INTERFACE

rsif.royalsocietypublishing.org

Research

Check for updates

Cite this article: Petersen AM. 2018 Multiscale impact of researcher mobility. *J. R. Soc. Interface* **15**: 20180580. http://dx.doi.org/10.1098/rsif.2018.0580

Received: 29 July 2018
Accepted: 3 September 2018

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
biogeography, computational biology, biocomplexity

Keywords:
human mobility, propensity score matching, network science, science of science, researcher careers, diversity

Author for correspondence:
Alexander M. Petersen
e-mail: petersen.xander@gmail.com

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4226732.

THE ROYAL SOCIETY PUBLISHING

Multiscale impact of researcher mobility

Alexander M. Petersen

Ernest and Julio Gallo Management Program, Management of Complex Systems Department, School of Engineering, University of California, Merced, CA 95343, USA

AMP, 0000-0002-0955-3483

International mobility facilitates the exchange of scientific, institutional and cultural knowledge. Yet whether globalization and advances in virtual communication technologies have altered the impact of researcher mobility is a relevant and open question that we address by analysing a broad international set of 26 170 physicists from 1980 to 2009, focusing on the 10-year period centred around each mobility event to assess the impact of mobility on research outcomes. We account for secular globalization trends by splitting the analysis into three periods, measuring for each period the effect of mobility on researchers' citation impact, research topic diversity, collaboration networks and geographical coordination. In order to identify causal effects we leverage statistical matching methods that pair mobile researchers with non-mobile researchers that are similar in research profile attributes prior the mobility event. We find that mobile researchers gain up to a 17% increase in citations relative to their non-mobile counterparts, which can be explained by the simultaneous increase in their diversity of co-authors, topics and geographical coordination in the period immediately following migration. Nevertheless, we also observe that researcher's completely curtail prior collaborations with their source country in 11% of the cross-border mobility events. As such, these individual-level perturbations fuel multiscale churning in scientific networks, e.g. rewiring the connectivity of individuals and ideas and affecting international integration. Together these results provide additional clarity on the complex relationship between human capital mobility and the dynamics of social capital investment, with implications for immigration and national innovation system policy.

1. Introduction

The dispersion of knowledge across institutional and national borders is fundamental to scientific progress. Historically, knowledge exchange has been mediated via the physical mobility of people and printed publications, however, it is progressively mediated by a combination of physical and virtual networks. While researcher relocation remains a prominent conduit for knowledge transfer, researchers can now choose from a variety of virtual alternatives to explore new research environments and collaborations. And while the professional prospects and call of adventure associated with relocation may be alluring to some, there are nevertheless risky trade-offs associated with physical relocation that require careful assessment of local versus non-local socio-economic, family, work and funding opportunities [1–4]. This common dilemma factors into the cost of human and social capital investment in science, which is rather substantial and continues to grow with the globalization of the scientific endeavour [1,5].

To stay or leave? We provide guidance on this pivotal question by quantifying the impact of researcher mobility from four perspectives—scientific impact, research topic diversity, collaboration and international integration—together providing insights into the socially mediated network of networks connecting individuals and knowledge production [6]. As such, our results contribute to the literature on how knowledge flows [7,8] and how careers grow [9] following pivotal events—e.g. winning the Nobel Prize [10], initiating a dedicated

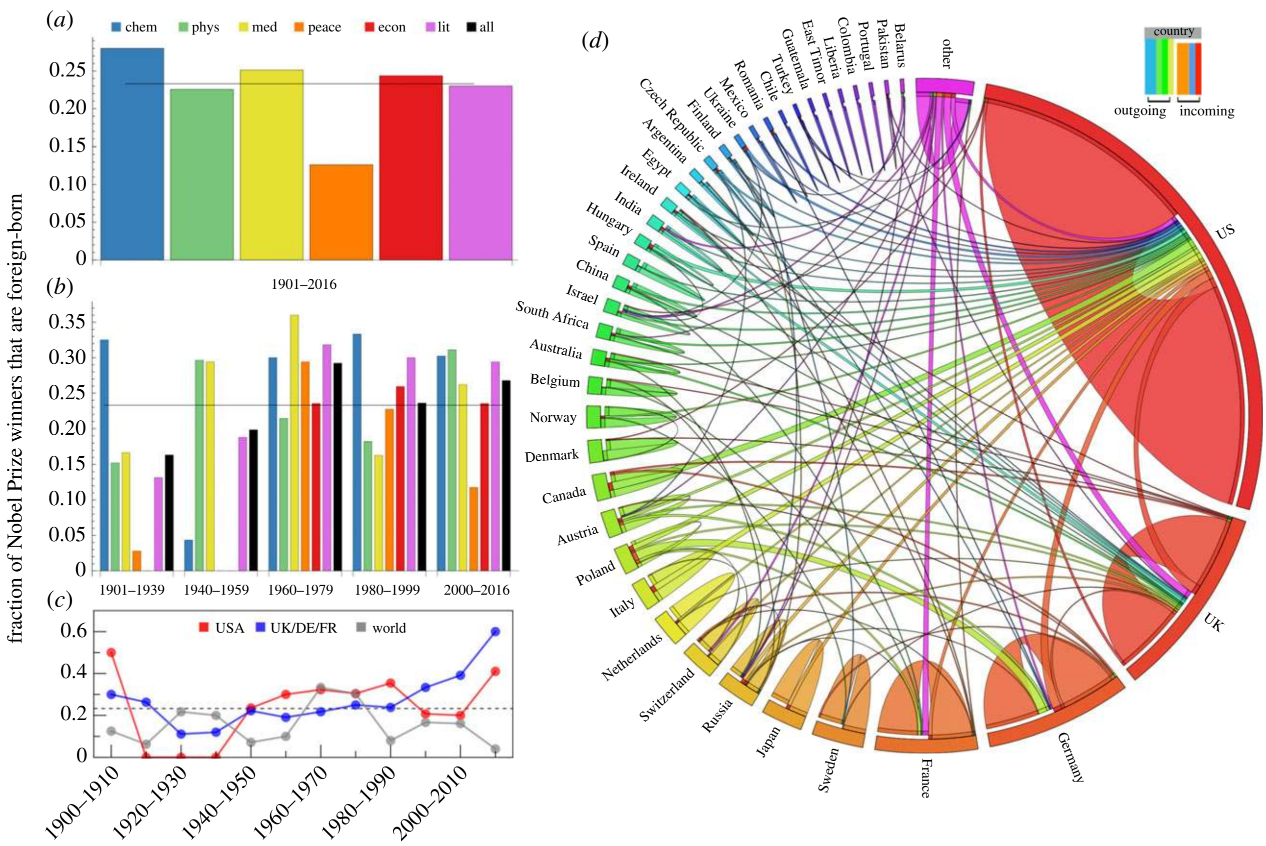

Figure 1. International mobility of Nobel laureates. (*a*) Fraction of Nobel Prize winners who received their award for work done in a country different than their birth country—'foreign-born'; 23% of Nobel Prize winners from 1901 to 2016 belong to the foreign-born category, indicated by the horizontal line. (*b*) Fraction by award and time period. (*c*) Fraction by region of achievement over 10-year intervals. (*d*) The Nobelist birth-discovery network: countries are represented along the arc, with arc-length proportional to the number of laureates born in a given country. Link width represents the number of individuals born in country c_b that performed award research in country c_a. The link direction is denoted by the gap, which differentiates incoming and outgoing links: e.g. the majority of the international links associated with the US terminate with a gap, indicating foreign-born. The majority of links are intra-country (e.g. 29% of all laureates have the US as their country of birth and their country of achievement), but the international links represent the nearly 1 in 4 Nobelists who migrated internationally. Data from *Wolfram Alpha* [35]. (Online version in colour.)

partnership [11], obtaining tenure [12] and pursuing cross-disciplinary migration [13]—which in turn mediate shifts in collaboration [14–16] and research topic exploration [17–20].

In particular, we focus on the domain of physics, a broad research field characterized by relatively high levels of geographical mobility, which contributes to growing efforts to measure the impact of high-skilled migration in both industry and academia [1–5,7,8,14–16,21–33]. We employ a data-driven approach, starting with author disambiguation of 355 808 research articles published in American Physical Society (APS) journals over the period 1980–2009, resulting in a final dataset of 26 170 prominent researcher profiles that are broadly distributed both in terms of geography, productivity and scientific impact.

We leverage the size and breadth of this researcher dataset to address common reverse causality and selection bias problems that render identifying causal mobility effects rather challenging [34]. For example, many studies to date focus on select subsets of migrant researchers—e.g. elite scientists, winners of prestigious international fellowships or participants in select national exchange programmes—in which measured shifts in performance may endogenously depend on the selection of migrant researchers being analysed. Instead, following a data-driven approach we are able to estimate the mobility effect within a diverse population, accounting for variation in time period, geographical region, and social and cognitive aspects of research activity.

Our results demonstrate the impact of researcher mobility across scales—from individual careers, to collaboration network churning, to international integration. As motivation, the Nobel Prize offers a prime example of multiscale impact, highlighting how countries with attractive academic and industrial R&D environments that support objective and impartial immigration policy are positioned to benefit greatly from the immigration of even just a single elite scientist capable of prosecuting ground-breaking and transformative research.

To provide insights into the global immigration patterns of elite scientists, we analyse the international pathways of all Nobel Prize winners through 2016 (figure 1). We found that roughly 1 in 4 are foreign-born—having performed their prize-winning research in a country different than their birth country. This finding is rather stable across the different award categories, with the exception of the Peace Prize, which has lower frequencies since there is a propensity for human rights activists to lead movements in their native countries. The UK, Germany and France, the three countries in Europe with the highest R&D spending, have seen a recent surge since the 1990s in the fraction of Nobelists from their country who are foreign-born. The rest of the world, not including the US, shows the opposite trend since the 1960s. There is also significant disparity in the flow counts from source to destination country, as figure 1*d* illustrates how research hubs like the USA, the UK and Germany attract a disproportionate share of foreign-born Nobelists.

While winning the Nobel Prize is likely to remain out of reach for most scientists, the bulk of scientists will nevertheless encounter an opportunity and be faced with the decision of whether to relocate. Based on our data, we estimate the rate of international mobility among all researchers in our dataset, independent of productivity level, to be roughly 2 in 5 (40%); and including US inter-state cross-border mobility this rate is closer to 3 in 5 (58%).

In what follows, we focus on several thousand relatively productive physicists across three decades, implementing statistical matching techniques that leverage the breadth of our dataset to compare researchers who migrate to similar researchers who do not. As such, we demonstrate the value of cross-border mobility on research outcomes, reflecting the pervasive and persistent value of physical mobility as a conduit for the exchange of intellectual, professional and cultural experience.

2. Results

2.1. Researcher data

We analysed the APS *Physical Review* datasets 'Citing article pairs' and 'Article metadata' [36], from which we extracted researcher profiles using a network-based author name disambiguation method [37]; see electronic supplementary material, appendix A.1 for additional details. Application of this algorithm identified 208 734 distinct researcher profiles over the 30-year period 1980–2009. Each researcher profile, indexed by i, is characterized by the year of researcher's first APS publication, y_i^0, and total number of APS publications N_i. We concentrate on the 26 170 researchers with $N_i \geq 10$ publications who also meet additional career longevity and productivity criteria. Electronic supplementary material, figure S1(A–C) provides a statistical summary of the publication and researcher profile data.

Despite the large sample size, there are several limitations common to large-scale data-driven approaches, which derive from missing contextual information at the researcher and geographical level. First, we lack data for physics and non-physics research published by these researchers outside of the set of APS journals. Yet because of the reputation and prominence of the APS journal family, we are able to make the assumption that researchers characteristic of physics-related domains are likely to publish in one or more of these journals rather frequently over the course of his/her career. Thus, as long as their publications in the APS journals are characteristic of their contemporaneous research output in other journals (albeit excluding the exceptionally novel and infrequent research published in high-impact multidisciplinary letters journals) then we do not expect this to significantly bias our estimates. In other words, our reported estimates measure the impact of mobility on typical research outcomes.

Second, and relevant to this previous point, in order to facilitate the precision of the statistical matching method used here that leverages the large size of our researcher sample, we only include relatively prolific APS researchers. This is achieved by thresholding on publication rates during the 10-year window around each mobility event, which also ensures that the publication measures are aggregated from sufficient sample size to be a reasonably robust measure of researcher attributes in the immediate (5-year) period before and after the event. However, this also means that we likely exclude many early stage (doctoral) and postdoctoral researchers who commonly migrate for brief periods, but do not have significant publication output to enter our study before they exit the academic career path. For example, according to US National Science Foundation Statistics, by 2008 the rate of US PhD recipients (postdoctoral fellows) in science and engineering who were foreign-born had grown to roughly 1 in 2 (3 in 5) [5]. Thus, our results do not capture the mobility premium for this highly mobile portion of the academic workforce, which deserves additional attention given shifts in the early and mid-career labour market [12,38–41] and the uncertainty associated with academic career trajectories [24,29,42–44].

And third, in addition to missing bibliometric information, we also lack other important information such as gender, which is an important factor relevant to career longevity and productivity in science [42,45–48], and possibly also migration decisions. In a broader sense, we lack contextual information on what *push and pull* factors underly researchers' decisions to migrate. For example, a 2011 survey analysis based on more than 15 000 researchers' responses reports that destination countries with competitiveness-oriented national science policy (i.e. oriented around 'prestige' and 'research excellence') is a major attractive factor [31]. A different study based on more than 10 000 biomedical researchers, which incorporates social and familial factors in additional to the traditional professional factors, also shows that researchers are attracted by the prospects associated with a competitive professional and peer environment; however, more accomplished researchers are less likely to move if they recently obtained NIH funding or if their children are in high school [3]. Countries with relatively high R&D expenditure levels tend to have higher rates of elite physicist immigration [2]. And researcher mobility between two countries is positively correlated with international student mobility in the opposite direction, indicative of the formation of cross-generational brain circulation channels [49]. Complementing these findings is an analysis of high-skilled migration within Europe over the period 1997–2014, showing that countries with higher government expenditure on education also make significantly more attractive destination countries [4].

2.2. Researcher mobility framework

The aim of our study is to measure shifts in researcher profiles before versus after an cross-border mobility event occurring in year $t^*_{i,T}$. Electronic supplementary material, figure S2 illustrates our framework for measuring the reconfiguration of research attributes before and after such a mobility event; see electronic supplementary material, A.1 for further details on the calculation of $t^*_{i,T}$. In particular, our framework relies on the following subscript notation: we split into three time periods (indexed by T), drawing on numerous researcher profiles (indexed by i), which are then split into three mobility groups (indexed by G). Moreover, in order to assess individual researcher profile attributes, we also analysed the collaborators, keywords, and countries associated with each researcher profile, hereafter generically indexed by j.

For two main reasons, we split the analysis into three mobility observation periods denoted by T, each defined by non-overlapping lower and upper year limits, t_T^- and t_T^+: $T_1 \equiv [1990-1997]$, $T_2 \equiv [1998-2003]$ and $T_3 \equiv [2004-2007]$. First, this separation facilitates drawing period-specific conclusions; considered together, they facilitate identifying trends in the mobility effect over time. Second, this choice provides a compromise on the issue of how to treat multiple mobility events for a given researcher. In principle, a researcher could migrate multiple times over a short period (e.g. characteristic of a short postdoctoral or sabbatical period). However, we are primarily interested in the shifts in research activity associated with the first mobility event within a reasonable time frame. Thus, for a given analysis period T, we use just the first $t^*_{i,T}$ associated with a sequence of mobility events. By partitioning our analysis into multiple T, we then allow for subsequent mobility events by the same i to also contribute to our analysis. Electronic supplementary material, figure S1(D) provides a schematic of the separation of the analysis into three observation periods: R is the number of researcher profiles analysed per T: $R_1 = 4124$, $R_2 = 9362$ and $R_3 = 13\,457$. The variable window size accounts for the growth of the profession, and the latter period is chosen to occur after 2004 so to account for the increased international mobility following the EU enlargement [4,16].

Electronic supplementary material, figure S1(E) provides a schematic description of our procedure for classifying each researcher i that was active in a given period T according to three groups: (i) $G_{i,T} = 1$ identifies researchers who did not migrate before or during period T; $G_{i,T} = 2$ identifies a 'placebo' mobility group comprised of researchers who were mobile prior to T but not during T; and (iii) $G_{i,T} = 3$ identifies researchers who were mobile during T, with specific mobility year $t^*_{i,T}$. Researchers in a given T belonging to either group $G_{i,T} = 1$ or 2 are prescribed $t^*_{i,T} \equiv \mathrm{Median}[t^-_T, t^+_T]$, the midpoint of T (e.g. $t^*_{i,1} \equiv 1994$). We then aggregated each researcher's publications in the $\Delta t \equiv 5$-year window before $t^*_{i,T}$, i.e. over the interval $[t^*_{i,T} - \Delta t, t^*_{i,T} - 1])$; likewise, we aggregated the publications after $t^*_{i,T}$ over the interval $[t^*_{i,T}, t^*_{i,T} + \Delta t - 1])$. This framework facilitates measuring research patterns over a balanced observation period for each i, thereby leveraging the longitudinal dimension of the researcher profiles. As a final dataset refinement, we excluded researcher profiles with fewer than three publications in either the period before or after $t^*_{i,T}$ and fewer than four distinct years of publication activity in total.

