Unsupervised Embedding of Trajectories Captures the Latent Structure of Mobility

Dakota Murray
Indiana University

Jisung Yoon
Pohang University of Science and Technology  https://orcid.org/0000-0002-8463-6094

Sadamori Kojaku
Indiana University

Rodrigo Costas
Leiden University

Woo-Sung Jung
Pohang University of Science and Technology

Staša Milojević
Indiana University

Yong-Yeol Ahn (yyahn@iu.edu)
indiana university bloomington

Article

Keywords: Human mobility, geographic distance, mathematical equivalence, granularity.

Posted Date: November 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1062258/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Unsupervised embedding of trajectories captures the latent structure of mobility

Dakota Murray$^{1,\dagger}$, Jisung Yoon$^{1,2,\dagger}$, Sadamori Kojaku$^1$, Rodrigo Costas$^{3,4}$, Woo-Sung Jung$^{2,5,6}$, Staša Milojević $^\ast 1$, and Yong-Yeol Ahn $^{1,7,8}$

$^1$Center for Complex Networks and Systems Research, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47408, USA
$^2$Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang 37673, Republic of Korea.
$^3$Centre for Science and Technology Studies (CWTS), Leiden University, P.O. Box 905, 2300 AX Leiden, The Netherlands
$^4$DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, Stellenbosch, South Africa
$^5$Department of Physics, Pohang University of Science and Technology, Pohang 37673, Republic of Korea.
$^6$Asia Pacific Center for Theoretical Physics, Pohang 37673, Republic of Korea.
$^7$Network Science Institute, Indiana University, Bloomington (IUNI), IN 47408, USA
$^8$Connection Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
$\dagger$These authors contributed equally: Jisung Yoon, Dakota Murray.

November 9, 2021

*smilojev@indiana.edu
†yyahn@iu.edu
Human mobility drives major societal phenomena including epidemics, economies, and innovation. Historically, mobility was constrained by geographic distance, however, in the globalizing world, language, culture, and history are increasingly important. Here, we show a mathematical equivalence between word2vec model and the gravity model of mobility and demonstrate that, by using three human trajectory datasets, word2vec encodes nuanced relationships between locations into a systematic and meaningful vector-space, providing a functional distance between locations, as well as a representation for studying the many dimensions of mobility. Focusing on the case of scientific mobility, we show that embeddings implicitly learn cultural, linguistic, and hierarchical relationships at multiple levels of granularity. Connecting neural embeddings to the gravity model opens up new avenues for the study of mobility.

Introduction

How far apart are two places? The question is surprisingly hard to answer when it involves human mobility. Although geographic distance has historically constrained human movements, it is becoming less relevant in a world connected by rapid transit and global airline networks. For instance, a person living in Australia is more likely to migrate to the United Kingdom, a far-away country with similar language and culture, than to a much closer country such as Indonesia. Similarly, a student in South Korea is more likely to attend a university in Canada than one in North Korea. Although geographic distance has been used as the most prominent basis for models of mobility, such as the Gravity and Radiation models, the decreasing relevance of geography calls for alternative ways of conceptualizing “distance”. Yet, functional distances are often low-resolution, computed at the level of countries rather than regions, cities, or organizations, and have focused on only a single facet of mobility at
a time, whereas real-world mobility is multi-faceted, influenced simultaneously by geography, language, culture, history, and economic opportunity. Low dimensional distance alone cannot represent the multitude of inter-related factors that drive mobility. Networks offer a solution to representing many dimensions of mobility, yet edges only encode simple relationships between connected entities. Capturing the complexity of mobility requires moving beyond simple functional distances and networks, to learning high-dimensional landscapes of mobility that incorporate many facets of mobility into a single fine-grained and continuous representation.

Here, we demonstrate that the word2vec model (Skip-Gram Negative Sampling) is equivalent to the gravity law of mobility, a fundamental mobility framework. We then empirically test the resulting representation by its ability to derive from mobility trajectories the functional distances between locations. After validating that it coherently represents real-world data, we leverage unique embedding-based methods to demonstrate word2vec’s capacity for encoding rich information relating to geography, culture, language, and even prestige, at multiple scales of analysis. We embed trajectories from three datasets: U.S. passenger flight itinerary records, Korean accommodation reservations, and a dataset of scientists’ career mobility between organizations captured in bibliometric records (Detailed descriptions are available in the Methods).

We focus in most detail on scientific mobility due to its richness and importance. Scientific mobility—which is a central driver of the globalized scientific enterprise and strongly related to innovation, impact, collaboration, and the diffusion of knowledge—is not only an important topic in the Science of Science but also ideal for our study thanks to its well-known structural properties such as the centrality of scientifically advanced countries and the strong prestige hierarchy. In spite of its importance, understandings of scientific mobility have been limited by the sheer scope and complexity of the phenomenon, being further confounded by the diminishing role of geography in shaping the landscape of scientific mobility. Trajectories of scientific mobility are constructed using more than three million name-
disambiguated authors who were mobile—having more than one affiliation—between 2008 and 2019, as evidenced by their publications indexed in the Web of Science database (see Methods). As a scientist’s career progresses, they move between organizations or pick up additional (simultaneous) affiliations forming affiliation trajectories (Fig. 1a). Thus, the trajectories encode both migration and co-affiliation—the holding of multiple simultaneous affiliations involving the sharing of time and capital between locations—that is typical of scientific mobility (see Supporting Information).

Here, we study the skip-gram negative sampling (SGNS), or word2vec neural-network architecture (see Methods). This neural embedding model, originally designed for learning models of language, has been making breakthroughs by revealing novel insights into texts, networks, and trajectories. It works under the notion that a good representation should facilitate prediction, learning a mapping between words that can predict a target word based on its context (surrounding words). The model is also computationally efficient, easy to use, robust to noise, and can encode relations between entities as geometric relationships in the vector space. When applied to the trajectory data, each location is encoded into a single vector representation, and vectors relate to one another based on the likelihood of locations appearing adjacent to one another in the same trajectory. Also, word2vec can be interpreted as a metric recovery, which recovers the underlying metric of the semantic manifold.

Our study builds on the gravity model framework, widely-used, fundamental mobility model that connects the expected flux, $\hat{T}_{ij}$, between locations based on their populations and distance:

$$\hat{T}_{ij} = C m_i m_j f(r_{ij}),$$

where $m_i$ is the population of location $i$, $f(r_{ij})$ is a decay function with respect to distance between locations, and $C$ is a constant estimated from data (see Methods). For the flight itinerary
data, we use population $m_i$ as the total number of unique passengers who passed through each
airport, for the Korean accommodation reservation data, we use the total number of unique
customers who booked with each accommodation, and for scientific mobility, we use the mean
annual number of unique mobile and non-mobile authors who were affiliated with each orga-
nization. $\hat{T}_{ij}$, which is often referred to as “expected flux”\textsuperscript{4}, is the expected frequency of the
co-occurrence of location $i$ and $j$ in the trajectory in the gravity model.

The gravity model dictates that the expected flow, $\hat{T}_{ij}$, ($\hat{T}_{ij} = \hat{T}_{ji}$), is proportional to the loca-
tions’ population, $\hat{T}_{ij} \propto m_i m_j$, and decays as a function of their distance, $f(r_{ij})$. Traditionally,
the decay function has been defined in terms of geographic distance, due to its intuitiveness
and availability. Here, we also consider the embedding distance, calculated as the cosine dis-
tance between location vectors modeled by word2vec, to test the ability to encode the complex
relationships in the mobility data. The decay function $f(r_{ij})$ defines the effect of distance,
and different decay functions can model fundamentally different mechanisms\textsuperscript{42} such as the cost
functions for a given distance and the spatial granularity of the observation. For geographic dis-
tance, we define $f(r_{ij})$ as the standard power-law function, and for the embedding distance, we
use the exponential function, selected as the best performing for each case (Fig. S7 and Fig. S8).

