Reproducing scientists' mobility:
A data-driven model

Giacomo Vaccario[a], Luca Verginer[a,b], and Frank Schweitzer[a]

[a] *ETH Zürich, Chair of Systems Design, CH-8092 Zürich, Switzerland*
[b] *IMT School for Advanced Studies Lucca, AXES Lab, IT-55100 Lucca, Italy*

September 23, 2020

Abstract

High skill labour is an important factor underpinning the competitive advantage of modern economies [5, 6, 3]. Therefore, attracting and retaining scientists has become a major concern for migration policy [9]. In this work, we study the migration of scientists on a global scale, by combining two large data sets covering the publications of 3.5 Mio scientists over 60 years. We analyse their geographical distances moved for a new affiliation and their age when moving, this way reconstructing their geographical "career paths". These paths are used to derive the world network of scientists mobility between cities and to analyse its topological properties. We further develop and calibrate an agent-based model, such that it reproduces the empirical findings both at the level of scientists and of the global network. Our model takes into account that the academic hiring process is largely demand-driven and demonstrates that the probability of scientists to relocate decreases both with age and with distance. Our results allow interpreting the model assumptions as micro-based decision rules that can explain the observed mobility patterns of scientists.

keywords: high-skilled labor | mobility network | data-driven model

Introduction

Scientists are highly mobile individuals. This has been true in the past and is even more true today [19]. Therefore, an increasing number of works analyses the mobility of scientists and their motivation to relocate [13, 26]. Many publications have focused on the relationship between movements and scientific impact [18, 15, 17, 34, 30]. Other works have analysed scientists mobility within and across *countries*, to determine policy impacts [27, 12] or to study the brain circulation phenomenon [7, 33, 1, 43].

While most of these studies focus on the aggregated level, e.g., on bilateral flows between countries, there is a need to better understand scientific mobility at the individual level [2, 16]. Empirical works in this direction [20, 45] are often

based on survey data that provide only partial coverage of the global mobility of scientists. Theoretical works on scientist mobility [25], on the other hand, are rarely validated against data. Researches in complexity and network theory have mostly analyzed scientific collaborations [29, 46, 21] or hiring practices [11].

Our work addresses this research gap in a two-fold manner. First, we provide empirical insights into scientific mobility at the individual level, by reconstructing 3.5 million geographical career paths from large-scale data sets. Second, we provide an agent-based model that is calibrated against the available data and is capable of reproducing the distributions of relocation *distances* and relocation *age*. In developing our *data-driven* model, we follow the approach of [36, 35, 41]. Our model incorporates two factors that are known to affect academic mobility: (i) geographical distance [28, 26], (ii) prestige, and selectiveness of academic institutions [11, 40, 44]. We also contribute to the understanding of global mobility, by reconstructing and analysing the world network of scientists mobility at the level of *cities*, not countries. From this global network, we extract topological features such as the distributions of degrees, local clustering coefficients, path lengths, and assortativity, to demonstrate that these can also be reproduced by our agent-based model.

Results

Empirical findings

By combining two large-scale bibliographic datasets as described in the *Methods* we obtain for $N = 3'740'187$ scientists information about the sequence of cities they worked in their careers, between 1950 and 2009. The merged dataset contains $M = 5'485$ unique cities. The data allows us to construct the geographical "career path" of these scientists. An illustrative example is given in Table S1.

Statistics of geographical career paths. From the career paths we compute the *relocation distances* scientists moved when changing their affiliation, using the Haversine formula for geodesic distances. The distribution obtained from 62 465 scientists relocating between 2000 and 2008 is shown in Figure 1(a). We note that it is a left-skew distribution with a median of 1 000 km, i.e., most scientists find a new affiliation within a radius of 1 000 km around their current affiliation. However, relocations toward cities that are more than 6 000 km away are also quite frequent.

The data also allows us to relate the frequency of such moves to the age of scientists. Because the physical age of scientists is unavailable, we rely on their *academic age*, t_i^a, also measured in years. $t_i^a = 0$ when the scientist publishes his/her first paper, according to our database. The frequency of any recorded moves over the academic age t^a is shown in Figure 1(b). Again, it is a left-skew distribution with a median of 7 years. This matches the known fact that the mobility of scientists drastically decreases with age [10, 42]. However, we also identify that some scientists change their working location after been active for 40 years.
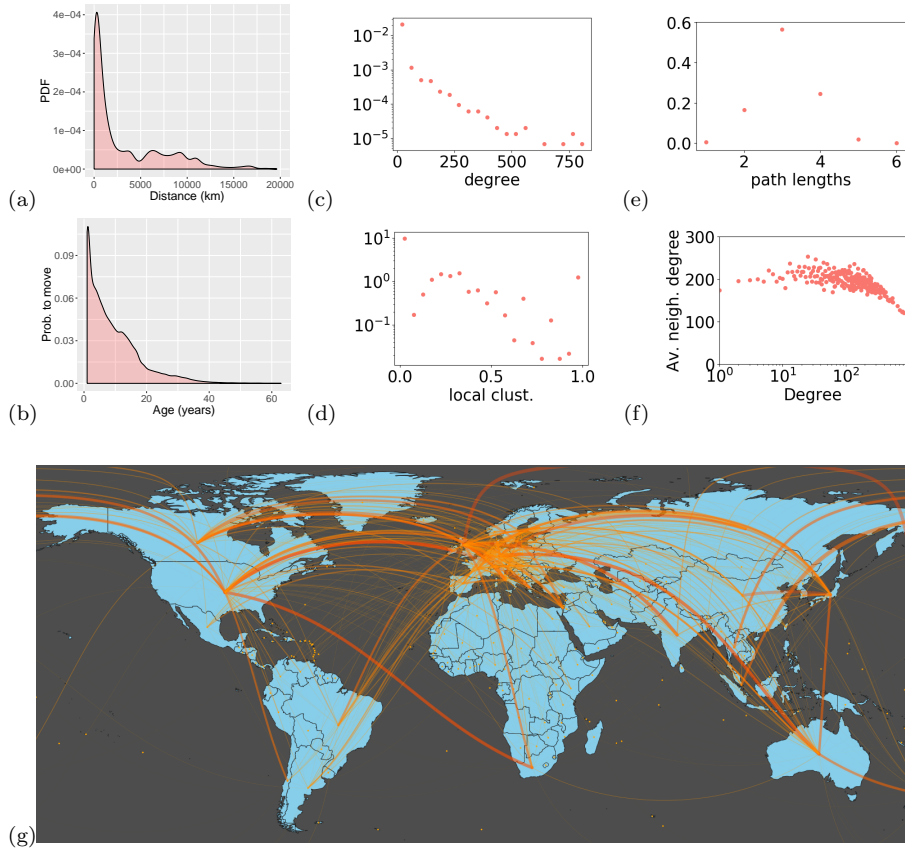
Figure 1: *Characterization of the empirical academic mobility.* At the individual level, we have the distribution of relocation distances of scientists (a) and the distribution of moves dependent on the (academic) age of scientists (b). At the global level, we reconstruct the mobility network for which, we plot the distribution of degrees (c), local clustering coefficients (d) and path lengths (e). In (d) we plot the average degree of neighbors of a node with degree k in function of k. Finally, the mobility network of scientists in between 1990 and 2008 in (g). The link width and the color indicate the magnitude of the *total flow* between any two countries. For visualization purpose, the total flows have been aggregated at country level and logarithmically scaled.