Electronic supplementary material, figure S1(F) shows the annual distribution of the total 31 075 mobility events, including in this count the multiple movements by a single researcher within a single T. Refining to just the first mobility event per researcher per period, we observe 6498 profiles belonging to the $G_{i,T} = 3$ group (i.e. roughly 21% of the total mobility events). These select researcher profiles are rich in data and sufficient in number to implement a matching method approach to estimate the impact of researcher mobility on research outcomes by comparing groups 1 and 3; moreover, we perform a robustness check by comparing researchers in groups 1 and 2 and researchers in groups 2 and 3. Intuitively, we expect that the mobility effect for the groups 1 and 3 comparison should be larger than the other group comparison estimates if our specification is to be considered consistent.

2.3. Research activity measures

We provide four complementary perspectives on (i) scientific impact, (ii) collaboration, (iii) research topics and (iv) geographical coordination, defined as follows:

2.3.1. Citation impact

We normalized the standard integer citation count $n_{i,p,t}$ for each publication p published in year t to account for temporal bias. The result is a normalized z-score $z_{i,p}$, which is normally distributed for all t. We then calculate the average citation impact value $Z^{+,-}_i \equiv \langle z^{+,-}_{p,i} \rangle$ across the $N^{+,-}_i$ publications in each $\Delta t \equiv 5$-year interval, i.e. before $(-)$ or after $(+)$ $t^*_{i,T}$. We also use a measure of total citation impact, $\Sigma^{+,-}_i$, useful for assessing the magnitude of the mobility effect in real terms by applying a 'citation deflator' that accounts for 'citation inflation' [50].

2.3.2. Co-author diversity

For each $\Delta t \equiv 5$-year interval we count the number of articles published by i with co-author j, given by k_{ij}. We then calculate the Shannon entropy for the distribution of k_{ij} across the set of K_i co-authors, defined as $E^{+,-}_{K,i}$. Higher levels of variation correspond to larger entropy values, with $E \geq 0$; the limiting case of no variation, $k_{ij} =$ const. for all j, corresponds to $E = 0$.

2.3.3. Research topic diversity

Similar to (§2.3.2), we aggregate the instances of *Physics and Astronomy Classification Scheme* (PACS) codes associated with each publication to calculate a Shannon entropy measuring the variation in research topics, $E^{+,-}_{Q,i}$.

2.3.4. Geographic reach diversity

We aggregate the countries indicated in the affiliation bylines and again calculate a Shannon entropy measuring the variation in geographical coordination, $E^{+,-}_{C,i}$.

See the Material and methods for further details on each measure. In what follows, we use the quantities Z^+_i, Σ^+_i, $E^+_{K,i}$, $E^+_{Q,i}$ and $E^+_{C,i}$ as dependent variables in five separate models, and the corresponding Z^-_i, Σ^-_i, $E^-_{K,i}$, $E^-_{Q,i}$ and $E^-_{C,i}$, along with five other covariate data measured before $t^*_{i,T}$, in order to match i between $G_{i,T}$ groups. Electronic supplementary material, figures S3–S4 show the distribution of the quantities $Z^{+,-}_i$, $E^{+,-}_{K,i}$, $E^{+,-}_{Q,i}$ and $E^{+,-}_{C,i}$; see electronic supplementary material, A.2 and figure S5 for the analysis of the difference in these variables around $t^*_{i,T}$.

2.4. Reorganization of social and geographical links

When a researcher relocates, there is an immediate impact on his/her proximity to former collaborators. This reorganization of collaboration networks at the individual level can have far-reaching implications at the national and global scale. A case example is how East–West migration imbalance within Europe following the 2004 enlargement of the European Union (EU) [16] negatively impacted rates of international collaboration; it is not unlikely that high-skilled migration in response to Brexit will have a similar effect on international collaboration patterns in Europe [4].

Against this background, in this section we investigate the extent to which mobility mediates collaboration-based shifts in social and geographical integration, providing insights into the formation and disintegration of social capital. As above, C^- (C^+) denotes the list of countries extracted from the publication affiliation byline (dropping the subscript i for brevity), with list length denoted by $|C^-|$ $(|C^+|)$. We also define the set of distinct countries as $\hat{C}^-$ $(\hat{C}^+)$, with lengths $|\hat{C}^-|$ and $|\hat{C}^+|$, thereby disregarding the multiplicity

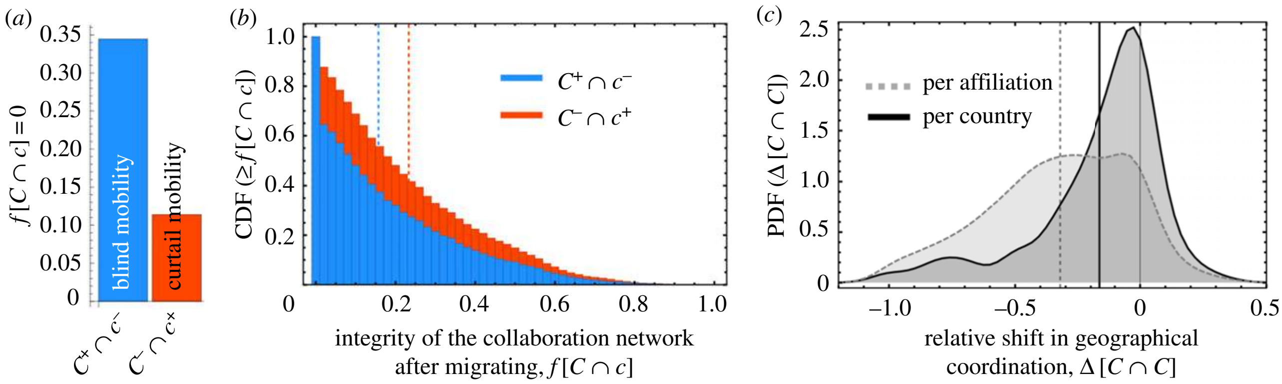

Figure 2. The impact of mobility on geographical collaboration networks. $C^+ \cap c_i^-$ ($C^- \cap c_i^+$) represents the geographical overlap between a researcher's location before (after) migrating, denoted by c_i^- (c_i^+), and the set of collaborator locations after (before) migrating, denoted by C^+(C^-). (*a*) Shown is the fraction of mobility events corresponding to: (i) 'blind mobility' (blue): in which there is no overlap between a researcher's prior location and the locations of future collaborators ($f[C^+ \cap c_i^- = 0] = 0.34$); and (ii) 'curtail mobility' (red): in which there is no overlap between the locations of a researcher's prior collaborators and his/her destination country ($f[C^- \cap c_i^+ = 0] = 0.11$). (*b*) The distributions of collaboration network integrity, measured by $f[C^+ \cap c_i^-]$ and $f[C^- \cap c_i^+]$, are right-skewed: on average there is only a 16% overlap between C^+ and c_i^- and a 23% overlap between C^- and c_i^+. (*c*) The difference $\Delta[C \cap C]$ measures the change in the amount of geographical overlap (see equation (2.2)), measured in two ways: per affiliation and per country. Negative values indicate more overlap before as compared to after. Both methods indicate relatively high levels of collaboration network disintegration following a mobility event: 89% (70%) of the values are negative when measuring per affiliation (per country). Shown are calculations on data aggregated across all three periods, T_{123}; for analogous plots specific to a given period T see electronic supplementary material, figure S6. Vertical lines indicate distribution mean values. (Online version in colour.)

of the countries in the original lists. By way of example, consider the hypothetical list of countries associated with the affiliations derived from an arbitrary set of publications, $C^- \equiv \{\mathrm{CA, CA, FR, FR, AU, AU}\}$; then the corresponding unweighted country list is $\hat{C}^- = \{\mathrm{CA, FR, AU}\}$. These two definitions provide complementary perspectives on the degree to which geographical collaborations reorganize: the first measure ($C^{+,-}$) is weighted proportional to the number of publications (i.e. per author affiliation), and the second is more aggregate (i.e. per country).

As such, we seek to quantify the degree to which the source country c_i^- and destination country c_i^+ of a mobile researcher are related to $C^{+,-}$. Thus, we define the 'per-affiliation' geographical association of the source (destination) country c_i^- (c_i^+) with C^- (or C^+) as

$$\left.\begin{aligned} & f[C^+ \cap c_i^-] = \frac{|C^+ \cap c_i^-|}{|C^+|} \\ \text{and} \quad & f[C^- \cap c_i^+] = \frac{|C^- \cap c_i^+|}{|C^-|}, \end{aligned}\right\} \tag{2.1}$$

where $|X \cap Y|$ denotes the number of elements in the intersection of the two sets X, Y. Similarly, the 'per-country' measure of geographical association is defined using $\hat{C}^{+,-}$ in equation (2.1) instead.

Figure 2*a* shows that roughly 34% of the mobility events were 'blind', corresponding to the value $f[C^+ \cap c_i^-] = 0$, i.e. the destination country was not in the sphere of prior collaborations. Similarly, roughly 11% of the migrations corresponded to the scenario in which the sphere of destination collaborations did not intersect with the source country ($f[C^- \cap c_i^+] = 0$), corresponding to the maximal curtailing of prior collaborations.

In addition to these extreme cases, we find that most other mobility events are followed by significant collaboration reorganization. Figure 2*b* shows the cumulative distribution function (CDF) CDF($\geq f[C^- \cap c_i^+]$) and CDF($\geq f[C^+ \cap c_i^-]$), with both distributions more concentrated around small values. In the case of $f[C^- \cap c_i^+]$), this indicates a relatively small likelihood that researchers maintain prior collaborations when they move.

We also analysed the dynamics of geographical coordination, conditional on the source country, defined as the difference

$$\Delta[C \cap C] = f[C^+ \cap c_i^-] - f[C^- \cap c_i^-], \tag{2.2}$$

which is negative if there is more overlap between c_i^- and C^- than between c_i^- and C^+. Figure 2*c* shows that most $\Delta[C \cap C]$ are negative—i.e. migration is associated with significant churning within collaboration networks. Electronic supplementary material, figure S6 shows the same distributions for data disaggregated by T, demonstrating the stability of these observations over time.

2.5. Estimating the mobility effect using propensity score matching

To what degree does researcher mobility affect scientific impact, topical direction and collaboration? In this section, we describe the Rubin causal inference framework [51] as it applies to estimating the impact of cross-border mobility on various quantitative researcher career metrics. Using potential outcome notation, we estimate the *average treatment effect on the treated* (ATET),

$$\begin{aligned} \tau_{W=1}[Y] &\equiv E[Y_1 - Y_0 | W = 1] \\ &\approx N_{W=1}^{-1} \sum_{i:W_i=1} (Y_i - \hat{Y}_i(0)), \end{aligned} \tag{2.3}$$

where Y_i indicates the outcome variable of interest, defined as one of the variables Z_i^+, Σ_i^+, $E_{K,i}^+$, $E_{Q,i}^+$, or $E_{C,i}^+$ and $E[Y_1 - Y_0]$ denotes the expected difference between the two counterfactual outcome measures. The indicator $W_i = 1$ denotes 'treatment', i.e. mobility by individual i in period T; conversely, $W_i = 0$ corresponds to no mobility during T. Of course, a researcher is observed with either $W_i = 0$ or $W_i = 1$ in a given T, but not both; thus, the challenge is to estimate

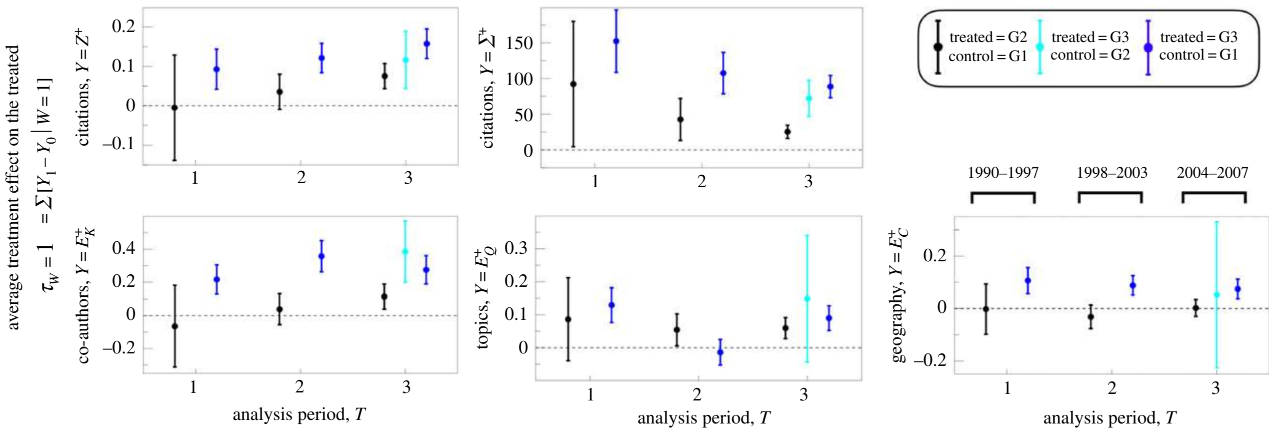

Figure 3. Estimation of the effect of mobility on various career measures. For a given variable Y and period of analysis T, we applied the PSM method to compare a 'treated' group (either $G_i = 2$ or 3) to a 'control' group (either $G_i = 1$ or 2), resulting in three comparison: G1–G2, G2–G3 and G1–G3. Each panel reports the results for a different dependent variable Y; shown are point estimates with error bar indicating 95% CI. (Online version in colour.)

the counterfactual outcome, i.e. what would have happened had the researcher i not migrated. Hence, we use the propensity score matching (PSM) method [52] to identify researcher pairs (i, i'), where i' is as similar as possible to i in terms of likelihood to belong to the treated group based on measured researcher characteristics (denoted by $\boldsymbol{X}_i^-$) prior to $t^*_{i,T}$. Equation (2.3) defines the mobility-effect estimate $\tau_{W=1}[Y]$ according to the PSM method, which approximates the counterfactual outcome $\hat{Y}_i(0)$ for researcher i with $Y_{i'}(0)$ by identifying the closest match i' from the pool of $N_{W=0}$ researchers with $W = 0$. In this way, $\tau_{W=1}$ is approximated as the average researcher-level effect across the sample of $N_{W=1}$ researchers with $W = 1$. For a practical review of the matching estimators and their implementation see [53].

We estimated the treatment effect associated with each dependent variable Y_i^+ using a vector $\boldsymbol{X}_i^-$ comprised of six covariates. We specified the logit model to calculate the likelihood of treatment for each i, which is used by PSM to match i with the closest i' (results are reported for single matches, which we found to be consistent with nearest-neighbour matching up to the five closest matches). To be specific, for a given dependent variable Y_i^+, the six covariates we used to match are: (i) the outcome variable calculated for the period before the mobility, Y_i^-; (ii) the number of distinct co-authors, $|k_{ij}^-|$; (iii) the number of publications, N_i^-; (iv) the average citation impact, Z_i^-; (v) the researcher age $s_i^* = t^*_{i,T} - y_i^0 + 1$ in the year of the mobility event; and (vi) a factor variable F_i^- which maps the country of residence c_i^- to one of five geographical regions (N. America, S. & C. America, Europe, Asia & Australasia, Africa; see electronic supplementary material, A.3). Electronic supplementary material, figure S7 shows the distribution of each model variable and the corresponding correlation matrix.

Electronic supplementary material, table S1 shows the estimates of a logit model for the dependent variable $1_{G_i} = 3$, with value 1 if a researcher i belongs to the mobile group $G_i = 3$ and 0 otherwise, thereby estimating the mobility likelihood depending on a given researcher's attributes. See electronic supplementary material, appendix A.4 for analysis of the factors that correlate with cross-border mobility, along with refs. [1,3] which describe additional employment and other factors not feasible to include within our data-driven approach.

2.6. Impact of mobility on research outcomes

We applied the PSM method to five different dependent variables $(Z^+, \Sigma^+, E_K^+, E_Q^+, E_C^+)$ calculated for three non-overlapping analysis time periods and for three different control-treated subgroup comparisons. More specifically, the matching between the groups G3 (treated, comprised of researchers with $G_{i,T} = 3$) and G1 (control, comprised of researchers with $G_{i,T} = 1$) provides the core estimate of the mobility effect; the matching between G2 (placebo-treated) and G1 (placebo control) provides a robustness check, as we do not expect there to be significant differences between the researchers in these two groups; and the matching between G3 (treated) and G2 (control), serves as an additional robustness check, since we hypothesize the mobility effect to be present, but to a lesser degree than the G3–G1 estimation.