Our results show that embedding distance deriving from the word2vec significantly improves
the ability to predict the mobility compare to geographical distance no matter decay function
while providing the useful representation of locations as well (see Tables S3, S4, S5). However,
we note that the geographic distance best models real-world mobility with the power-law de-
cay function, likely resulting from the function’s suitability for large, complex, and scale-free
spatial systems\textsuperscript{43}; the embedding distance, in contrast, models our mobility data best for the
exponential decay function, which stems from an underlying connection between the function
and word2vec.
Results

word2vec and the gravity model

We first demonstrate the mathematical equivalence between the SGNS model and the gravity model. The word2vec model takes a location trajectory, denoted by \((a_1, a_2, \ldots, a_T)\), as input. A target location \(a_t = i\) is considered to have a context location \(a_{t'} = j\) that appears in the previous or subsequent \(w\) locations in the trajectory, i.e., \(j \in [a_{t-w}, \ldots, a_{t-1}, a_{t+1}, \ldots, a_{t+w}]\). word2vec learns an embedding by estimating the probability that location \(i\) has context \(j\):

\[
P(j \mid i) := \frac{\exp(u_j \cdot v_i)}{Z_i},
\]

where the denominator \(Z_i = \sum_{j' \in \mathcal{A}} \exp(u_{j'} \cdot v_i)\) is a normalization constant, and \(\mathcal{A}\) is the set of all locations. Although word2vec generates two embedding vectors \(v_i\) and \(u_i\)—referred to as the in-vector and out-vector, respectively—we follow convention to use the in-vector \(v_i\) as an embedding of location \(i\).

When \(w = 1\) (Fig. S4), the flow can be written as

\[
\hat{T}_{ij} \propto P(i)P(j \mid i) = \frac{P(i) \exp(u_j \cdot v_i)}{Z_i},
\]

where \(P(j)\) is the fraction of location \(j\)'s frequency in the data. In general, calculating \(Z_i\) is computationally expensive and there are two common approximations: hierarchical softmax and negative sampling. Due to its simplicity and performance, negative sampling is the most widely used strategy, which we also adopt in our study.

Although negative sampling is the most common approximation, it is a biased estimator and fits a different probability model. When taking into account this bias, word2vec with skip-gram and negative sampling fits a probability model given by

\[
P(j \mid i) := \frac{P^\gamma(j) \exp(u_j \cdot v_i)}{Z'_i},
\]
where we redefine the normalization constant as $Z'_i = \sum_{j' \in \mathcal{A}} P(j') \exp(u_{j'} \cdot v_i)$. (See the Methods and Supporting Information for the full derivation).

Parameter $\gamma = 1$ is a special choice that ensures that, when the embedding dimension is sufficiently large, there exists optimal in-vectors and out-vectors such that $v_i = u_i^{35}$. Setting $\gamma = 1$ and substituting $v_i = u_i$ into Eq. 4, the flow predicted by word2vec is given by

$$\hat{T}_{ij} \propto N(i)P(j \mid i) = \frac{NP(i)P(j) \exp(v_j \cdot v_i)}{Z_i'^{}}.$$ (5)

where $N$ is the sum of frequencies in all locations and $N(i) = NP(i)$ is the frequency of location $i$ in the data.

The flow $\hat{T}_{ij}$ is symmetric (i.e., $\hat{T}_{ij} = \hat{T}_{ji}$) because the skip-gram model neglects whether the context $j$ appears before or after the target $i$ in the trajectory. If we swap $i$ and $j$ in Eq. 5, the numerator remains the same but the denominator can be different. Therefore, to ensure $\hat{T}_{ij} = \hat{T}_{ji}$, the denominator $Z_i$ should be a constant.

Taken together, the word2vec model with the negative sampling predicts a flow in the same form as in the gravity model:

$$\hat{T}_{ij} = CP(i)P(j) \exp(v_j \cdot v_i).$$ (6)

In other words, word2vec with skip-gram negative sampling is mathematically equivalent to the gravity model, with the mass given by the location’s frequency $P(i)$, and the distance measured by their dot similarities. While the gravity model describes mobility flows from the given mass and locations, word2vec estimates proper locations in the vector space which best explain the given mobility flow. This makes word2vec a powerful tool for learning models of mobility.
Embeddings provide functional distance between locations

To ensure that word2vec learns an systematic representation of mobility that encodes meaningful functional distances, we devise an empirical validation task. Past studies have utilized a range of validating tasks for assessing the accuracy word2vec, including analogy completion and human surveys, in the case of natural language, or link prediction, in the case of network embedding. Here, we expect that an extracted representation from the mobility data should provide a functional distance that better models the flux between institutions than does geographic distance, which is at present the most widely measure used in the gravity model. We test this notion using three human mobility datasets, showing that word2vec consistently offers a better representation of actual mobility flows than geographic distance, as well as alternative network and direct optimization approaches.

In the case of flight itineraries, the embedding distance explains more than twice the expected flux between airports ($R^2 = 0.51$, Fig. 1b) than does geographic distance ($R^2 = 0.22$), which has traditionally been used to quantify distance for the gravity model. Also, the embedding distance produces better predictions of actual flux between airports than does the geographic distance (Fig. 1c). In the case of Korean accommodation reservations, embedding distance better explains the expected flux ($R^2 = 0.57$, Fig. 1d) than does geographic distance ($R^2 = 0.25$), and predictions made using the embedding distance outperform those made with geographic distance (Fig. 1e). This performance is consistent in the case of scientific mobility: the embedding distance explains more than twice the expected flux ($R^2 = 0.48$, Fig. 1f) than does the geographic distance ($R^2 = 0.22$), and predictions made using the embedding distance outperform those using the geographic distance (Fig. 1g). These patterns hold for the subsets of only domestic (within-country organization pairs, Fig. S7 and Fig. S9c) and only international mobility flows (across-country organization pairs, Fig. S9d).

The embedding distance also out-performs alternative diffusion-based network distance
Figure 1: Neural embedding provides functional distance that improves predictive power of the gravity model of mobility best across three distinct human trajectory datasets. **a.** A unique identifier is assigned to each organization and they are assembled into an affiliation trajectory ordered by year of publication (top). If an author lists multiple organization affiliations within the same year, we shuffle the order within that year in each training iteration (bottom, see Supporting Information). **b.** Embedding distance (top) better explains the expected flux of passengers between U.S. airports than does geographic distance (bottom). The red line is the line of the best fit. Black dots are mean flux across binned distances. 99% confidence intervals are plotted for the mean flux in each bin. Correlation is calculated on the data in the log-log scale ($p < 0.0001$ across all fits). The lightness of each hex bin indicates the frequency of organization pairs within it. **c.** Predictions of flux between airport pairs made using embedding distance (top) outperform those made using geographic distance (bottom). Box-plots show the distribution of actual flux for binned values of predicted flux. Box color corresponds to the degree to which the distribution overlaps with $y = x$. “RMSE” is the root-mean-squared error between the actual and predicted values. Embedding distance consistently produce powerful functional distance for Korean accommodation reservations (**d,e**), and global scientific mobility (**f,g**).
measures including the personalized Page Rank scores calculated from the underlying mobility network (Fig. S5, Fig. S11, Fig. S12). The embedding distance derived from neural embedding also explains more of the flux and better predicts mobility flows than simpler embedding baselines, such as distances derived from a singular-value decomposition and a Laplacian Eigenmap embedding\(^47\) of the underlying location co-occurrence matrix, Levy’s symmetric \textit{word2vec}\(^35\), and even direct optimization of the gravity model (Fig. S5 and Tables S3, S4, S5). In sum, our results demonstrate that, consistently and efficiently, the embedding distance better captures patterns of actual mobility than does the geographic distance.

In practice, because we only have limited amounts of noisy data and the optimization may not find the true optimum, the mathematical result may only approximately hold. Indeed, we find that the in- and out-vectors tend to be different and that the cosine similarity tends to better capture real-world mobility than the inner product similarity. This result echoes other applications of word embedding, such as word analogy testing\(^48\), in which cosine distance also outperformed the inner product similarity. Nevertheless, a model with the inner product similarity has the second-best performance after cosine similarity (Tables S3, S4, S5), and the embedding distance still outperforms all alternatives we considered.