Reconstructing the mobility network of scientists. While the career paths and their statistics refer to individual scientists, we can also analyse the network that results from aggregating all of the career paths of a given year. This aggregation changes the unit of analysis to the *city level*. For each year, we obtain the number of scientists $N_K(t)$ in a given city K from their publications by taking unique geo-located scientists into account.

We further calculated for each year t the number of scientists $\Delta N_{K \leftarrow L}(t)$ moving into city K from another city L, i.e., the inflow, and the number of scientists $\Delta N_{L \leftarrow K}(t)$ moving out of city K to another city L, i.e. the outflow. For any given pair (K, L) of cities, we then calculate the *total flow* of scientists between these two cities, $\Delta N_{L \leftarrow K} + \Delta N_{K \leftarrow L}$. This flow allows us to visualise the mobility network of scientists at the world level, shown in Figure 1(g). The links are undirected but weighted according to the total flow.

Topological properties of the mobility network. To obtain the topological properties common in network analysis, we aggregate the mobility networks for the period 2000 – 2008. On this *aggregated network*, we calculate standard measures, such as the *degree distribution* $P(d)$, where d is the number of cities scientists in a given city either move to or come from. Figure 1(c) shows that this is a broad distribution. Some cities act as hubs, with a large degree, most cities, however, only have a small degree.

The distribution of *path lengths*, shown in Figure 1(e), measures how many steps are needed to reach, on the network, any city from a given starting point. The small number of steps indicates that the network is dense in a *topological sense*, not necessarily in a geographical one.

The *local clustering coefficient*, on the other hand, measures whether three neighbouring cities (with respect to their geographical proximity) form closed triangles, i.e., whether there is an exchange of scientists among all three of them. Figure 1(d) shows the distributions of these values, and we find that most cities have a small local clustering coefficient.

The *neighbor connectivity*, eventually, measures to what extent cities with a certain degree are connected to other cities with a similar degree. Figure 1(f) shows a non-monotonous dependency. Cities with a low degree tend to show an *assortative* pattern, i.e., they are connected to cities that have a similar number of neighbours. Cities with a high degree, which are characterised as *hubs* above, are rather connected to cities with a lower degree, i.e., they are *dissortative*. This result gives us already on the topological level important information about the origin of scientists coming to the hubs and the destination of scientists leaving the hubs. They do not hop between hubs, which would have been indicated by an assortative pattern.

Modeling the mobility of scientists

We now develop a model to reproduce the characteristic empirical properties of the scientists' mobility network discussed above. Precisely, we want to reproduce features both at *scientist* and *network* level. These are, at the scientist level, (i) the distribution of moved distances, Figure 1(a) and (ii) the "age at move" distributions, Figure 1(b). At the network level, we want to reproduce (iii) the distributions of the topological features shown in Figure 1(c-f), i.e., degrees, local clustering coefficients, path lengths, and average neighbour degree.

We develop an *agent-based model* because we want to model the relocation of *scientists*, as opposed to a system dynamics model that would merely reproduce the flows between *cities*. This choice implies that *macroscopic features* describing the system, such as the topological properties already discussed, must be *emergent properties* arising from the agent dynamics.

Our model is composed of two entities, `agents` and `locations`. `Agents` represent scientists. Each agent i is characterized by three properties that change over time: its position, $r_i(t)$, its fitness, $f_i(t)$, and its years of activity $y_i(t)$. Time is measured in discrete simulation steps, each step representing one year. When we start our simulations at time $t = 0$, which is chosen as the year 2000 below, we have to consider that many agents have already published before 2000, which is included in $y_i(t)$. For instance, an agent that published its first paper in 1995 will have a $y_i(2000) = 5$. This information is essential to measure an agent's fitness, $f_i(t = 0)$, which we do below.

`Locations` represent cities and host agents. Each location K is characterised by three properties that can also partly change over time: its position R_K defined in real geographical space by means of longitude and latitude, its fitness, $F_K(t)$, and the number of agents it hosts, $N_K(t)$. Note that R_K and $N_k(t)$ are taken from the available empirical data. For the position $r_i(t)$ of an agent, we assume that at each time step the agent can be found in one of the available locations. So $r_i(t) = R_K$ where K is the location, where agent i is located at time t.

Agent and location fitness. The individual agent fitness $f_i(t)$ represents the academic impact of a scientist. We proxy this impact by the papers that he/she has co-authored. Precisely, we assign to each paper a score equal to the impact factor of the journal (taken from SCImago) where it was published divided by the number of co-authors. Then, for each scientist, we aggregate the scores of his/her co-authored papers in the last two years of activity. By this, we obtain a distribution of fitness values that we can assign to agents.

