





Charting mobility patterns in the scientific knowledge landscape

Chakresh Kumar Singh^{1,2} , Liubov Tupikina^{2,3} , Fabrice Lécuyer⁴ , Michele Starnini^{5,6}  and Marc Santolini^{1,2*} 

*Correspondence:

marc.santolini@cri-paris.org

¹ Université Paris Cité, Inseam, System Engineering and Evolution Dynamics, F-75004, Paris, France

² Learning Planet Institute, F-75004, Paris, France

Full list of author information is available at the end of the article

Abstract

From small steps to great leaps, metaphors of spatial mobility abound to describe discovery processes. Here, we ground these ideas in formal terms by systematically studying mobility patterns in the scientific knowledge landscape. We use low-dimensional embedding techniques to create a knowledge space made up of 1.5 million articles from the fields of physics, computer science, and mathematics. By analyzing the publication histories of individual researchers, we discover patterns of scientific mobility that closely resemble physical mobility. In aggregate, the trajectories form mobility flows that can be described by a gravity model, with jumps more likely to occur in areas of high density and less likely to occur over longer distances. We identify two types of researchers from their individual mobility patterns: interdisciplinary *explorers* who pioneer new fields, and *exploiters* who are more likely to stay within their specific areas of expertise. Our results suggest that spatial mobility analysis is a valuable tool for understanding the evolution of science.

Keywords: Science of science; Human mobility; Social dynamics; Knowledge exploration

1 Introduction

Quantifying the evolution of scientific knowledge is crucial to understanding the past and predicting future innovations [1], which ultimately lead to societal progress. At the forefront of scientific innovation are researchers recombining ideas to push the boundaries of the known [2, 3]. With the exponential growth in the number of authors and publications [4, 5], novel methods are needed to represent and provide insights into knowledge development.

The increased access to large-scale publication data sets has provided opportunities to quantify the choices made by researchers and examine the factors governing the evolution of scientific knowledge. By studying the citation patterns of researchers in their publications, studies have measured how conflicting ideas are pursued by researchers before they converge to a common consensus [6] or give way to new ideas [7]. Other studies have focused on identifying ‘hot topics’ in research [8], quantifying knowledge flow patterns [9] and memory effects in the evolution of knowledge [10, 11], or predicting the ultimate impact of a researcher [12, 13]. Similarly, keywords and phrases from publications can be

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

leveraged to track the evolution of scientific ideas and fields [14, 15] or quantify how scientists choose and shift their research focus over time [16, 17]. For example, the Physics and Astronomy Classification Scheme (PACS) used in articles published by the American Physical Society can be exploited to study the “essential tension” between exploring the boundaries of a research area and exploiting previous work [18]. Finally, scientific credit among researchers and their mutual scientific interest (quantified by citations between papers and keywords, respectively) can be combined to improve the prediction of new scientific collaborations [19].

Therefore, studying the publication trajectories of researchers can help identify the complex processes underlying the evolution of science. Such trajectories are often talked about metaphorically, for example when referring to scientific advances as ‘great leaps’ [20]. Here, we aim to explore the parallel between scientific and human mobility more formally, by leveraging insights from human mobility studies. Using large-scale real-world data on human trajectories, previous studies have indeed uncovered several laws underlying human mobility. Despite heterogeneity in their movement, humans exhibit recurring patterns in their mobility [21, 22]. These patterns have been shown to give rise to scaling laws for the travel distance distribution [23]. At the macroscopic level, the resulting flows between two locations follow a gravity model [24], mimicking the Newtonian law of attraction between two masses at a given distance. Beyond jump distance, individuals show reproducible properties at the whole trajectory level. For example, individuals can be categorized into two classes, *returners* and *explorers*, depending on their propensity to come back to the same location or explore new ones [25]. More generally, studies on both individual and collective mobility datasets have proposed various quantitative models explaining the dynamics of human mobility [23, 26–30]. Crucially, these reproducible patterns are not unique to human mobility [31]. Multiple studies across disciplines have found striking similarities between human mobility in geographic space and animals foraging [32], insects swarms [33], and search methods in abstract environments such as memory space [34], organizational learning [35], and cyberspace [36–38]. For example, human mental search processes have been described by Lévy flights typical of a random search in an unknown environment, suggesting its use as a unifying concept to materialize the cognitive-spatial search analogy [39, 40].

In the context of scientific evolution, tools and data sources now abound for spatial representations. Natural language processing and embedding methods with metadata from publications such as citations, keywords, or abstracts, can be combined to exploit similarities between research publications and derive a low-dimensional representation of the scientific landscape. Such representations have been used to quantify the cognitive extent of ideas explored by researchers [41, 42], describe the disciplinary structure of science [43], predict future research topics for scholars [44, 45] or analyze and predict the evolution over time of the research capacity and specialization of specific geographical areas [46].

In this work, we solidify these intuitions into a quantitative framework to represent scientific knowledge and exploit metrics derived from human mobility to describe research trajectories. To this aim, we leverage the unique features of papers published on the arXiv. By utilizing embedding methods on publication metadata from arXiv pre-prints [47], we create a lower-dimensional *knowledge space* that captures the essence of the scientific landscape, by quantifying the relations between different topics and research fields. Using the derived lower dimensional embedding we then track the mobility of disambiguated

authors in this space using their publication records. We find that scientific exploration shows striking similarities with human mobility in physical space. First, we show that scientific mobility in the knowledge space follows a gravity model, with jumps more likely to occur in areas of high density and less likely to occur over longer distances. Second, we retrieve a dichotomy in exploration between interdisciplinary scientific *explorers* – more likely to disrupt and pioneer new fields – and *exploiters*, who tend to exploit a particular area of expertise, mirroring what is observed in spatial mobility between explorers and returners. Finally, we discuss the usefulness of scientific mobility analyses for the study of science and innovation, and discuss limitations and implications for future works.