**Embeddings capture global structure of mobility**

In the remainder of the paper, we focus on scientific mobility and interrogate the geometric space generated by the neural embedding to shed light on the multi-faceted relationships between organizations. To explore the topological structure of the embedding, we use a topology-based dimensionality reduction method (UMAP\(^49\)) to obtain a two-dimensional representation of the embedding space (Fig. 2a). By leveraging the unique characteristics of representation learning approach, we are able to show the relationships between individual organizations, rather than aggregates such as nations or cities, this projection constitutes the largest and highest
resolution “map” of scientific mobility to date.

Globally, the geographic constraints are conspicuous; organizations tend to form clusters based on their national affiliations and national clusters tend to be near their geographic neighbors. At the same time, the embedding space also reflects a mix of geographic, historic, cultural, and linguistic relationships between regions much more clearly than alternative network representations (Fig. S13) that have been common in studies of scientific mobility\textsuperscript{8,50}.

The embedding space also allows us to zoom in on subsets and re-project them to reveal local relationships. For example, re-projecting organizations located in Western, Southern, and Southeastern Asia with UMAP (Fig. 2b) reveals a gradient of countries between Egypt and the Philippines that largely corresponds to geography, but with some exceptions seemingly stemming from cultural and religious similarity; Malaysia, with its official religion of Islam, is nearer to Middle Eastern countries in the embedding space than to many geographically-closer South Asian countries. We validate this finding quantitatively with the cosine distance between nations (the centroids of organizations vectors belonging to that country). Malaysia is nearer to many Islamic countries such as Iraq ($d = 0.27$), Pakistan ($d = 0.32$), and Saudi Arabia ($d = 0.41$) than neighboring but Buddhist Thailand ($d = 0.43$) and neighboring Singapore ($d = 0.48$).

Linguistic and historical ties also affect scientific mobility. We observe that Spanish-speaking Latin American nations are positioned near Spain (Fig. 2c), rather than Portuguese-speaking Brazil ($d = 0.35$ vs. $d = 0.54$ for Mexico and $d = 0.39$ vs. $d = 0.49$ for Chile) reflecting linguistic and cultural ties. Similarly, North-African countries that were once under French rule such as Morocco are closer to France ($d = 0.32$) than to similarly geographically-distant European countries such as Spain ($d = 0.39$), Portugal ($d = 0.52$), and Italy ($d = 0.52$). Comparable patterns exist even within a single country. For example, organizations within Quebec in Canada are located nearer France ($d = 0.37$) than the United States ($d = 0.51$).
Mirroring the global pattern, organizations in the United States are largely arranged according to geography (Fig. 2d). Re-projecting organizations located in Massachusetts (Fig. 2e) reveals structure based on urban centers (Boston vs. Worcester), organization type (e.g., hospitals vs. universities), and university systems (University of Massachusetts system vs. Harvard & MIT). For example, even though UMass Boston is located in Boston, it clusters with other universities in the UMass System \(d = 0.29\) rather than the other typically more highly-ranked and research-focused organizations in Boston \(d = 0.39\), implying a relative lack of mobility between the two systems. Similar structures can be observed in other states such as among New York’s CUNY and SUNY systems (Fig. S14), Pennsylvania’s state system (Fig. S15), Texas’s Agricultural and Mechanical universities (Fig. S16), and between the University of California and State University of California systems (Fig. S17).

Just as the embedding space makes it possible to zoom in on subsets of organizations, it is also possible to zoom out by aggregating organizational vectors. In doing so, we can examine the large-scale structure that governs scientific mobility. We define the representative vector of each country as the average of their organizational vectors and, using their cosine similarities, perform hierarchical clustering of nations that have at least 25 organizations represented in the embedding space (see Fig. 3a). The six identified clusters roughly correspond to countries in Asia and North America (orange), Northern Europe (dark blue), the British Commonwealth and Iran (purple), Central and Eastern Europe (light blue), South America and Iberia (dark green), and Western Europe and the Mediterranean (light green). The cluster structure shows that not only geography but also linguistic ties and cultural between countries are related to scientific mobility.

We quantify the relative importance of geography (by region), and language (by the most widely-spoken language of each country) using the element-centric clustering similarity\(^5\), a method that can compare hierarchical clustering and disjoint clustering (geography, language...)
Figure 2: Projection of embedding space reveals complex multi-scale structure of organizations. 

a. UMAP projection of the embedding space reveals country-level clustering. Each point corresponds to an organization and its size indicates the average annual number of mobile and non-mobile authors affiliated with that organization from 2008 to 2019. Color indicates the region. The separation of organizations in Quebec and the rest of Canada is highlighted. 

b.Zooming into (re-projecting) the area containing countries in Western, South, and Southeast Asia shows a geographic and cultural gradient of country clusters.

c. Similarly, zooming into the area containing organizations in Spain, Portugal, South, and Central America shows clustering by most widely-spoken majority language group: Spanish and Portuguese.

d. Doing the same for organizations in the United States reveals geographic clustering based on state, roughly grouped by Census Bureau-designated regions.

e. Zooming in further on Massachusetts reveals clustering based on urban center (Boston, Worcester), organizational sector (hospitals vs. universities), and university systems and prestige (UMass system vs. Harvard, MIT, etc.).
at the different level of hierarchy by explicitly adjusting a scaling parameter $r$, acting like a *zooming lens*. If $r$ is high, the similarity is based on the lower levels of the dendrogram, whereas when $r$ is low, the similarity is based on higher levels. Fig. 3b demonstrates that regional relationships play a major role at higher levels of the clustering process (low $r$), and language (family) explains the clustering more at the lower levels (high $r$). This suggests that the embedding space captures the hierarchical structure of mobility.

**Embeddings capture latent prestige hierarchy**

The embedding space can also encode more fine-grained relationships between entities. For examples, prestige hierarchies are known to underpin the dynamics of scientific mobility, in which researchers tend to move to similar or less prestigious organizations\textsuperscript{16,17}. Could the embedding space, to which no explicit prestige information is given, encode a prestige hierarchy? This question is tested by exploiting the geometric properties of the embedding space with SemAxis\textsuperscript{36}. Here, we use SemAxis to operationalize the abstract notion of academic prestige, defining an axis in the embedding space where poles are defined using known high- and low-ranked universities. As an external proxy of prestige, we use the Times Ranking of World Universities (we also use research impact from the Leiden Ranking\textsuperscript{53}, see Supporting Information); the high-rank pole is defined as the average vector of the top five U.S. universities according to the rankings, whereas the low-rank pole is defined using the five bottom-ranked (geographically-matched by U.S. census region) universities. We derive an embedding-based ranking for universities based on the geometrical spectrum from the high-ranked to low-ranked poles (see Data and Methods).

The embedding space encodes the prestige hierarchy of U.S. universities that are coherent with real-world university rankings. The embedding-based ranking is strongly correlated with the Times ranking (Spearman’s $\rho = 0.73$, Fig. 4a). For reference, the correlation be-
Figure 3: Geography, then language, conditions international mobility. a. Hierarchically clustered similarity matrix of country vectors aggregated as the mean of all organization vectors within countries with at least 25 organizations. Color of matrix cells corresponds to the cosine similarity between country vectors. Color of country names corresponds to their cluster. Color of three cell columns separated from the matrix corresponds to, from left to right, the region of the country, the language family, and the dominant language. b. Element-centric cluster similarity reveals the factors dictating hierarchical clustering. Region better explains the grouping of country vectors at higher levels of the clustering. Language family, and then the most widely-spoken language, better explain the fine-grained grouping of countries.
tween the Times ranking and the publication impact scores from the Leiden Ranking, a bibilometrically-based university ranking, is 0.87 (Spearman’s $\rho$, Fig. 4b). The correlation between the embedding-based ranking and the Times ranking is robust regardless of the number of organizations used to define the axes (Fig. S18), such that even using only the single top-ranked and bottom-ranked universities produces a ranking that is significantly correlated with the Times ranking (Spearman’s $\rho = 0.46$, Fig. S18a). The correlation is also comparable to more direct measures such as node strength (sum of edge weights, Spearman’s $\rho = 0.73$) and eigenvector centrality (Spearman’s $\rho = 0.76$, see Supporting Information) from the mobility network. The strongest outliers that were ranked more highly in the Times ranking than in the embedding-based ranking tend to be large state universities such as Arizona State University and the University of Florida. Those ranked higher in the embedding-based ranking tend to be relatively-small universities near major urban areas such as the University of San Francisco and the University of Maryland Baltimore County, possibly reflecting exchanges of scholars with nearby high-ranked institutions at these locations. In sum, our results suggest that the embedding space is capable of capturing information about academic prestige, even when the representation is learned using data without explicit information on the direction of mobility (as in other formal models$^{16}$), or prestige.

