 and city B also appears in that year. City L would have been chosen as a destination setting $t = 1998$ as the candidate move year and a buffer time of one year ($b = 1$). However, we intentionally applied a longer time buffer (five years) to increase the sensitivity of our method in the presence of multiple affiliations in a short period of time, that might be considered as false positives in the presence of double affiliations and publication lag times.

Fig. 4. The global mobility network of researchers. The map shows the network of scientist mobility in 2004 with a five-year time buffer. Only most common routes, with 50 or more moves between two cities, are reported.

and buffer *b*, where the direction reflects the direction of travel of researchers and the weight the number of individuals who moved between every pair of cities. Since disambiguated author names are available only up to 2009, and we need to have a buffer time of five years after the observation year, we limited our analysis up to the year 2004. As a total, we analyze 2,239,357 disambiguated author names for which geolocation data is available in the period 1990–2009.

Fig. 2 shows the evolution of the number of disambiguated authors over time. Fig. 3 shows the distribution of the number of cities, countries and number of publications a given author has been to or authored. All distributions are highly skewed (hence plotted in log-log scale) with a sharp decline for all values beyond 1. Around one-fourth of authors have been active in at least two different cities. Only one-tenth of researchers have been affiliated with institutions in two or more countries or three or more different cities in their careers or published at least eight papers. Similarly, only about 1% of authors have worked in at least three different countries or five cities, or published at least 38 papers.

In general, we find that the propensity to change cities and countries has been increasing. The proportion of mobile authors has increased from one in five in 1996 to one in four in 2004 (see Fig. A2 in the Appendix). Moreover the probability to leave the country (globally) increased from 10% in 1996 to more than 12% in 2004 (see Fig. A3 in the Appendix).

3.2. The global scientist mobility network

Based on the author moves, we traced with the method described above, we reconstructed the mobility network between cities from 1996 to 2004, with a time buffer of 5 years. The network experienced a constant growth of the number of cities (nodes) and mobility routes (links). An illustration of the mobility network for the year 2004 is shown in Fig. 4 (see also Table A2 in the Appendix for detailed statistics).

As robustness checks, we can increase the number of papers required in any given location before and after a move. This restriction might reduce the chance that a move is spurious (e.g. visiting periods or double affiliations). Similarly, we can reduce the buffer, thus requiring authors to have fewer gaps in their publication sequences. This restriction would, however, drop scientists not publishing at least once in the two periods before and after the move.

3.3. Defining global cities

In this work, we define as global cities, those cities which are most central in the global scientist mobility network. We argue that the defining feature of a global city is that it is a "hub", i.e., a central city in the global network of urban areas. A simple network measure capturing the number of cities a city is connected to is the degree centrality. In our case, this corresponds to the number of cities scientists move to or from a given city.[3]

The working definition of a global city in this work is a city belonging to the top 10% of cities by degree centrality (about 200 locations). Fig. 5 shows the relative ranking of the top 50 cities by degree centrality as measured in 2004, along with changes in rank over the period. We note that the ranking among the top 10 cities is relatively stable and that most of the top 50 locations are US cities. The remaining 20 cities in Fig. 5 are located in the UK (3), Canada (3), Germany (3), Japan (2), Australia (2), Sweden (1), France (1), Korea (1), the Netherlands (1), Spain (1), Switzerland (1) and China (1). The most relevant change in the ranking is the rise of Beijing since 2002 (indicated in red in Fig. 5).

More sophisticated centrality measures to identify hubs are available to take into account directionality, the weight of links and non-local network properties. Measures such as PageRank, betweenness centrality and K-Core would be viable alternatives to the chosen degree centrality measure. However, these network centrality measures have high Pearson and Kendall rank correlations (above 0.9). Therefore, in the interest of clarity, we use in this work the more straightforward degree centrality measure.[4]

3.4. Measures of scientific output

To estimate the quality of the scientific production, we augment the publication history with forward citation data obtained by merging Microsoft Academic Graph, Pubmed Central and AMiner. These datasets combined contain the reference lists of 15,541,158 Medline papers.

We are aware that the number of citations is not a perfect measure of scientific relevance or importance. Still, citations are widely used for performance evaluations of research centres and universities. Therefore citation-based impact measures are also likely to be relevant for hiring

[3] Specifically, in a directed network, there are two types of degrees: in-degree and out-degree. The degree centrality of a city is the total number of directly connected cities in whichever direction.

[4] As a robustness check in the Appendix, we report the regression analysis results for the PageRank centrality measure (Table A.6). Results for betweenness and k-core centrality are also available upon request. Not surprisingly, when used in the regressions analysis, alternative centrality measures yield very similar results.

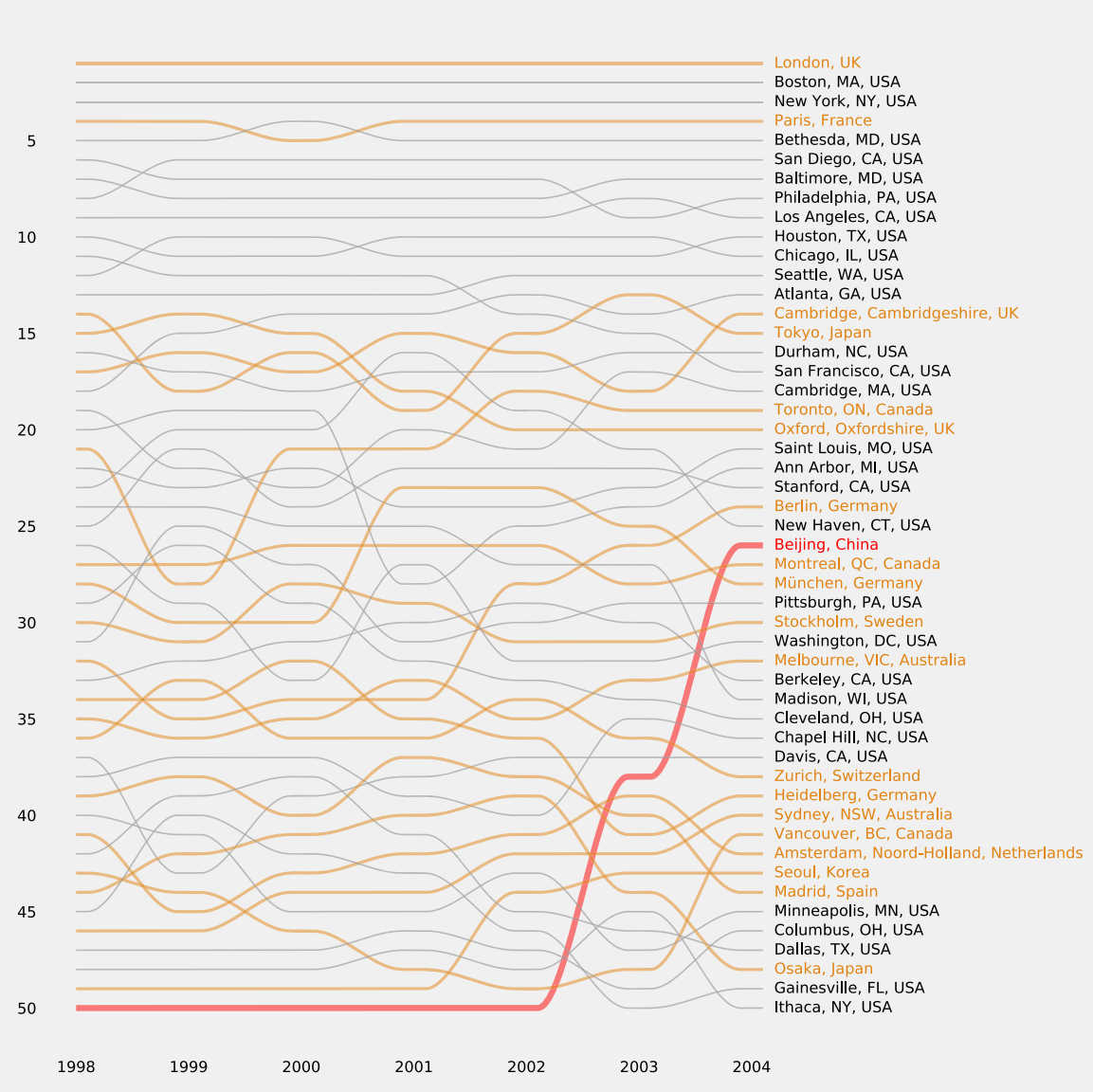

Fig. 5. Ranking plot of most central cities from 1998 to 2004 as measured by their Degree Centrality on the global mobility network. The plot shows the evolution of the relative ranks for the 50 most central cities as measured in 2004. US cities are listed in Black, non-US cities in orange and the rise of Beijing as a global city is highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and other research evaluations.