Figure 3 reports the resulting 35 $\tau_{W=1}[Y]$ mobility-effect estimations; insufficient samples sizes for the G2 and G3 groups limited the PSM matching performance for the periods T_1 and T_2 (see electronic supplementary material, figure S2(D) for the size of G1, G2, G3 for each T). To demonstrate model robustness, electronic supplementary material, figure S8 shows the estimates using the nearest-neighbour method (*nnmatch* in STATA13) instead of the PSM method (*psmatch* in STATA13); the important difference is that with the *nnmatch* method one can force an exact match on the source region F_i^-. The results using *nnmatch* are consistent with the results of PSM, with the exception of the ATET for E_K^+, which are significantly smaller in significance and magnitude. The Material and methods section summarizes additional robustness checks to further assess the statistical significance of our PSM estimates.

The PSM results indicate that researcher mobility has a significant positive effect on citation impact and increases diversity in research topics and collaboration at the individual and geographical level. Notably, the effect size for mean citation impact, $\tau_{W=1}[Z^+]$, increased over each period T. Because $z_p \equiv (\ln[n_p + 1] - \mu_t)/\sigma_t$ is a logarithmic transform of the citation count variable n_p (see Material and methods), the mobility effect for the average publication, in terms of the per cent increase in n_p, is approximately $100 \times \langle\sigma\rangle \times \tau_{W=1}[Z^+] = 9\%$ to 17% depending on T.

We also observe a decreasing trend for $\tau_{W=1}[\Sigma^+]$, which at first appears inconsistent with the trend for $\tau_{W=1}[Z^+]$. However, the Σ^+ measure suffers from right-censoring bias

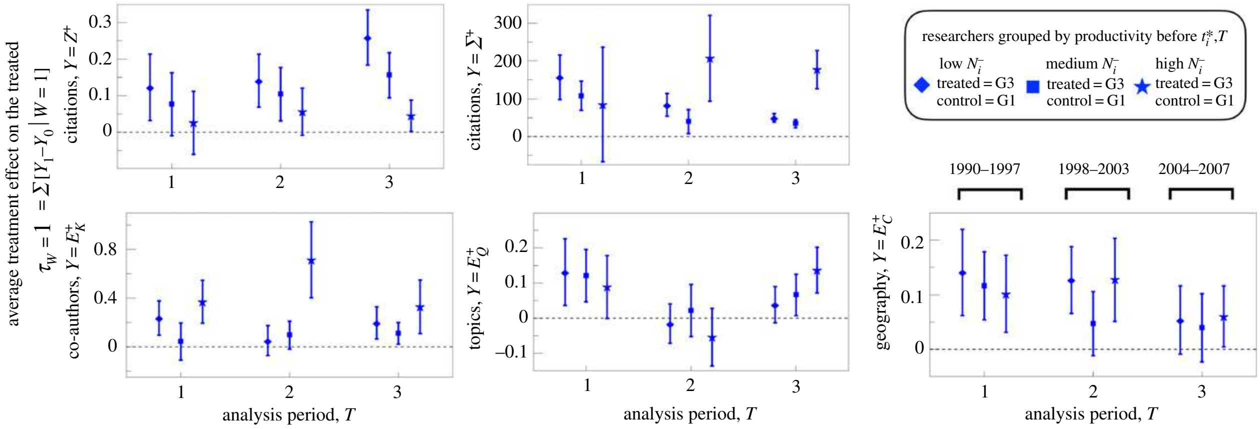

Figure 4. Estimation of the mobility effect by productivity subgroup. Researchers were separated into three groups according to the number of publications in the pre-treatment period, N_i^-: low (bottom tercile, triangles), medium (middle tercile, squares) and high (upper tercile, stars). For a given variable and period of analysis T, we applied the PSM method to compare the mobile 'treated' group ($G_i = 3$) to the not-mobile 'control' group ($G_i = 1$) corresponding to the group comparison G1 – G3 in figure 3. Each error bar is a point estimate with 95% confidence interval. (Online version in colour.)

(publications analysed for later T have not the same amount of time to accrue citations as for publications from previous T) and so the negative trend is confounded by this bias; the value of the citation premium is nevertheless statistically significant and relevant in magnitude: $\tau_{W=1}[\Sigma^+] \approx 90–150$ citations depending on T, which provides an estimate for the aggregate mobility effect tallying over all research produced in the 5-year period after mobility. This is a lower-bound estimate because the APS dataset does not include citations from articles published in other journals, nor does Σ^+ account for the citation tally from researchers' publications in non-APS journals. Thus, it is likely that $\tau_{W=1}[\Sigma^+]$ would be significantly larger if non-APS publications and citations were included.

Among the entropy variables, the mobility effect is largest for $\tau_{W=1}[E_K^+]$ (diversity of co-authors) and $\tau_{W=1}[E_C^+]$ (diversity of geographical coordination). The effect of mobility on E_Q^+ (diversity of research topics) is statistically significant for periods T_1 and T_3, but not for T_2. These results quantify the extent to which mobile researchers facilitate a valuable interface for intellectual exchange. A prime example was the cross-disciplinary interface formed by computing and biology researchers in the genomics revolution, in which this constructive configuration facilitated the cross-pollination of methods and institutional know-how that were crucial for the success of the Human Genome Project [13].

We further estimated the mobility effect using OLS regression on matched researcher pairs (i, i'), which facilitates incorporating covariates $\boldsymbol{X}^-$ to explain the outcome variable Y^+ associated with each $\tau_{W=1}[Y]$. See electronic supplementary material, A.5 and tables S2–S6 for the description and results of this additional analysis. For example, we find that increasing researcher age correlates with lower research topic diversity, E_Q^+, suggesting that the ability to diversify into new research topics decreases with age ($\beta_{s*} < 0$; electronic supplementary material, table S5). We also observe a negative relation ($\beta_{s*} < 0$) between researcher age and citation impact variables, e.g. Z^+ and Σ^+, consistent with findings from previous studies on researcher careers [11,13].

Are the observed mobility effects merely a feature of elite scientists in our dataset? To address this question, we separated the researchers into three terciles according to N_i^- in order to further analyse the degree to which variation in $\tau_{W=1}[Y]$ is mediated by researcher productivity, a variable that is highly correlated with a researcher's characteristic team size, among other factors. Figure 4 reports the PSM estimates after matching researchers only within each subgroup, showing that the mobility effect on mean citation impact is largest for researchers belonging to the low- and medium-N_i^- tercile groups; the mobility effect on co-author diversity is dominated by the high N_i^- tercile group; and no significant variation is observed for topic diversity or geographical diversity.

3. Discussion

The globalization of science and decreasing costs of migration have democratized the opportunities for international mobility [2], thereby becoming an increasingly relevant topic for national innovation system policy [1], the economics of science [5], and the multiscale modelling of the scientific system [6]. Against this backdrop, we assembled and analysed a relatively comprehensive longitudinal dataset of researchers in physics in order to accurately quantify how researcher mobility affects career growth and intellectual exploration at the micro-level, and the churning of the collaboration network facilitating international integration at the macro-level. We used statistical matching techniques to measure the differences between mobile researchers and similar non-mobile researchers, thereby addressing the reverse causality and sample selection biases that hinder estimating causal effects. To this end, we developed a methodical framework for comparing the shifts in a researcher's publication profile in the 10-year period centred around each mobility event (see electronic supplementary material, figure S2 for a schematic of this framework).

We proceeded by separating our analysis into three time periods (T) in order to be able to identify significant temporal trends. Qualitatively, our quantitative results are not sensitive to T, indicating that the mobility premium identified for each variable ($Z^+, \Sigma^+, E_K^+, E_Q^+, E_C^+$) has not diminished as virtual avenues for exploring new collaborations proliferate. The results that are relevant to individual researchers are as follows. We measure a 9–17% increase in the number of citations received by articles published by mobile researchers, relative to the matched non-mobile control set. Aggregating this differential across all publications in the 5 years after migrating, on average this citation premium tallies up to the order of 100 citations (figure 3). A hypothetical mechanism that emerges directly from our analysis of the research diversity variables (E^+_K, E^+_Q, E^+_C) is that mobility increases the likelihood of drawing together different intellectual capacities and professional experience distributed across researchers, which when combined, provide a valuable perspective on how to best advance research efforts [17]. As such, this recombinant collaboration emerges as an important factor promoting recombinant innovation [54].

We complement this analysis based upon a broad cross-section of physicists with an alternative perspective based upon the extreme tail of scientific success—the 884 recipients of the Nobel Prize from 1901 to 2016. Analysis of their international trajectories reveals that 23% of these Nobelists were foreign-born, having received the award for work done in a country different than their birth country (figure 1). As such, these success stories provide ample evidence that just a single immigrant can have a monumental positive impact on a destination country, explaining to some degree the substantial international competition for elite scientists [2,3,5,21].

Motivated by recent work connecting migration imbalance (i.e. East–West 'brain drain') following the 2004 European Union enlargement to the stalled integration of national innovation systems across Europe [16,55], we also analysed the coevolution of mobility and collaboration from a systems perspective. In this regard, we provide additional micro-level evidence for such a mobility-mediated disintegration mechanism by showing that researchers completely curtail all former international collaborations roughly 11% of the time they migrate (figure 2). Conversely, we also found that 34% of the mobile researchers in our analysis moved to a location that was not in their sphere of prior collaborations. In addition to these exceptional cases, we observed remarkably high levels of churning following most mobility events. In this way, migration affords a prime opportunity to phase out old (possibly stagnant) collaborations and research topics in order to make way for exploring new (potentially lucrative) avenues. Other scholars have also noted that having migrated once likely increases the likelihood to migrate again, not merely due to additional experience and openness in regards to future opportunities, but also because recently shuffled collaboration networks are likely to be less restricting [31].

In conclusion, we analysed the dynamic interpersonal and international interface mediated by mobile researchers, one that facilitates the exchange of knowledge, as well as other valuable professional and cultural perspectives. While our focus was on mobile researchers in physics, our results are generalizable to other scientific and intellectual domains where cross-border experience confers a considerable comparative advantage by virtue of increased exposure to group diversity [56] and opportunities to broker knowledge [30], thereby promoting career development along an otherwise uncertain career path [24,29,42–44].

Against this background, recent scholarly efforts have provided considerable new insights into the professional, social and familial factors influencing scientists' decisions to stay or go [3,31,49]. For example, survey responses from more than 15 000 respondants indicated that 'opportunities to improve future career prospects' and 'prestige/excellence of the institution' were the two most important factors underlying their decision to move abroad [31]; however, a separate large study found that researchers were significantly less likely to relocate if they recently obtained competitive funding or if their children were in high school [3]. Related empirical work has also sought to determine whether other types of mobility—i.e. institutional, social, or inter-sectoral—produce a measurable effect on subsequent productivity and citation impact [28,57,58]; altogether these studies consistently report positive shifts in productivity, but a little if not insignificant effect on citation impact, when moving to a better institution.

International recruitment and home-return policy are typically geared around elite scientists [2,21,32], yet our results suggest that the career benefits of mobility are common to all ranks. Another important consideration is immigration policy, which can have a significant impact on the attraction and retention of talented researchers, thereby affecting the development of national innovation systems. An important example is the 2000 change in USA H-1B visa policy, which nullified the cap on available visas for non-profit sponsors, thereby eliminating the competition between industry and academia over this critical type of visa that facilities high-skilled immigration [5]. By analysing the annual changes in the number of H-1B visa recipients in industrial science and engineering, scholars found that higher rates of Chinese and Indian H-1B recipients correlated with higher levels of employment and patenting in innovation hubs that depend on immigrant high-skilled labour; moreover, the scholars report limited effects of the increase in H-1B immigrant population on native inventors' employment and patenting productivity [59]. These are important results showing that immigration of high-skilled labour does not appear to crowd out native employment opportunities and innovation capacity. And finally, countries with travel visa restrictions are negatively associated with international researcher mobility, in particular because they significantly increase the cost of travel, thereby inhibiting short-term collaboration visits that may precede the opportunity and ultimate decision to migrate long-term [49].

Considered in this way, our results contribute to these discussions by suggesting that national innovation systems and high-skilled immigration policy should not be so exclusively focused on the attraction and retention of elite scientists, but rather, should extend strategies to develop competitiveness by fostering international community and cultural diversity that serve all ranks of scientists. Moving forward on these issues, especially in consideration of the inherent difficulties—both ethical and statistical—in testing and empirically measuring the impacts of immigration policy [60], there is a need to develop better data-driven analytical and systems modelling methods to inform science policy on these issues. Indeed, anticipating the multiscale impacts of science policy is a formidable challenge calling for continued trans-disciplinary efforts to better understand the scientific enterprise [5,6,13,61].

4. Material and methods

4.1. Network-based author disambiguation method

We use the rich publication metadata in the APS dataset as input for a network-based author disambiguation method [37] that groups publications into researcher profiles. The Helbing algorithm leverages three key pieces of information available for each publication: (i) the co-author names, (ii) the publications listed in the reference list, i.e. outgoing citations, and (iii) the list of incoming citations from other publications (note that incoming and outgoing citations are restricted to publications within the APS dataset). Importantly, we did not provide any geographical affiliation information to the disambiguation algorithm, and so the resolved researcher profiles are free of geographical biases, and thus, well-suited for the study of cross-border mobility.

In an effort to assess statistically significant changes in publication profile attributes, and to reduce the frequency of spurious fluctuations arising from small sample size, we only analysed profiles with $N_i \geq 10$ publications in the APS dataset and initial publication year $y_i^0 \geq 1985$, resulting in 26 170 researcher profiles; electronic supplementary material, figure S1(C) shows the distribution y_i^0 for our final dataset, with 1995 as the median starting-year value.

4.2. Specification of research activity measures

4.2.1. Citation impact

In order to compare the citation impact of publications (indexed by p) from different years (t), we apply a normalization method that maps the integer citation count $n_{i,p,t}$ to a z-score using a logarithmic transform. To be specific, $z_{i,p} \equiv (\ln(n_{i,p,t}+1) - \mu_t)/\sigma_t$, where $\mu_t = \langle \ln(n_t+1)\rangle$ and $\sigma_t = \sigma[\ln(n_t+1)]$ denote the mean and standard deviation calculated over the set of publications from a specific year cohort (t). By mapping citations to their log value (adding $+1$ to avoid the divergence associated with ln 0), and accounting for the age-cohort-specific mean and standard deviation, $z_{i,p}$ follows a normal distribution $N(0,1)$ that is stable across t [11,13]. Publications with $z_{i,p} > 0$ are thus above the average log citation impact, μ_t, and since they are measured in units of σ, standard intuition and statistics of z-scores apply. As such, $z_{i,p}$ is well-suited for cross-temporal analysis, e.g. OLS regression, as well as averaging and summing within profiles. For this reason we calculate the average citation impact value $Z_i^{+,-} \equiv \langle z_{p,i}^{+,-}\rangle = (1/N_i^{+,-})\sum_{p\in N} z_{p,i}^{+,-}$ across the $N_i^{+,-}$ publications in each $\Delta t \equiv 5$-year interval, before $(-)$ or after $(+)$ $t_{i,T}^*$. Moreover, due to the properties of logs, the per cent difference in n_t associated with the mobility effect is given by $100\Delta n_p/n_p = 100 \times \sigma_t \times (\partial z/\partial W) \approx 100 \times \langle\sigma\rangle \times \tau_{W=1}[Z^+]$, which follows because σ_t is approximately constant over time, and so we approximate σ_t with the average value $\langle\sigma\rangle = 1.05$.

For comparison, we also considered an 'extensive' citation impact measure, as opposed to z_p, which is an 'intensive' impact measure. The deflated citation count $\hat{n}_{i,p,t}$ accounts for the fact that the total number of references produced by science is steadily growing with time (electronic supplementary material, figure S1(A)), the result of which is a 'citation inflation' measurement bias associated with the nominal citation count $n_{i,p,t}$ [50]. However, the 'deflated' variable $\hat{n}_{i,p,t}$ is well-suited for comparisons of citation counts for articles published in different years, and so we instead tally $\hat{n}_{i,p,t}$ for each interval, defining the total as $\Sigma_i^{+,-} \equiv \sum_{p\in +,-} \hat{n}_{i,p}^{+,-}$. As such, $\Sigma_i^{+,-}$ is amenable to modelling differences in total citation impact before and after $t_{i,T}^*$.

4.2.2. Co-authors

Within each $\Delta t \equiv 5$-year interval, K_i denotes the number of distinct co-authors we count for each i. Similarly, the number of publications including central researcher i and co-author j is k_{ij}; the total number of co-author instances across all N_i publications is $A_i = \sum_j k_{ij}$. We then use the Shannon entropy, a diversity index measuring the variety across the K_i co-authors, defined as $E_{K,i} = -\sum_{j=1}^{K_i} (k_{ij}/A_i)\ln(k_{ij}/A_i)$. More specifically, we calculate the entropy $E_{K,i}^-$ ($E_{K,i}^+$) using data in the Δt-year interval before (after) $t_{i,T}^*$. Higher levels of variation correspond to larger entropy values $E \geq 0$; the limiting case of no variation, $k_{ij} =$ const. for all j, corresponds to $E = 0$.