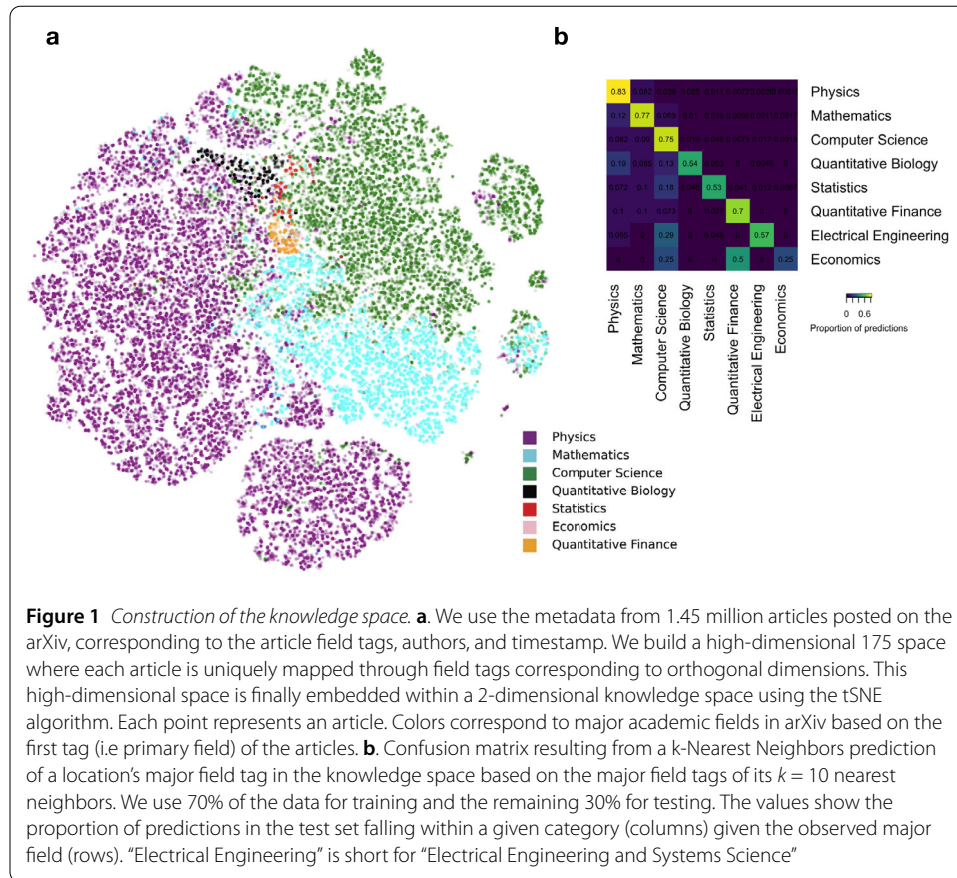
2 Results

2.1 Scientific trajectories in the arXiv knowledge space

To build the knowledge space, we leverage the arXiv dataset, encompassing 1,456,403 scientific articles published online between 1992 and 2018 (see Methods, Fig. S1 in Additional file 1 and ref. [48]). Our interest in this dataset is two-fold. First, it has a clear and stable ontology for field tags, which are used by authors to specify the relevant research area(s) covered by their articles. There is a strong incentive for authors to document these tags as precisely as possible, in order for their article to appear in the right arXiv section searched by the target scientific community, and in the relevant daily email digest that interested scientists can subscribe to. Second, as a pre-print server, it has no editorial barrier or publication cost, creating a low threshold for publication. This allows us to track the publication history of an author in a fine-grained manner, at the time they are considered finished, and irrespective of their perceived novelty. As such, arXiv pre-prints can be thought of as tracking knowledge steps to a high resolution, without requirements for novelty thresholds to be met.

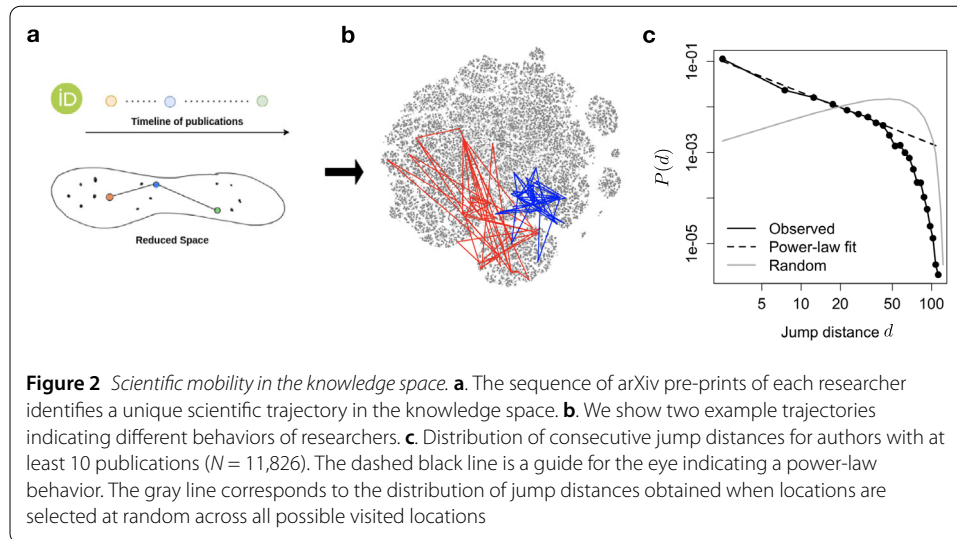
We first build a spatial representation of the knowledge space formed by arXiv pre-prints. The structure of this space is determined by the 175 tags used by submitting authors to assign scientific sub-fields to articles. Articles can be assigned with one or more tags. For instance, an article can be tagged with Social and Information Networks (cs.si) and Physics and Society (physics.soc-ph). An article can thus be represented as a binary vector $X = (0, 0, 1, 0, 1, \dots, 0)$ in the high dimensional 175 sub-fields space, with $X_i = 1$ if the article is assigned with the tag corresponding to scientific field i .