We assign to each location K a fitness value $F_K(t)$ reflecting the quality of the academic institutions hosted in a city. To make this idea explicit and measurable, we take $F_K(t)$ to be the mean fitness of the agents located in K. Note that this approach is in line with how rankings of academic institutions are created. Indeed, university rankings are determined considering the academic impact and quality of the scientists working there. In our model, we assume that the $F_K(t)$ is public information, and thus, agents may use this information in their decision rule.

Relocation preferences. Our central modelling assumption is that agents prefer to work in locations that provide a higher fitness than the one they are currently based. These locations, however, can be distant from the current location, which implies higher relocation costs. Therefore, an agent i takes into account the fitness $F_K(t)$ of locations and its geodesic distance $\Delta_{i,K}(t)$. Agents combine this information in a *re-scaled fitness score* $\tilde{F}_{i,K}(t) = F_K(t)/(\Delta_{i,K}(t))^b$ for each location K. b is a model parameter, used to weight the impact of spatial distances. The bigger b, the more important any spatial distance becomes.

Ranking the values $\tilde{F}_{i,K}(t)$ from high to low, each agent then obtains an *individual* ranking that reflects its preferences where to move to. Agents in L

will consider only those locations where $F_K(t) > F_L(t)$, i.e., where the average fitness of scientists is larger than the average fitness of scientists in their city. Hence, each agent i assigns to a location K the score:

$$R(i, K) = \Theta\left[F_K(t) - F_L(t)\right] \frac{F_K(t)}{(\Delta_{i,K}(t))^b}$$

where $\Theta\left[F_K(t) - F_L(t)\right]$ is equal to 1 when $F_K(t) > F_L(t)$ and equal to 0, otherwise.

Relocation decisions. Agents only come up with a ranked list of possible locations they would consider to move to (and we assume that they send applications to the academic institutions in these locations). However, agents do not decide where to move. This decision, whether or nor to accept the agent, is taken by the location.

A location K will accept new agents only if it has sufficient capacity. The capacity $N_K(t)$ for a given city K, is estimated from the number of scientists empirically observed in city K in year t. Dependent on the individual ranking of agents, some locations obtain more applications than the capacity allows them to accept. So each location ranks the qualified agents according to their fitness $f_i(t)$. Available slots are filled starting from agents with higher fitness values until the capacity $N_K(t)$ is reached. Precisely, if $f_i(t) > F_K(t)$, location K considers agent i with probability $p = 1$ because this allows location K to increase its fitness $F_K(t)$. If $f_i(t) \leq F_K(t)$, location K considers agent i only with a probability $p = (f_i(t)/F_K)^s$ where s is our second model parameter. This parameter s represents the *selectiveness* of locations, the higher s, the more difficult it is to be hired. Hence, if a location K has some openings, its probability to accept agent i is:

$$p(K, i) = \begin{cases} 1 & f_i(t) > F_K(t) \\ \left(\frac{f_i(t)}{F_K(t)}\right)^s & \text{otherwise} \end{cases}$$

In Figure 2, we summarise and visualise the basic rules of our model.

Matching agents to locations. In our model agents rank locations, while locations rank agents. To match locations and agents, we have to solve a matching problem similar to the stable marriage problem. However, our problem is slightly different as a location can accept more than one agent until the capacity $N_K(t)$ is reached. To solve this matching problem, we use an applicant proposing algorithm, similar to the NRMP-algorithm [32]. The details are given in the Methods section.

Fitness dynamics. To model those agents not accepted at a new location, we consider that agent which stay at their current location, i.e., $r_i(t+1) = r_i(t)$, use the time step to further improve their fitness, $f_i(t)$. For this we assume a stochastic dynamics, precisely an additive stochastic process with a variance proportional to the fitness of the current location:
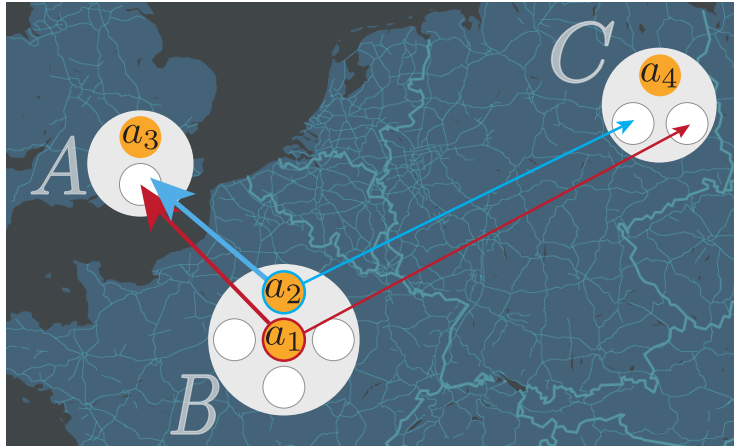
$$f_i(t + 1) = f_i(t) + \alpha_{\text{loc}}\eta$$

Figure 2: *Example of relocation rules.* Four agents (a_1, a_2, a_3 and a_4) are hosted in three locations, A, B or C, representing London, Paris and Berlin. Each location has a maximum number of available positions illustrated by small circles: $N_A = 2$, $N_B = 4$ and $N_C = 3$. In this image, agents a_1 and a_2 compute the rescaled fitness of the available locations (A and C) and rank these location accordingly. Here, we have assumed that A and C have the same fitness ($F_A(t) = F_C(t)$), but A is closer to B than C ($\Delta_{i,A} < \Delta_{i,C}$ for $i = 1, 2$). For this reason, both a_1 and a_2 express a preference for A over C. Since location A has $N_A = 2$ and one position is already taken, A must decide to accept either a_1 or a_2, depending on their fitness.

where α_{loc} is a parameter proportional to the quality of the agent location, and η is a normally distributed stochastic variable with 0 mean and variance equal to 1 (i.e.,$\eta \approx \mathcal{N}(0, 1)$). By this, we assume that the change in fitness of an agent depends on its location. Also, it is not guaranteed that agents will improve their fitness; they can also lower it.

At the end of each time step, we update the fitness of locations, $F_K(t)$, by averaging over the fitness $f_i(t)$ of all those agents that are currently based there.