To compute the citation weighted publications *for a scientist* we first collect her publications at a given location. We then look at the citations each of these papers has collected in the five years after publication dividing it by the number of authors (i.e., fractional count) and summing them up. This sum becomes the *citation weighted scientific output* of individual i in city c.[5]

Similarly, at *city level* we compute the citation-weighted output by obtaining all publications listing that city as an affiliation in the relevant period. This period is five years as it is the buffer for the "before" and "after" periods we used to reconstruct the mobility network. We apportion the citations proportional to the contributing authors working in that city. For example, for a paper with 9 citations obtained within 5 years after publication and 2 scientists active in city A and 1 scientist in city B, we apportion $9/3*2 = 6$citations to A and $9/3*1 = 3$ to B.

[5] Note that this impact is only fully revealed at the time of the publication, thus at t the impact of the scientific production of a scientist might not be fully known to the scientific community and the hiring institutions.

3.5. Additional variables and controls

For the sake of completeness, below, we list all the variables we will be using in our analysis with a short description. In the panel regression at the city level, we will be using the following variables.

MedianCitations The median citations per scientist of city c at time t, computed according to Section 3.4, i.e., the median of the fractional count of citations obtained up to 5 years after publication by all scientists active in city c in the period $[t, t+b)$.
DegreeQuantile The quantile of the degree centrality of city c in year t. In the regression Table 1 we use deciles (10% steps) and for the marginal effects in Fig. 6 we use ventiles (5% steps).

For the analysis of the determinants of the location choice of scientists, we will be using the following variables and controls.

ln(Citations) Natural log of citations received at most five years after publication for papers published before the move. The citations are

Table 1
Quantile regression for the median citations per scientist as a function of the centrality of a city.

	Median citations per scientist ($MedianCitations_{ct}$)			
	All	Stayers	Arrivals	Leavers
City degree centrality deciles (basline 1–10%):				
11-20%	0.0155	–0.0150	–0.0788	–0.0521
	[–0.0539,0.0849]	[–0.112,0.0822]	[–0.205,0.0477]	[–0.216,0.111]
21-30%	0.103**	0.0666	-0.0829	–0.107
	[0.0261,0.179]	[–0.0399,0.173]	[–0.211,0.0451]	[–0.278,0.0632]
31-40%	0.191***	0.140*	-0.00277	–0.0976
	[0.103,0.280]	[0.0231,0.258]	[–0.142,0.136]	[–0.267,0.0721]
41-50%	0.196***	0.175**	-0.0622	–0.103
	[0.102,0.290]	[0.0582,0.293]	[–0.208,0.0838]	[–0.276,0.0710]
51-60%	0.326***	0.275***	0.0627	–0.0443
	[0.223,0.429]	[0.148,0.401]	[–0.0812,0.207]	[–0.214,0.125]
61-70%	0.393***	0.312***	0.171*	0.173*
	[0.285,0.500]	[0.182,0.443]	[0.0253,0.317]	[0.000687,0.344]
71-80%	0.505***	0.422***	0.315***	0.213*
	[0.393,0.617]	[0.285,0.559]	[0.161,0.468]	[0.0469,0.380]
81-90%	0.723***	0.667***	0.619***	0.429***
	[0.583,0.864]	[0.507,0.827]	[0.455,0.784]	[0.256,0.602]
91-100%	1.140***	1.119***	1.157***	0.815***
	[0.977,1.303]	[0.932,1.307]	[0.974,1.341]	[0.640,0.990]
Constant	0.885***	0.925***	1.343***	1.541***
	[0.808,0.962]	[0.831,1.019]	[1.225,1.460]	[1.401,1.681]
Year (t)	Yes	Yes	Yes	Yes
City (c)	Yes	Yes	Yes	Yes
Num. of obs.	13,687	13,687	13,687	13,687

95% confidence intervals in brackets. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

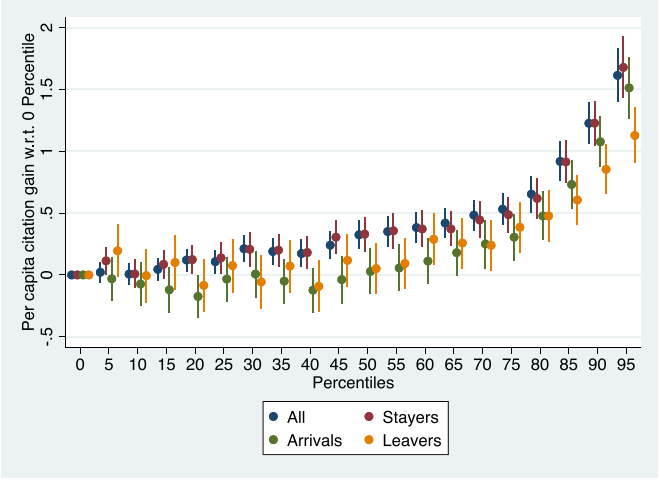

Fig. 6. Marginal effects on the average citations per scientist by percentiles of the degree centrality of a city in the scientist mobility network (baseline is the bottom percentile). Scientists are grouped by type: *Stayers* are scientists who were active in the city in two consecutive 5-year time windows, *Arrivals* are scientists who were not present in the city in the first five year period but arrived in the second period, *Leavers* are scientists active in a city in the first period who moved elsewhere in the second period, *All* are all scientists active in a city in the second period (Stayers and Arrivals). Scientists in the top 5% of the city centrality distribution collect around 1.5 more citations than colleagues in the bottom 5% of the distribution.

fractional, i.e., a paper with two citations and two authors means that each author gets one citation.
ln(Papers) Natural log of papers published before the move.
Age Group The Age group to which the author belongs, these are cohorts of 1 year up to 11 and larger cohorts after that. Age is measured as the number of years since the first publication by the author.
Field The Field of research the author is active in. The fields are the classification of journals in SCImago, and in case of ambiguity, the most common field is chosen.
ln(km distance) Natural log of distance in km between the origin city and destination city. The distance is computed using the Haversine formula.
SameCountry If the move is domestic, i.e., within the same country, then the value is 1 and 0 otherwise.
SameLanguage If the move was to a different country, i.e. SameCountry=0, then this value is 1 if the two countries share their official language (data from CEPII).
P(Others Move) Probability of authors in a given Field to move, excluding all authors in the same city working in the same field as the focal author.
BornThere If the author has his first affiliation in this location, then the value is 1.
Observation Gap The number of years between the last publication before the move and the first after the move.
Year The year the move took place.
Origin Country The country the origin city is located in.
OC Degree The degree centrality of the origin city (OC).
TC Degree The degree centrality of the target city (TC).
OC Size Number of scientists observed in the origin city (OC).

In the appendix we report the main summary statistics (Table A3) and the correlation Table A4 for all the variables.

4. Results

To test our research hypotheses, we run two sets of regressions, one at the city level (Section 4.1) and another one at the individual scientist level (Section 4.2). First in Section 4.1, we test, through a panel regression approach, if the scientific output by researchers in global

cities attracts more citations (H1). Then in Section 4.2, we analyze the determinants of the decision by individual scientist to move to a global city (H2.1) and to stay there (H2.2).

4.1. Scientists in global cities attract more citations

We analyze the effect of the centrality of a city in the global mobility network on local scientific productivity through a fixed effect panel quantile regressions of the form.

$$\text{MedianCitations}_{ct} = \beta_0 + \beta_1 \text{DegreeQuantile}_{ct} + \delta_t + \gamma_c + u_{ct} \quad (1)$$

where δ_t and γ_c are year and city fixed effects, respectively.

We measure city centrality through degree centrality in the mobility network ($\text{DegreeQuantile}_{ct}$), as defined in Section 3.3, and scientific productivity as the median citations per scientist ($\text{MedianCitations}_{ct}$).[6] Citations are counted as described in Section 3.4. In the quantile regressions, the dependent variable $\text{MedianCitations}_{ct}$ is the median number of the fractional citations received in the five years after publication by the papers published in the period from t to $t+b$ by scientists in the city c, with t ranging from 1990 to 2004 and $b = 5$.

More specifically, We estimate four separate models for different groups of scientists:

1. *Stayers* are scientists who have been in a given city c in the five years before t and during the five years after t.
2. The *Arrivals* group encompasses scientists who arrived in city c in the five years after t.
3. The *Leavers* group, on the other hand, includes all the scientists, who left city c after t. Since $\text{MedianCitations}_{ct}$ is the median number of citations received by the scientific production of researchers in the period from t to $t+b$, the $\text{MedianCitations}_{ct}$ of *Leavers* measures the impact of their scientific production in the destination city, which is different from c.
4. The last group includes *All* scientists active in city c in the period from t to $t+b$. This group consists of *Stayers* and *Arrivals* and does not include the *Leavers* who are no longer active in city c in the period from t to $t+b$.