Since articles rarely combine more than a few tags (see Fig. S2 in Additional file 1), the knowledge space is sparsely populated. Moreover, some tags co-occur frequently, creating redundant information [48]. Following these observations, we reduce the dimensionality of this initial space by embedding it into a low-dimensional space via the tSNE algorithm [49, 50] (see Methods). In this study, we focus on a two-dimensional embedding to match traditional studies of human geographical mobility. In addition, we discuss the stability of the results with other embedding approaches in the Methods section. Figure 1a shows the resulting *knowledge space*, where articles are represented as points colored according to their primary (first) field tag. We observe that articles belonging to the major fields from arXiv cluster into distinctive regions of the space, with interdisciplinary fields such as Quantitative Biology (q-bio) or Quantitative Finance (q-fin) located at the interface between related disciplines. In order to quantitatively validate the observed clustering, we ran a k-Nearest Neighbors classification task on 70% of the data to measure the accuracy with which the field associated with a location is predicted by the most frequent field



among its k nearest neighbors (see Methods). We show in Fig. 1b the confusion matrix obtained on the remaining 30% of the data. Overall, we find an average prediction accuracy of 77.8%. We note that some interdisciplinary fields have fuzzier predictions. For example, Quantitative Biology is often confused for Computer Science (for bioinformatics articles) or Physics (for biophysics articles). We note that this is consistent with Quantitative Biology being located at the interface of Physics and Computer Science when investigating the tag co-occurrence network (see Fig. 5a of [48]).

The chronological sequence of articles published by an author defines a sequence of locations in the knowledge space, tracing their *scientific trajectory* (Fig. 2a-b). In order to obtain high-quality trajectories, we select a sample of 11,826 from a total of 50,402 disambiguated researchers for which we have a unique ORCID identifier, and who published at least 10 articles. Within a trajectory, two consecutive articles constitute a jump, with a length equal to the Euclidean distance computed in the embedding, and duration equal to the number of days elapsed between the two articles. If the authors were randomly jumping across all possible locations in the space, the jump distribution would follow a bounded distribution around a typical, large step size (Fig. 2c, gray line), according to a pure diffusive process. Instead, Fig. 2c shows that the jump distance distribution is compatible with a power-law functional form, with a cut-off at large distances due to the finite size of the space, differing significantly from a diffusive process. Importantly, we observe that this feature is robust with respect to different embedding techniques (Fig. S4 in Additional file 1). This indicates that, while the majority of jumps are small, with researchers orbiting

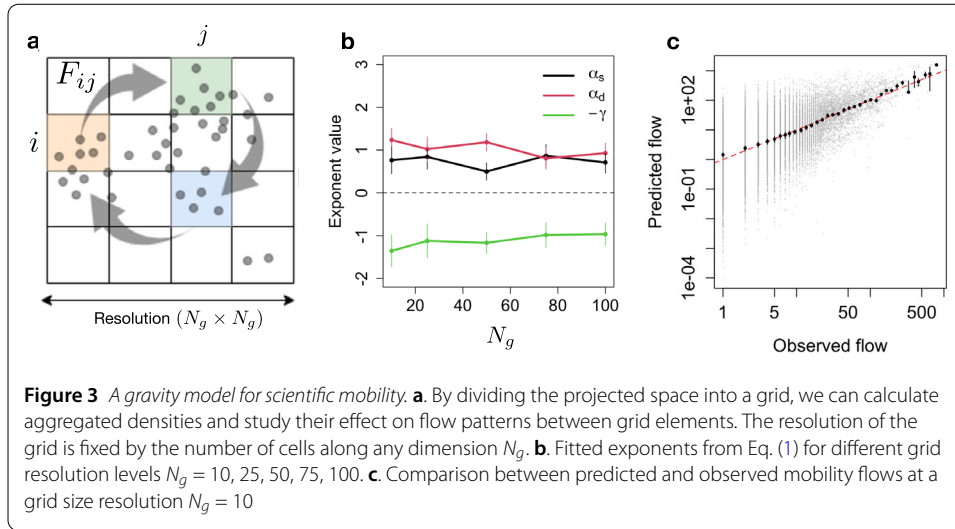


relatively close to a particular research interest, a small fraction of jumps extend far into the knowledge space, standing for researchers crossing fields. In the next section, we investigate whether simple models of human mobility can be compatible with the observed behavior.

2.2 A gravity model of scientific mobility

The observed heavy-tail form (with a cutoff) of the jump size distribution is reminiscent of the inverse relation with distance observed in human mobility flows between two locations. This observation led to a simple and intuitive model in human mobility studies, the gravitation model, where the flux F_{ij} between two locations i and j is proportional to the population sizes at i and j and inversely proportional to the distance d_{ij} between them. Earlier works on spatial distribution models and urban modelling [51–53] have shown that such a model can be functionally derived from statistical mechanics insights and empirical laws such as Zipf’s law [54]. When considering population-scale mobility in an origin-destination setting such as ours, the gravity model naturally emerges as the expectation of the distribution maximizing the entropy of mobility between two locations [30].

Much like the urban vs rural landscape, where populations conglomerate into a few, dense regions corresponding to urban areas, there are denser regions in the low-dimensional knowledge space, corresponding to more investigated areas. However, unlike cities and administrative areas, we do not have a clear definition of boundaries in the knowledge space. Here, we use a simple box/container model by defining a grid of size $N_g \times N_g$ covering the knowledge space, where N_g is a parameter quantifying the resolution level, and population counts are aggregated at the grid level (Fig. 3a). The grid resolution influences the type of behavior we observe: while large grid cells mostly capture long jumps and topic-switching behavior, small grid cells will capture more local jumps to nearby topics. An overly fine grid will end up capturing jumps between very close cells in a manner reminiscent of the ping-pong observed in mobile phone data [55]. To mitigate this ping-pong effect, we need to assess what we consider the typical size of a region with the same topic. For this, we assume that two consecutive articles of an author fall within the same topic if they share at least one field tag. We show in Fig. S5b in Additional file 1



that the probability to jump to the same topic decreases non-linearly as a function of jump distance, following a slow decrease at very small distances down to a sharp decline at a distance of ~ 5 , until a distance of 20 where the probability saturates. We consider this early fall at distance 5 to represent the characteristic spatial scale of a topic. We chose grid sizes N_g that surround this threshold, while still being located at the upper end of the “local” region of similar topic, where the probability to share a topic is close to 1. As such we can expect that the jumps between grid cells that are measured consist of *bona fide* exploration, rather than mere artificial ping-pong.

We then define a gravity model to predict the observed flow F_{ij} between two grid locations i and j in the knowledge space, defined as the number of scientists jumping from grid location i to location j , by using a rolling time window of 5 years:

$$\tilde{F}_{ij} = G \frac{V_i^{\alpha_s} V_j^{\alpha_d}}{d_{ij}^{\gamma}}, \quad (1)$$

where \tilde{F}_{ij} is the predicted flow between locations i and j , G is a normalization constant, d_{ij} is the distance between locations i and j , and V_i and V_j (visits) are the numbers of authors who have published an article in locations i and j during the 5 previous years. The exponents α_s , α_d , and γ introduce non-linear scalings, such as crowding effects for the number of visits, where higher densities lead to sublinear ($\alpha < 1$) or superlinear ($\alpha > 1$) increase in flow.