4.2.3. Research topics

Similar to (§4.2.2), we aggregate the PACS codes associated with each publication, a system used by the American Institute of Physics and implemented broadly in physics journals since 1975; see https://publishing.aip.org/publishing/pacs/pacs-2010-regular-edition. This five-level classification is comprised of more than 5000 individual PACS codes, which authors self-assign to their publications (e.g. '89.75.-k' corresponds to 'Complex systems'); we observe on average 2.5 unique PACS per publication with only 1% of publications having five or more PACS. We aggregated the PACS codes from all publications in each observation period into the two lists denoted by $q_j^{+,-}$. We then define the variation $E_{Q,i}^{+,-}$ in each $q_j^{+,-}$ list of PACS codes using the same Shannon entropy measure in (ii) above.

4.2.4. Geographical reach

As above, we aggregate the country codes associated with each publication affiliation into two lists, $C_i^{+,-}$. We define the variation $E_{C,i}^{+,-}$ in the categorical country code lists using the Shannon entropy (e.g. the list $C = \{\mathrm{NL, NL, NL, UK, IT, JP, AU}\}$ has entropy $E_C = 1.475$).

4.3. Robustness check of propensity score matching method

We randomized the treatment/control group classification variable $1_{G_i} = 3$ to test the likelihood that we could obtain an effect size as large as the observed $\tau_{W=1}[Y]$ by chance. To be specific, we shuffled the treatment indicator $1_{G_i} = 3$ within each $G3$–$G1$ estimation, without replacement, thereby conserving the total number of observations (researchers) classified as being mobile during a specific period (i.e. belonging to group $G_i = 3$). We applied this shuffling procedure 10 000 different times, each time recording the value of the 'placebo' estimate for $\tau_{W=1}[Y]$. Figure S9 shows the distribution of placebo estimates, $P(\tau_{W=1}[Y])$, for each Y; in estimates, with one exception (panel D for $\tau_{W=1}[E_Q^+]$), we find the observed treatment effect to be significantly larger and outside the 99% CI bounds of the placebo estimate, thereby demonstrating the robustness of our PSM specification to spurious correlations.

Data accessibility. We used two principal open-source datasets provided by the American Physical Society [36] and WolframAlpha [35]; raw mobility data will be made available upon request.

Competing interests. We declare we have no competing interests.

Funding. We received no funding for this study.

Acknowledgements. We thank the A.P.S. for making their data available. We are also grateful to two anonymous referees and F. Arroyave Bermudez for helpful comments, and acknowledge the developers of the open-source *Circos* circular network layout Table Viewer software [62] used to produce the Nobelist mobility network.

References

1. Geuna A (ed.). 2015 *Global mobility of research scientists*. Oxford, UK: Academic Press.
2. Hunter RS, Oswald AJ, Charlton BG. 2009 The elite brain drain. *Econ. J.* **119**, F231–F251. (doi:10.1111/j.1468-0297.2009.02274.x)
3. Azoulay P, Ganguli I, Zivin JG. 2017 The mobility of elite life scientists: professional and personal determinants. *Res. Policy* **46**, 573–590. (doi:10.1016/j.respol.2017.01.002)
4. Petersen AM, Puliga M. 2016 High-skilled labour mobility in Europe before and after the 2004 enlargement. *J. R. Soc. Interface* **14**, 20170030. (doi:10.1098/rsif.2017.0030)

5. Stephan P. 2012 *How economics shapes science*. Cambridge, MA: Harvard University Press.
6. Fortunato S *et al.* 2018 Science of science. *Science* **359**, eaaao0185. (doi:10.1126/science.aao0185)
7. Agrawal A, Cockburn I, McHale J. 2006 Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships. *J. Econ. Geogr.* **6**, 571–591. (doi:10.1093/jeg/lbl016)
8. Breschi S, Lissoni F. 2009 Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *J. Econ. Geogr.* **9**, 439–468. (doi:10.1093/jeg/lbp008)
9. Petersen AM, Fortunato S, Pan RK, Kaski K, Penner O, Rungi A, Riccaboni M, Stanley HE, Pammolli F. 2014 Reputation and impact in academic careers. *Proc. Natl Acad. Sci. USA* **111**, 15 316–15 321. (doi:10.1073/pnas.1323111111)
10. Mazloumian A, Eom YH, Helbing D, Lozano S, Fortunato S. 2011 How citation boosts promote scientific paradigm shifts and Nobel prizes. *PLoS ONE* **6**, e18975. (doi:10.1371/journal.pone.0018975)
11. Petersen AM. 2015 Quantifying the impact of weak, strong, and super ties in scientific careers. *Proc. Natl. Acad. Sci. USA* **112**, E4671–E4680. (doi:10.1073/pnas.1501444112)
12. Franzoni C, Rossi-Lamastra C. 2017 Academic tenure, risk-taking and the diversification of scientific research. *Ind. Innovation* **24**, 691–712. (doi:10.1080/13662716.2016.1264067)
13. Petersen AM, Majeti D, Kwon K, Ahmed ME, Pavlidis I. 2018 Cross-disciplinary evolution of the genomics revolution. *Sci. Adv.* **4**, eaat421. (doi:10.1126/sciadv.aat4211)
14. Jonkers K, Cruz-Castro L. 2013 Research upon return: the effect of international mobility on scientific ties, production and impact. *Res. Policy* **42**, 1366–1377. (doi:10.1016/j.respol.2013.05.005)
15. Scellato G, Franzoni C, Stephan P. 2015 Migrant scientists and international networks. *Res. Policy* **44**, 108–120. (doi:10.1016/j.respol.2014.07.014)
16. Doria Arrieta OA, Pammolli F, Petersen AM. 2017 Quantifying the negative impact of brain drain on the integration of European science. *Sci. Adv.* **3**, e1602232. (doi:10.1126/sciadv.1602232)
17. Fleming L, Sorenson O. 2004 Science as a map in technological search. *Strategic Manage. J.* **25**, 909–928. (doi:10.1002/smj.384)
18. Fleming L, Mingo S, Chen D. 2007 Collaborative brokerage, generative creativity, and creative success. *Adm. Sci. Q.* **52**, 443–475. (doi:10.2189/asqu.52.3.443)
19. Foster JG, Rzhetsky A, Evans JA. 2015 Tradition and innovation in scientists' research strategies. *Am. Sociol. Rev.* **80**, 875–908. (doi:10.1177/0003122415601618)
20. Jia T, Wang D, Szymanski B. 2017 Quantifying patterns of research-interest evolution. *Nat. Hum. Behav.* **1**, 0078. (doi:10.1038/s41562-017-0078)
21. Laudel G. 2005 Migration currents among the scientific elite. *Minerva* **43**, 377–395. (doi:10.1007/s11024-005-2474-7)
22. Hoisl K. 2007 Tracing mobile inventors—the causality between inventor mobility and inventor productivity. *Res. Policy* **36**, 619–636. (doi:10.1016/j.respol.2007.01.009)
23. Marx M, Strumsky D, Fleming L. 2009 Mobility, skills, and the Michigan non-compete experiment. *Manage. Sci.* **55**, 875–889. (doi:10.1287/mnsc.1080.0985)
24. Cruz-Castro L, Sanz-Menendez L. 2010 Mobility versus job stability: assessing tenure and productivity outcomes. *Res. Policy* **39**, 27–38. (doi:10.1016/j.respol.2009.11.008)
25. Edler J, Fier H, Grimpe C. 2011 International scientist mobility and the locus of knowledge and technology transfer. *Res. Policy* **40**, 791–805. (doi:10.1016/j.respol.2011.03.003)
26. Franzoni C, Scellato G, Stephan P. 2014 The mover's advantage: the superior performance of migrant scientists. *Econ. Lett.* **122**, 89–93. (doi:10.1016/j.econlet.2013.10.040)
27. Baruffaldi SH, Landoni P. 2012 Return mobility and scientific productivity of researchers working abroad: the role of home country linkages. *Res. Policy* **41**, 1655. (doi:10.1016/j.respol.2012.04.005)
28. Deville P, Wang D, Sinatra R, Song C, Blondel VD, Barabasi AL. 2014 Career on the move: geography, stratification, and scientific impact. *Sci. Rep.* **4**, 4770. (doi:10.1038/srep04770)
29. Baker A. 2015 Non-tenured post-doctoral researchers' job mobility and research output: an analysis of the role of research discipline, department size, and coauthors. *Res. Policy* **44**, 634–650. (doi:10.1016/j.respol.2014.12.012)
30. Wang D. 2015 Activating cross-border brokerage: interorganizational knowledge transfer through skilled return migration. *Adm. Sci. Q.* **60**, 133–176. (doi:10.1177/0001839214551943)
31. Franzoni C, Scellato G, Stephan P. 2015 International mobility of research scientists: lessons from globsci. In *Global mobility of research scientists*, pp. 35–65. Amsterdam, the Netherlands: Elsevier.
32. Kahn S, MacGarvie M. 2016 Do return requirements increase international knowledge diffusion? Evidence from the Fulbright program. *Res. Policy* **45**, 1304–1322. (doi:10.1016/j.respol.2016.02.002)
33. Sugimoto CR, Robinson-García N, Murray DS, Yegros-Yegros A, Costas R, Larivière V. 2017 Scientists have most impact when they're free to move. *Nature* **550**, 29–31. (doi:10.1038/550029a)
34. Fernandez-Zubieta A, Geuna A, Lawson C. 2015 What do we know of the mobility of research scientists and impact on scientific production? In *Global mobility of research scientists*, pp. 1–33. Amsterdam, the Netherlands: Elsevier.
35. Wolfram Alpha LLC. 2016 Nobel Prize in physics. See https://www.wolframalpha.com/input/?i=Nobel+Prize+in + Physics.
36. American Physical Society. 2009 APS data sets for research. See http://journals.aps.org/datasets.
37. Schulz C, Mazloumian A, Petersen AM, Penner O, Helbing D. 2014 Exploiting citation networks for large-scale author name disambiguation. *EPJ Data Sci.* **3**, 11. (doi:10.1140/epjds/s13688-014-0011-3)
38. Chait R. 2009 *The questions of tenure*. Cambridge, MA: Harvard University Press.
39. Kaplan K. 2010 Academia: the changing face of tenure. *Nature* **468**, 123–125. (doi:10.1038/nj7320-123a)
40. Cyranoski D, Gilbert N, Ledford H, Nayar A, Yahia M. 2011 Education: the PhD factory. *Nature* **472**, 276–279. (doi:10.1038/472276a)
41. Powell K. 2015 The future of the postdoc. *Nature* **520**, 144–147. (doi:10.1038/520144a)
42. Kaminski D, Geisler C. 2012 Survival analysis of faculty retention in science and engineering by gender. *Science* **335**, 864–866. (doi:10.1126/science.1214844)
43. Petersen AM, Riccaboni M, Stanley HE, Pammolli F. 2012 Persistence and uncertainty in the academic career. *Proc. Natl Acad. Sci. USA* **109**, 5213–5218. (doi:10.1073/pnas.1121429109)
44. Way SF, Morgan AC, Clauset A, Larremore DB. 2017 The misleading narrative of the canonical faculty productivity trajectory. *Proc. Natl Acad. Sci. USA* **114**, E9216–E9223. (doi:10.1073/pnas.1702121114)
45. Duch J, Zeng XHT, Sales-Pardo M, Radicchi F, Otis S, Woodruff TK, Amaral LAN. 2012 The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLoS ONE* **7**, e51332. (doi:10.1371/journal.pone.0051332)
46. Jadidi M, Karimi F, Lietz H, Wagner C. 2017 Gender disparities in science? Dropout, productivity, collaborations and success of male and female computer scientists. *Adv. Complex Syst.* **21**, 1750011. (doi:10.1142/s0219525917500114)
47. Way SF, Larremore DB, Clauset A. 2016 Gender, productivity, and prestige in computer science faculty hiring networks. In *Proc. of the 25th Int. Conf. on World Wide Web*, pp. 1169–1179.
48. Lerchenmueller MJ, Sorenson O. 2018 The gender gap in early career transitions in the life sciences. *Res. Policy* **47**, 1007–1017. (doi:10.1016/j.respol.2018.02.009)
49. Appelt S, van Beuzekom B, Galindo-Rueda F, de Pinho R. 2015 Which factors influence the international mobility of research scientists? In *Global mobility of research scientists*, pp. 177–213. Amsterdam, the Netherlands: Elsevier.
50. Petersen AM, Pan RK, Pammolli F, Fortunato S. 2018 Methods to account for citation inflation in research evaluation. SSRN e-print: 3193712.
51. Rubin DB. 2005 Causal inference using potential outcomes. *J. Am. Stat. Assoc.* **100**, 322–331. (doi:10.1198/016214504000001880)
52. Rosenbaum PR, Rubin DB. 1983 The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55. (doi:10.1093/biomet/70.1.41)
53. Abadie A, Drukker D, Herr JL, Imbens GW. 2004 Implementing matching estimators for average treatment effects in STATA. *STATA J.* **4**, 290–311.

54. Fleming L. 2001 Recombinant uncertainty in technological search. *Manage. Sci.* **47**, 117–132. (doi:10.1287/mnsc.47.1.117.10671)
55. Chessa A, Morescalchi A, Pammolli F, Penner O, Petersen AM, Riccaboni M. 2013 Is Europe evolving toward an integrated research area? *Science* **339**, 650–651. (doi:10.1126/science.1227970)
56. Page SE. 2008 *The difference: how the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.
57. Allison PD, Long JS. 1990 Departmental effects on scientific productivity. *Am. Sociol. Rev.* **55**, 469–478. (doi:10.2307/2095801)
58. Fernández-Zubieta A, Geuna A, Lawson C. 2015 Mobility and productivity of research scientists. In *Global mobility of research scientists*, pp. 105–131. Amsterdam, the Netherlands: Elsevier.
59. Kerr WR, Lincoln WF. 2010 The supply side of innovation: H-1B visa reforms and US ethnic invention. *J. Labor. Econ.* **28**, 473–508. (doi:10.1086/651934)
60. Simon M, Schwartz C, Hudson D, Johnson SD. 2018 A data-driven computational model on the effects of immigration policies. *Proc. Natl Acad. Sci. USA* **115**, E7914–E7923. (doi:10.1073/pnas.1800373115)
61. Fealing KH (eds) 2011 *The science of science policy: a handbook*. Stanford, CA: Stanford Business Books.
62. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009 Circos: an information aesthetic for comparative genomics. *Genome. Res.* **19**, 1639–1645. (doi:10.1101/gr.092759.109)

Supplementary Material:

Multiscale impact of researcher mobility

Appendix, Figures S1-S9, and Tables S1-S6

Alexander M. Petersen[1]

[1] Ernest and Julio Gallo Management Program
Management of Complex Systems Department,
School of Engineering, University of California, Merced, CA 95343

S1. The American Physical Society dataset

We analyzed the 2009 American Physical Society (APS) *Physical Review article and citations* dataset, which is openly available in well-documented XML data files that record article-level author-byline data and PACS classification information for publications from the Physical Review journal family: *Physical Review A* (PRA), *Physical Review B* (PRB), *Physical Review C* (PRC), *Physical Review D* (PRD), *Physical Review E* (PRE), *Physical Review Letters* (PRL), and *Reviews of Modern Physics* (RMP). The publication metadata is homogenized and stable over time, and includes: (i) author name(s), (ii) affiliations with pointers to particular authors, (iii) citation data between APS articles, and (iv) Physics and Astronomy Classification Scheme (PACS) codes.

Data elements (i) and (iii) are inputs for the disambiguation of authors, detailed in the next subsection. It is important to note that the disambiguation algorithm we used does not use the affiliation metadata.. If the disambiguation algorithm *did* use the affiliation data, then there would be an increased likelihood of splitting researcher profiles according to intra-region publication clusters, which would not only increase the splitting (false negative) rate of researcher profiles into 2 or more clusters, but would under-represent the rate of researcher mobility. Instead, the clustering algorithm is not biased by geographic information contained in the affiliation data (ii). As a result, the publication clusters produced by the disambiguation algorithm – corresponding to disambiguated researcher profiles – are particularly amenable to geographic mobility analysis.

Author disambiguation method leveraging the collaboration and citation network. Because the PACS system was initiated in 1975, we include a 5-year buffer period before this year and the start year of our refined dataset. Hence, we analyzed 355,808 publications from 1980 – 2009. We then implemented the Helbing disambiguation algorithm [36]. This algorithm uses the citation network and the collaboration network to cluster publications into groups that are likely to correspond to an individual researcher. To be specific, the algorithm calculates a similarity score between any two given publications based on the overlap of (a) coauthor names, (b) the list of references cited by each publication, (c) the list of publications citing each publication, and (d) the particular scenario of direct citations between the two publications. This method was developed for large-scale data using the complete Web of Science dataset; Google Scholar profiles were used as a gold standard to obtain the algorithm parameters based on precision and recall error, in addition to several additional validation methods, including a theoretical model of the $h-$index distribution. Given the generality of this algorithm to scenarios in which the citation and collaboration network data are available, we applied it to the APS dataset using the optimal parameters reported in Shulz et al. [36].