If indeed scientists in global cities, i.e., cities in the top degree centrality percentiles, attract more citations than scientists in more peripheral cities, we should find that the median citations per scientists are higher for the top quantiles of the city centrality distribution. Moreover, by comparing stayers and movers, we can explore the contribution of incoming, and out-going mobility flows to the scientific productivity of cities.

From the available 7159 locations in the period 1990–2009 only 2292 have more than five active scientists in any five years, implying that there are a lot of small cities. These locations, beyond being very small, rarely appear in the panel data, i.e., they lead to a strongly unbalanced panel. To address the problem of excessively small locations and the lack of observability, we set a minimum size for inclusion equal to 5. This number strikes a fair balance between not dropping locations which are relevant for the analysis, but on the other hand, does not lead to a highly unbalanced panel.[7]

Table 1 shows our estimates for the quantile regressions. Overall, we find that scientists in the top 10% most central cities (i.e., global cities) receives, on average, 1.14 more citations than scientists in peripheral cities (bottom 10% of the centrality distribution), providing support to H1. Not only are stayers in central cities more productive, but also scientists arriving in global cities are more productive than peers who moved to more peripheral cities. Even though the productivity of leavers is also increasing with the centrality of cities, scientists leaving global cities tend to be less productive than new hires.

To show the effect of centrality on citations visually, Fig. 6 reports the change in citations per scientist by degree centrality. In this plot, we consider 20 quantiles of the degree centrality distribution. We note that in line with H1, the most central cities, i.e., global cities, attract more citations per scientists than less central cities, with a positive net contribution by mobility flows.

4.2. The mobility of scientists

In the previous Section, we have seen that the citation weighted scientific output depends positively on the centrality of a city. Moreover, the results of the quantile regressions suggest that the superior scientific performance of global cities might depend, at least partially, on the location choices of prolific scientists, which tend to be attracted by global cities.

To formally test hypotheses H2.1 and H2.2, in this Section we proceed to estimate a model of the location choice of scientists, as a function of their productivity and the centrality of cities in the global mobility flows. A potential issue with estimating the location choice of mobile researchers is selection bias. Scientists with specific characteristics (e.g. productivity) might be more likely to move, thus introducing a bias in our estimations. A possible bias in the propensity to move affects our analysis since we cannot observe a change in location for a non-mobile scientist and by extension, a change in the centrality of the target city (TC). To address this issue, we use a Heckman selection model. The Heckman model consists of two stages. The first stage (the selection stage), estimates the propensity of a scientist to move. This regression is then used in the second stage to correct for the likelihood of inclusion. In the second stage, we then estimate a scientists' relocation choice conditional on observing a move. Summing up, in the first stage (the selection equation), we control for the probability of observing a move at any given moment in the scientist's career. In the second stage (the regression equation), individual-level scientific productivity at origin is used to predict the centrality of the destination city.

The two stages of the Heckman model allow us to test H2.1 and H2.2. More precisely, in the first stage, we test explicitly for the probability of moving. By adding an interaction term between the centrality of the origin city (OC) and the citations of the author, we test if prolific scientists in central locations are less likely to move (H2.2). In the second stage, we test whether, conditional on observing a move, more prolific scientists relocate to more central cities, i.e., global cities (H2.1).

We use a Heckman two-stage regression model to control for the selection bias of mobile scientists since most scientists do not move in any given period, as shown in Fig. A2. To correctly define the Heckman model, we need to specify an *exclusion restriction*. That is, we need to include a variable in the selection equation, which affects the probability to move but does not influence the destination. We use *mobility of other fields* as the exclusion restriction (named *P(Others Move)* in the regression). The *mobility of other fields* is the probability to leave the focal city for all scientist not belonging to the scientific field the focal scientist belongs to. For example, for an author predominantly publishing in "Biochemistry", the probability of moving is computed as the fraction of scientists leaving the city in the same period, but are not biochemists themselves. The rational to use this variable as an exclusion restriction is that if we observe a high proportion of mobile authors 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 try to reduce the likelihood that the focal author is influenced by competition, imitation of peers working in the same field, and other labour market-specific effects.

We carry out this analysis on a repeated cross-section of scientist location choices in the period 1990 to 2004. Specifically, we look every

[6] Similar results for alternative measures of the centrality of cities in the global mobility network are available upon request.

[7] The results do not critically depend on the choice of the minimum inclusion size. The analysis has also been carried out with 3 and 10 as the minimum size yielding very similar results.

year at the relocation choices, recording several individual level and location-specific features.

The main individual-level variables of interest are the number of papers (*ln(Papers)*) and the number of citations (*ln(Citations)*) at origin. These measures are computed *before* the move in order to prevent the actual move to play a role.

We control for the field of research of the scientist, that is to say, the field in which the author has the majority of her publications. We proxy the research field of an author through the classifications of the journals the author publishes most in according to the SCImago thematic areas, employing a majority rule. For example, if three papers in the period are published in biochemistry journals and 1 in immunology, the author would be classified as a biochemist.[8] We also control for age group fixed effects. Precisely we include a dummy for age cohorts of similar size (i.e., one year up to 11 and larger from there). Career progression (i.e., age) is measured as years from the first publication. As further controls in the regression, we use the *Observation Gap* which is the number of years around the alleged mobility year in which we do not have any publication. This gap signals low activity and should negatively affect the probability to move. *Born there* controls if the focal scientist has published his first paper in the current city. The "Alma Mater" ought to be a unique location for the scientist thus influencing his willingness to leave. Location-specific effects, i.e., variables marked with TC and OC, are measured at the city level, for the target city (TC) and the origin city (OC), respectively. We also consider the dyadic distance between TC and OC, named *ln(km distance)*. As additional controls in all regressions we have year dummies and origin country.[9] As for the origin city, the main variables of interest are its size in terms of the number of active scientists in a given year (*ln(OC Size)*) and its centrality in the mobility network.

As in any non-experimental causal inference exercise, there are likely omitted variables that might affect the size, significance and interpretability of the results. We cannot exclude that our specification does not suffer from omitted variable bias. However, to reduce the chances of omitted variable bias, we include along with the variables mentioned above also controls for moves within the same country (SameCountry) and moves to countries that have the same official language (SameLanguage). These controls should alleviate concerns that the results are driven by high mobility within countries with many global cities (e.g., USA).

As a robustness check to verify that our results do not crucially depend on the Heckman specification, we estimate two additional OLS regressions. In the first OLS specification, we simply drop all non-mobile scientists. In the second alternative OLS specification, we estimate if the target city (TC) has a higher centrality than the origin city (OC). We also include this last specification to highlight that we observe a move to *more central* cities.

To test H2.1, we are interested to know if prolific scientists, the ones with a high citation weighted scientific output (*ln(Citations)*), are more likely to move to central cities, as measured by degree centrality. A positive coefficient for *ln(Citations)* in the second stage would support H2.1. To test H2.2, we estimate in the first stage the interaction of the centrality of the origin city (*ln(OC Degree)*) with the productivity of the scientist (*ln(Citations)*). A negative coefficient for this interaction would suggest that prolific scientists are less likely to move away if they are in central cities. Summing up, if it is true that more productive scientists move preferentially to central cities (H2.1), we expect *ln(Citations)* in the second stage to be positive. Similarly to support H2.2, prolific scientists stay in global cities, we expect *ln(OC Degree)* × *ln(Citations)* in the first stage to be negative. On the one hand, we investigate if more prolific scientists, as proxied by their citation weighted output before they move, are more likely to relocate to central cities in the mobility network. On the other hand, we test if more prolific scientists located in global cities have a lower propensity to leave. If these two hypotheses are jointly verified, after controlling for several other factors, it supports the claim that prolific scientists tend to move to global cities and to remain there.

The results of the three models are shown in Table 2.[10] Specifically, in the first column, we show the Pooled OLS model considering only mobile authors. In the second column, we show the OLS model where the dependent variable is not the degree of the destination but a binary

Table 2
Individual mobility regression results.