Figure 3b shows the values of the exponents obtained by fitting Eq. (1) to the empirical flows F_{ij} at different resolution levels N_g (see Methods). We find a remarkable stability across grid sizes, with $\alpha_s = 0.74 \pm 0.15$, $\alpha_d = 1.04 \pm 0.18$, and $\gamma = 1.12 \pm 0.16$ close to 1. We also note a trend where γ decreases with increasing resolution, though within error bars. This effect suggests a lower cost of distance for local jumps, creating a “flatter” distance distribution when considering a higher resolution. The quality of fit is shown in Fig. 3c, comparing predicted flows with observed flows, with a Pearson correlation value of $r = 0.58$, indicating that the model explains $r^2 = 33.6\%$ of the variance of the mobility flows in the knowledge space. We note that the observed correlation is larger than the ones observed for real-world mobility (with r between 0.03 – 0.49, see [26]). Finally, we find

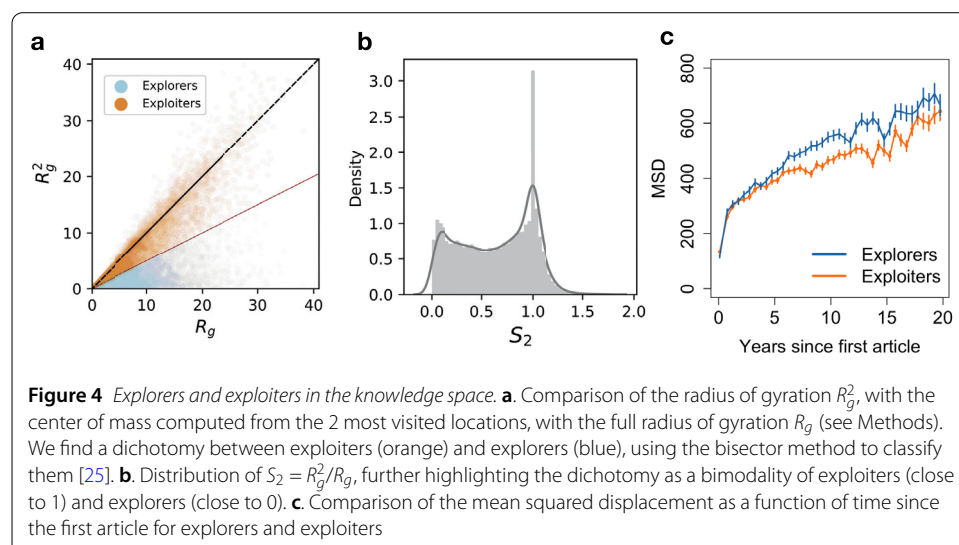
that beyond tSNE, the gravity model is able to represent flows of scientific mobility for different embedding techniques, with qualitatively similar exponents (see Figs. S7, S8 in Additional file 1).

2.3 Scientific explorers vs exploiters

When jumping to their next article, researchers can move to a novel region of the space, or return back a previous one. That is, in our framework researchers choose between exploring a new scientific field or exploiting the previous knowledge they built. While such behaviors can lead to similar jump distribution patterns, they will impact more general statistics about the full trajectory, such as the extent of spatial territory covered. Previous studies have uncovered such a heterogeneity between individual trajectories in human mobility patterns, highlighting a dichotomy between returners, who gravitate around a small number of locations, and explorers, who rather move to new locations. These results have been found to hold both for spatial [25], as well as virtual [23] contexts. Here, we explore whether such a heterogeneity exists in the context of scientific exploration.

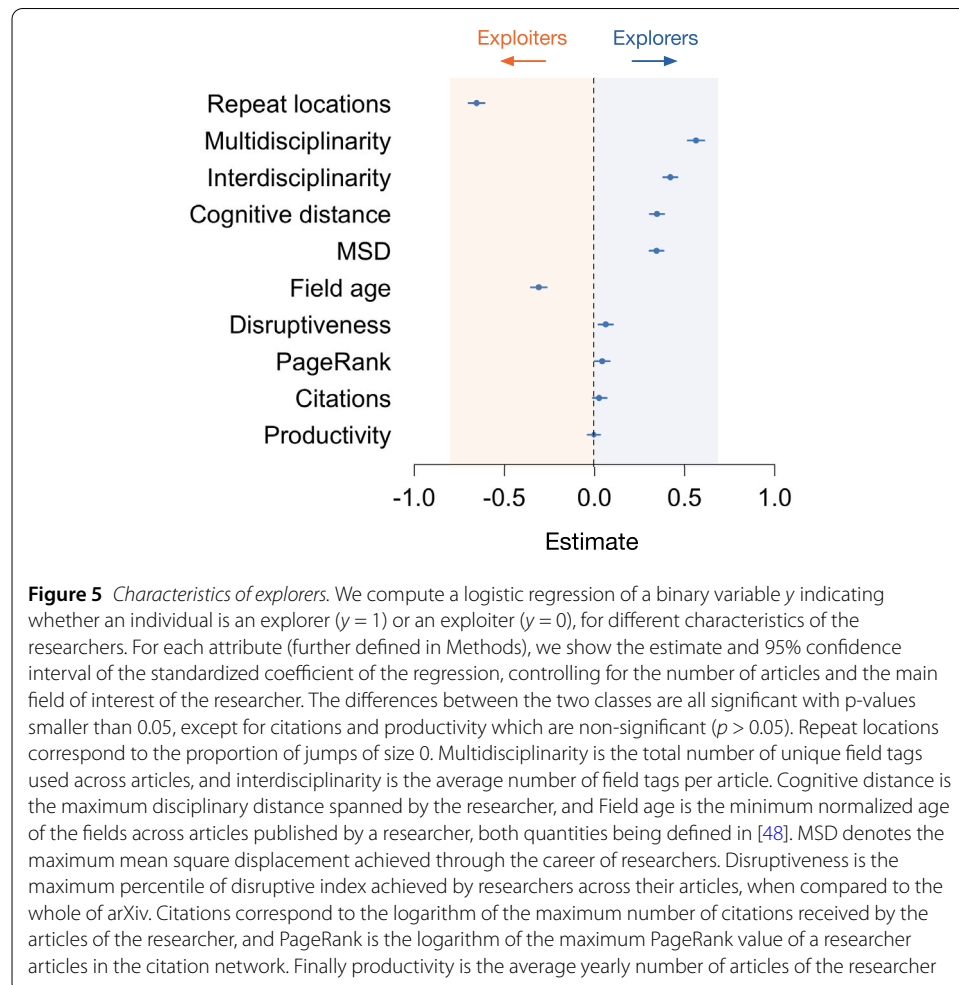
To assess the extent of territory covered by a trajectory, we study the radius of gyration R_g , defined as the average distance of visited locations to their center of mass (see Eq. (3) in Methods). By limiting to the top k most visited locations, one can define the corresponding radius of gyration R_g^k (Eq. (4)) and compare it with the full R_g to evaluate the extent to which the trajectory returns to a few (k) locations. Figure 4a shows the comparison of the total radius of gyration R_g and R_g^2 across researchers. We find that researchers can be roughly grouped into two main classes: *exploiters*, whose R_g^2 value is comparable to R_g (points along the diagonal), and *explorers* whose R_g^2 is considerably smaller than total R_g (points closer to the x-axis). The two classes are more evident when considering the distribution of $S_2 = R_g^2/R_g$, showing two peaks corresponding to the two populations of explorers and exploiters (Fig. 4b). This bimodality disappears when considering larger values of k (Fig. S9 in Additional file 1), supporting the use of $k = 2$ to distinguish the two classes.