More specifically, the algorithm works as follows. The starting point is the set of N_x publications that all list a given coauthor name, e.g. corresponding to the concatenated string A_x="LastName_FirstNameInitial" (e.g. Smith_A). We then calculate a similarity score between every pair of publications p and p' using a linear combination of weights for 4 factors: (i) the similarity

[1] Please send correspondence to: Alexander M. Petersen
E-mail: apetersen3@ucmerced.edu

in the set of coauthors, (ii) the self-citation scenario where on of p or p' cites the other, (iii) similarity in the reference lists of either publication, and (iv) similarity in the set of publications citing each publication. The algorithm first clusters the N_x publications into subgroups, and then merges subgroups into researcher profiles in a multi-step procedure.

Application of this algorithm produced A =208,734 publication clusters, indexed by $i = 1...A$, with each cluster containing N_i unique publications corresponding to the researcher profile of the author "LastName_FirstNameInitial$\#i$" (e.g. Smith_A#5). Figure S1(B) shows the distribution $P(N_i)$ of the number of publications per researcher profile.

Author selection procedure. We restricted our analysis to authors with greater than $N_i \geq 10$ publications spread over ≥ 3 distinct years and first publication $y_i^0 \geq 1985$; we implemented the last threshold to account for left censoring bias, i.e. to reduce the number of researchers in our analysis whose first publication was actually prior to 1980. As a final refinement, we excluded researcher profiles with fewer than three publications in either the period before or after $t^*_{i,T}$ and fewer than four distinct years of activity. The result of this additional selection is 26,170 APS researcher profiles corresponding to 206,272 distinct publications, which were cited 2,184,619 times altogether over their collective 986,287 years of citation activity. The total number of author career-year observations is 388,079, or roughly 15 career years per researcher profile.

Estimation of mobility year from raw publication data. The APS data has remarkably "clean" author affiliation data, which we used to geo-locate the individual articles by using string matches for country names, ISO2 and ISO3 country codes, and also the full names and 2-letter codes of US states which were used to classify affiliations that did not include "USA" but did include US State codes. Because the APS metadata has specific tags to link each researcher with one or more specific affiliations, we were able to link an individual i to specific countries and US states. When an author was affiliated with 2 or more countries in a given year, we tallied up these affiliations and assigned the primary location as the most common country within that year. In the case of a tie, we instead aggregated the affiliation data for the previous 3-year period and then used the most common country, which resolved 100% of the ties. Applying this geolocation method, we obtained an annual primary location time series for each author over the 30-year period 1980-2009; in the years in which the author did not publish in the APS dataset we denoted the primary affiliation as "blank". We then filled in the "blank" years in which the primary location before and after the blank period matched.

When the primary locations differ, before and after a period of $\delta y (\geq 0)$ "blank" years, this points to a mobility event. We estimate the mobility year t^*_i by first defining $y^+ \equiv y^- + \delta y$, where y^+ (y^-) is the first year after (before) the gap of "blank" years. If $\delta y = 1$ then we define the mobility transition year $t^*_i \equiv y^-$, and if $\delta y > 1$ then $t^*_i \equiv y^+ - \lceil \delta y/2 \rceil$.

S2. Research activity measures

Figures S3-S4 show the distribution of each model variable during the pre-mobility period $t \in [t^*_T - 5, t^*_T - 1]$ and post-mobility period $t \in [t^*_T, t^*_T + 4]$, respectively. We also calculated the change in the dependent variables for each researcher, between the pre- and post-treatment periods, as follows:

(i) *Citation impact*: We define the 2-period change in mean normalized citation impact as $\Delta Z_i \equiv Z_i^+ - Z_i^-$.

(ii) *Coauthors*: As a measure of the 2-period change in the coauthor list, we calculate the similarity between the two lists using a variant of the cosine similarity, $S_{K,i} \equiv S[k^+_{ij}, k^-_{ij}] = (|k^+_{ij}||k^-_{ij}|)^{-1} \sum_j k^+_{ij} k^-_{ij}$, where $|k_{ij}| = \sqrt{(\sum_j k^2_{ij})}$ is the euclidian norm of the list in which the order of the categories (j) are matched so that they correspond to the same entity (e.g. coauthor) in k^+_{ij} and k^-_{ij}. Since $k_{ij} \geq 0$, $S[k^+_{ij}, k^-_{ij}] \in [0, 1]$, with maximum correspondence only when $k^+_{ij} = k^-_{ij}$ for all j.

(iii) *Research topics*: As in (ii) we measure the change in the PACS lists using the similarity distance $S_{PACS,i} \equiv S[q^+_j, q^-_j]$.

(iv) *Geographic reach*: As above we measure the change in the list of country codes drawn from the affiliation lists of each publication using the similarity distance $S_{C,i} \equiv S[C^+_j, C^-_j]$.

Figure S5 shows the distribution of ΔZ_i, $S_{K,i}$, $S_{PACS,i}$, and $S_{C,i}$, measuring the characteristic scale of research profile shifts, before and after $t^*_{i,T}$.

S3. Country classification

We classified countries into geographic regions as follows (2-letter ISO codes followed in parenthesis by the number of affiliations recorded for a corresponding country):

- **[Africa]**: ZA (1758), MA (204), EG (199), DZ (99), CM (61), TN (41), ET (17), NG (12), NA (5), LY (5), KE (4), MG (2), ZW (2), TZ (1), BI (1), GA (1), LS (1), GN (1), BW (1)
- **[Asia & Australasia]**: JP (155447), CN (38023), IN (30066), KR (26298), TW (18053), AU (14441), HK (4249), SG (2226), NZ (1661), AM (1256), IR (598), PK (325), UZ (318), PH (231), SA (224), VN (219), KZ (192), BD (143), TH (118), ID (78), MY (77), LB (53), JO (45), QA (44), KW (43), MN (43), AE (36), AZ (31), GE (29), OM (9), MO (5), BH (4), KG (3), IQ (3), PS (2), NP (1), SY (1)
- **[Europe]**: DE (151210), IT (147885), FR (109531), UK (98944), ES (31843), NL (25129), SE (18397), BE (10623), DK (9144), AT (8702), FI (8376), GR (4330), PT (2914), IE (1911), LU (18), PL (16132), HU (5424), CZ (4935), SI (3898), RO (2155), SK (1288), BG (1087), LV (275), LT (247), EE (238), CY (136), MT (5), CH (96970), RU (54597), IL (17797), NO (3935), UA (2939), HR (1913), YUGO (1260), TR (1240), CS (688), BY (418), RS (384), ME (109), IS (104), MD (84), MK (42), JE (8), AL (2)
- **[North America]**: USA (1,360,653), CA (63645), MX (6628)
- **[South America, Central America, and Carribean]**: BR (24304), AR (7553), CO (2022), CL (1573), VE (568), EC (235), CU (230), UY (187), PE (51), CR (14), BO (13), JM (12), PA (5), GD (4), BB (2), GY (1)

We classify the origin country (c_i^-) according to 5 broad regions, denoted by the factor variable F_i^- in the Propensity Score Matching and regression model specifications: (a) Europe; (b) N. America; (c) Central America, South America and the Caribbean; (d) Asia/Australia and (e) Africa. Because there were not many researchers from Africa with sufficiently large publication profile to meet our pruning criteria, observations associated with this region were excluded from our model estimates.

S4. Modeling mobility with the Logit model

We analyzed the factors that correlate with mobility in period T by modeling the dependent binary indicator variable $1_{G_i=3}$ – which takes the value 1 if $G_i = 3$ and 0 otherwise – by applying Logistic regression. This Logit model is specified within the Propensity Score Matching method to identify matched pairs [51]. We focus on just two sets of researchers for a given period, those researchers with $G_i = 1$ (not mobile up to and including the upper limit year t_T^+ of the period T) and $G_i = 3$ (mobile in T). Thus, we model the likelihood $P(G_{i=3})$ that a researcher is mobile given his/her research profile information, and so the binary outcomes follow the simple relation $P(G_i = 3) + P(G_i = 1) = 1$. For each i we included 5 variables measured, as previously, for the $\Delta t \equiv 5$-year period before $t_{i,T}^*$: the number of distinct coauthors, $|k_{ij}^-|$, the number of publications, N_i^-, the mean citation impact Z_i^-, the researcher age, s_i^*, and a factor variable representing the researcher's geographic region, F_i^-.

We model the odds $O \equiv P(G_i = 3)/P(G_i = 1)$ according to the Logit regression model specified as

$$\log\Big(\frac{P(G_i = 3)}{P(G_i = 1)}\Big) = \beta_1 |k_{ij}^-| + \beta_2 N_i^- + \beta_3 Z_i^- + \beta_4 s_i^* + \beta_0 + F_i^- + \epsilon \,, \tag{S1}$$

which we estimate using robust standard errors. Table S1 reports the exponentiated coefficient, $\exp(\beta)$, which is the odds ratio, or factor by which the odds O changes for each 1-unit increase in the corresponding independent variable, i.e. $O_{+1}/O = \exp(\beta)$; put another way, $100(\exp(\beta)-1)$ is the percent change in O corresponding to a 1-unit increase in the corresponding independent variable. As a result, reported $\exp(\beta)$ values that are less than (greater than) unity indicate variables that negatively (positively) correlate with cross-border mobility.

The results of the model show that more coauthors correlate with a marginally smaller likelihood of migration for all T. Higher productivity (N_i^-) and citation impact (Z_i^-) correlate with a statistically significant higher likelihood of migration for T_1 and T_2 but not T_3, suggesting that mobility is becoming less contingent on researcher prestige. The most significant correlate is researcher age, which indicates a strong and statistically significant negative relation between increasing research age and likelihood of migration, observed for all T. The factor variables capturing the geographic region of residence prior to mobility (F_i^-) indicate that, relative to N. America (the most likely to migrate), a researcher residing in S. & C. America is the second most likely to migrate, followed by researchers from Europe, and then Asia & Australasia, in that order.

S5. Matched regression

While Propensity Score Matching is suitable for estimating the impact of treatment on post-treatment outcomes, it does not provide guidance as to the causal link between certain pre-treatment factors and the differential outcome. In order to estimate the degree to which certain researcher variables prior to $t^*_{i,T}$ correlate with the same set of variables after $t^*_{i,T}$, we used the set of matched researcher pairs (i, i') identified by the Propensity Score Matching method to regress each outcome (dependent) variable Y_i^+ against the set of pre-treatment matching variables denoted by $\vec{X}$. For example, in the first case where $Y_i^+ \equiv Z_i^+$, we regressed the post-migration average citation impact, Z_i^+, against the pre-migration variables $\vec{X} = (Z_i^-, |k_{ij}^-|, N_i^-, s_i^*, F_i^-, 1_{G_i=3})$, where F_i^- indicates a factor variable for the researcher's geographic sub-region (determined by the home country c_i^- prior to $t^*_{i,T}$), and $1_{G_i=3}$ is a binary indicator variable equal to 1 if the researcher migrated in period T and 0 otherwise. We performed OLS regression on the set of matched observations (i, i') according to the linear model

$$Z_i^+ = \beta_1 |k_{ij}^-| + \beta_2 N_i^- + \beta_3 Z_i^- + \beta_4 s_i^* + \beta_5 1_{G_i=3} + \beta_0 + F_i^- + \epsilon \,. \tag{S2}$$

Table S2 shows the results in columns (1,3,5) for each sample period T, respectively. The coefficient $\beta_5 \approx \tau_{W=1}[Y \equiv Z]$ reported in Fig. 3 for each T. That is, the treatment effect calculated by estimating the mean pairwise difference $Y_i - Y_{i'}$ between the matched researcher pairs (see Eq. [3]) is consistent with the difference in Z_i^+ between the two groups, controlling for $\vec{X}$. That is, the PSM treatment effect estimate is not confounded by $\vec{X}$.

Similarly, the model estimates reported in columns (2,4,6) of Table S2 correspond to the same model but including additional interaction terms between the scalar variables and the mobility indicator variable,

$$Z_i^+ = (\beta_1 |k_{ij}^-| + \beta_2 N_i^- + \beta_3 Z_i^- + \beta_4 s_i^*) \times 1_{G_i=3} + \beta_5 1_{G_i=3} + \beta_0 + F_i^- + \epsilon \,. \tag{S3}$$

As a result, this model specification yields two coefficients for each interacted covariate, one coefficient ($\beta_{x,G_i=3}$) derived from observations with $G_i = 3$ and a second coefficient ($\beta_{x,G_i=1}$) derived from those with $G_i = 1$. Tables S2-S6 report the coefficient $\beta_{x,G_i=1}$ followed by the difference in the two coefficients $\delta_3(x) \equiv \beta_{x,G_i=3} - \beta_{x,G_i=1}$, which facilitates identifying covariates that distinguish the mobile/treated (G3) and not-mobile/untreated (G1) groups.

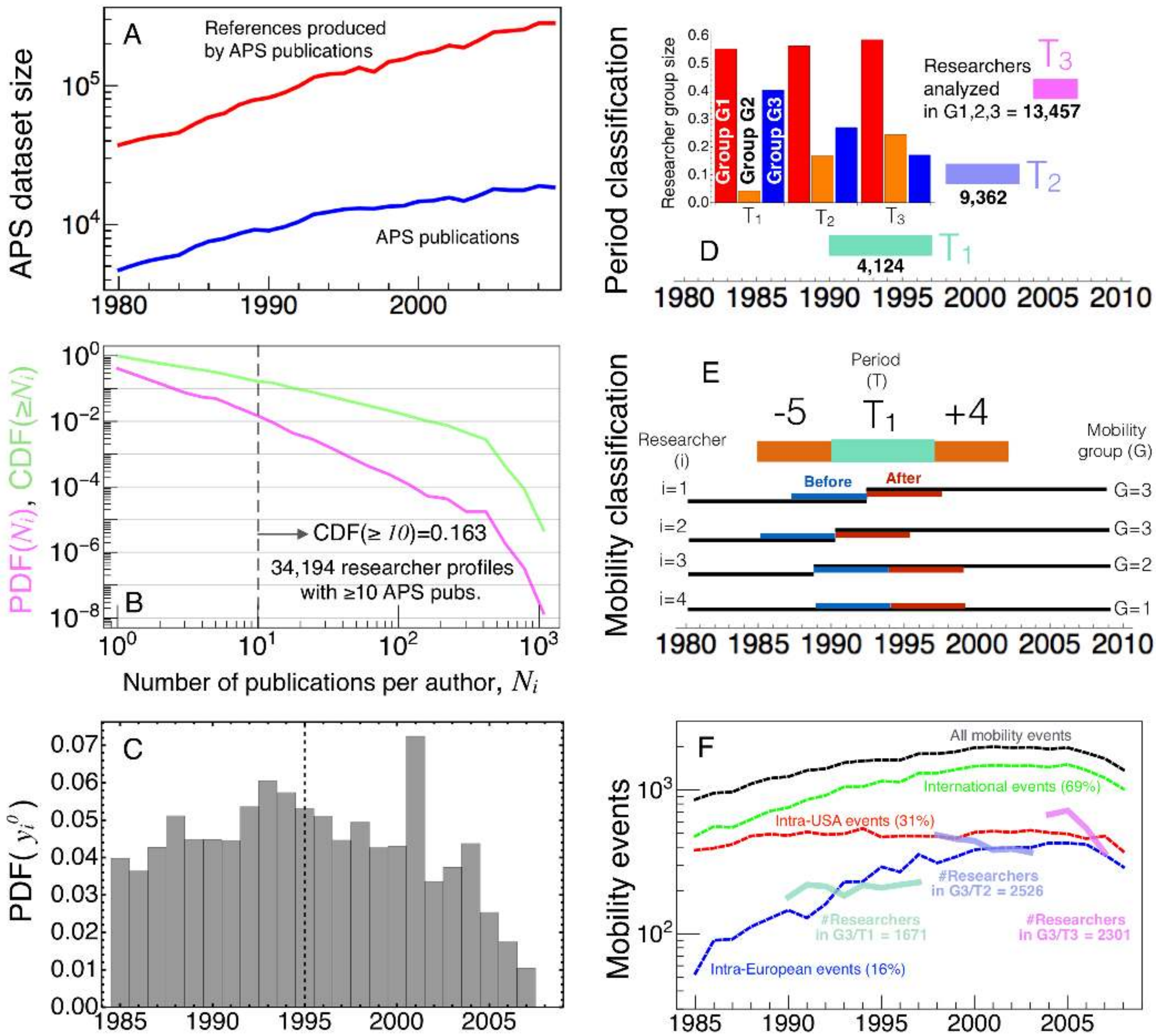

FIG. S1: **Data summary and 3-period observational framework**. **(A)** The total number of publications per year across the APS journals PRA, PRB, PRC, PRD, PRE, PRL, and RMP (blue), and the total number of references made by these publications that cite other APS publications within this journal set (red). Combined, the total number of publications is growing at roughly a 4.6% annual rate, and the total number of references made is growing at roughly a 7.2% annual rate over 1980–2009. **(B)** The distribution $P(N_i)$ of APS publications per researcher profile; 16.3% of the disambiguated researcher profiles have $N_i \geq 10$ corresponding to 34,194 profiles. **(C)** Distribution of researcher profiles according to their first APS publication year. We only analyzed researcher profiles with $y_i^0 \geq 1985$ and $N_i \geq 10$ publications spread across at least 3 distinct years, resulting in a total of 26,170 profiles. **(D)** We separated the mobility analysis into 3 non-overlapping observation periods, denoted by T, ensuring that each researcher contributes to the analysis of each T just once. (inset) Shown are the fraction of researchers belonging to a given mobility group G_T for a given T. The total number of researcher profiles by period are: 4,124 in T_1; 9,362 in T_2; 13,457 in T_3. Researchers (indexed by i) from the same T but different G are paired in the PSM analysis in order to estimate counterfactual outcomes. **(E)** Schematic of the classification process for 4 researcher profiles with respect to the observation period T_1: researchers 1 and 2 were mobile (indicated by the disjoint line) within the T_1 interval – thus they both belong to group $G3$, and so we aggregate the publication data in the 5-year window before and after the mobility event specific to each i; researcher 3 was mobile prior to T_1 but not during T_1, and so we use the midpoint of T_1 as a placebo mobility year and aggregate his/her publication data before and after the midpoint year of T_1 and assign this researcher to the placebo group $G2$; researcher 4 was neither mobile prior to nor during T_1, and thus belongs to the group $G1$. **(F)** Dashed lines correspond to the number of mobility events observed per year, allowing for multiple events per researcher profile: (blue) Intra-European32 mobility (e.g. DE to FR; EU32 corresponds to 28 EU members and CH, NO, LI, and IS); (red) Intra-USA state mobility (e.g. MA to CA); (green) International mobility (e.g. IT to USA); (black) All cross-border mobility, including both international and also inter-US state. Solid lines correspond to the number of mobile researchers in $G3$ by each period T. Note that even if a mobile researcher moved two or more times in a given T (i.e. multiple mobility events), this latter $G3$ researcher tally only counts these researchers once.