Entry and exit of agents. At every time step, agents can exit and enter the system. This dynamic simulates the fact that academia is an open system, i.e., every year, scientists exit the systems, but also new ones enter it. Precisely, agents are removed with a small probability of p_e at the end of every time step. Hence, the probability that an agents is removed from the simulation after n steps is $(1 - p_e)^{n-1} p_e$. This process allows us to replicate the observed (academic) survival probability function of scientists (see Figure S2 in SM).

Model calibration. We use the empirical data not only as an *input* to our model, to determine the *initial conditions*, but also for *calibration*. For this, we use only a subset of the available data (see the *Methods* section). A major effort was spent to determine the optimal values of the two free parameters of our model, b and s. $b^{\mathrm{opt}} = 0.5$ means that a location A which is twice as far away as B must have a fitness $F_A = F_B * \sqrt{2}$ to be as attractive as B. Moreover, $s^{\mathrm{opt}} = 0.5$ means that agents are accepted by locations with a probability larger
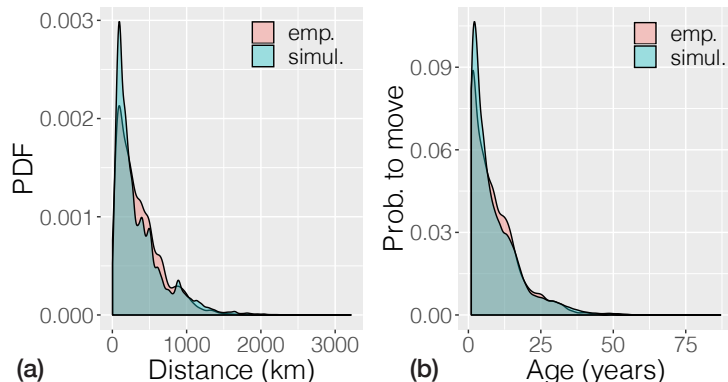
Figure 3: *Comparison of simulated and empirical individual scientist features.* (a) Distribution of relocation distances of scientists. (b) Distribution of moves dependent on the (academic) age of scientists. (red) indicates the empirical distribution, (blue) the distributions that are obtained from our agent-based simulations.

than their fitness ratio. For example, an agent with fitness ratio $f_i/F_k = 0.5$ is accepted with probability 0.71 and an agent with $f_i/F_k = 0.25$ is accepted with $p = 0.5$.

Model validation.   The calibrated agent-based model has to prove its evidence in that it can reproduce the whole set of empirical findings that have *not* been used during the calibration procedure. We *validate* the model by comparing two distributions on the level of scientists, and four distributions on the level of the mobility network. To simulate a large number of realisations, we focus on three neighbouring countries in Europe, namely Germany, France and the UK. Furthermore, we restrict the simulation to the period 2000 to 2006. The upper limit 2006 is given by the fact that the last publication in `Author-ity` is in 2009, and we require a 3-year window to identify moves.

Results of agent-based simulations.   The results of the validation are shown in Figure 3 and 4. To allow for a direct comparison, we plot the empirical data in red and the simulation in blue. We can report a good match of all distributions both on the level of scientists and on the network level. Specifically, on the scientists' level, we are able to reproduce the two distributions of *relocation distances* and of *age* when moving, see Figure 3(a,b).

On the network level, we are able to reproduce the four distributions of *clustering coefficients*, *path lengths*, *degree* and *average neighbor degree*, see Figures 4(a-d). We emphasise that these results are far from being trivial. As we start with an agent-based perspective, the results of our simulations refer to *career paths* of individual agents. From these, we have to reconstruct an aggregated network of mobility. Our simulation results for the network topology are reported for these simulated networks.

In conclusion, we report that our agent-based model captures the different features of the empirical data well, both on the scientists' and the network level, without using direct information from these for the calibration.
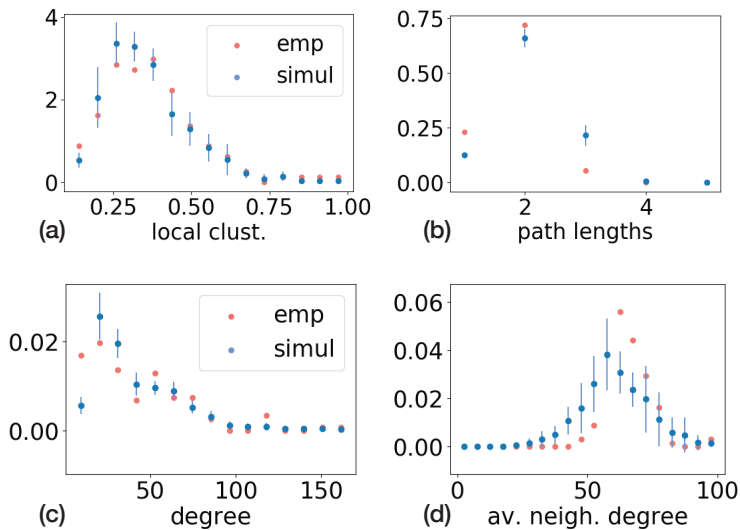
8

Figure 4: *Comparison of empirical and simulated topological properties of the mobility network.* Distributions of (a) local clustering coefficients, (b) path lengths, (c) degrees, and (d) degrees of neighbors. (red) indicates the empirical distribution, (blue) the distributions that are obtained from our agent-based simulations. The error bars correspond to the standard deviations of the measures computed on the 10 different realisations of the simulated mobility network.

Discussion

This paper provides several results with relevance for both the empirical and the theoretical understanding of the global mobility of scientists. As a novel contribution, we introduce the concept of a *geographical career path* of an individual scientist, which can be extracted from data. Using records of 3.5 Mio scientists, we provide a statistical analysis of such career paths, that later form the basis for comparison with our model, on the scientists' level. Aggregating over these career paths, we are further able to reconstruct the *world network of scientists' mobility*, with cities as nodes and inflow/outflow of scientists as links. With this, we reveal the patterns of scientists' mobility on two levels: the level of an individual scientist (age, relocation distance), and the level of cities forming a global network, which is a new empirical insight.