By design, the difference between explorers and exploiters will affect the research space spanned by scientific trajectories over time. While the radius of gyration considers the gravitation of a researcher around a particular center of attraction (the center of gravity),



other measures focus on the dynamics of departure from an original starting. In mobility analysis, this is typically quantified by the mean squared displacement (MSD) [56], a quantity that tracks the average distance traveled from the starting location over time (see Methods). The particular interest in MSD stems from the fact that simple diffusion processes in homogeneous spaces observe a functional scaling with time, $\text{MSD}(t) \sim t^\beta$, with the exponent β indicating a super- or sub-diffusive process. In our case, we find that, while both classes make jumps of similar size and duration (Fig. S10 in Additional file 1) and have a similar sublinear MSD growth, explorers span a larger fraction of the knowledge space early in their career, as indicated by a faster MSD growth between 5 and 15 years ($p < 0.01$, Fig. 4c and Fig. S11 in Additional file 1). This difference decreases in the later phase of their career (around 20 years), indicating that researchers tend to explore mostly in the middle of their academic life, while senior scientists tend to exploit more their previous research. This finding supports prior observations that scientists become less disruptive and more critical of emerging work as they age [57].

Lastly, beyond differences in mobility patterns, we ask whether there are other characteristics that distinguish exploiters and explorers. To answer this question, we perform a logistic regression to predict if an individual researcher is an explorer as a function of several attributes. To account for different trajectory lengths and field-specific behavior, we



control for the total number of articles published and the area of interest (Fig. 1) in which the author has published the most. Figure 5 shows that, in line with the previous observation and consistent with their definition, explorers navigate broader regions of space, as measured by their maximum MSD achieved throughout their career, while exploiters tend to remain at the same location, measured by the proportion of their jumps being of distance 0. Moreover, explorers cover more disciplines both within and across articles, and these disciplines tend to be cognitively distant, i.e. they are far in the field tag co-occurrence network [48]. When considering the association with specific developmental stages of scientific fields [48], we find that explorers publish in the earlier stages of a field's evolution, a marker of pioneering activity and innovative work. Supporting this observation, we find a higher disruptiveness for explorers ($p = 9.9\text{e-}4$), a quantitative marker of innovative works quantifying the extent to which articles citing an article of interest also cite its sources (low disruptiveness) or not (high disruptiveness) [58], measured here by the percentile of their most disruptive article. Finally, we observe that while explorers and exploiters show similar productivity and local impact, measured by the maximum citations obtained in one of their articles ($p = 0.16$), explorers are more central in the global citation network, as measured by the maximum PageRank centrality across their articles ($p = 0.026$). This might reflect their ability to tie down several disciplines together and thereby access a wider span of citation pools. We note however that results for the citation-based metrics are to be taken with care, as the citation network is only considering within-arXiv citations [47], and is therefore incomplete and subject to field-specific habits.

3 Discussion

In this study, we show that methods from mobility analysis applied to a low-dimensional representation of a knowledge space can help understand the scientific mobility of researchers. Using data from 1.5M articles from the pre-print repository arXiv across 30 years, we find that the mobility patterns of researchers resemble those found in human mobility studies. Flows between different regions of the knowledge space follow a gravity model, with an inverse relation to distance. This result is not an artefact from a particular representation, as it holds across various embedding parameters and methods (see Figs. S4, S7, S8 in Additional file 1). Furthermore, the model accuracy outperforms empirical results from human mobility studies [26], showing that despite its simplicity, this model is a promising foundation to build on future work. In addition, by analyzing individual trajectories, we find that researchers can be categorized into *exploiters*, whose trajectories are bound to a particular area of the knowledge space, and *explorers*, who jump across boundaries and pioneer novel fields. This dichotomy is reminiscent of the “essential tension” between tradition and innovation in scientific research, where the desire to explore new promising areas is counterbalanced by the need to capitalize on the work done in the past [18, 59].

Here we identify this tension by uncovering two types of scientific mobility patterns through the bimodality observed in R_g^2/R_g . We found that the peak of R_g^2/R_g signalling the presence of explorers disappears with $k > 2$, whereas in mobility this effect happens after $k = 8$ [25]. We note however that the dichotomy observed in the original paper by Pappalardo et al. [25] disappears after $k = 2$ when using GPS data, which is limited to private vehicles and only consists of a subset of all individual trips. In our case, arXiv

represents a subset of the entire knowledge space consisting of 3 main fields (Computer Science, Mathematics, and Physics). The fact that we find a lower k value therefore seems compatible with this restriction, and the comparison with the real-world mobility data (restricted GPS vs high-resolution GSM) suggests that having a wider research landscape, for example leveraging Web of Science data, would allow future research to reach and investigate higher values of k .