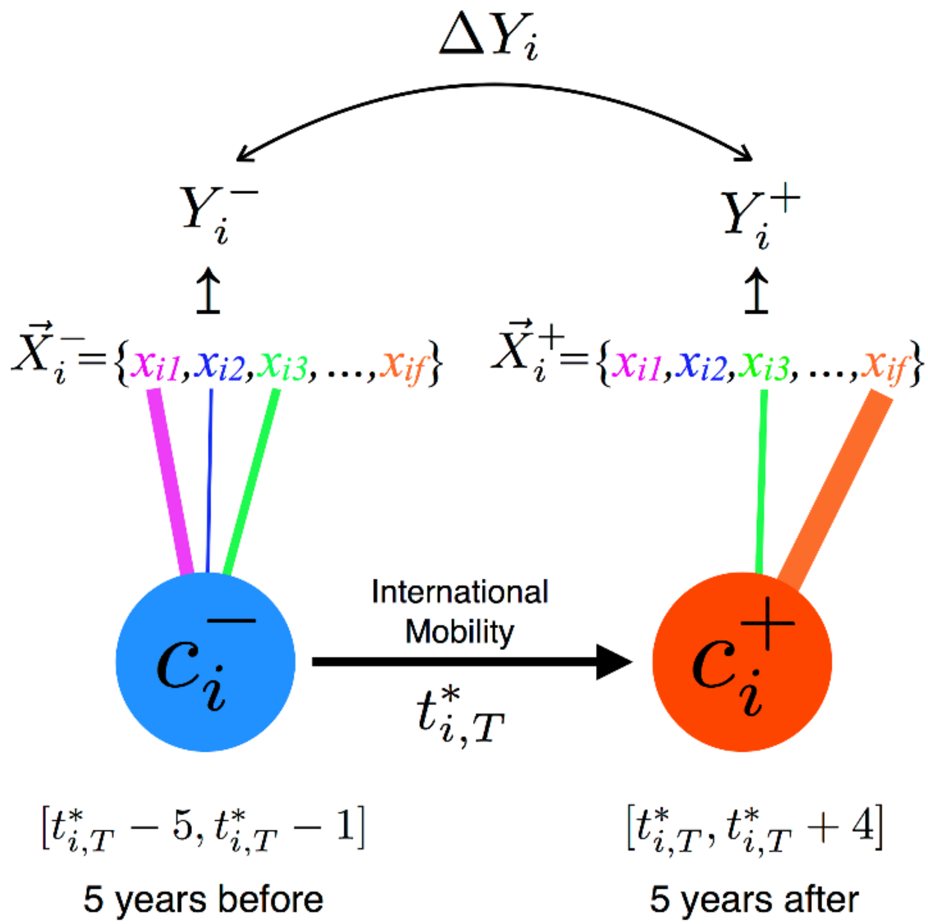

FIG. S2: **Schematic of researcher mobility framework**. For each researcher i we record their attributes $\vec{X}_i^{+,-}$ during the 5-year periods before and after the mobility event occurring in year $t^*_{i,T}$, from country c_i^- to country c_i^+. The weighted element x_{ij} represents a particular attribute, which by way of example, may be the number of publications with a particular collaborator, the number of instances of a particular PACS "keyword" capturing research topics, or other attributes of a single publication such as its citation count n_p or the set of countries C_p listed in the affiliation byline. We define a summary outcome variable $Y_i^{+,-}$, determined by particular information contained in $\vec{X}_i^{+,-}$, which facilitates: (a) measuring the change ΔY_i in the researcher profile attribute; (b) matching mobile and non-mobile researchers according to $\vec{X}_i^-$ and Y_i^- in order to obtain a causal estimate of the impact of mobility on Y_i^+.

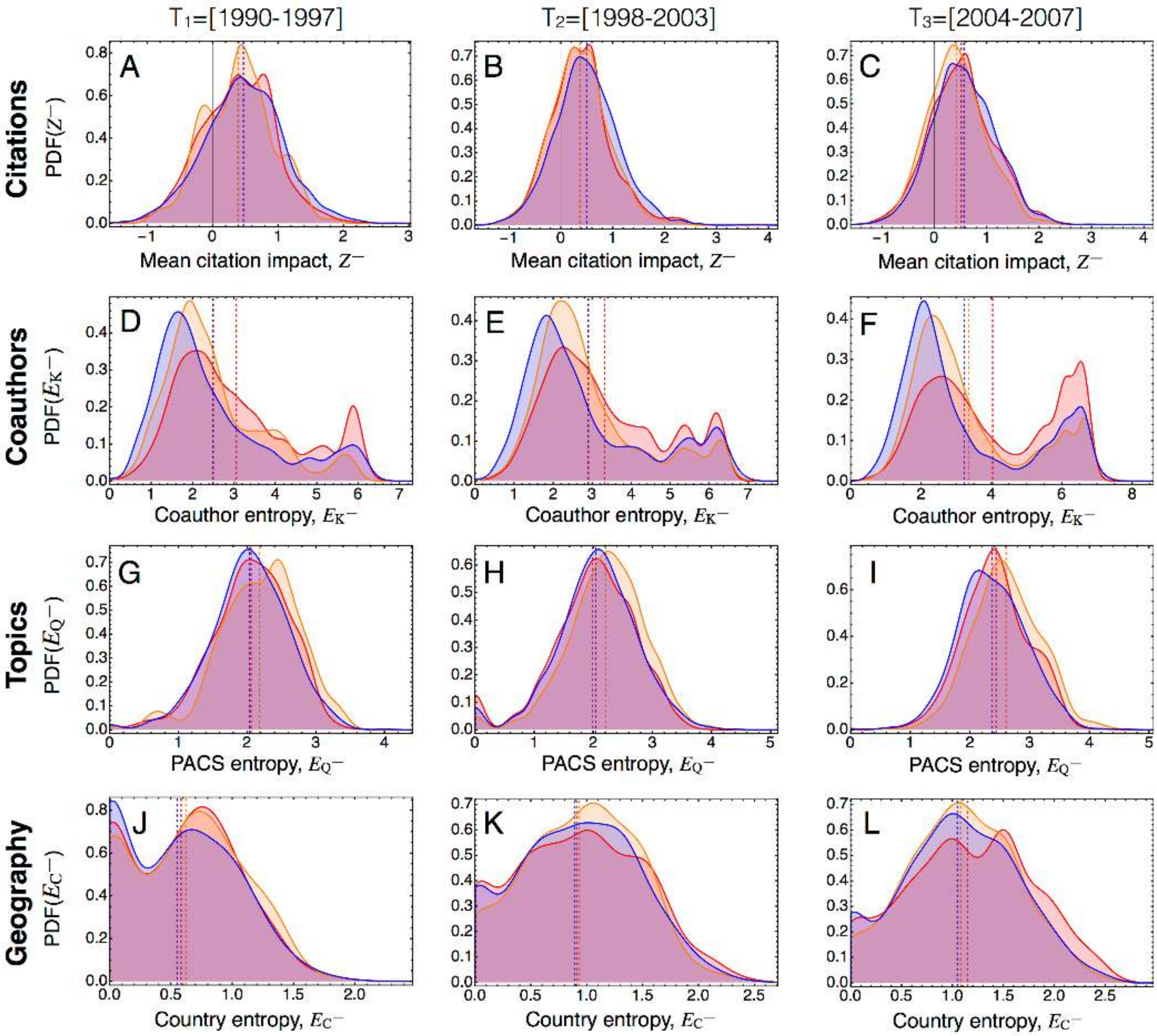

FIG. S3: **Distribution of PSM dependent variable values – before t^* by period**. Distributions demonstrate a high degree of stability between the three subgroups $G1$ (red: not mobile prior to the end of T), $G2$ (orange: mobile prior to the beginning of T but not mobile during T), and $G3$ (blue: mobile during T).

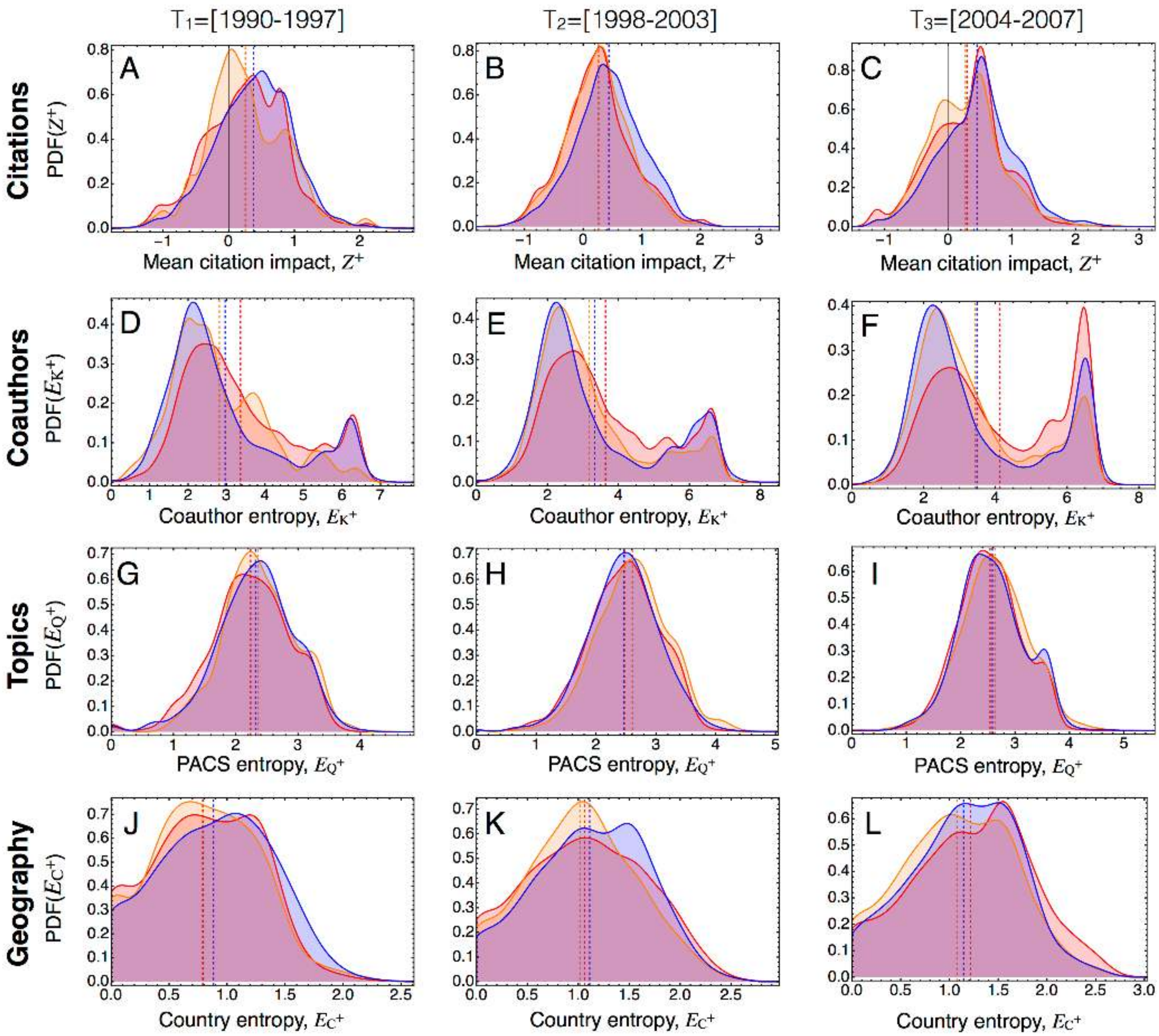

FIG. S4: **Distribution of PSM dependent variable values – after t^* by period**. Distributions demonstrate a high degree of stability between the three subgroups $G1$ (red: not mobile prior to the end of T), $G2$ (orange: mobile prior to the beginning of T but not mobile during T), and $G3$ (blue: mobile during T).

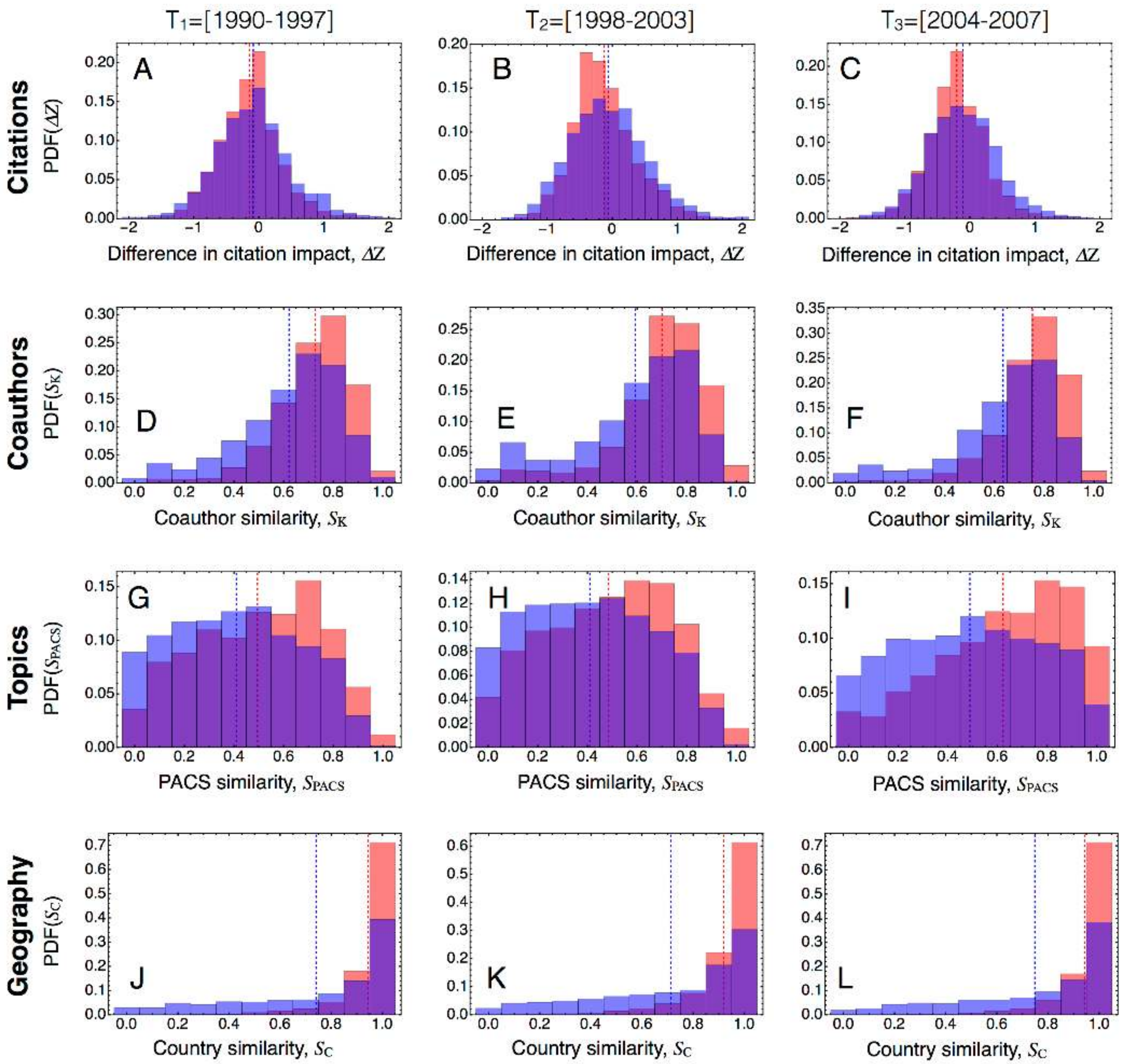

FIG. S5: **Distribution of change in career measures – before and after t^* – by period**. Each panel shows the probability distribution of a given quantity by mobility group and observation period. Comparison between groups $G1$ (no prior mobility, red) and $G3$ (mobility in period T, blue) provides an unconditional estimate of the impact of mobility on researcher trajectories in a given T. All variables measure the change in a given variable *after minus before* t^*_T. **(A-C)** Change in the citation impact: on average, researchers in the mobile group have slightly more positive change in citation impact. **(D-F)** Change in the collaborator network. **(G-I)** Change in the PACS research topics. **(J-L)** Change in the geographic network. For (D-L), one average, the mobile researchers have less similarity between their coauthors/topics/geography after migrating as compared to before migrating, than researchers from the control group.