	(1) ln(TC Degree)	(2) TC Higher Degree	(3) ln(TC Degree)
Individual Level Variables			
ln(Citations)	0.102***	0.484***	0.103***
	(7.57)	(6.95)	(7.60)
ln(Papers)	–0.0480***	–0.0598***	–0.0496***
	(–9.18)	(–4.71)	(–9.33)
ln(km distance)	0.108***	0.169***	0.108***
	(8.48)	(4.68)	(8.48)
SameLanguage	0.108*	0.170	0.108*
	(2.24)	(1.63)	(2.24)
SameCountry	0.138*	0.248*	0.138*
	(2.22)	(2.12)	(2.22)
ln(OC Degree) × ln (Citations)	–0.00340	–0.0687***	–0.00351
	(–1.31)	(–5.30)	(–1.35)
City Level Variables			
ln(OC Degree)	0.0636	–1.003***	0.0686
	(1.65)	(–6.74)	(1.79)
ln(OC Size)	–0.0406	–0.388***	–0.0447
	(–1.48)	(–3.49)	(–1.65)
Constant	4.277***	6.812***	4.254***
	(24.47)	(14.95)	(24.39)
moved			
Individual Level Variables			
ln(Citations)			0.0920***
			(7.97)
ln(Papers)			–0.0126*
			(-2.43)
P(Others Move)			2.394***
			(29.52)
BornThere			–1.115***
			(–145.58)
Observation Gap			–0.181***
			(–66.86)
ln(OC Degree) × ln (Citations)			–0.0163***
			(–6.95)
City Level Variables			
ln(OC Degree)			–0.0411*
			(–2.55)
ln(OC Size)			0.0458***
			(3.65)
Constant			–1.448***
			(–17.63)
arth(ρ)			0.0148*
			(2.05)
ln(σ)			0.115***
			(18.76)
Year	Yes	Yes	Yes
Origin Country	Yes	Yes	Yes
Age Group	Yes	Yes	Yes
Field	Yes	Yes	Yes
Observations	505,550	505,550	2,239,357

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

[8] The propensities to move by field are listed in Fig. A.5 in the Appendix. There are significant differences across disciplines, in line with the findings of Laudel and Bielick (2019).

[9] We reports the propensity to leave specific countries in Fig. A.4 of the Appendix.

[10] Similar results for PageRank centrality are available in the Appendix (see Table A.6).

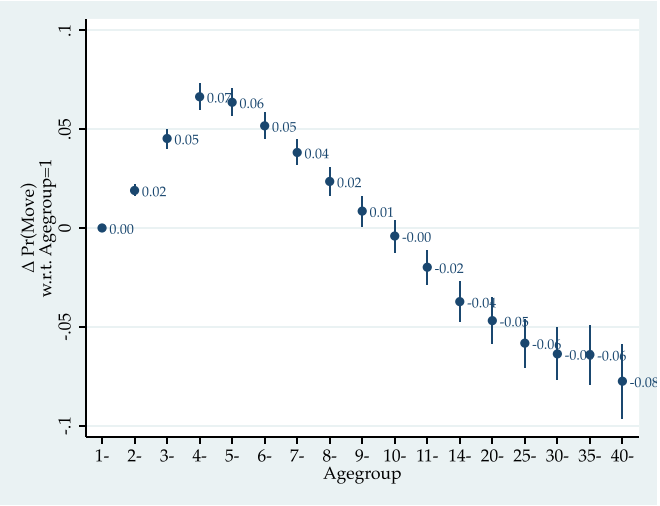

Fig. 7. This figure shows the marginal effects in probability to move compared to an author with age=1 (a year after the first publication in Medline). An author with a 9 to 10 years career has the same probability of being observed moving as an author at the beginning of her career. Error bars indicate the 95% confidence interval.

outcome variable indicating if the move was to a location with a higher degree. In the third column, we show our main model, i.e., the Heckman selection model.

Concerning H2.1, that more prolific scientists move preferentially to central cities, we find that *ln(Citations)* has a significant and robust effect on the centrality of the target cities (TC Degree) in all our regressions. Note also that it is citations and not publications to have a positive impact. Taken together, this means that scientists with more citations per publication are more likely to move to more central locations, but not scientists who merely publish a lot. We also confirm that more prolific scientists are more likely to move, as we see in the first stage of the Heckman regression. Interestingly the number of papers (*Papers*) has a negative effect in all specifications. This result highlights the fact that the "quality" signal matters, whereas the number of papers published per se has a negative effect if it does not translate into more citations. Overall, we find evidence in support of H2.1, that prolific scientists gravitate towards central cities, suggesting that mobility positively contributes to the superior performance of global cities. On top of this, central cities attract scientists on a global scale. In all regressions, the distance between locations is higher for moves to central destination cities (see positive *ln(km distance)*).

By looking at the effect of the interaction between the centrality of the origin city and the citations (*ln(OC Degree)* × *ln(Citations)*) we find also support for H2.2, that prolific scientists tend to remain in central cities. Note also that, generally speaking, prolific scientists are more likely to move (positive *ln(Citations)* in the first stage). However, the centrality of the origin cities decreases their probability to leave (*ln(OC Degree)* in the first stage). Similarly, scientists are more likely to stay in the city where they published their first paper (i.e., the "BornThere" effect is negative).

In addition to the above results, we also find that during a scientific career, the propensity to move varies considerably. Fig. 7 shows the marginal effects by age group on the propensity to move, obtained from the first stage of the Heckman model shown in Table 2. We find that in the early phase of their career scientists are significantly more likely to move. At the same time, the probability decreases for senior scientists active more than ten years after their first publication. As shown in the Appendix (see Fig. A4) the propensity to move varies considerably also across countries (US and UK scientists are more likely to relocate) and across disciplines (see Fig. A6 Fig. A5): physicists and biologists are more likely to move than physicians.

All in all, we find that global cities attract and retain prolific scientists. This effect is likely to positively contribute to the superior scientific production of global cities we documented in the previous section.

Scientists in global cities might experience a positive network effect since their scientific output will become more visible, thus attracting a higher potential number of citations. Moreover, by working in global cities, scientists might boost their career prospects, beyond and above their scientific output, with possibly better access to job opportunities. On average, the position of the source city should increase upward mobility by improving the chances that scientists will end up working in a place with higher scientific impact. In this work, this visibility effect cannot be cleanly separated from "innate ability" or talent. Therefore we do not claim that working in a central and large city is sufficient per se to boost citations and possibly career prospects. However, we observe that working in a global city will positively contribute to the citation weighted output of scientists.

As a further robustness check, in the Appendix, we provide a similar table in which we selected a 10% random sample of scientists (see Appendix Table A5)[11]. The sub-sampling serves two purposes: (1) to show that the significant effect is not merely an artefact of the size of the dataset, and (2) by sub-sampling we pick up fewer scientists moving together (e.g., the move of a prominent scientist and his group). Thus the assumption that the moves are independent is more likely. Results are not affected by the reduction of the sample size and the random selection of the observations. Also, the results are very similar across models (1) to (3) in Table 2 and when we consider an alternative centrality measure (i.e. PageRank, see the Appendix Tables A6). We observe that the parameters in models (1) and (3) are practically identical, suggesting that the selection into mobility does not introduce a substantial bias, which we could have corrected for with our first stage selection formulation.

5. Final discussion

Cities are critical loci of innovation, culture and economic activity: they are home to a large portion of the world's population and function as melting pots and cauldrons of creativity and human interactions. Despite declarations that distance is irrelevant in a globalized world, geography is very much alive. Geography has been found by various authors to be an essential dimension to understand and appreciate the modern knowledge economy.

Against this background, this work explores the impact the mobility of scientists on knowledge production. Due to data limitations, most of the literature so far has analyzed single regions, countries, disciplines or selected samples of researchers. Traditionally in this field, there has been a big divide between "micro" studies about the mobility of scientists between regions and institutions in a single country/domain, and the "macro" analysis of the migration of researchers between countries. To the best of our knowledge, this is the first analysis of global scientist mobility that takes a broader view, combining big data and network methodologies, to reconstruct and analyze the pattern of mobility between cities both within and across national borders (Verginer and Riccaboni, 2018). Thanks to this integrated data-driven approach, we highlight the crucial role of global cities. First, we show that scientists in global cities are more productive in terms of median citation weighted number of publications. Second, we find that global cities are central in the network of brain circulation, both within and across national borders. In the paper, we show that these two findings are related: global cities, defined as central urban areas in the network of scientists' mobility, take advantage of their privileged position in terms of spatial and social proximity to offer higher levels of individual citation-weighted scientific output. When we compare stayers and

[11] We run a similar analysis for the top 10% of scientists in terms of citations (the so-called star scientists). The results are available upon request.

movers, we notice that part of the advantage of global cities stems from the higher productivity of incoming scientists and the lower productivity of leavers, pointing to a net "brain gain" effect of knowledge hubs. To further investigate this mechanism, we estimate an individual level location choice model to find that more prolific scientists move preferentially to global cities and stay there. Therefore, we contribute to the literature on the mobility of scientists (Azoulay et al., 2017) by shedding more light on the crucial role of spatial mobility as a driver of the superior productivity of scientists in global cities. This phenomenon is likely a combined effect of social and geographical proximity. On the one hand, movers (i.e., mobile scientists) are likely to stay connected with their former colleagues in the origin cities. On the other hand, they will take advantage of the localized knowledge spillovers and social networks of global cities.