When considering the properties of scientific trajectories in the knowledge space, we observe that the mobility patterns of explorers and exploiters show sub-diffusive regimes. Theoretically, when considering the mobility of an individual in a homogeneous space, such as the initial hypercube or a regular lattice, the MSD follows a linear regime if the second moment of the step size distribution and the first moment of the waiting-time probability distribution are finite [56]. In our case, the observed deviation from a linear MSD (Fig. 4) may be due to the heavy-tailed waiting time probability distribution of the two groups of researchers [56] (Fig. S10 in Additional file 1). Another possibility is that the multi-scale nature of the knowledge space, as a complex and evolving cognitive construct, may be responsible for this trapped-like behavior. Further investigation is needed to determine which of these approaches is more suitable for explaining the observed non-linearity in MSD behavior.

A limitation of our analysis is that a researcher might have several active topics on which they are working in parallel, and would therefore need to be located not just in one location, but in several locations “at the same time”. This resembles the real-world mobility case where an individual would have a primary and a secondary home, and would frequently visit both, in contrast to moving from one location to the next. Here, we focus on publications as a marker of the position of an author. This could result in a “ping-pong” effect [55] where the author would frequently oscillate between two or more locations in a manner dictated by the chronological ordering of the publications in each domain. Here we have shown that such an effect is limited: the characteristic timescale of topic-switching is 1.5 years, significantly larger than the inter-article time of 0.53 years (Fig S5a). Still, in 20% of cases, consecutive articles do not share a field tag, which might be indicative of a switch of research interest, or of an active work in different topics. In the future, it would be interesting to study the heterogeneity of mechanisms behind the exploration behavior.

We assumed stable categories of exploiters and explorers using the full trajectory of researchers, yet there can be variation throughout their career. For example, we observed aging patterns within trajectories, with MSD of explorers and exploiters showing similar values within the first 5 years after their first publication, after which MSD values for explorers are significantly larger (Fig. 4c, Fig. S11 in Additional file 1). This could indicate that explorers go through two phases: a first phase where they are staying within a few most visited locations, followed by an exploratory behaviour towards other locations. Such a behavior could be formally captured using the concept of “intermittent behavior” from stochastic processes [60]. Future work could investigate such temporal patterns across research trajectories, for example by using time windows or the convex hull method to analyse dynamic profiles at a finer scale, and assess whether phases might correspond to institutional constraints, with some environment fostering the individual development towards more exploratory patterns.

Our framework relies on an embedding method to define the knowledge space. There is no ground truth in the use of embedding methods, and each can bias results towards

specific idiosyncratic properties. In our case we focused on tSNE for our analyses. The adoption of the tSNE method has been particularly central to the field of cellular biology for representing temporal or developmental trajectories of cell differentiation, a question of a similar nature as ours. The use of UMAP has also been growing among biologists as a popular alternative for a better preservation of the global structure of the data. However, Kobak and Linderman highlighted that the t-SNE and UMAP methods have comparable performance for preserving pairwise distances and reproducing the large-scale structures of the initial space when choosing appropriate initializations [61]. In particular, they showed that, when using informative initialisation methods for the algorithms, tSNE and UMAP reach comparatively high accuracies in preserving pairwise distance (Pearson's $r \simeq 0.5$) and large-scale structures ($r \simeq 0.9$). This motivated our choice of a PCA initialisation for the tSNE method to produce a reliable embedding.

In addition, the tSNE representation in low dimension has been shown to be of relevance to represent a knowledge space. In a recent study leveraging a 2D tSNE embedding of 21 million Pubmed articles [62], the accuracy of a k -Nearest-Neighbors prediction in the 2D tSNE embedding (63%) was found to be very close to the one obtained with a 768-dimensional BERT embedding (69.7%), and a 4,679,130-dimensional TF-IDF method (65.2%). This finding reflects the high information preservation of 2D tSNE embeddings in the context of representing scientific landscapes and the good performance of tSNE in comparison to alternative and higher-dimensional methods. Finally, we have shown that a variety of methods (tSNE, UMAP, PaCMap) yield qualitatively similar results over a range of hyperparameters, both in terms of long-tailed jump distribution (Fig. S4 in Additional file 1) and gravity model fit (Figs S7-S8). This indicates that, despite some variations coming from the structure of the space itself, the general mobility patterns uncovered here are not space representation artifacts.

In our study we make use of the human-annotated tags to position the articles within a high-dimensional space, thereby relying on the intrinsic clarity these tags provide. In the future, implementing dense and continuous vector-space representations for the entities, such as neural embeddings could enhance the representation of scientific articles in a multidimensional space, offering a complementary method to our current approach in contexts lacking such metadata. Research, including work with SciBert and BERT [63], has indeed demonstrated the utility of these embeddings in classifying and identifying similarities. In addition, the role of unsupervised pre-training in natural language processing, especially for texts that are unclassified or not tied to specific categories, has become increasingly significant. By using text such as abstracts and titles for the computation of the embedding, these methods could not only offer future utility for datasets lacking tags, but also allow researchers to disambiguate trajectories that have limited self-reported information. We note however some limitations. While advanced models like deep learning and transformer-based ones can decipher complex, non-linear data relationships, they present interpretational challenges [64], leading to a reduction in the explainability that our method using annotated tags offers. Moreover, while potentially gaining in predictive power, such embeddings generally reduce very high dimensional data (eg, text, images, etc) to not-so-low-dimensional spaces (dozens of dimensions), which would be of less interest for the comparison with mobility data that is our focus here.

Our work is focused on a dataset of arXiv pre-prints. This dataset provides a high precision for the identification of subfields, which is useful for both the construction of the

knowledge space and the computation of features such as field age [48]. Yet, it is limited in overall size, fields covered, and incompleteness of the citation network [47]. For example, while about 64% of all arXiv e-prints are published in a Web of Science (WoS)-indexed journal, only 3.6% of 2010 WoS papers (all disciplines combined) were submitted to arXiv, a proportion rising to 20% when considering Physics and Mathematics [65]. It is therefore yet unclear how our findings generalize to other disciplines, for example when considering the humanities or social sciences. Future work should explore the reproducibility of our findings across larger and diverse datasets, leveraging other field identification methods, such as the ones using Natural Language Processing [63].