The axes can be visualized to examine the relative position of organizations along the prestige axis, and along a geographic axis between California and Massachusetts. Prestigious universities such as Columbia, Stanford, MIT, Harvard, and Rockefeller are positioned towards the top of the axis (Fig. 4c). Universities at the bottom of this axis tend to be regional universities with lower national profiles (yet still ranked by Times Higher Education) and with more emphasis on teaching, such as Barry University and California State University at Long Beach. By projecting other types of organizations onto the prestige axis, SemAxis offers a new way of representing a continuous spectrum of organizational prestige for which rankings are often
low-resolution, incomplete, or entirely absent, such as for regional and liberal arts universities (Fig. 4d), research institutes (Fig. 4e), and government organizations (Fig. 4f). Their estimated prestige is speculative, though we find that it significantly correlates with their citation impact (Fig. S22).

We also find that the size (L2 norm) of the organization embedding vectors provides insights into the characteristics of organizations (Fig. 5). Up to a point (around 1,000 researchers), the size of U.S. organization’s vectors tends to increase proportionally to the number of researchers (both mobile and non-mobile) with published work; these organizations are primarily teaching-focused institutions, agencies, and hospitals that either are not ranked or have a low ranking. However, at around 1,000 researchers, the size of the vector decreases as the number of researchers increases. These organizations are primarily research-intensive and prestigious universities with higher rank, research outputs, R&D funding, and doctoral students (Fig. S23). A similar pattern has been observed in applications of neural embedding to natural language, in which the size of word vectors were found to represent the word’s specificity, i.e., the word associated with the vector frequently co-appears with particular context words. If the word in question is universal, appearing frequently in many different contexts, it would not have a large norm due to a lack of strong association with a particular context. Likewise, an organization with a small norm, such as Harvard, appears in many contexts alongside many different organizations in affiliation trajectories—it is well-connected. The concavity of the curve emerges in part from the relationship between the size of the vector and the expected connectedness of the organization, given its size ($R^2 = 0.17$). Large, prestigious, and well-funded research universities such as Princeton and Harvard have smaller vector norms because they appear in many different contexts compared to more teaching-focused organizations such as NY Medical College, and the University of Michigan at Flint. Some universities, such as the University of Alaska at Fairbanks, have considerably small vectors, which may be a result of their remote
Figure 4: Embedding captures latent geography and prestige hierarchy. a. Comparison between the ranking of organizations in the Times ranking and the embedding ranking derived using SemAxis. Un-filled points are those top and bottom five universities used to span the axis. Even when considering only a total of ten organization vectors, the estimate of the Spearman’s rank correlation between the embedding and Times ranking is $\rho = 0.73$ ($n = 145$, $p < 0.0001$), which increases when more top- and bottom-ranked universities are included (Fig. S18). b. The Times ranking is correlated with Leiden Ranking of U.S. universities with Spearman’s $\rho = 0.87$ and $p < 0.0001$. c-f. Illustration of SemAxis projection along two axes; the latent geographic axis, from California to Massachusetts (left to right) and the prestige axis. Shown for U.S. Universities (c), Regional and liberal arts colleges (d), Research institutes (e), and Government organizations (f). Full organization names are listed in Table S1.
locations and unique circumstances.

We report that this curve is almost universal across many countries. For instance, China’s curve closely mirrors that of the United States (Fig. 5b). Smaller but scientifically advanced countries such as Australia and other populous countries such as Brazil also exhibit curves similar to the United States (Fig. 5b, inset). Other nations exhibit different curves which lack the portions with decreasing norm, probably indicating the lack of internationally-prestigious institutions. Similar patterns can be found across many of the 30 countries with the most total researchers (Fig. S24; see Supporting Information for more discussion).

Figure 5: Size of organization embedding vectors captures prestige and size of organizations. a. Size (L2 norm) of organization embedding vectors compared to the number of researchers for U.S. universities. Color indicates the rank of the university from the Times ranking, with 1 being the highest ranked university. Uncolored points are universities not listed on the Times ranking. A concave-shape emerges, wherein larger universities tend to be more distant from the origin (large L2 norm); however, the more prestigious universities tend to have smaller L2 norms. b. We find a similar concave-curve pattern across many countries such as the United States, China, Australia, Brazil, and others (inset, and Fig. S24). Some countries exhibit variants of this pattern, such as Egypt, which is missing the right side of the curve. The loess regression lines are shown for each selected country, and for the aggregate of remaining countries, with ribbons mapping to the 99% confidence intervals based on a normal distribution. Loess lines are also shown for organizations in Australia, Brazil, and Egypt (inset).
Conclusion

Neural embedding approaches offer a novel, data-driven solution for efficiently learning an effective and robust representation of locations based on trajectory data, encoding the complex and multi-faceted nature of mobility. Its unique strength stems from our discovery that word2vec is equivalent to a gravity model, making it a natural and theoretically-grounded tool for modeling mobility. By virtue of this equivalency, word2vec learns systematic representations of mobility across disparate domains, including U.S. flight itineraries, Korean hotel accommodation reservations, and global scientific mobility. Focusing on the case of scientific mobility, we leverage the unique topological structure of the embedding space to reveal how it encodes nuanced aspects of mobility, including global and regional geography, shared languages, and prestige hierarchies, all of which are learned using only the mobility trajectories of individuals.

In revealing the correspondence between neural embeddings and the gravity model, the study of human mobility can move beyond geographic and network-based models of mobility, and instead leverage the high-order structure from individuals’ mobility trajectories using these robust and efficient methods. While we focus on three domains of human mobility, this approach could be applied to many different domains, such as animal migration, transit-network mobility, international trade, and more. Once learned, functional distances between locations, such as countries, cities, or organizations, or the embedding model itself, can be published to facilitate re-use, and support reproducibility and transparency in cases when the underlying data is too sensitive to make available. Moreover, this approach can be used to learn a functional distance even between entities for which no geographic analog exists, such as between occupational categories based on individuals’ career trajectories. In addition to providing a functional distance that supports modeling and predicting mobility patterns, the topology of the
embedding space is amenable to a range of unique applications for studying mobility. As we have shown, the embedding space allows the visualization of the complex structure of scientific mobility at high resolution across multiple scales, providing a large and detailed map of the landscape of global scientific mobility. Other operations such as comparing entities or calculating aggregates, which could be complex and computationally-expensive for other methods, are here reduced to simple vector arithmetic. Embeddings also allows us to quantitatively explore abstract notions such as academic prestige, and can potentially be generalized to other abstract axes. Investigation of the structure of the embedding space, such as the vector norm, reveals universal patterns based on the organization’s size and their vector norm that could be leveraged in future studies of mobility.

This approach, and our study, also have several limitations. First, the skip-gram *word2vec* model does not leverage directionality, meaning that embedding will be less effective at capturing mobility for which directionality is critical. Future studies may consider bi-directional embeddings, such as BERT\(^{55}\), to incorporate directionality, as well as their correspondance to asymmetric mobility models, such as the radiation model\(^4\). Second, the neural embedding approach is most useful in cases of mobility between discrete geographic units such as between countries, cities, and businesses; this approach is less useful in the case of mobility between locations represented using geographic coordinates, such as in the modeling of animal movements. Third, neural embeddings are an inherently stochastic procedure, and so results may change across different iterations. However, in this study we observe all results to be robust to stochasticity, likely emerging from the limited “vocabulary” of scientific mobility, airports, and accommodations (several thousand) and the relatively massive datasets used to learn representations (several million trajectories). Applications of *word2vec* to problem domains where the ratio of the vocabulary to data is smaller, however, should be implemented with caution to ensure that findings are not the result of random fluctuations. Finally, the case of scientific
mobility presents domain-specific limitations. Reliance on bibliometric metadata means that we capture only long-term mobility such as migration, rather than the array of more frequent short-term mobility such as conference travel and temporary visits. The kinds of mobility we do capture—migration and co-affiliation—although conceptually different, are treated identically by our model. Also, our data might further suffer from bias based on publication rates: researchers at prestigious organizations tend to have more publications, leading to these organizations appearing more frequently in affiliation trajectories.