Our methodology has several strengths, chief among which is the processing of massive bibliographic data to quantify the individual-level and location-specific scientific production and to trace scientist mobility at the city level on a global scale. However, we are aware that author name disambiguation is an imperfect process, i.e., confusion of author names, and that the definition of a *city* is still a matter of active research and debate. Efforts such as the ORCID system to assign open access unique identifiers to researchers is a step in the right direction and will allow future research to be even more convincing. Nevertheless, this work should serve as a step towards understanding more deeply the importance of international scientific mobility. Understanding its impact in a world where human capital and innovation are paramount for success represents an opportunity to highlight actionable research policies. The results presented here are relevant in understanding the possible causes of growth differentials across regions and cities. While there are undoubtedly positive effects for both sending and receiving countries as shown by Agrawal et al. (2011), our results still suggest that there is a rich-get-richer effect fueled by global mobility. This result corroborates previous findings on the crucial role of central hubs in networks of innovators to better target national innovation policies (Chessa et al., 2013; Morescalchi et al., 2015). This observation is particularly important for Europe in a globalized world for advanced human capital with the emergence of new innovation hubs like Beijing in the Far East and the still dominant role of US global cities. A promising avenue for future work is to leverage our database to extend the analysis to the location of citing papers, as a proxy of knowledge flows in the scientific community, in analogy to the research tradition initiated by Jaffe et al. (1993) for patents. Moreover, in future work, we plan to update our global network data to analyze the impact of exogenous shocks (such as the COVID-19 crisis) and targeted policies (like Singapore's innovation policy since the early 2000s) on the mobility of researchers and to quantify the contribution of brain circulation to innovation and knowledge diffusion.

CRediT authorship contribution statement

Luca Verginer: Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Data curation, Visualization, Software, Formal analysis. **Massimo Riccaboni:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Formal analysis.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknoledgments

We would like to thank Vetle Torvik and Neil Smalheiser for granting us access to Author-ity and MapAffil and the participants of workshops and presentations at INET, Università Cattolica, Queen Mary University of London and the GeoInno 2018 Conference. We would like to thank specifically Prof. Frank Schweitzer, Prof. Pietro Panzarasa and Prof. Marco Vivarelli who offered valuable and important feedback on this work as well as the two anonymous reviewers whose thoughtful feedback was crucial in sharpening the message of this paper.

Appendix A

A1. Data sources

The central and primary source of information is Medline. This dataset provides open access to more than 26 million records of scientific publications, with most of the corpus covering research related to the life sciences. The earliest publication in the dataset is from 1987 and Medline is updated continuously. In this work we analyze scientific publications from 1990 to 2009. The reason to restrict our analysis to this period is to guarantee the best coverage in the various datasets. This ensures that we have adequate coverage of geo-referenced scientific production at the city level and disambiguated individual level data. Moreover, this choice allows enough time to observe forward citations after this period. To track authors across publications and identify locations we rely on MapAffil and Author-ity (Torvik, 2015; Torvik and Smalheiser, 2009). MapAffil lists for a large part of papers in Medline the disambiguated city of the affiliation as listed on the paper (ca. 37,396, 671 author-locations). Author-ity created by Torvik and Smalheiser (2009) contains the disambiguations of 61,658,514 names in Medline papers (author-name instances). These author-name instances are mapped to 9,300,182 disambiguated authors. With this dataset it is possible to map the affiliation string to a city. By merging Medline with Author-ity we are able to trace an author across publications. The ability to reconstruct mobility comes from merging the previous two datasets with MapAffil. Without this last step, affiliations would not be disambiguated, and we would have hundreds of different versions of "Boston University" in our dataset. Fortunately, MapAffil can accurately[12] 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 illustration showing the various stages in the data processing pipeline are shown in Fig. A1.

We carried out a manual check of 50 randomly chosen individual scientists (with at least five publications in our dataset) using publicly available information. Specifically, we searched for personal web-pages and faculty websites to obtain the author's CV. We were able to locate 32 out of 50 CVs successfully. We compared then these CVs to the extracted city sequences. Comparing the moves in our dataset with CV entries revealed that the moves we have identified happened indeed most of the time. The year of the move we have identified, on the other hand, was often too late. In other words, the move if it took place, happened 1 or 2 years earlier than our approach suggests. This time discrepancy, we argue stems from the fact that publications are a delayed signal of production/presence since submission-to-publication times, especially in specific disciplines, can range from months to years. As for the reliability of the geo-location of the affiliation, Torvik (2015) have carried out their own ground-truth comparison reporting state of the art accuracy. For a detailed discussion of their validation procedure see Torvik (2015).

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

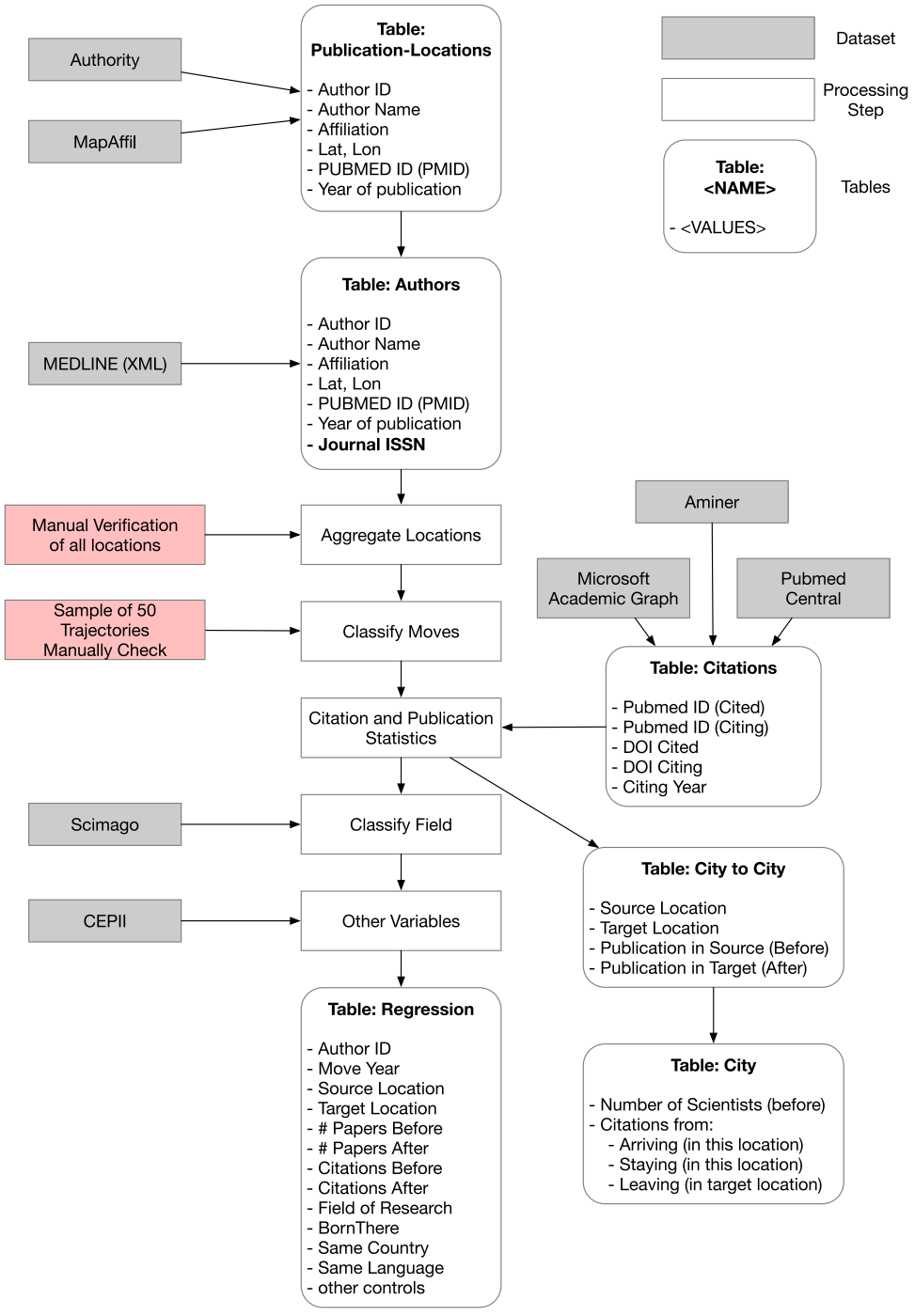

Fig. A1. Steps in the data processing from raw to regression tables.

A2. Example affiliation record

Table A1
Example of career path of a specific author. For each record we have the year of publication, the city of the affiliation and the relative PubMed ID (PMID) identifying the paper. The algorithm would record a move from Los Angeles to New Orleans in 2010 and no moves in any other year.