The most important contribution, however, is an agent-based model that allows reproducing these empirical findings on both the scientist and the city level. In our model, we assume as most relevant factors geographical distances, academic importance, and selectiveness of cities. The model uses as input for the initial conditions only variables that can be proxied by the available data. In particular, academic importance, denoted as the fitness of agents, is proxied from the available publications of scientists. The fitness of locations, another ingredient of the model, is then obtained by averaging over the fitness of agents at the particular location.

The agent-based model succeeds with simple assumptions for the relocation of agents. Agents rank all locations according to their fitness and their distance

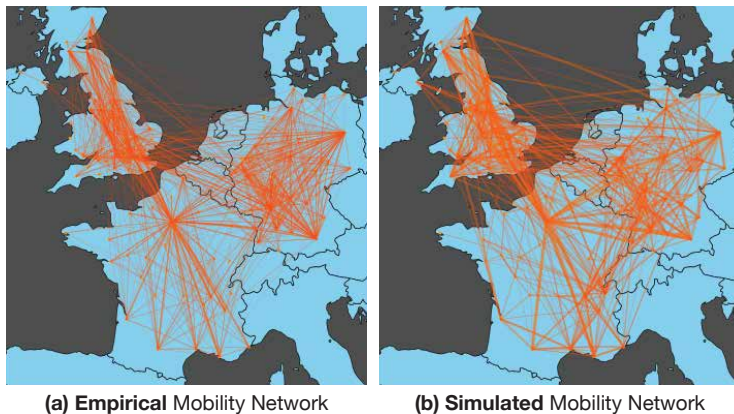**(a) Empirical** Mobility Network   **(b) Simulated** Mobility Network

Figure 5: *Empirical and Simulated mobility networks for France, Germany and UK.* The empirical network (a) depicts the flows between cities, the thickness of links indicates their magnitude. The map in (b) depicts one realization of the ABM with optimal parameters.

to the current location. However, they do not decide whether to move. This choice is made by the locations using the information on the fitness of the agents and capacity constraints. In essence, this poses a matching problem and can be related to similar problems discussed in the literature.

Our agent-based model only considers two free parameters, which need to be calibrated against the available data: b weights the spatial distance between the current location of an agent and any other location, s weights the selectiveness of locations when accepting agents that have a fitness below the location's fitness. We find as *optimal parameters* $(s^{\mathrm{opt}}, b^{\mathrm{opt}}) = (0.5, 0.5)$. These parameters are maximally different from 0 or 1 and indicate that both selectiveness and distances are essential to reproduce the empirical mobility patterns. $b^{\mathrm{opt}} = 0.5$ characterizes the *supply side*, i.e., the ranking of locations by the *agents*. A location A needs to have $\sqrt{n}$ times the fitness of another location B if it is n times further away, to be equally attractive for an agent. $s^{\mathrm{opt}} = 0.5$ characterizes the *demand side*, i.e., the ranking of agents by *locations*. Provided there are sufficient openings available, an agent with a fitness ratio $f_i/F_k = 0.5$ is accepted more than 2 out of 3 times, and an agent with $f_i/F_k = 0.25$ still has a 50% chance to be accepted.

Using the model calibrated with the optimal parameters, our simulations match the available empirical data quite well. This is remarkable because the model does *not* include many factors which arguably play a role in the relocation decision. In other words, only using minimal assumptions about the supply (scientists) and demand (cities), and a simple matching mechanism, we are able to capture emergent features of the scientist mobility network. Some minor differences between the simulated and the empirical distributions become only noticeable if we plot the network of scientists' mobility on the European scale, as shown in Figure 5. We observe that the empirical network in Figure 5(a) has more pronounced hubs than the simulated network shown in Figure 5(b). Specifically, in the empirical network, significantly more French cities are linked to Paris than in the simulation.

10

Finally, we stress that more factors are influencing the relocation choices of scientists than explicitly covered in our model. For example, quality of life, better networking opportunities or higher salaries might be relevant factors here. The more remarkable is the fact that our model, even at this level of detail, works considerably well.

In summary, we have provided the first agent-based model reproducing the mobility of scientists. In a data-driven approach, our model has been calibrated and validated against data, and we have found a remarkably good match between simulations and empirics. We show that minimal decision rules capture many complex features of the observed mobility of scientists. Besides, we have quantified the relative importance between geographical distances and academic attractiveness from the perspective of a scientist trying to relocate.

Materials and Methods

Extracting individual career paths of scientists. For our work, we use the MEDLINE database, the largest open-access bibliographic database in the life sciences. Our analysis is based on two datasets provided by Torvik and co-authors, namely `MapAffil` [37] and `Author-ity` [38], which have been extracted from MEDLINE. `MapAffil` lists for each MEDLINE paper and each scientist the *disambiguated city names* of the listed affiliation (37,396,671 city-name instances). It further gives a unique identifier as well as the geo-coordinates of each city. `Author-ity` contains the *disambiguated scientist names*, linking them to their respective publications. Combining these two sources of information about geo-coordinates and time allows us to construct the geographical "career path" of scientists, using the approach of [43].

Formally we denote a career path of a scientist $i \in N$ as a sequence p_i, for example, $p_i = \{A_{t_0}, A_{t_1}, A_{t_2}, B_{t_2}, B_{t_4}, C_{t_5}, C_{t_6}\}$. A denotes the city as defined by its geo-location $R_A = (X, Y)$ where X gives the latitude and Y the longitude according to the data from `MapAffil`. The subscript t_0 refers to the year scientist i was affiliated in the respective city, according to the career path data obtained. For more information about the data used, see the SM.

Determining locations in geographical space. Defining the boundaries of a city is a central problem in urban studies. A standard definition available for US cities is the "Metropolitan Statistical Area" (MSAs) [39]. However, as the name suggests, this definition is *not* available outside the US. Therefore, to identify cities, we rely on the definition of "location" provided by Google Maps. This definition reflects administrative boundaries, which are not perfect substitutes. As argued by [14, 4] natural and administrative definitions follow different size distributions. However, because we do not argue about the size distribution of cities, this is not a crucial concern.