Mobility and migration are at the core of human nature and history, driving societal phenomena as diverse as epidemics \(^{41,56}\) and innovation \(^{11–15}\). However, the paradigm of scientific migration may be changing. Traditional hubs of migration have experienced many politically-motivated policy changes that affect scientific mobility, such as travel restrictions in the U.S. and U.K. \(^{57}\). At the same time, other nations, such as China, are growing into major scientific powers and attractors of talent \(^{58}\). Unprecedented health crises such as the COVID-19 pandemic threaten to bring drastic global changes to migration by tightening borders and halting travel. By revealing the correspondence between neural embedding and the gravity model and revealing their utility and efficacy, our study opens a new avenue in the study of mobility.

**Methods**

**U.S. flight itinerary data**

We source U.S. airport itinerary data from the Origin and Destination Survey (DB1B), provided by the Bureau of Transportation Statistics at the United States Department of Transportation. DB1B is a sample of 10 percent of domestic airline tickets between 1993 and 2020, comprising 307,760,841 passenger itineraries between 828 U.S. airports. A trajectory is constructed for each passenger flight itinerary, forming an ordered sequence of unique identifiers of the origin and destination airports. Each itinerary is associated with a trajectory of airports including the
origin, destination, and intermediary stops.

Korean accommodation reservation data

We source Korean accommodation reservation data from collaboration with Goodchoice Company LTD. The data contains customer-level reservation trajectories spanning the period of August 2018 through July 2020 and comprising 1,038 unique accommodation locations in Seoul, South Korea. A trajectory is constructed for each customer, containing the ordered sequences of accommodations they reserved over time.

Scientific mobility data

We source co-affiliation trajectories of authors from the Web of Science database hosted by the Center for Science and Technology Studies at Leiden University. Trajectories are constructed from author affiliations listed on the byline of publications for an author. Given the limitations of author-name disambiguation, we limit our analyses to papers published after 2008, when the Web of Science began providing full names and institutional affiliations that improved disambiguation (see Supporting Information). This yields 33,934,672 author-affiliation combinations representing 12,963,792 authors. Each author-affiliation combination is associated with the publication year and an ID linking it to one of 8,661 disambiguated organizational affiliations (see Supporting Information for more detail). Trajectories are represented as the list of author-affiliation combinations, ordered by year of publication, and randomly ordered for combinations within the same year. The most fine-grained geographic unit in this data is the organization, such as a university, research institute, business, or government agency.

Here, authors are classified as mobile when they have at least two distinct organization IDs in their trajectory, meaning that they have published using two or more distinct affiliations between 2008 and 2019. Under this definition, mobile authors constitute 3,007,192 or 23.2%
of all authors and 17,700,095 author-affiliation combinations. Mobile authors were associated
with 2.5 distinct organizational affiliations on average. Rates of mobility differ across countries.
For example, France, Qatar, the USA, Iraq, and Luxembourg had the most mobile authors
(Fig. S2c). However, due to their size, the USA, accounted for nearly 40 % of all mobile
authors worldwide (Fig. S2a), with 10 countries accounting for 80 % of all mobility (Fig. S2b).
The countries with the highest proportion of mobile scientists are France, Qatar, the United
States, and Iraq, whereas those with the lowest are Jamaica, Serbia, Bosnia & Herzegovina, and
North Macedonia (Fig. S2c). In most cases, countries with a high degree of inter-organization
mobility also have a high degree of international mobility, indicating that a high proportion of
their total mobility is international (Fig. S2d); However, some countries such as France and
the United States seem to have more domestic mobility than international mobility. While the
number of publications has increased year-to-year, the mobility and disciplinary makeup of the
dataset has not notably changed across the period of study (Fig. S1).

**Embedding**

We embed trajectories by treating them analogously to sentences and locations analogously to
words. For U.S. airport itinerary, trajectories are formed from the flight itineraries of individual
passenger, in which airports correspond to unique identifiers. In the case of Korean accom-
modation reservations, trajectories comprise a sequence of accommodations reserved over a
customer’s history. For scientific mobility, an “affiliation trajectories” is constructed for each
mobile author, which is built by concatenating together their ordered list of unique organization
identifiers, as demonstrated in Fig. 1a (top). In more complex cases, such as listing multiple
affiliations on the same paper or publishing with different affiliations on multiple publications
in the same year, the order is randomized within that year, as shown in Fig. 1a (bottom).

These trajectories are used as input to the standard skip-gram negative sampling word em-
bedding, commonly known as word2vec\(^19\). word2vec constructs dense and continuous vector representations of words and phrases, in which distance between words corresponds to a notion of semantic distance. By embedding trajectories, we aim to learn a dense vector for every location, for which the distance between vectors relates to the tendency for two locations to occur in similar contexts. Suppose a trajectory, denoted by \((a_1, a_2, \ldots, a_T)\), where \(a_t\) is the \(t\)th location in the trajectory. A location, \(a_t\), is considered to have context locations, \(a_{t-w}, \ldots, a_{t-1}, a_{t+1}, \ldots, a_{t+w}\), that appear in the window surrounding \(a_t\) up to a time lag of \(w\), where \(w\) is the window size parameter truncated at \(t-w \geq 0\) and \(t+w \leq T\). Then, the model learns probability \(p(a_{t+\tau}|a_t)\), where \(-w \leq \tau \leq w\) and \(\tau \neq 0\), by maximizing its log likelihood given by

\[
J = \frac{1}{T} \sum_{t=1}^{T} \sum_{-w \leq \tau \leq w, \tau \neq 0} \log p(a_{t+\tau}|a_t), \tag{7}
\]

where,

\[
p(j \mid i) = \frac{\exp(u_j \cdot v_i)}{Z_i}, \tag{8}
\]

where \(v\) and \(u\) are the “in-vector” and “out-vector”, respectively, \(Z_i = \sum_{j' \in \mathcal{A}} \exp(u_{j'} \cdot v_i)\) is a normalization constant, and \(\mathcal{A}\) is the set of all locations. We follow the standard practice and only use the in-vector, \(v\), which is known to be superior to the out-vector in link prediction benchmarks\(^20-25,28\).

We used the word2vec implementation in the python package gensim. The skip-gram negative sampling word2vec model has several tunable hyper-parameters, including the embedding dimension \(d\), the size of the context window \(w\), the minimum frequency threshold \(f_{\text{min}}\), initial learning rate \(\alpha\), shape of negative sampling distribution \(\gamma\), the number of the negative samples should be drawn \(k\), and the number of iterations. For main results regarding scientific mobility, we used \(d = 300\) and \(w = 1\), which were the parameters that best explained the flux between
locations, though results were robust across different settings (Fig. S4). Although the original word2vec paper uses $\gamma = 0.75^{19}$, here we set $\gamma = 1.0$, though results are only trivially different at different values of $\gamma$ (Fig. S5). We used $k = 5$, which is suggested default of word2vec. We also use same setting for U.S. airport itinerary and Korean accommodation reservation data.

To mitigate the effect of less common locations, we set $f_{\text{min}} = 50$, limiting to locations appearing at least 50 times across the training trajectories; 744 unique airport for U.S. airport itinerary, 1004 unique accommodations for Korean accommodation reservation data, and 6,580 unique organizations for scientific mobility appear in the resulting embedding. We set $\alpha$ to its default value of 0.025 and iterate five times over all training trajectories. For scientific mobility, across each training iteration, the order of organizations within a single year is randomized to remove unclear sequential order.