Year	City	PMID
2007	Los Angeles, CA, USA	17382381
2009	Los Angeles, CA, USA	18996587
2009	Los Angeles, CA, USA	19321991
2009	Los Angeles, CA, USA	18701812
2009	Los Angeles, CA, USA	19236004
2009	Los Angeles, CA, USA	19518912
2010	New Orleans, LA, USA	20160068
2011	New Orleans, LA, USA	21521360
2011	New Orleans, LA, USA	21256987
2012	New Orleans, LA, USA	22153326
2012	New Orleans, LA, USA	22447582
2012	New Orleans, LA, USA	23338820
2013	New Orleans, LA, USA	23635887
2013	New Orleans, LA, USA	23288544
2014	New Orleans, LA, USA	25319365
2014	New Orleans, LA, USA	24748612

A3. Definition of city

A major problem in urban studies is finding a good definition for the city boundaries (Bee et al., 2019; Rozenfeld et al., 2011; Zipf, 1949). Metropolitan Statistical Areas (MSAs) are commonly used as a standard units of analysis in the US (US Census Bureau, 2018). However, given our objective to study global mobility, we would require such metropolitan boundaries to be defined globally, which unfortunately is not the case. We, therefore, rely on the definition provided by Google Maps of a "location". This definition reflects more closely administrative rather than natural boundaries which are not perfect substitutes. A notable difference between the natural and administrative boundary definitions is that they follow different size distribution (Bee et al., 2013; Eeckhout, 2004). Since we do not explicitly rely on the size distribution of cities, we will relegate this consideration to future refinements. Moreover, given that the underlying locations are scientific affiliations, we argue that their addresses reflect the location of research institutions and not necessarily the precise location of scientists anyway. From MapAffil we obtain as location the centre of a city (low resolution). However, these are mixed with locations at a higher resolution, which identifies a suburb or part of a city. For example for "London, UK" we have the location (lat=51.5, lon=-0.13) 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. In a first pass, all locations within 100 km from each other with similar names (Levenshtein Distance) have been grouped into candidate clusters. Then we have manually reviewed 16,000 clusters, most of which were singletons. So for example "Bethnal Green, London, UK" and "Goodmayes, Ilford, Redbridge, London, UK" have been mapped to "London, UK" at position (lat=51.5, lon=-0.13). Similarly, the Boston neighbourhoods "Jamaica Plain, Boston, MA, USA" and "Roslindale, Boston, MA, USA" are mapped to the lower resolution city centre "Boston, MA, USA" (lat=42.36, lon=-71.06). By applying this method, we obtain 7159 urban areas in the period 1990 to 2009. After removing any location with less than 5 active authors in a five year period, we are left with 2292 cities worldwide.

A4. Probability to move

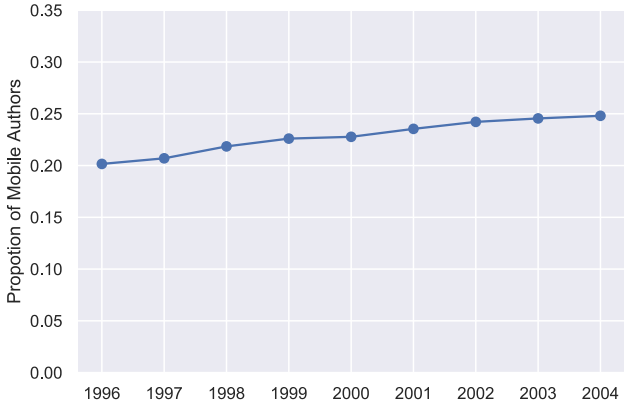

Fig. A2. Yearly proportion of authors being observed to move. Note that a move in one year means that the location 5 years prior and 5 years after are different, thus the probability is not the probability to move in a given year but the probability to move in this interval.

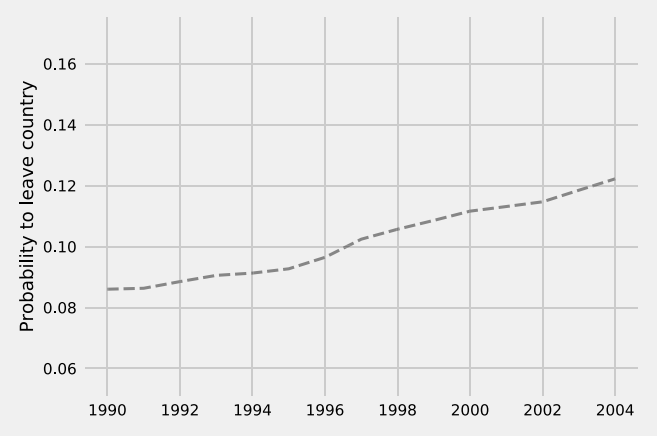

Fig. A3. Probability to leave country.

A5. Mobility network descriptive statistics

The Giant Connected Component (GCC) is the largest sub-graph among which there is at least one possible connection among the nodes. The percentage of all nodes that fall in the GCC is a common proxy statistic to highlight how connected the network is. We see that from 1996 to 2004 it increased from 77.3% to 79.4%. The number of scientists correspond to the number of active authors in that year. Authors may be repeated across years if they were active over a longer period.

Table A2
Mobility Network Statistics.

Year	Cities	Edges	Density	GCC %	Number of scientists
1996	4217	38,422	0.22%	77.3%	324,545
1997	4261	38,692	0.21%	77.1%	321,113
1998	4391	42,657	0.22%	77.5%	341,879
1999	4559	45,535	0.22%	78.6%	358,415
2000	4691	46,923	0.21%	78.1%	368,493
2001	4847	51,290	0.21%	78.4%	397,775
2002	5070	55,343	0.21%	78.4%	425,058
2003	5317	59,764	0.21%	79.0%	455,696
2004	5531	63,587	0.21%	79.4%	488,419

A6. Descriptive statistics for scientists relocation

Table A3
Descriptive Statistics for Scientist Level variable used in the Heckman 2 stage regression.

	Stayed		Moved		All	
	mean	sd	mean	sd	mean	sd
ln(Citations)	0.22	1.44	0.21	1.40	0.21	1.43
Age	9.34	8.16	7.36	6.54	8.90	7.87
P(Others Move)	0.23	0.07	0.26	0.08	0.23	0.08
Born There	0.79	0.41	0.34	0.47	0.69	0.46
Observation Gap	1.85	0.99	2.18	1.10	1.92	1.02
ln(km distance)	0.00	0.00	6.84	1.90	1.55	3.00
Same Language	0.00	0.00	0.12	0.32	0.03	0.16
Same country	1.00	0.00	0.60	0.49	0.91	0.29
OC degree	299.87	252.22	304.17	251.52	300.84	252.07
TC degree	299.88	252.22	294.80	255.62	298.73	253.00
OC size	1591.09	1898.68	1490.55	1832.54	1568.39	1884.42
TC size	1591.11	1898.67	1393.72	1754.86	1546.55	1868.99
Observations	1,733,807		505,550		2,239,357	

Table A4
Cross-correlation table, t-statistics in parenthesis.

Variables	1	2	3	4	5	6	7	8	9	10	11	12
1) Moved	1.000											
2) ln(Citations)	–0.002	1.000										
	(0.012)											
3) ln(Papers)	–0.099	0.345	1.000									
	(0.000)	(0.000)										
4) ln(OC Size)	–0.039	0.136	0.063	1.000								
	(0.000)	(0.000)	(0.000)									
5) ln(OC Degree)	0.005	0.219	0.087	0.941	1.000							
	(0.000)	(0.000)	(0.000)	(0.000)								
6) P(Others Move)	0.168	0.167	0.038	–0.239	0.029	1.000						
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)							
7) BornThere	–0.408	–0.066	0.036	0.045	–0.008	–0.172	1.000					
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)						
8) Observation Gap	–0.136	0.311	0.262	0.030	0.034	0.001	0.019	1.000				
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.135)	(0.000)					
9) ldistance	0.954	0.005	–0.097	–0.017	0.024	0.137	–0.393	–0.129	1.000			
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)				
10) SameCountry	–0.585	0.004	0.067	–0.010	–0.014	–0.027	0.237	0.075	–0.714	1.000		
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)			
11) SameLanguage	0.306	0.014	–0.031	0.002	0.016	0.036	–0.128	–0.040	0.367	–0.523	1.000	
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)		
12) Age	–0.106	0.147	0.756	0.035	0.058	0.035	0.058	0.074	–0.107	0.073	–0.036	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	

A7. Supplementary regression results for scientist relocation

Table A5
Individual Mobility Regression Results for 10% Sample for Degree Centrality.