Determining the free model parameters. Parameter b weights the impact of spatial distances on the individual rankings of agents. $b = 0$ would imply that distances do not play a role in relocation preferences; with $b = 1$, a location A which is twice as far away as location B needs to have twice the fitness of B ($F_A = 2 * F_B$) to be equally attractive. In general, a location A which is n times as far away as B must have a fitness $F_A = F_B * n^b$ to be as attractive as B.

11

Parameter s weights the flexibility of locations to still accept agents with a fitness less than the fitness of the location. For $s = 0$, the probability of a location to accept agents is independent of their fitness and always equal to 1 ($p(K, i) = 1$); with $s = 1$ the probabilty to be accepted is equal to the ratio between the agent's and location's fitness (f_i/F_k), e.g., an agent with $f_i = 1/2F_k$ will be accepted with probability $1/2$. In the case of $s > 1$, an agent is accepted by a location with a probability smaller than their fitness ratio ($p(K, i) < f_i/F_k$).

Calibration procedure. To *calibrate* the model parameters b, s, we use an established approach in agent-based modeling [41], machine learning [31, 24, 8] and computer simulations in general [23]. It combines two elements: (a) a grid search and (b) a performance score.

The grid search consists of an exploration of the (low dimensional) parameter space through computer simulations. For b the values $\{0.005, 0.01, 0.05, 0.1, 0.5, 1.0, 5.0\}$ are considered, for s the values $\{0.05, 0.1, 0.5, 1.0, 5.0, 1.0, 10.0\}$. For each parameter combination, we obtain from the simulations two distributions for the *inflow* and *outflow*. To determine the optimal combination of (b, s), we compare the simulated and the empirical inflow and outflow distributions. For this comparison, we use a performance score based on the Kolmogorov-Smirnov(KS) statistic [22]. Precisely, for each combination of parameters (b, s), we compute the KS-statistic between the empirical and simulated distributions of inflow, $D_1(b, s)$, and of outflow $D_2(b, s)$. We then define the performance score as $1/(D_1(b, s) \times D_2(b, s))$, such that the optimal combination $(b^{\mathrm{opt}}, s^{\mathrm{opt}})$ maximizes this score.

From the calibration procedure, we find *optimal parameters* $(s^{\mathrm{opt}}, b^{\mathrm{opt}}) = (0.5, 0.5)$. The comparison between the empirical and the simulated distributions is shown in Figure S6(a,b) in SM. The close match demonstrates that our model is correctly calibrated.

Simulation initialisation. At the beginning of the simulation, we populate the cities at only 80% capacity, which means that we initiated the simulation with 22,000 agents, and ca. 300 locations. As the starting year $t = 0$, we take 2000. From each city, we take its geographical position and the number of scientists in the year 2000. We assign these quantities to locations to characterize their R_K and $N_K(t = 0)$. The initial fitness value of a location, $F_K(t = 0)$, is determined by averaging over the fitness values of those agents based in the given city in 2000.

From each scientist, we obtain its geographical position (in a given city), his/her academic impact, and the years of activity as of the year 2000. We assign these quantities to agents to characterize their $r_i(t = 0)$, $f_i(t = 0)$ and $y_i(t = 0)$. The academic impact is proxied by the papers that a scientist has authored in the two years prior. As above described, we assign to each paper a score equal to the impact factor of the journal where it was published divided by the number of co-authors. Then, for each scientist, we sum the scores of the papers he/she has co-authored between 1998 and 2000. This defines the starting fitness of agents, i.e., $f_i(t = 0)$. We then run the agent-based model using parallel updates of all agents per time step.

Simulating the entry and exit dynamics of agents. Our empirical analysis finds that the number of scientists is almost constant in six years time windows (see Figure S3 in SM). Thus, the total number of agents is almost constant during our simulations. Specifically, the number of new agents n_n is proportional to the number of removed agents n_r at the previous time step. We sample n_n from a Gaussian distribution with mean n_r and standard deviation $\sigma = n_r(0.1/2)$.

Simulating the matching problem. To match agents with locations, we first create a ranking of the agents according to their fitness. Starting from the agent with higher fitness, we look at its top five preferred locations. If one of these locations accept the agent, we move it there. When an agent i has moved to a new location K, we update its position vector, $r_i(t+1) = R_K$, and keep its fitness constant, $f_i(t+1) = f_i(t)$. Then, we consider the second agent in the ranking and keep trying to match it to a new location. With this approach, we ensure that agents relocate to their preferred locations if they are accepted. Also, since we first try to match agents with higher fitness, locations obtain agents with higher fitness, i.e., their preferred ones.

Data Availability

The raw XML data on all MEDLINE articles are available for download from the NIH at

- https://www.nlm.nih.gov/databases/download/pubmed_medline.html

- ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline

The disambiguation of authors (Authority) and affiliations (MapAffil) has been obtained from

- http://abel.lis.illinois.edu/downloads.html

Access to this resource can be requested for free from the maintainers through the online form on the same page. Note that due to an agreement with the providers of Authority and MapAffil, these datasets may only be shared by requesting access through the previously mentioned online form.

We make the aggregated mobility network at city level available with no individual identifying information through figshare after publication.

References

[1] Ajay Agrawal et al. "Brain drain or brain bank? The impact of skilled emigration on poor-country innovation". In: *Journal of Urban Economics* 69.1 (2011), pp. 43–55. ISSN: 00941190. DOI: 10.1016/j.jue.2010.06.003. arXiv: arXiv:1011.1669v3. URL: https://ac.els-cdn.com/S0094119010000367/1-s2.0-S0094119010000367-main.pdf?%7B%5C_%7Dtid=8c0509de-cfbe-11e7-99c0-00000aab0f27%7B%5C&%7Dacdnat=1511380476%7B%5C_%7D83ccbac53763a07ffd8df3ca8e577a2f.

[2]   Silvia Appelt et al. "Chapter 7 - Which Factors Influence the International Mobility of Research Scientists?" In: *Global Mobility of Research Scientists*. Ed. by Aldo Geuna. San Diego: Academic Press, 2015, pp. 177–213. ISBN: 978-0-12-801396-0. DOI: https://doi.org/10.1016/B978-0-12-801396-0.00007-7. URL: http://www.sciencedirect.com/science/article/pii/B9780128013960000077.