**An implicit bias in the negative sampling**

Negative sampling trains word2vec using a binary classification task as follows. For each target word $i$, we sample a context word $j$ from the given data and label it as positive, denoted by $Y_j = 1$. Then, we sample $k$ words $\ell$ from a noise distribution $p_0(\ell)$ and label them as negative, denoted by $Y_\ell = 0$. In word2vec, the noise distribution is given by $p_0(\ell) \propto P^\gamma(\ell)$, where $P(j)$ is the fraction of $j$ in the data, and $\gamma$ is a hyper-parameter. Then, for the sampled words, we fit a logistic regression model

$$P^{NS}(Y_j = 1; v_i, u_j) = \frac{1}{1 + \exp(-u_j \cdot v_i)}, \quad (9)$$

by maximizing the log-likelihood:

$$J^{NS} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{D}} [Y_j \log P^{NS}(Y_j = 1; v_i, u_j) + (1 - Y_j) \log P^{NS}(Y_j = 0; v_i, u_j)], \quad (10)$$

where $\mathcal{D}$ is the set of all sampled context words.
This procedure does not guarantee that the embedding optimally converges, even when increasing the training samples and iterations. To make this bias explicit, let us consider the unbiased variant of negative sampling, i.e., the noise contrastive estimation (NCE). NCE is an unbiased estimator for a probability model $P_m$ of the form:

$$P_m(x) = \frac{f(x)}{\sum_{x' \in \mathcal{X}} f(x')}.$$  \hspace{1cm} (11)

where $f$ is a non-negative likelihood function of data $x$, and $\mathcal{X}$ is the set of all data. NCE fits a logistic regression model:

$$P_{\text{NCE}}(Y_j = 1|j) = \frac{1}{1 + \exp \left[ - \ln f(u_j \cdot v_i) + \ln p_0(j) + c \right]},$$  \hspace{1cm} (12)

where $c = \ln k + \ln \sum_{x' \in \mathcal{X}} f(x')$ is a constant and maximizes the log-likelihood

$$J_{\text{NCE}} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{D}} \left[ Y_j \log P_{\text{NCE}}(Y_j = 1|j) + (1 - Y_j) \log P_{\text{NCE}}(Y_j = 0|j) \right].$$  \hspace{1cm} (13)

by calculating the gradients for embedding vectors $u_j, v_i$ and iteratively updating them (see Supporting Information for full derivation). Note that NCE is an unbiased estimator that has convergence to the optimal embedding in terms of the original word2vec’s objective function, $J_{\text{NCE}}$, if we increase the number of words to sample and the training iterations.

Let us revisit negative sampling from the perspective of NCE. We rewrite the logistic regression model in negative sampling (Eq. 9) in form of the posterior probability:

$$P_{\text{NS}}(Y_j = 1|j) = \frac{1}{1 + \exp \left[ - (u_j \cdot v_i + \ln p_0(j) + c) + \ln p_0(j) + c \right]},$$  \hspace{1cm} (14)

$$= \frac{1}{1 + \exp \left[ - \ln f(u_j \cdot v_i) + \ln p_0(j) + c \right]},$$  \hspace{1cm} (15)

where we define the likelihood function $f$ by

$$f(u_j \cdot v_i) = \exp (u_j \cdot v_i + \ln p_0(j) + c),$$  \hspace{1cm} (16)
which is the unbiased estimator for the probability model

\[
P_{m}^{NS}(u_j \cdot v_i) = \frac{f(u_j \cdot v_i)}{\sum_{j' \in A} f(u_{j'} \cdot v_i)}, \tag{17}
\]

\[
= \frac{p_0(j) \exp(u_j \cdot v_i)}{\sum_{j' \in A} p_0(j') \exp(u_{j'} \cdot v_i)}, \tag{18}
\]

\[
= \frac{P^\gamma(j) \exp(u_j \cdot v_i)}{\sum_{j' \in A} P^\gamma(j') \exp(u_{j'} \cdot v_i)} \quad (\because p_0(\ell) \propto P^\gamma(\ell)). \tag{19}
\]

Taken together, the conditional probability that SGNS word2vec actually optimizes is

\[
P^{NS}(j \mid i) = P_{m}^{NS}(u_j \cdot v_i) = \frac{P^\gamma(j) \exp(u_j \cdot v_i)}{Z'_i}, \tag{20}
\]

where \(Z'_i = \sum_{j' \in A} P^\gamma(j') \exp(u_{j'} \cdot v_i)\).

**Distance**

We calculate \(T_{ij}\) as the total number of co-occurrence between two locations \(i\) and \(j\) across the data-set. In scientific mobility, \(T_{ij} = 10\) indicates that the number of co-occurrence between both organization \(i\) and \(j\) between 2008 and 2019 is 10, as evidenced from their publications. Here, we treat \(T_{ij} = T_{ji}\) for the sake of simplicity and, in the case of scientific mobility, because directionality cannot easily be derived from bibliometric records, or may not be particularly informative (see Supporting Information).

We calculate two main forms of distance between locations. The geographic distance, \(g_{ij}\), is the pairwise geographic distance between locations. Geographic distance is calculated as the great circle distance, in kilometers, between pairs of locations. In the case of U.S. flight itinerary and scientific mobility, we impute distance to 1 km when their distance is less than one kilometer. In the case of Korean accommodation reservation data, because this data is intra-city mobility trajectory at a much smaller scale, we impute distance to 0.01 km when their distance is less than 0.01 km. The embedding distance with the cosine distance, \(d_{ij}\), is calculated as \(d_{ij} = 1 - \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|}\), where \(v_i\) and \(v_j\) are the embedding vectors for locations \(i\) and \(j\), respectively. Note
that $d_{ij}$ is not a formal metric because it does not satisfy the triangle inequality. Nevertheless, cosine distance is often shown to be useful in practice\cite{6,7,60}. We compare the performance of this cosine-based embedding distance against those derived using inner product similarity and euclidean distance.

We compare the performance of the embedding distance to many baselines. These include distances derived from simpler embedding approaches, such as Singular Value Decomposition (SVD) and a Laplacian Eigenmap embedding performed on the underlying location co-occurrence matrix. We also use network-based distances, calculating vectors using a Personalized Page Rank approach and measuring the distance between them using cosine distance and Jensen-Shannon Divergence (see Supporting Information). Finally, we compare the embedding distance against embeddings calculated through direct matrix factorization, following the approach that word2vec implicitly approximates\cite{35}.

**Gravity Law**

We model co-occurrences $T_{ij}$ for locations $i$ and $j$ (referred to as flux), using the gravity law of mobility\cite{3}. The gravity law of mobility, which was inspired by Newton’s law of gravity, postulates that attraction between two locations is a function of their population and the distance between them. This formulation and variants have proven useful for modeling and predicting many kinds of mobility\cite{38-41}. In the gravity law of mobility, the expected flux, $\hat{T}_{ij}$ between two locations $i$ and $j$ is defined as,

$$\hat{T}_{ij} = C m_i m_j f(r_{ij}), \quad \text{(21)}$$

where $m_i$ and $m_j$ are the population of locations, defined as the total number of passenger who passed through each airport for U.S. airport itineraries, the total number of customer who booked with each accommodation for Korean accommodation reservations, and the yearly-
average count of unique authors, both mobile and non-mobile, affiliated with each organization for scientific mobility. $f(r_{ij})$ is a decay function of distance $r_{ij}$ between locations $i$ and $j$.

Here, we used the most basic gravity model which assumes symmetry of the flow $\hat{T}_{ij} = \hat{T}_{ji}$ and distance $r_{ij} = r_{ji}$, while there are four proposed variants\textsuperscript{61}. There are two popular forms for the $f(r_{ij})$: one is a power law function in the form $f(r_{ij}) = r_{ij}^{-\alpha}$ ($\alpha > 0$), and the other is an exponential function in the form $f(r_{ij}) = e^{-\beta r_{ij}}$ ($\beta > 0$)\textsuperscript{43}. The parameters for $f(r_{ij})$ and $C$ are fit to given mobility data using a log-linear regression\textsuperscript{4,38–41}.