	(1) ln(TC Degree)	(2) TC Higher Degree	(3) ln(TC Degree)
main			
ln(OC Size)	–0.0489	–0.397**	–0.0566
	(–1.42)	(–3.26)	(–1.66)
ln(OC Degree)	0.0763	–0.995***	0.0858
	(1.58)	(–5.94)	(1.78)
ln(OC Degree) × ln (Citations)	–0.00660	–0.0627***	–0.00683
	(–1.50)	(–4.08)	(–1.56)
ln(Citations)	0.122***	0.457***	0.123***
	(5.22)	(5.42)	(5.28)
ln(Papers)	–0.0582***	–0.0742**	–0.0609***
	(–5.38)	(–3.09)	(–5.61)
ln(km distance)	0.109***	0.177***	0.109***
	(7.82)	(4.57)	(7.83)
CommonLanguage	0.0996	0.147	0.0995
	(1.89)	(1.29)	(1.89)
SameCountry	0.146*	0.307*	0.145*
	(2.17)	(2.42)	(2.16)
Constant	4.390***	7.148***	4.341***
	(8.75)	(5.99)	(8.62)
ln(OC Size)			0.0518*
			(2.37)
ln(Citations)			0.0949***
			(5.48)
ln(Papers)			–0.00665
			(–0.85)
P(Others Move)			2.398***
			(20.58)
BornThere			–1.111***
			(–115.22)
Observation Gap			–0.179***
			(–42.80)
ln(OC Degree)			–0.0461
			(–1.66)
ln(OC Degree) × ln (Citations)			–0.0175***
			(–5.09)
Constant			–1.777***
			(–9.61)
arth(ρ)			0.0287*
			(2.40)
ln(σ)			0.114***
			(16.11)
Year	Yes	Yes	Yes
Origin Country	Yes	Yes	Yes
Age Group	Yes	Yes	Yes
Field	Yes	Yes	Yes
Observations	50 922	50 883	224 428

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

Table A6
Individual Mobility Regression Results for PageRank Centrality.

	(1) ln(TC PageRank)	(2) TC Higher PageRank	(3) ln(TC PageRank)
main			
ln(OC Size)	–0.0627**	–0.139*	–0.0674**
	(–2.65)	(–1.98)	(–2.88)
ln(OC PageRank)	0.0856**	–1.165***	0.0903**
	(2.84)	(–11.70)	(3.04)
ln(OC PageRank) × ln (Citations)	–0.000745	–0.0197*	–0.00104
	(-0.32)	(–2.14)	(–0.45)
ln(Citations)	0.0845***	–0.0105	0.0824***
	(5.50)	(–0.18)	(5.35)
ln(Papers)	–0.0504***	–0.0624***	–0.0529***
	(–8.71)	(–5.17)	(–9.01)
ln(km distance)	0.0953***	0.137***	0.0954***
	(6.24)	(3.79)	(6.24)
ComlangOff	0.104	0.154	0.104
	(1.75)	(1.44)	(1.76)
SameCountry	0.111	0.164	0.111
	(1.50)	(1.37)	(1.50)
Constant	–5.957***	–7.338***	–5.936***
	(–12.61)	(–5.28)	(–12.58)
ln(OC Size)			0.0552***
			(6.45)
ln(Citations)			–0.0793***
			(–5.40)
ln(Papers)			–0.0129*
			(–2.49)
P(Others Move)			2.412***
			(38.59)
BornThere			–1.115***
			(–145.79)
Observation Gap			–0.181***
			(–67.01)
ln(OC PageRank)			–0.0468***
			(–4.32)
ln(OC PageRank) × ln (Citations)			–0.0135***
			(–6.31)
Constant			–2.023***
			(–13.08)
arth(ρ)			0.0213**
			(2.77)
ln(σ)			0.209***
			(40.17)
Year	Yes	Yes	Yes
Origin Country	Yes	Yes	Yes
Age Group	Yes	Yes	Yes
Field	Yes	Yes	Yes
Number of observations	505 550	505 537	2 239 357

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

A8. Propensities to move by country and field

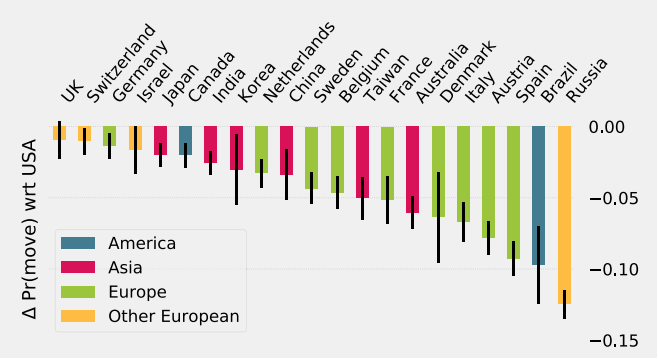

Fig. A4. The figures shows the marginal effect on the probability to move compared to the base case USA and the 95% confidence interval (black bars). A negative value such as Taiwan (–5%) means that keeping everything else fixed, a scientist in Taiwan is 5% less likely to move than a colleague in the US. Only countries are shown here for which we have observed at least 3000 scientists in the country in the period 2000–2004.

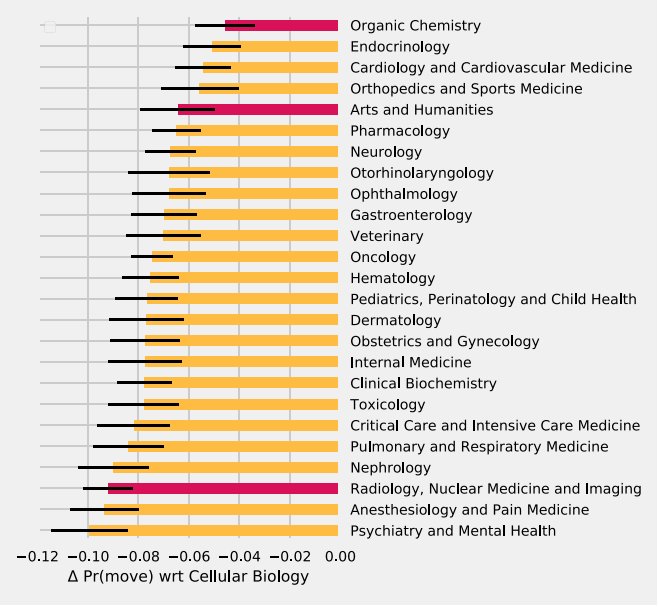

Fig. A5. Marginal probability to move compared to Cellular Biology (i.e. the largest field) for the 25 most mobile fields. The 95% confidence interval is illustrated as black bars.

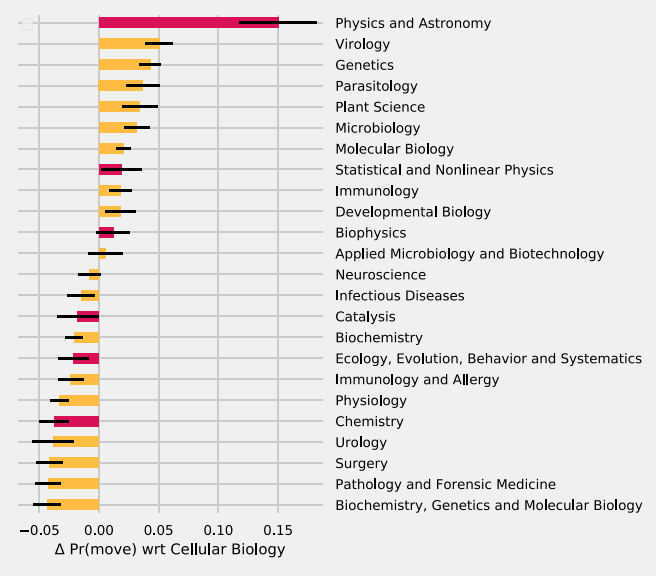

Fig. A6. Marginal probability to move compared to Cellular Biology (i.e. the largest field) for the 25 least mobile fields. The 95% confidence interval is illustrated as black bars.

References

Agrawal, A., Cockburn, I., McHale, J., 2006. Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships. Journal of Economic Geography 6 (5), 571–591. https://doi.org/10.1093/jeg/lbl016.

Agrawal, A., Kapur, D., McHale, J., 2008. How do spatial and social proximity influence knowledge flows? evidence from patent data. J Urban Econ 64 (2), 258–269.

Agrawal, A., Kapur, D., McHale, J., Oettl, A., 2011. Brain drain or brain bank? The impact of skilled emigration on poor-country innovation. J Urban Econ 69 (1), 43–55. https://doi.org/10.1016/j.jue.2010.06.003.

Almeida, P., Kogut, B., 1999. Localization of knowledge and the mobility of engineers in regional networks. Manage Sci 45 (7), 905–917.