[3]   Dany Bahar, Ricardo Hausmann, and Cesar A Hidalgo. "International Knowledge Diffusion and the Comparative Advantage of Nations". In: *HKS Faculty Research Working Paper Series and CID Working Papers* (2012), pp. 001–11. URL: i.dont.know.yet.

[4]   Marco Bee, Massimo Riccaboni, and Stefano Schiavo. "The size distribution of US cities: Not Pareto, even in the tail". In: *Economics Letters* 120.2 (2013), pp. 232–237.

[5]   Schon Beechler and Ian C. Woodward. "The global "war for talent"". In: *Journal of International Management* 15.3 (Sept. 2009), pp. 273–285. ISSN: 1075-4253. DOI: 10.1016/J.INTMAN.2009.01.002. URL: https://www.sciencedirect.com/science/article/pii/S1075425309000453.

[6]   Michel Beine, Frederic Docquier, and Hillel Rapoport. "Brain Drain and Economic Growth: Theory and Evidence". In: *Journal of Development Economics* 64.1 (2001), pp. 275–289. ISSN: 03043878. DOI: 10.1016/S0304-3878(00)00133-4. URL: www.elsevier.comrlocatereconbase%20http://www.sciencedirect.com/science/article/B6VBV-41TWJDK-D/2/66b8ed88481e52a55bd112c79008b24c.

[7]   Jean Pascal Bénassy and Elise S Brezis. "Brain drain and development traps". In: *Journal of Development Economics* 102 (2013), pp. 15–22. ISSN: 03043878. DOI: 10.1016/j.jdeveco.2012.11.002. URL: https://ac.els-cdn.com/S030438781200096X/1-s2.0-S030438781200096X-main.pdf?%7B%5C_%7Dtid=5a9ed9fe-cfa7-11e7-9793-00000aab0f02%7B%5C&%7Dacdnat=1511370514%7B%5C_%7Da398dcb3fed8589319af9bd9e26c0012.

[8]   James Bergstra and Yoshua Bengio. "Random search for hyper-parameter optimization". In: *Journal of Machine Learning Research* 13.2 (2012), pp. 281–305.

[9]   Anna Boucher and Lucie Cerna. "Current Policy Trends in Skilled Immigration Policy". In: *International Migration* 52.3 (June 2014), pp. 21–25. ISSN: 00207985. DOI: 10.1111/imig.12152. URL: http://doi.wiley.com/10.1111/imig.12152.

[10]   Carolina Cañibano, F. Javier Otamendi, and Francisco Solís. "International temporary mobility of researchers: a cross-discipline study". In: *Scientometrics* 89.2 (July 2011), p. 653. ISSN: 1588-2861. DOI: 10.1007/s11192-011-0462-2. URL: https://doi.org/10.1007/s11192-011-0462-2.

[11]   Aaron Clauset, Samuel Arbesman, and Daniel B Larremore. "Systematic inequality and hierarchy in faculty hiring networks". In: *Science advances* 1.1 (2015), e1400005.

[12]   Mathias Czaika and Christopher R. Parsons. "The Gravity of High-Skilled Migration Policies". In: *Demography* 54.2 (Apr. 2017), pp. 603–630. ISSN: 0070-3370. DOI: 10.1007/s13524-017-0559-1. URL: http://link.springer.com/10.1007/s13524-017-0559-1.

[13]   Michael S Dahl and Olav Sorenson. "The migration of technical workers". In: *Journal of Urban Economics* 67.1 (2010), pp. 33–45.

[14]   Jan Eeckhout. "Gibrat's law for (all) cities". In: *American Economic Review* 94.5 (2004), pp. 1429–1451.

[15]   Ana Fernandez-Zubieta, Aldo Geuna, and Cornelia Lawson. "Chapter 1 - What Do We Know of the Mobility of Research Scientists and of its Impact on Scientific Production". In: *Global Mobility of Research Scientists*. Ed. by Aldo Geuna. San Diego: Academic Press, 2015, pp. 1–33. ISBN: 978-0-12-801396-0. DOI: https://doi.org/10.1016/B978-0-12-801396-0.00001-6. URL: https://www.sciencedirect.com/science/article/pii/B9780128013960000016.

[16]   Santo Fortunato et al. "Science of science". In: *Science* 359.6379 (2018), eaao0185.

[17]   Chiara Franzoni, Giuseppe Scellato, and Paula Stephan. "Chapter 2 - International Mobility of Research Scientists: Lessons from GlobSci". In: *Global Mobility of Research Scientists*. Ed. by Aldo Geuna. San Diego: Academic Press, 2015, pp. 35–65. ISBN: 978-0-12-801396-0. DOI: https://doi.org/10.1016/B978-0-12-801396-0.00002-8. URL: http://www.sciencedirect.com/science/article/pii/B9780128013960000028.

[18]   Chiara Franzoni, Giuseppe Scellato, and Paula Stephan. "The mover's advantage: The superior performance of migrant scientists". In: *Economics Letters* 122.1 (Jan. 2014), pp. 89–93. ISSN: 01651765. DOI: 10.1016/j.econlet.2013.10.040. URL: www.elsevier.com/locate/ecolet%20http://linkinghub.elsevier.com/retrieve/pii/S0165176513004874.

[19]   Aldo Geuna. *Global mobility of research scientists: The economics of who goes where and why*. Academic Press, 2015.

[20]   John Gibson and David McKenzie. "Scientific mobility and knowledge networks in high emigration countries: Evidence from the Pacific". In: *Research Policy* 43.9 (2014), pp. 1486–1495. ISSN: 0048-7333. DOI: https://doi.org/10.1016/j.respol.2014.04.005. URL: http://www.sciencedirect.com/science/article/pii/S0048733314000663.

[21]   Charles J Gomez and David MJ Lazer. "Clustering knowledge and dispersing abilities enhances collective problem solving in a network". In: *Nature communications* 10.1 (2019), pp. 1–11.

[22]   Andrey Kolmogorov. "Sulla determinazione empirica di una legge di distribuzione". In: *Inst. Ital. Attuari, Giorn.* 4 (1933), pp. 83–91.