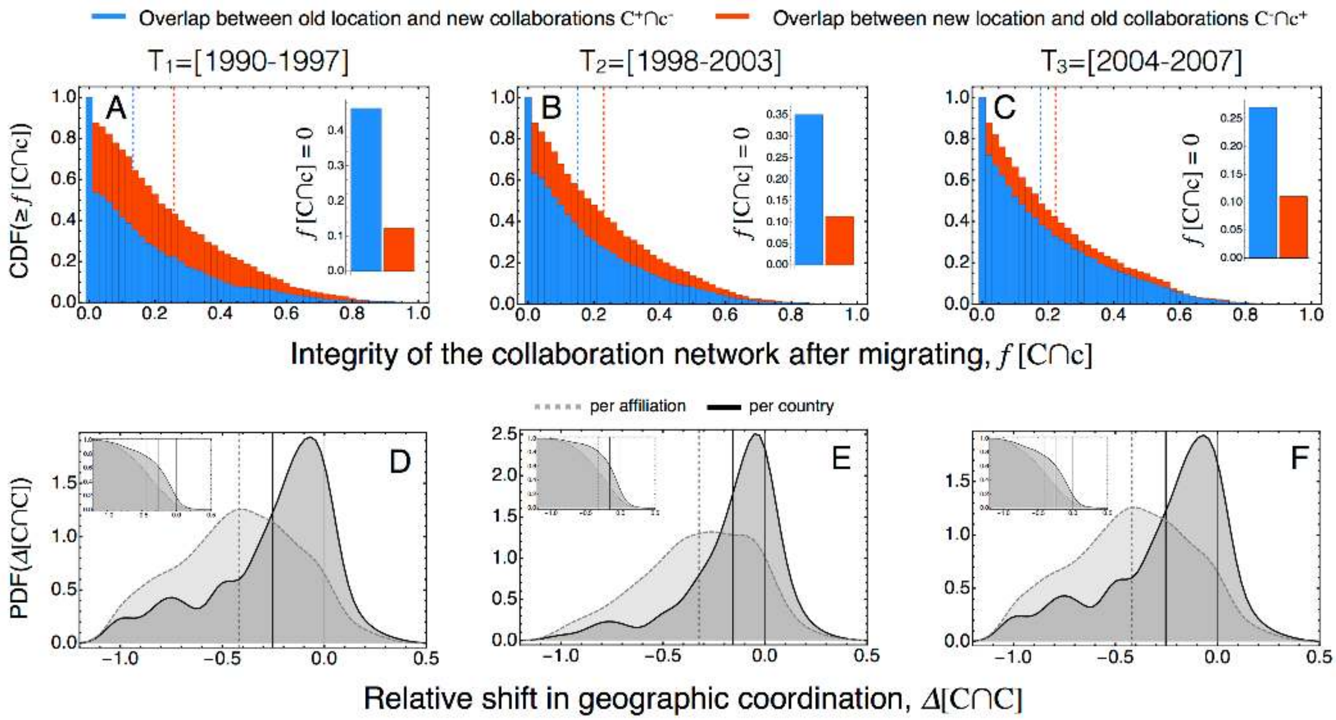

FIG. S6: **The impact of mobility on the geographic projection of collaboration networks – by period**. These results are calculated using data specific to each indicated period, thereby demonstrating the robustness of the distributions over time; compare with Fig. 2. **(A,B,C)** The degree of collobaration-mediated mobility measures the similarity between source and destination country of each i and the geographic distribution of his/her collaborators, before and after t^* – small values indicate the relatively low levels of similarity. (inset) Comparison of the "blind mobility" and "curtail mobility" rates. **(D,E,F)** Probability distribution of $\Delta[C \cap C]$ which measures the change in the geographic association between the collaborators before and after with respect to the source country of mobility, c_i^-. Negative values indicate that there is less overlap between c_i^- and the collaborators after the mobility event. For robustness, we calculate the geographic overlap in two ways: using distinct country lists (per country) and allowing for multiplicity due to multiple affiliations per publication (per affiliation). (inset) Cumulative probability distribution indicating that the majority of $\Delta[C \cap C]$ values are negative. Vertical lines indicate mean values.

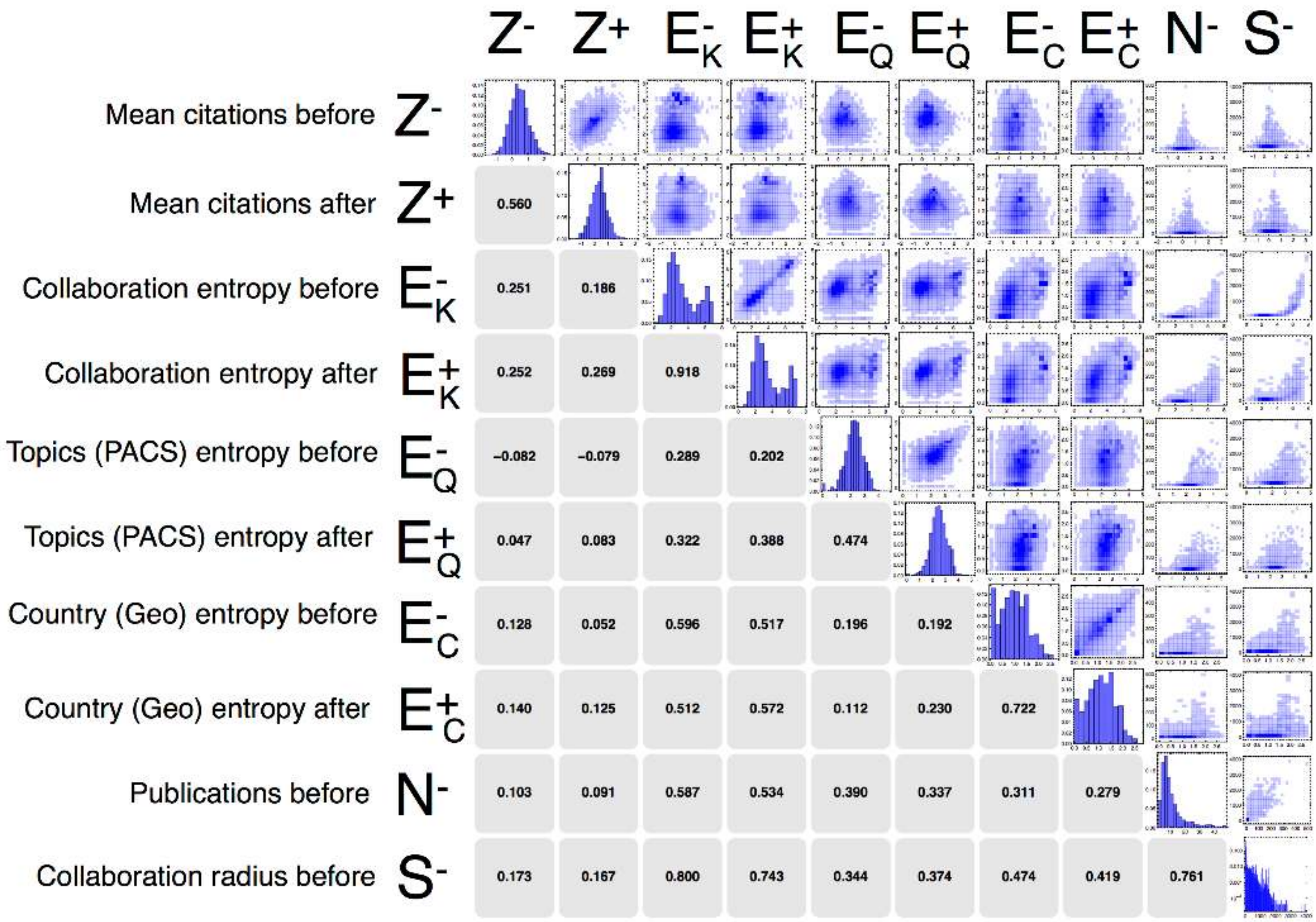

FIG. S7: **Model variables – distribution and covariation.** Shown is the correlation matrix calculated using the variables included in the PSM model; data are combined across the three periods (T). The diagonal elements show the distribution of the variable quantities; the upper-diagonal elements show the density-weighted scatter plots of any given pair of data observations; the lower-diagonal elements list the Pearson correlation coefficient between the corresponding variable pairs.

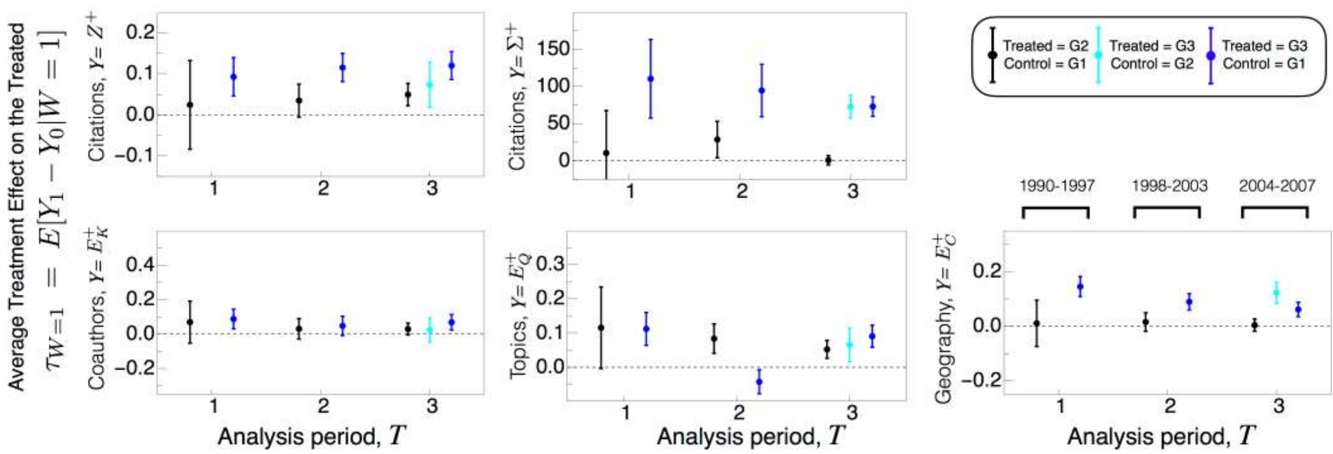

FIG. S8: **Estimation of the mobility effect using the nearest-neighbor *nnmatch* matching method**. Robustness check for the propensity score matching results reported in Fig. 3. The *teffects nnmatch* routine differs from the *teffects psmatch* in that the former calculates a single distance between multi-variate observations using a the Mahalanobis metric, and then matches to the nn closest observations (we used nn=1) [52]. One particular advantage of the *teffects nnmatch* method is that it allows the option to force a match on specified variables (using the ematch option); hence, we forced matches on the geographic region factor variable F_i^- representing one of the 5 geographic (continental) regions that the researcher primarily resided in prior to $t^*_{i,T}$ (see "Country classification" in Section S3). In this capacity, the *teffects nnmatch* estimate appropriately matches mobile individuals to un-mobile individuals from the same geographic region, thereby controlling for variation in regional migration opportunity. Despite this key difference, each set of estimates are robust with respect to the *teffects psmatch* estimates with the exception of the coauthor analysis (bottom row, left panel). Each error bar is a point estimate with 95% confidence interval.

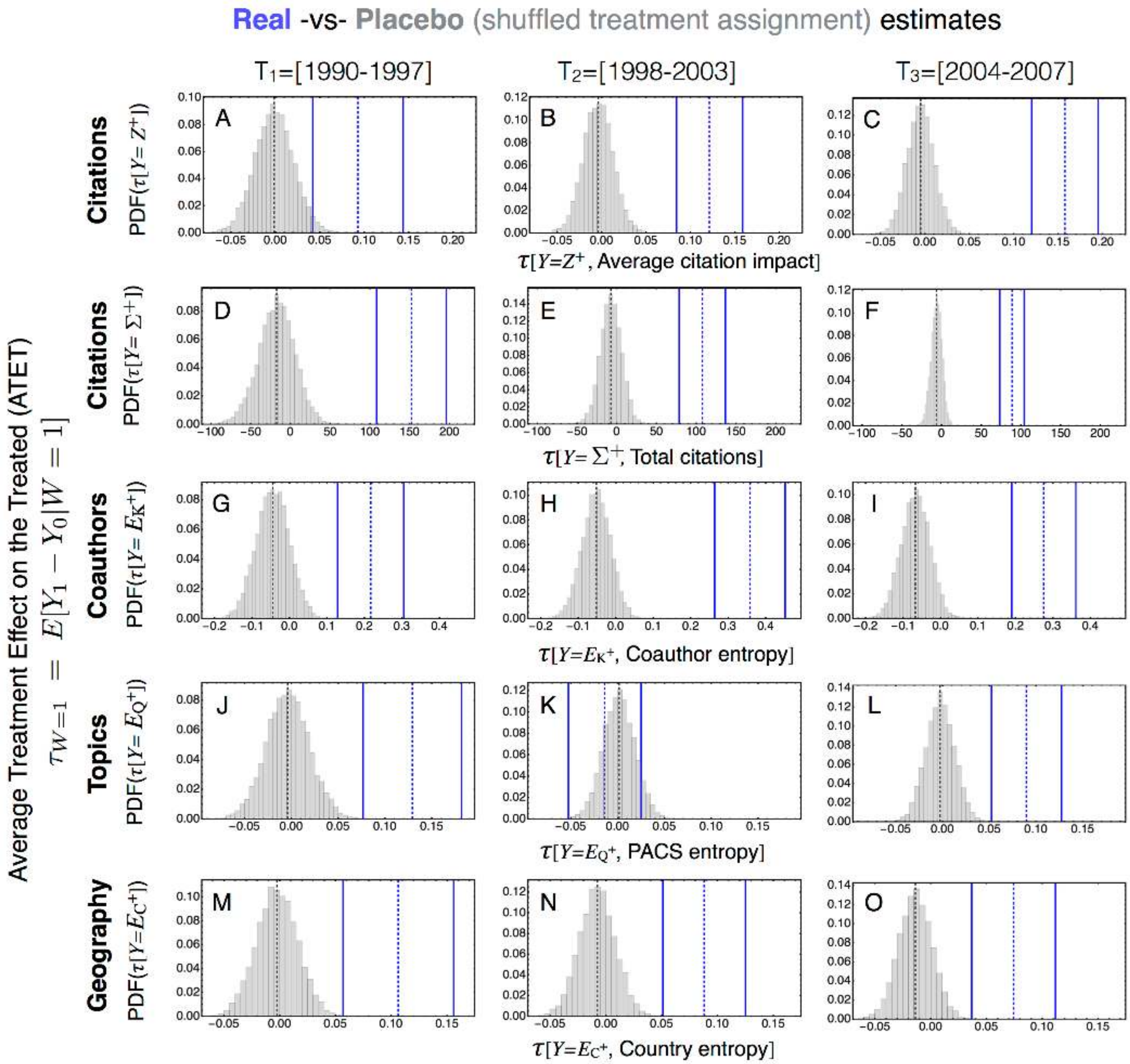

FIG. S9: **Testing the statistical significance of** $\tau_{W=1}$. It is possibility that spurious correlations could give rise to the statistically significant PSM estimations for $\tau_{W=1}[Y]$ reported in Figs. 3 and S8. We explored this possibility for the PSM models comparing $G1$ (control) and $G3$ (mobility) groups by randomizing the group assignments, implemented by shuffling G_i without replacement so that the total number of researchers in each group is conserved relative to the unshuffled (real) data. Thus, for each dependent variable (Y), we produced $N = 10,000$ shuffled datasets ('placebo model'), calculating $\tau_{W=1}[Y]$ for each. Shown for each specification is the probability distribution $P(\tau_{W=1})$ of the placebo estimates for $\tau_{W=1}[Y]$; the solid vertical blue line indicates the real $\tau_{W=1}[Y]$, and the dashed lines indicate the corresponding 95% confidence interval. In all cases except for in panel K, in which $\tau_{W=1}[E_{Q,T2}]$ is not statistically significant in the first place, we can rule out the possibility that $\tau_{W=1}[Y]$ estimations are statistically significant due to chance.