Alnuaimi, T., Opsahl, T., George, G., 2012. Innovating in the periphery: the impact of local and foreign inventor mobility on the value of indian patents. Res Policy 41 (9), 1534–1543.

Azoulay, P., Ganguli, I., Graff Zivin, J., 2017. The mobility of elite life scientists: professional and personal determinants. Res Policy 46 (3), 573–590. https://doi.org/10.1016/j.respol.2017.01.002.

Bathelt, H., Malmberg, A., Maskell, P., 2004. Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. Prog Hum Geogr 28 (1), 31–56.

Bee, M., Riccaboni, M., Schiavo, S., 2013. The size distribution of us cities: not pareto, even in the tail. Econ Lett 120 (2), 232–237.

Bee, M., Riccaboni, M., Schiavo, S., 2019. Distribution of City size: gibrat, pareto, zipf. The Mathematics of Urban Morphology. Springer, pp. 77–91.

Belderbos, R., Benoit, F., Edet, S., Lee, G.H., Riccaboni, M., 2020. Global cities' innovation network. In: Castellani, D., Perri, A., Scalera, V., Zanfei, A. (Eds.), Cross-border Innovation in a Changing World. Players, Places and Policies. Oxford University Press, Oxford. forthcoming

Bettencourt, L.M., Lobo, J., Strumsky, D., 2007. Invention in the city: increasing returns to patenting as a scaling function of metropolitan size. Res Policy 36 (1), 107–120. https://doi.org/10.1016/J.RESPOL.2006.09.026.

Bettencourt, L.M.A., 2013. The origins of scaling in cities. Science 340 (6139), 1438–1441. https://doi.org/10.1126/science.1235823.

Bohannon, J., Doran, K., 2017. Introducing ORCID. Science 356 (6339), 691–692. https://doi.org/10.1126/science.356.6339.691.

Boschma, R., 2005. Proximity and innovation: a critical assessment. Reg Stud 39 (1), 61–74.

Breschi, S., Lenzi, C., 2016. Co-invention networks and inventive productivity in us cities. J Urban Econ 92, 66–75.

Breschi, S., Lissoni, F., 2009. Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. Journal of economic geography 9 (4), 439–468.

Breschi, S., Lissoni, F., Miguelez, E., 2017. Foreign-origin inventors in the USA: testing for diaspora and brain gain effects. Journal of Economic Geography 17 (5), 1009–1038. https://doi.org/10.1093/jeg/lbw044.

Cantwell, J., Piscitello, L., 2005. Recent location of foreign-owned research and development activities by large multinational corporations in the european regions: the role of spillovers and externalities. Reg Stud 39 (1), 1–16.

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. Res Policy 44 (9), 1749–1762. https://doi.org/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. https://doi.org/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 (6120), 650–651. https://doi.org/10.1126/science.1227970.

Culotta, E., 2017. People on the move: the science of migrations. Science. https://doi.org/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. Sci Rep 4, 4770.

Eeckhout, J., 2004. Gibrat's law for (all) cities. American Economic Review 94 (5), 1429–1451.

Feldman, M.P., 1999. The new economics of innovation, spillovers and agglomeration: a review of empirical studies. Economics of innovation and new technology 8 (1–2), 5–25.

Fink, C., Miguelez, E., Raffo, J., 2017. Determinants of the international mobility of knowledge workers. The International Mobility of Talent and Innovation: New Evidence and Policy Implications 162–190. https://doi.org/10.1017/9781316795774.006.

Florida, R., 2005. Cities and the creative class. Routledge. https://doi.org/10.4324/9780203997673.

Franzoni, C., Scellato, G., Stephan, P., 2012. Foreign-born scientists: mobility patterns for 16 countries. Nat. Biotechnol. 30 (12), 1250.

Franzoni, C., Scellato, G., Stephan, P., 2014. The mover's advantage: the superior performance of migrant scientists. Econ Lett 122 (1), 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 (1), 27–59. https://doi.org/10.1111/joms.12279.

Ganguli, I., 2015. Who leaves and who stays? evidence on immigrant selection from the collapse of soviet science. Global Mobility of Research Scientists. Elsevier, pp. 133–154.

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

Glaeser, E.L., 1999. Learning in cities. J Urban Econ 46 (2), 254–277. https://doi.org/10.1006/juec.1998.2121.

Graf, H., Kalthaus, M., 2018. International research networks: determinants of country embeddedness. Res Policy 47 (7), 1198–1214. https://doi.org/10.1016/j.respol.2018.04.001.

Jacobs, J., 1969. The Economy of Cities. New York - Random House.

Jacobs, J., 1984. Cities and the Wealth of Nations. New York - Random House.

Jaffe, A.B., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. Q J Econ 108 (3), 577–598.

Laudel, G., Bielick, J., 2019. How do field-specific research practices affect mobility decisions of early career researchers? Res Policy 48 (9), 103800.

Li, G.-C., Lai, R., D'Amour, A., Doolin, D.M., Sun, Y., Torvik, V.I., Amy, Z.Y., Fleming, L., 2014. Disambiguation and co-authorship networks of the us patent inventor database (1975–2010). Res Policy 43 (6), 941–955.

Mayer, T., Zignago, S., 2011. Notes on CEPII's distances measures: The GeoDist database. Working Papers. CEPII.

Mokyr, J., 2016. A culture of growth: The Origins of the Modern Economy. Princeton University Press.

Morescalchi, A., Pammolli, F., Penner, O., Petersen, A.M., Riccaboni, M., 2015. The evolution of networks of innovators within and across borders: evidence from patent data. Res Policy 44 (3), 651–668.

Moretti, E., Wilson, D.J., 2017. The effect of state taxes on the geographical location of top earners: evidence from star scientists. American Economic Review 107 (7), 1858–1903.
Morrison, G., Riccaboni, M., Pammolli, F., 2017. Disambiguation of patent inventors and assignees using high-resolution geolocation data. Sci Data 4, 170064.
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.
Pan, R.K., Kaski, K., Fortunato, S., 2012. World citation and collaboration networks: uncovering the role of geography in science. Sci Rep 2, 902.
Rozenfeld, H.D., Rybski, D., Gabaix, X., Makse, H.A., 2011. The area and population of cities: new insights from a different perspective on cities. American Economic Review 101 (5), 2205–2225. https://doi.org/10.1257/aer.101.5.2205.
Sassen, S., 2016. The Global City: Strategic Site, New Frontier. Managing Urban Futures. Routledge, pp. 89–104.
Schlapfer, M., Bettencourt, L.M.A., Grauwin, S., Raschke, M., Claxton, R., Smoreda, Z., West, G.B., Ratti, C., 2014. The scaling of human interactions with city size. Journal of The Royal Society Interface 11 (98). https://doi.org/10.1098/rsif.2013.0789.20130789–20130789
Scholl, T., Garas, A., Schweitzer, F., 2018. The spatial component of r&d networks. Journal of Evolutionary Economics 28 (2), 417–436.
Serafinelli, M., Tabellini, G., 2017. Creativity over time and space. SSRN.
Solimano, A., 2008. The international mobility of talent: Types, causes, and development impact, 394. Oxford University Press Oxford.
Taylor, P.J., Derudder, B., 2015. World City Network: A Global Uurban 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 (11/12). https://doi.org/10.1045/november2015-torvik.
Torvik, V.I., Smalheiser, N.R., 2009. Author name disambiguation in medline. ACM Transactions on Knowledge Discovery from Data (TKDD) 3 (3), 1–29.
UN, 2018. World urbanisation prospects: Key findings.
US Census Bureau, 2018. Metropolitan and Micropolitan. www.census.gov/programs-surveys/metro-micro/about.html. Accessed: 2020-02-06.
Vaccario, G., Verginer, L., Schweitzer, F., 2020. The mobility network of scientists: analyzing temporal correlations in scientific careers. Applied Network Science 5 (1), 36. https://doi.org/10.1007/s41109-020-00279-x.
Verginer, L., Riccaboni, M., 2018. Brain-Circulation Network: The Global Mobility of the Life Scientists. IMT Institute for Advanced Studies Lucca - Working Papers.
Verginer, L., Riccaboni, M., 2020. Cities and countries in the global scientist mobility network. Applied Network Science 5 (1), 38. https://doi.org/10.1007/s41109-020-00276-0.
Zacchia, P., 2018. Benefiting colleagues but not the city: localized effects from the relocation of superstar inventors. Res Policy 47 (5), 992–1005.
Zipf, G.K., 1949. Human Behavior and the Principle of Least Effort. Aaddison-Wesley Press.
Zucker, L.G., Darby, M.R., 2007. Star scientists, innovation and regional and national immigration. Technical Report. National Bureau of Economic Research.