[23]   Averill M Law, W David Kelton, and W David Kelton. *Simulation modeling and analysis*. Vol. 2. McGraw-Hill New York, 1991.

[24]   Brian K Lee, Justin Lessler, and Elizabeth A Stuart. "Improving propensity score weighting using machine learning". In: *Statistics in medicine* 29.3 (2010), pp. 337–346.

[25] Sami Mahroum. "Scientific Mobility". In: *Science Communication* 21.4 (June 2000), pp. 367–378. ISSN: 1075-5470. DOI: 10.1177/1075547000021004003. URL: http://journals.sagepub.com/doi/10.1177/1075547000021004003.

[26] Ernest Miguélez and Rosina Moreno. "What attracts knowledge workers? The role of space and social networks". In: *Journal of Regional Science* 54.1 (2014), pp. 33–60. ISSN: 00224146. DOI: 10.1111/jors.12069.

[27] Enrico Moretti and Daniel J Wilson. "State incentives for innovation, star scientists and jobs: Evidence from biotech". In: *Journal of Urban Economics* 79 (2014), pp. 20–38.

[28] Kevin Morgan. "The exaggerated death of geography: Learning, proximity and territorial innovation systems". In: *Journal of Economic Geography* 4.1 (2004), pp. 3–21. ISSN: 14682702. DOI: 10.1093/jeg/4.1.3.

[29] Mark EJ Newman. "The structure of scientific collaboration networks". In: *Proceedings of the national academy of sciences* 98.2 (2001), pp. 404–409.

[30] Alexander M Petersen. "Multiscale impact of researcher mobility". In: *Journal of The Royal Society Interface* 15.146 (2018), p. 20180580.

[31] Lior Rokach and Oded Maimon. "Top-down induction of decision trees classifiers-a survey". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 35.4 (2005), pp. 476–487.

[32] Alvin E Roth and Elliott Peranson. "The redesign of the matching market for American physicians: Some engineering aspects of economic design". In: *American economic review* 89.4 (1999), pp. 748–780.

[33] AnnaLee Saxenian. "From Brain Drain to Brain Circulation: Transnational Communities and Regional Upgrading in India and China". In: *Studies in Comparative International Development* 40.2 (June 2005), pp. 35–61. ISSN: 0039-3606. DOI: 10.1007/BF02686293. URL: http://link.springer.com/10.1007/BF02686293.

[34] Giuseppe Scellato, Chiara Franzoni, and Paula Stephan. "A mobility boost for research". In: *Science* 356.6339 (May 2017), pp. 694.2–694. ISSN: 0036-8075. DOI: 10.1126/science.aan4052. URL: http://science.sciencemag.org/content/sci/356/6339/696.2.full.pdf%20http://www.sciencemag.org/lookup/doi/10.1126/science.aan4052.

[35] Mario V Tomasello, Giacomo Vaccario, and Frank Schweitzer. "Data-driven modeling of collaboration networks: a cross-domain analysis". In: *EPJ Data Science* 6.1 (2017), p. 22.

[36] Mario V Tomasello et al. "The role of endogenous and exogenous mechanisms in the formation of R&D networks". In: *Scientific reports* 4 (2014), p. 5679.

[37] Vetle I. Torvik. "MapAffil: A Bibliographic Tool for Mapping Author Affiliation Strings to Cities and Their Geocodes Worldwide". In: *D-Lib Magazine* 21.11/12 (Nov. 2015). ISSN: 1082-9873. DOI: 10.1045/november2015-torvik. URL: http://www.dlib.org/dlib/november15/torvik/11torvik.html.

[38] Vetle I. Torvik and Neil R. Smalheiser. "Author name disambiguation in MEDLINE". In: *ACM Transactions on Knowledge Discovery from Data* 3.3 (July 2009), pp. 1–29. ISSN: 15564681. DOI: 10.1145/1552303.1552304. URL: http://portal.acm.org/citation.cfm?doid=1552303.1552304.

[39] US Census Bureau. *Metropolitan and Micropolitan.* https://www.census.gov/programs-surveys/metro-micro/about.html. [Online; 2020-02-06]. 2018.

[40] Giacomo Vaccario, Luca Verginer, and Frank Schweitzer. "The mobility network of scientists: Analyzing temporal correlations in scientific careers". In: *Applied Network Science* 5.1 (July 2020), p. 36. ISSN: 2364-8228. DOI: 10.1007/s41109-020-00279-x. eprint: 1905.06142. URL: https://doi.org/10.1007/s41109-020-00279-x.

[41] Giacomo Vaccario et al. "Quantifying knowledge exchange in R&D networks: A data-driven model". In: *Journal of Evolutionary Economics* 28.3 (Aug. 2018), pp. 461–493. DOI: 10.1007/s00191-018-0569-1. URL: https://link.springer.com/content/pdf/10.1007/s00191-018-0569-1.pdf.

[42] Luca Verginer and Massimo Riccaboni. "Brain–Circulation Network: The Global Mobility of the Life Scientists". In: *(Working Papers)* 10/2018 (Oct. 2018).

[43] Luca Verginer and Massimo Riccaboni. "Cities and countries in the global scientist mobility network". In: *Applied Network Science* 5.1 (July 2020), p. 38. DOI: 10.1007/s41109-020-00276-0.

[44] Luca Verginer and Massimo Riccaboni. "Talent Goes to Global Cities: The World Network of Scientists' Mobility". In: *Research Policy* In Press (2020). DOI: 10.1016/j.respol.2020.104127.

[45] Reinhilde Veugelers and Linda Van Bouwel. "Chapter 8 - Destinations of Mobile European Researchers: Europe versus the United States". In: *Global Mobility of Research Scientists.* Ed. by Aldo Geuna. San Diego: Academic Press, 2015, pp. 215–237. ISBN: 978-0-12-801396-0. DOI: https://doi.org/10.1016/B978-0-12-801396-0.00008-9. URL: http://www.sciencedirect.com/science/article/pii/B9780128013960000089.

[46] Caroline S Wagner and Loet Leydesdorff. "Network structure, self-organization, and the growth of international collaboration in science". In: *Research policy* 34.10 (2005), pp. 1608–1618.