When comparing explorers and exploiters in terms of research impact, we found that while having similar levels of productivity, explorers produce more disruptive work and enjoy a more influential position in the citation network, measured by the PageRank centrality. On the one hand, this result is consistent with the observation that the diversity in research interests is associated with increased impact, while showing no association with productivity [66]. However, the small effect size we obtained on the local measure of impact in the citation network (citation count) might reflect the fact that, while interdisciplinary researchers achieve relatively lower impact with their publications in the short run, they eventually outperform their specialized counterparts in long-term success, measured by funding performance [67]. This observation is consistent with our finding that explorers produce more disruptive works and are more likely to contribute to innovative fields, both of which might be more important selection criteria for successfully securing funding than short term citation counts. In addition, when considering PageRank centrality as a measure of “long-term” impact in the citation network, i.e. the degree to which an article is cited by other articles that are themselves central in the citation network, we found that explorers have significantly more influential papers. This supports the idea of a more long-term impact for exploratory behavior. We note however that the citation network in arXiv is very sparse [47] and does not provide a comprehensive view of the actual citation network, which would require access to a more comprehensive dataset, for example from the Web of Science.

While our study is focused on the description of individual trajectories, most articles are team-authored [68], and chaperoning patterns are fundamental to scientific careers [69]. Therefore, future studies could study the couplings between individual trajectories, leading to correlated patterns and ultimately collective flows. In addition, the gravity model could be extended to incorporate variables corresponding to local attributes, such as impact (e.g. through citations), field age, devoted funding, etc. These features might act as biasing forces shaping collective flows towards certain areas of the knowledge space. On a macroscopic level, these fields can affect mobility, in the same way that force fields affect the trajectories of particles in physics. Novel methods based on deep learning, such as a Deep Gravity Model [26], coupled with more extensive data on citations and funding, could help extend our work. Overall, the insights gained from leveraging a mobility analysis in the knowledge space could help study the effect of policies on scientific exploration and exploitation, with applications for funding agencies and more generally the evaluation of research.

4 Methods

4.1 Overview of the arXiv dataset

In our study we use a previously published dataset consisting of article metadata from the arXiv preprint repository [48].

The dataset consists of 1,456,403 articles published between 1992 and 2018, covering mainly the fields of physics, mathematics and computer science, and to a lesser extent Quantitative Biology, Statistics, Finance, Economy, and Engineering. We note in particular the important rise of Computer Science articles in the past decade, with the number of articles published bound to soon outweigh those from the physics field, which was at the core of the early arXiv usage (Fig. S1 in Additional file 1).

When uploading an article, the submitting author selects a main, primary tag identifying the core discipline, along with secondary tags if needed. In most cases, arXiv require authors who are submitting papers to a subject category for the first time to get an endorsement from an established arXiv author, as a quality control mechanism. The tags span 175 predefined subfields, such as Quantum Algebra (math.QA) or Signal Processing (eess.SP), all indicated on the website's main page. There is a strong incentive for authors to select appropriate fields, as arXiv proposes a subscription service to a daily digest email system to automatically receive novel submitted articles containing a specific field tag. As such, the tag system is directly tied to a relevant audience for the publishing individual, incentivizing for an accurate self-report.

The tags have been mostly stable across the existence of the arXiv, but some tags have been discontinued in its early years, as the classification system evolved into broader categories and subcategories (see Fig. 1 of [48]). We find 14 such discontinued tags: “adap-org”, “alg-geom”, “chao-dyn”, “cmp-lg”, “comp-gas”, “solv-int”, “dg-ga”, “patt-sol”, “funct-an”, “q-alg”, “bayes-an”, “acc-phys”, “ao-sci” and “mtrl-th”. These tags constitute a minor portion of our dataset, with 7762 articles featuring at least one of them, accounting for 0.53% of the total articles analyzed. The impacted articles always contain at least one other tag that has not been discontinued. We therefore expect these changes to have a negligible impact on our primary findings.

4.2 Low dimensional embeddings

To reduce the dimensionality of the initial 175-dimensional field space, we use the tSNE algorithm, an unsupervised, parametric dimensionality reduction technique that retains the local data structure in the latent space [49, 50]. The tSNE method captures much of the local structure of the high-dimensional data, while also revealing global structure such as the presence of clusters at different scales. The visualisation of the resulting embedding of the arXiv knowledge space into a two dimensional space is shown in Fig. 1. Each point corresponds to one of the 49,575 observed combinations of field tags within arXiv articles. We note that permutations of tags map to the same point, so that our analysis does not depend on the order of tags.

For the implementation of the tSNE algorithm we use the `scikit-learn` package in Python [70]. The dimension of the embedded space is set to 2. The main parameters of the embedding method, such as learning rate, number of iterations and early exaggeration parameters are set to default values. The tSNE initialisation is set to PCA (Principal Component Analysis), a setting known to drastically improve the recovery of both local and global structures for tSNE [61]. In order to test the robustness of the tSNE embedding to

varying parameters, we generated tSNE mappings for different perplexity levels p , a parameter related to the effective number of neighbors considered during the algorithm's optimization process, and learning rate parameters LR of the algorithm, and plotted the pairwise distance distribution between randomly sampled points across different settings (Fig. S3 in Additional file 1). We find remarkable stability across various parameters of the tSNE suggested in [71], including perplexity levels, indicating that choosing different tSNE parameters would not strongly affect the results.

Finally, in order to quantitatively assess the information loss after the dimensionality reduction of the initial space, we ran a k -Nearest Neighbor algorithm with $k = 50$ neighbors to compute the accuracy of the 2D tSNE embedding in preserving the local neighborhoods of the original space. We find a high average recall (overlap of nearest neighbors) of 66%. This compares to a value of 6.2% obtained in a recent study exploring the 2D tSNE embedding of 21 million Pubmed articles using $k = 10$ neighbors [62]. The high preservation of neighbors we observe might be due to the quality of self-reported tags which provide a solid ground on which to assess the proximity between two articles.