We consider separate variants of $f(r_{ij})$ for the geographic distance, $g_{ij}$, and the embedding distance, $d_{ij}$, report the best-fit model of each distance. For the geographic distance, we use the power-law function of the gravity law, $f(g_{ij}) = g_{ij}^{-\alpha}$ (Eq. 22). For the embedding distance, we use the exponential function, with $f(d_{ij}) = e^{-\beta d_{ij}}$ (Eq. 23).

\begin{equation}
\ln \frac{T_{ij}}{m_i m_j} = \ln C - \alpha \ln g_{ij},
\end{equation}

\begin{equation}
\ln \frac{T_{ij}}{m_i m_j} = \ln C - \beta d_{ij},
\end{equation}

where $T_{ij}$ is the actual flow from the data. The gravity law of mobility is sensitive to $T_{ij} = 0$, or zero movement between locations. In our dataset, non-zero flows account for only 4.2 % of all possible pairs of the 6,580 organizations for scientific mobility, while 76.4% of all possible pairs of the 744 airports for U.S. airport Itinerary and 62.5 % of all possible pairs of the 1,004 accommodations for Korean accommodation reservation data. This value is comparable to other common applications of the gravity law, such as phone calls, commuting, and migration\textsuperscript{4}. We follow standard practice and exclude zero flows from our analysis.
Element-centric clustering similarity

Element-centric clustering similarity is a similarity measure that can produce disjoint, overlapping, and hierarchically-structured clusterings. Element-centric clustering similarity captures cluster-induced relationships between elements through a cluster affiliation graph where one vertex set is the original element \( V = \{ v_1, ... v_N \} \) and the other corresponds to the cluster \( C = \{ c_1, ... c_M \} \) as a bipartite graph \( B(V \cup C, R) \). An undirected edge \( a_{i\beta} \in R \) denotes element \( v_i \) as a member of cluster \( c_\beta \). For hierarchically structured clustering, each cluster \( c_\beta \) is assigned a hierarchical level \( l_\beta \in [0, 1] \) by re-scaling the dendrogram according to the maximum path length from its roots. The weight of the cluster affiliation edge is given by the hierarchy weight function \( a_{i\beta} = e^{rl_\beta} \) with scaling parameter \( r \) which determines the relative importance of membership at different levels of hierarchy. Here, lower \( r \) focuses on the higher levels of the hierarchy, and lower levels are treated as a refinement of the higher level. Larger \( r \) emphasizes lower-level cluster structure, and higher levels of the hierarchy are seen as an aggregation of the lower-level structure. If \( r = 0 \), every cluster is treated equally.

The cluster affiliation graph is projected onto a cluster-induced element graph which is a weighted, directed graph summarizing the relationship induced by common cluster memberships. In the cluster-induced element graph, each edge between element \( v_i \) and \( v_j \) has weight \( w_{ij} = \sum_y \frac{a_{iy}a_{jy}}{\sum_x a_{ix}\sum_m a_{my}} \). Given a cluster-induces element graph with weighted matrix \( W \), the personalized PageRank vector \( p_i \) is used as membership-aware similarity between element \( i \) and other elements in the graph. Then, the element-wise similarity of an element \( v_i \) in two clusters \( A \) and \( B \) is calculated with \( S_i(A, B) = 1 - L_1(p_i^A, p_j^B) \), and the final element-centric similarity of two clustering \( A \) and \( B \) is found as the average of the element-wise similarities, \( S(A, B) = \frac{1}{N} \sum_{i=1}^{N} S_i(A, B) \).
SemAxis

SemAxis and similar studies demonstrated that “semantic axes” can be found from an embedding space by defining the “poles” and the latent semantic relationship along the semantic axis can be extracted with simple arithmetic. In the case of natural language, the poles of the axis could be “good” and “bad”, “surprising” and “unsurprising”, or “masculine” and “feminine”. We can use SemAxis to leverage the semantic properties of the embedding vectors to operationalize abstract relationships between organizations.

Let $S^+ = \{v_1^+, v_2^+ \cdots v_n^+\}$ and $S^- = \{v_1^-, v_2^- \cdots v_n^-\}$ be the set of positive and negative pole organization vectors respectively. Then, the average vectors of each set can be calculated as $V^+ = \frac{1}{n} \sum_{i=1}^{n} v_i^+$ and $V^- = \frac{1}{n} \sum_{i=1}^{n} v_i^-$. From these average vectors of each set of poles, the semantic axis is defined as $V_{\text{axis}} = V^+ - V^-$. Then, a score of organization $a$ is calculated as the cosine similarity of the organization’s vector with the axis,

$$\frac{v_a \cdot V_{\text{axis}}}{||v_a|| \cdot ||V_{\text{axis}}||}$$

where a higher score for organization $a$ indicates that $a$ is more closely aligned to $V^+$ than $V^-$. We define two axes to capture geography and academic prestige, respectively. The poles of the geographic axis are defined as the mean vector of all vectors corresponding to organizations in California, and then the mean of all vectors of organizations in Massachusetts. For the prestige axis, we define a subset of top-ranked universities according to either the Times World University Ranking or based on the mean normalized research impact sourced from the Leiden Ranking. The other end of the prestige axis is the geographically-matched (according to census region) set of universities ranked at the bottom of these rankings. For example, if 20 top-ranked universities are selected and six of them are in the Northeastern U.S., then the bottom twenty will be chosen to also include six from the Northeastern U.S. From the prestige axis, we derive a
ranking of universities that we then compare to other formal university rankings using Spearman rank correlation.

Acknowledgments

We thank the Center for Science and Technology Studies at Leiden University for managing and making available the dataset of scientific mobility. We also thank the Goodchoice Company LTD. for making available the dataset of Korean accommodation reservation data. For their comments, we thank Guillaume Cabanac, Cassidy R. Sugimoto, Vincent Larivière, Alessandro Flammini, Filippo Menczer, Lili Miao, Xiaoran Yan, Inho Hong, and Esteban Moro Egido. This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-19-1-0391. Rodrigo Costas is partially funded by the South African DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP).

Author Contributions

All authors contributed extensively to the work presented in this paper. D. M. and J.Y. were involved in all stages of conceptualization, analysis, and writing, S.K. developed the theoretical framework, R.C. assembled input data, and W.J., S.M., and Y.A. contributed to conceptualization. All authors discussed the results and commented on the manuscript at all stages.

Additional Information

Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed to Dr. Staša Milojević and Dr. Yong-Yeol Ahn. Reprints and permissions information is available at www.nature.com/reprints.
Data Availability

The U.S. airline itinerary dataset can be found at https://www.transtats.bts.gov/DataIndex.asp. The raw Korean accommodation reservation dataset, due to privacy concerns, cannot be shared publicly. Due to its proprietary nature, the global scientific mobility dataset, sourced from the Web of Science, cannot be provided; however, metadata and trained neural embeddings have been published at https://doi.org/10.6084/m9.figshare.13072790.v1

Code Availability

Code used in this analysis can be found at https://github.com/murrayds/sci-mobility-emb

References

1. Origins and Destinations of the World's Migrants, 1990-2017 tech. rep. (Pew Research Center, Washington D.C., USA, 2018).
2. Global Flow of Tertiary-Level Students tech. rep. (UNESCO Institute of Statistics, Paris, France, 2019).
3. Zipf, G. The P1 P2 / D hypothesis: On the intercity movement of persons. American Sociological Review 11, 677–686 (1946).
4. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. Nature 484, 96–100 (2012).
5. Boschma, R. Proximity and Innovation: A Critical Assessment. Regional Studies 39, 61–74 (2005).
6. Brown, L. A., Odland, J. & Golledge, R. G. Migration, Functional Distance, and the Urban Hierarchy. Economic Geography 46, 472–485 (1970).
7. Kim, J., Park, J. & Lee, W. Why do people move? Enhancing human mobility prediction using local functions based on public records and SNS data. PLOS ONE 13, e0192698 (2018).
8. Czaika, M. & Orazbayev, S. The globalisation of scientific mobility, 1970-2014. Applied Geography 96, 1–10 (2018).
9. Box, S. & Barsi, E. The Global Competition for Talent: Mobility of the Highly Skilled tech. rep. (OECD, 2008).
10. Braunerhjelm, P., Ding, D. & Thulin, P. Labour market mobility, knowledge diffusion and innovation. *European Economic Review* **123**, 103386 (2020).