TABLE S1: Logit model. The dependent variable of the model is the binary outcome variable $1_{G_i=3}$ with value 1 if researcher i migrated during T and value 0 if there was no migration during or before T. Reported are odds ratios, $\exp(\beta)$.

	$T_1 = [1990 - 1997]$		$T_2 = [1998 - 2003]$		$T_3 = [2004 - 2007]$	
Researcher variables						
coauthors, $\lvert k_{ij}^{-}\rvert$	0.996***	(0.000)	0.999***	(0.000)	0.999***	(0.000)
publications, N_i^-	1.023***	(0.000)	1.005*	(0.037)	1.001	(0.247)
citation impact Z_i^-	1.132*	(0.043)	1.155***	(0.001)	0.969	(0.449)
researcher age, s_i^*	0.781***	(0.000)	0.856***	(0.000)	0.899***	(0.000)
Researcher geographic region, F_i^-						
N. America	1	(.)	1	(.)	1	(.)
S. & C. America	0.871	(0.591)	0.400***	(0.000)	0.383***	(0.000)
Europe	0.455***	(0.000)	0.492***	(0.000)	0.452***	(0.000)
Asia & Australasia	0.373***	(0.000)	0.350***	(0.000)	0.361***	(0.000)
N	4117		9347		13446	

p-values in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE S2: **Results of OLS regression using matched researcher pairs** (i, i'). The dependent variable is $Y_i^+ \equiv Z_i^+$, the average citation impact after $t_{i,T}^*$.

	(1)		(2)		(3)		(4)		(5)		(6)	
	T_1 = [1990–1997]		T_1 = [1990–1997]		T_2 = [1990–1997]		T_2 = [1998–2003]		T_3 = [2004–2007]		T_3 = [2004–2007]	
			(w/ $1_{G_i=3}$ interaction)				(w/ $1_{G_i=3}$ interaction)				(w/ $1_{G_i=3}$ interaction)	
coauthors, $\lvert k_{ij}^{-}\rvert$	0.000***	(0.000)	0.000	(0.099)	0.000***	(0.000)	0.000***	(0.000)	0.000***	(0.000)	0.000***	(0.000)
# interaction difference, $\delta_3(\lvert k_{ij}^{-}\rvert)$			0.000	(0.052)			-0.000*	(0.047)			-0.000	(0.055)
publications, N_i^-	-0.000	(0.885)	0.004*	(0.022)	-0.003***	(0.000)	-0.004***	(0.000)	-0.001	(0.137)	-0.000	(0.515)
# interaction difference, $\delta_3(N_i^-)$			-0.009***	(0.000)			0.002	(0.105)			-0.000	(0.848)
citation impact, Z_i^-	0.486***	(0.000)	0.585***	(0.000)	0.460***	(0.000)	0.540***	(0.000)	0.465***	(0.000)	0.495***	(0.000)
# interaction difference, $\delta_3(Z_i^-)$			-0.195***	(0.000)			-0.160***	(0.000)			-0.069**	(0.004)
researcher age, s_i^*	-0.009	(0.062)	-0.005	(0.393)	-0.006**	(0.003)	0.003	(0.370)	-0.011***	(0.000)	-0.007***	(0.001)
# interaction difference, $\delta_3(s_i^*)$			-0.008	(0.379)			-0.017***	(0.000)			-0.007*	(0.032)
Mobile researcher indicator ($1_{G_i=3}$)	0.088***	(0.000)	0.286***	(0.000)	0.124***	(0.000)	0.325***	(0.000)	0.114***	(0.000)	0.231***	(0.000)
Constant	0.098***	(0.001)	0.004	(0.919)	0.162***	(0.000)	0.056*	(0.039)	0.170***	(0.000)	0.115***	(0.000)
Researcher geo. region fixed effect, F_i^-	Y		Y		Y		Y		Y		Y	
N	3342		3342		5048		5048		4600		4600	
adj. R^2	0.274		0.289		0.267		0.275		0.306		0.309	
F	158.903		114.212		230.241		160.577		254.016		172.004	

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE S3: **Results of OLS regression using matched researcher pairs** (i, i'). The dependent variable is $Y_i^+ \equiv \Sigma_i^+$, the total deflated citations after $t_{i,T}^*$.

	(1)		(2)		(3)		(4)		(5)		(6)	
	T_1 = [1990–1997]		T_1 = [1990–1997]		T_2 = [1990–1997]		T_2 = [1998–2003]		T_3 = [2004–2007]		T_3 = [2004–2007]	
			(w/ $1_{G_i=3}$ interaction)				(w/ $1_{G_i=3}$ interaction)				(w/ $1_{G_i=3}$ interaction)	
total deflated citations, Σ_i^-	0.353***	(0.000)	0.446***	(0.000)	0.099***	(0.001)	0.311***	(0.000)	0.175***	(0.000)	-0.007	(0.808)
# interaction difference, $\delta_3(\Sigma_i^-)$			-0.195*	(0.032)			-0.395***	(0.000)			0.307***	(0.000)
coauthors, $\lvert k_{ij}^-\rvert$	3.849***	(0.000)	3.995***	(0.000)	1.500***	(0.000)	1.185***	(0.000)	0.389***	(0.000)	0.264***	(0.000)
# interaction difference, $\delta_3(\lvert k_{ij}^-\rvert)$			-0.219	(0.259)			0.433***	(0.000)			0.192***	(0.000)
publications, N_i^-	-8.839***	(0.000)	-13.611***	(0.000)	-1.844*	(0.011)	-5.654***	(0.000)	0.413	(0.185)	3.794***	(0.000)
# interaction difference, $\delta_3(N_i^-)$			8.714*	(0.025)			7.853***	(0.000)			-5.508***	(0.000)
citation impact, Z_i^-	23.215	(0.181)	22.389	(0.362)	34.774**	(0.004)	8.535	(0.621)	-16.293*	(0.017)	25.608**	(0.009)
# interaction difference, $\delta_3(Z_i^-)$			7.748	(0.822)			49.320*	(0.036)			-68.474***	(0.000)
researcher age, s_i^*	-29.642***	(0.000)	-22.662***	(0.000)	0.679	(0.691)	1.005	(0.672)	-4.492***	(0.000)	-3.841***	(0.000)
# interaction difference, $\delta_3(s_i^*)$			-14.780	(0.082)			-2.855	(0.402)			-1.160	(0.402)
Mobile researcher indicator $(1_{G_i=3})$	107.441***	(0.000)	165.309***	(0.001)	68.059***	(0.000)	30.758	(0.298)	60.188***	(0.000)	84.689***	(0.000)
Constant	253.634***	(0.000)	229.265***	(0.000)	115.496***	(0.000)	147.993***	(0.000)	32.515***	(0.001)	15.183	(0.197)
Researcher geo. region fixed effect, F_i^-	Y		Y		Y		Y		Y		Y	
N	3342		3342		5048		5048		4600		4600	
adj. R^2	0.527		0.528		0.358		0.369		0.489		0.502	
F	415.097		267.964		313.401		212.108		489.345		331.721	

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE S4: **Results of OLS regression using matched researcher pairs** (i, i'). The dependent variable is $Y_i^+ \equiv E_{K,i}^+$, the coauthor entropy after $t_{i,T}^*$.

	(1)		(2)		(3)		(4)		(5)		(6)	
	T_1 = [1990–1997]		T_1 = [1990–1997]		T_2 = [1990–1997]		T_2 = [1998–2003]		T_3 = [2004–2007]		T_3 = [2004–2007]	
			(w/ $1_{G_i=3}$ interaction)				(w/ $1_{G_i=3}$ interaction)				(w/ $1_{G_i=3}$ interaction)	
coauthor entropy, $E_{K,i}^-$	0.848***	(0.000)	0.804***	(0.000)	0.865***	(0.000)	0.766***	(0.000)	0.844***	(0.000)	0.840***	(0.000)
# interaction difference, $\delta_3(E_{K,i}^-)$			0.087**	(0.003)			0.155***	(0.000)			0.009	(0.652)
coauthors, $\lvert k_{ij}^-\rvert$	0.001***	(0.000)	0.002***	(0.000)	0.001***	(0.000)	0.002***	(0.000)	0.000***	(0.000)	0.000***	(0.000)
# interaction difference, $\delta_3(\lvert k_{ij}^-\rvert)$			-0.001**	(0.008)			-0.002***	(0.000)			-0.000	(0.229)
publications, N_i^-	-0.008***	(0.000)	-0.009***	(0.001)	-0.003***	(0.001)	-0.008***	(0.000)	-0.001	(0.182)	-0.001	(0.090)
# interaction difference, $\delta_3(N_i^-)$			0.003	(0.378)			0.006**	(0.006)			0.001	(0.313)
citation impact, Z_i^-	-0.049*	(0.015)	-0.019	(0.509)	-0.114***	(0.000)	-0.110***	(0.000)	0.047**	(0.005)	0.023	(0.342)
# interaction difference, $\delta_3(Z_i^-)$			-0.056	(0.162)			0.008	(0.815)			0.048	(0.152)
researcher age, s_i^*	-0.036***	(0.000)	-0.030***	(0.000)	-0.027***	(0.000)	-0.027***	(0.000)	-0.022***	(0.000)	-0.022***	(0.000)
# interaction difference, $\delta_3(s_i^*)$			-0.016	(0.198)			0.006	(0.344)			-0.001	(0.736)
Mobile researcher indicator $(1_{G_i=3})$	0.084***	(0.000)	0.010	(0.907)	0.036	(0.102)	-0.355***	(0.000)	0.075***	(0.000)	0.040	(0.517)
Constant	0.966***	(0.000)	1.010***	(0.000)	1.039***	(0.000)	1.256***	(0.000)	0.807***	(0.000)	0.825***	(0.000)
Researcher geo. region fixed effect, F_i^-	Y		Y		Y		Y		Y		Y	
N	3342		3342		5048		5048		4600		4600	
adj. R^2	0.776		0.776		0.767		0.769		0.852		0.852	
F	1285.405		828.658		1847.981		1202.122		2946.382		1894.009	

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE S5: **Results of OLS regression using matched researcher pairs** (i, i'). The dependent variable is $Y_i^+ \equiv E_{Q,i}^+$, the PACS (research topic) entropy after $t_{i,T}^*$.

	(1)		(2)		(3)		(4)		(5)		(6)	
	T_1 = [1990–1997]		T_1 = [1990–1997]		T_2 = [1990–1997]		T_2 = [1998–2003]		T_3 = [2004–2007]		T_3 = [2004–2007]	
			(w/ $1_{G_i=3}$ interaction)				(w/ $1_{G_i=3}$ interaction)				(w/ $1_{G_i=3}$ interaction)	
PACS (research topic) entropy, $E_{Q,i}^-$	0.336***	(0.000)	0.397***	(0.000)	0.243***	(0.000)	0.312***	(0.000)	0.458***	(0.000)	0.481***	(0.000)
# interaction difference, $\delta_3(E_{Q,i}^-)$			-0.114**	(0.004)			-0.131***	(0.000)			-0.048	(0.106)
coauthors, $\lvert k_{ij}^-\rvert$	0.001***	(0.000)	0.002***	(0.000)	0.001***	(0.000)	0.001***	(0.000)	0.000***	(0.000)	0.000***	(0.000)
# interaction difference, $\delta_3(\lvert k_{ij}^-\rvert)$			-0.001**	(0.004)			-0.000***	(0.001)			0.000	(0.116)
publications, N_i^-	0.002	(0.136)	0.000	(0.970)	-0.001	(0.089)	-0.005***	(0.000)	0.000	(0.906)	0.000	(0.989)
# interaction difference, $\delta_3(N_i^-)$			0.004	(0.164)			0.006***	(0.000)			0.000	(0.899)
citation impact, Z_i^-	-0.038*	(0.021)	-0.059*	(0.011)	0.096***	(0.000)	0.170***	(0.000)	0.010	(0.400)	0.006	(0.740)
# interaction difference, $\delta_3(Z_i^-)$			0.047	(0.145)			-0.139***	(0.000)			0.011	(0.643)
researcher age, s_i^*	-0.039***	(0.000)	-0.025***	(0.000)	-0.002	(0.487)	-0.003	(0.384)	-0.006***	(0.000)	-0.009***	(0.000)
# interaction difference, $\delta_3(s_i^*)$			-0.032**	(0.002)			0.005	(0.235)			0.006	(0.069)
Mobile researcher indicator ($1_{G_i=3}$)	0.123***	(0.000)	0.501***	(0.000)	-0.024	(0.111)	0.246***	(0.000)	0.083***	(0.000)	0.121	(0.078)
Constant	1.629***	(0.000)	1.435***	(0.000)	1.847***	(0.000)	1.689***	(0.000)	1.363***	(0.000)	1.343***	(0.000)
Researcher geo. region fixed effect, F_i^-	Y		Y		Y		Y		Y		Y	
N	3342		3342		5048		5048		4600		4600	
adj. R^2	0.204		0.212		0.148		0.157		0.330		0.331	
F	96.397		65.195		98.686		67.912		252.875		163.279	

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE S6: **Results of OLS regression using matched researcher pairs** (i, i'). The dependent variable is $Y_i^+ \equiv E_{C,i}^+$, the country entropy after $t_{i,T}^*$.

	(1)		(2)		(3)		(4)		(5)		(6)	
	T_1 = [1990–1997]		T_1 = [1990–1997]		T_2 = [1990–1997]		T_2 = [1998–2003]		T_3 = [2004–2007]		T_3 = [2004–2007]	
			(w/ $1_{G_i=3}$ interaction)				(w/ $1_{G_i=3}$ interaction)				(w/ $1_{G_i=3}$ interaction)	
country entropy, $E_{C,i}^-$	0.466***	(0.000)	0.553***	(0.000)	0.536***	(0.000)	0.632***	(0.000)	0.597***	(0.000)	0.721***	(0.000)
# interaction difference, $\delta_3(E_{C,i}^-)$			-0.182***	(0.000)			-0.227***	(0.000)			-0.265***	(0.000)
coauthors, $\lvert k_{ij}^-\rvert$	0.001***	(0.000)	0.001***	(0.000)	0.001***	(0.000)	0.001***	(0.000)	0.000***	(0.000)	0.000*	(0.017)
# interaction difference, $\delta_3(\lvert k_{ij}^-\rvert)$			0.000	(0.107)			0.000	(0.969)			0.000***	(0.000)
publications, N_i^-	-0.003**	(0.005)	-0.001	(0.508)	0.001	(0.249)	0.000	(0.604)	0.000	(0.241)	0.000	(0.381)
# interaction difference, $\delta_3(N_i^-)$			-0.004	(0.080)			-0.000	(0.897)			-0.000	(0.673)
citation impact, Z_i^-	0.013	(0.293)	0.042*	(0.023)	-0.015	(0.147)	-0.008	(0.573)	0.086***	(0.000)	0.129***	(0.000)
# interaction difference, $\delta_3(Z_i^-)$			-0.059*	(0.019)			-0.011	(0.566)			-0.092***	(0.000)
researcher age, s_i^*	-0.005	(0.164)	-0.007	(0.168)	-0.006***	(0.001)	-0.008**	(0.002)	-0.008***	(0.000)	-0.009***	(0.000)
# interaction difference, $\delta_3(s_i^*)$			0.005	(0.544)			0.007	(0.059)			0.004	(0.167)
Mobile researcher indicator ($1_{G_i=3}$)	0.114***	(0.000)	0.237***	(0.000)	0.084***	(0.000)	0.244***	(0.000)	0.035**	(0.002)	0.293***	(0.000)
Constant	0.447***	(0.000)	0.382***	(0.000)	0.460***	(0.000)	0.372***	(0.000)	0.352***	(0.000)	0.229***	(0.000)
Researcher geo. region fixed effect, F_i^-	Y		Y		Y		Y		Y		Y	
N	3342		3342		5048		5048		4600		4600	
adj. R^2	0.328		0.335		0.426		0.438		0.536		0.552	
F	182.449		121.347		417.520		282.008		590.838		406.136	

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$