11. Kaiser, U., Kongsted, H. C., Laursen, K. & Ejsing, A.-K. Experience matters: The role of academic scientist mobility for industrial innovation. *Strategic Management Journal* **39**, 1935–1958 (2018).

12. Sugimoto, C. R. *et al.* Scientists have most impact when they’re free to move. *Nature* **550**, 29–31 (2017).

13. Petersen, A. M. Multiscale impact of researcher mobility. *Journal of The Royal Society Interface* **15**, 20180580 (2018).

14. Rodrigues, M. L., Nimrichter, L. & Cordero, R. J. B. The benefits of scientific mobility and international collaboration. *FEMS Microbiology Letters* **363** (2016).

15. Morgan, A. C., Economou, D. J., Way, S. F. & Clauset, A. Prestige drives epistemic inequality in the diffusion of scientific ideas. *EPJ Data Science* **7**, 40 (2020) (2018).

16. Clauset, A., Arbesman, S. & Larremore, D. B. Systematic inequality and hierarchy in faculty hiring networks. *Science Advances* **1**, e1400005 (2015).

17. Deville, P. *et al.* Career on the Move: Geography, Stratification and Scientific Impact. *Scientific Reports* **4**, 1–7 (2014).

18. Robinson-García, N. *et al.* The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics* **13**, 50–63 (2019).

19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. *Distributed Representations of Words and Phrases and Their Compositionality in Proceedings of the 26th International Conference on Neural Information Processing Systems* **2** (Curran Associates Inc., Alghero, Italy, 2013), 3111–3119.

20. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).

21. Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* **115**, E3635–E3644 (2018).

22. Kozlowski, A. C., Taddy, M. & Evans, J. A. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review* **84**, 905–949 (2019).

23. Hamilton, W. L., Leskovec, J. & Jurafsky, D. *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics, Berlin, Germany, 2016), 1489–1501.

24. Le, Q. & Mikolov, T. *Distributed Representations of Sentences and Documents* in. **32** (Beijing, China, 2014), 1188–1196.
25. Nakandala, S., Ciampaglia, G. L., Su, N. M. & Ahn, Y.-Y. Gendered Conversation in a Social Game-Streaming Platform in Proceedings of the Eleventh International AAAI Conference on Web and Social Media (Montreal, Canada, 2017), 10.

26. Grover, A. & Leskovec, J. Node2vec: Scalable Feature Learning for Networks in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Association for Computing Machinery, San Francisco, California, USA, 2016), 855–864.

27. Perozzi, B., Al-Rfou, R. & Skiena, S. DeepWalk: Online Learning of Social Representations in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (New York, USA, 2014), 701–710.

28. Linzhuo, L., Lingfei, W. & James, E. Social centralization and semantic collapse: Hyperbolic embeddings of networks and text. Poetics 78, 101428 (2020).

29. Liu, X., Liu, Y. & Li, X. Exploring the Context of Locations for Personalized Location Recommendations. in IJCAI (2016), 1188–1194.

30. Feng, S., Cong, G., An, B. & Chee, Y. M. Poi2vec: Geographical latent representation for predicting future visitors in Proceedings of the AAAI Conference on Artificial Intelligence 31 (2017).

31. Yao, Z., Fu, Y., Liu, B., Hu, W. & Xiong, H. Representing urban functions through zone embedding with human mobility patterns in Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18) (2018).

32. Cao, H., Xu, F., Sankaranarayanan, J., Li, Y. & Samet, H. Habit2vec: Trajectory semantic embedding for living pattern recognition in population. IEEE Transactions on Mobile Computing 19, 1096–1108 (2019).

33. Crivellari, A. & Beinat, E. From motion activity to geo-embeddings: Generating and exploring vector representations of locations, traces and visitors through large-scale mobility data. ISPRS International Journal of Geo-Information 8, 134 (2019).

34. Solomon, A., Bar, A., Yanai, C., Shapira, B. & Rokach, L. Predict demographic information using word2vec on spatial trajectories in Proceedings of the 26th conference on user Modeling, adaptation and personalization (2018), 331–339.

35. Levy, O. & Goldberg, Y. Neural word embedding as implicit matrix factorization in Advances in Neural Information Processing Systems 27 (Curran Associates, Inc., Montreal, Canada, 2014), 2177–2185.

36. An, J., Kwak, H. & Ahn, Y.-Y. SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics, 2018), 2450–2461.
37. Hashimoto, T. B., Alvarez-Melis, D. & Jaakkola, T. S. Word embeddings as metric recovery in semantic spaces. *Transactions of the Association for Computational Linguistics* **4**, 273–286 (2016).

38. Curiel, R. P., Pappalardo, L., Gabrielli, L. & Bishop, S. R. Gravity and scaling laws of city to city migration. *PLOS ONE* **13**, e0199892 (2018).

39. Jung, W.-S., Wang, F. & Stanley, H. E. Gravity model in the Korean highway. *EPL (Europhysics Letters)* **81**, 48005 (2008).

40. Hong, I. & Jung, W.-S. Application of gravity model on the Korean urban bus network. *Physica A: Statistical Mechanics and its Applications* **462**, 48–55 (2016).

41. Truscott, J. & Ferguson, N. M. Evaluating the Adequacy of Gravity Models as a Description of Human Mobility for Epidemic Modelling. *PLoS Computational Biology* **8** (2012).

42. Barthélemy, M. Spatial networks. *Physics Reports* **499**, 1–101 (2011).

43. Chen, Y. The distance-decay function of geographical gravity model: Power law or exponential law? *Chaos, Solitons & Fractals* **77**, 174–189 (2015).

44. Morin, F. & Bengio, Y. Hierarchical probabilistic neural network language model. in *Aistats* **5** (2005), 246–252.

45. Gutmann, M. & Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* **9** (PMLR, Chia Laguna Resort, Sardinia, Italy, 2010), 297–304.

46. Dyer, C. *Notes on Noise Contrastive Estimation and Negative Sampling* 2014. arXiv: 1410.8251 [cs.LG].

47. Belkin, M. & Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* **15**, 1373–1396 (2003).

48. Levy, O., Goldberg, Y. & Dagan, I. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* **3**, 211–225 (2015).

49. McInnes, L., Healy, J., Saul, N. & Grossberger, L. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software* **3**, 861 (2018).

50. Chinchilla-Rodríguez, Z. et al. A Global Comparison of Scientific Mobility and Collaboration According to National Scientific Capacities. *Frontiers in Research Metrics and Analytics* **3** (2018).

51. Gates, A. J., Wood, I. B., Hetrick, W. P. & Ahn, Y.-Y. Element-centric clustering comparison unifies overlaps and hierarchy. *Scientific reports* **9**, 8574 (2019).

52. *Ethnologue Global Dataset* [https://www.ethnologue.com/]. Accessed: 2019-10-30.
53. Waltman, L. et al. The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology* **63**, 2419–2432 (2012).

54. Schakel, A. M. J. & Wilson, B. J. Measuring Word Significance using Distributed Representations of Words. *arXiv:1508.02297* (2015).

55. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, 2019), 4171–4186.

56. Kraemer, M. U. G. et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020).

57. Chinchilla-Rodríguez, Z., Bu, Y., Robinson-García, N., Costas, R. & Sugimoto, C. R. Travel bans and scientific mobility: utility of asymmetry and affinity indexes to inform science policy. *Scientometrics* **116**, 569–590 (2018).

58. Cao, C., Baas, J., Wagner, C. S. & Jonkers, K. Returning scientists and the emergence of China’s science system. *Science and Public Policy* **47**, 172–183 (2020).

59. Caron, E. & van Eck, N. J. Large scale author name disambiguation using rule-based scoring and clustering in *Proceedings of the 14th Science and Technology Indicators Conference* (Leiden University, Leiden, Netherlands, 2014), 79–86.

60. Lerman, G. & Shakhnovich, B. E. Defining functional distance using manifold embeddings of gene ontology annotations. *Proceedings of the National Academy of Sciences* **104**, 11334–11339 (2007).

61. Wilson, A. *Entropy in urban and regional modelling* (Routledge, 2011).
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 20211108mobilitiesupp.pdf