4.3 k-nearest neighbor validation of the clustering

To quantitatively validate the clustering of scientific fields observed in the knowledge space, we ran a k -Nearest Neighbors (kNN) classification task to measure the accuracy with which the main field associated with a location is predicted by the majority of main fields associated with its k neighbors. To associate a location to a main field, we used the major arXiv field associated with the first tag in the article. The association between sub-fields and major fields can be found on the arXiv home page. We then used the `knn` function from the R `class` package, with $k = 10$. We built the predictor using 70% of our data for training, and computed the confusion matrix on the remaining 30% for testing. Finally, we computed the prediction accuracy as the proportion of locations in the test set that have been assigned the correct major field by the predictor.

4.4 Robustness with respect to the embedding method

In order to assess the robustness of our results, we tested the impact of different embedding methods, parameters, as well as subsamples of the data on the jump distribution (Fig. S4 in Additional file 1). Beyond tSNE, we evaluated the robustness of our analysis using PaCMAP and UMAP embeddings.

The Uniform Manifold Approximation and Projection (UMAP) [72] has its theoretical foundations in manifold theory and topological data analysis. At a macroscopic level, UMAP uses local manifold approximations and fuzzy simplicial sets to construct topological representations of data in high and low dimensions. It then minimizes the cross-entropy between the two topological representations to find an optimal lower-dimensional representation. UMAP can also be understood as a k -neighbour based graph learning algorithm that finds the best representation of weighted graphs in lower dimensions.

The Pairwise Controlled Manifold Approximation (PaCMAP) algorithm [71] is also a graph-based technique that identifies three sets of pairs namely – neighbor pairs, mid-near pairs and further pairs. It then systematically optimizes its loss function using a custom gradient descent algorithm to find a lower dimensional representation that preserves both local and global structures.

We find that the jump distance distributions shows a similar long-tail decay for both methods (Fig. S4a in Additional file 1). In addition, we find that the gravity model has stable results in UMAP (Fig. S7 in Additional file 1) and PaCMAP (Fig. S8 in Additional file 1) contexts, though we find an overall smaller exponent for the distance, closer to $\gamma \simeq 0.5$.

4.5 Fitting procedure for the gravity model

In order to fit the gravity model, we used a linear regression of log-transformed variables. For each resolution level, we first computed for each year starting in 1997 the number of jumps between a source cell i and a target cell j (with $i \neq j$) and number of articles published in each cell in the 5 preceding years. The resolution of 1 year is a sweet spot between the typical inter-article time of a researcher (0.53 years) and the characteristic time at which researchers switch major topics (1.5 years, see Fig. S5a in Additional file 1), ensuring that the movement observed is not dominated by a “ping-pong” [55] between several topics the research works in parallel, but rather *bona fide* scientific trajectories. In order to account for low sample size, we used a pseudo-count of 1 added to raw visit counts: $V_i \leftarrow V_i + 1$. We then computed the natural logarithm of all quantities, and used these log-transformed values for the regression analysis. Since there is a much larger number of small flow values compare with large flow values (Fig. 3), we used a binning technique to avoid overfitting our model to low flow values. For this, we cut the obtained log-flow values into 100 bins containing an equal number of points, and merged bins with the same breakpoints, resulting in 42 final bins. We then averaged the log-transformed features (flow, visits, distance) within these bins, and used these average values to fit the gravity model, using the `lm` function from R 4.2.2. Residuals of the model are shown in Fig. S6 in Additional file 1 and are normally distributed.

4.6 Radius of gyration

The gyration radius has originally been used as a measure in Physics and Engineering to describe the distribution of mass in an object or the spatial extent of a system of particles. Adapted to the context of mobility, the gyration radius of a trajectory is computed as follows [25]. We first define the center of mass R_{cm} of the trajectory across unique locations $i = 0, 1, \dots, n$:

$$R_{cm} = \frac{\sum_{i=1}^n M_i R_i}{\sum_{i=1}^n M_i}, \quad (2)$$

where M_i is the frequency of visitation of each location i , i.e. the number of times location i is visited by the individual, and R_i is the radius vector characterising the location in the knowledge space with respect to the chosen center of coordinates. The radius of gyration is then defined as the characteristic distance from the center of mass:

$$R_g = \sqrt{\frac{\sum_{i=1}^n M_i (R_i - R_{cm})^2}{\sum_{i=1}^n M_i}}. \quad (3)$$

In order to estimate the influence of a few locations over the trajectory, we define the k -th radius of gyration by considering only the top k most visited location by an author:

$$R_g^k = \sqrt{\frac{\sum_{i=1}^k M_i (R_i - R_{cm}^k)^2}{\sum_{i=1}^k M_i}}, \quad (4)$$

where R_{cm}^k is the center of mass using the top k most visited locations of an author.

4.7 Mean squared displacement

The mean squared displacement (MSD) at time t for a trajectory is defined as the deviation of the position of a walker (in our case, an author) with respect to a reference position over time:

$$\text{MSD}(t) = \langle |x(t) - x(0)|^2 \rangle, \quad (5)$$

where $x(t)$ stands for the position of researcher at time t since the first article, and $x(0)$ stands for the starting point of the trajectory.

4.8 Logistic regression for explorers vs. exploiters

In order to explore the characteristics associated with explorers in Fig. 5, we compute a logistic regression with dependent variable y_i , a binary variable indicating whether an individual i is an explorer ($y_i = 1$) or an exploiter ($y_i = 0$), and various individual features x_i . We control for the main field F_i in which the author has published (given by the most represented field tag across their articles), as well as the number of articles N_i of the researcher. The fields were encoded as factors. We used the `glm` function in R to fit the model $y_i \sim x_i + N_i + F_i$, with parameter family = `binomial` set to a logistic regression. Regression summaries were obtained using the `summ` function from the `jtools` package in R, with parameters `scale=T` to standardize the regression coefficients by scaling and mean-centering input data, and `confint=T` to obtain 95% confidence intervals.

4.9 Innovation, disruptiveness and impact

To measure the innovative level of a work, we used two methods. First, we computed for each article how “early” (in terms of developmental stage) it occurs within the fields that it mentions. To do so, we computed the minimum *rescaled time* (RT) across its associated field tags, using the method described in [48]. The rescaled time is a normalized quantity that allows us to associate an article to a developmental stage of a field (early, peak, or late phase) even when fields have drastically different rise and fall durations. Denoting by t the time since the first article was published in the field, we fitted the share of articles in the field $\hat{G}(t)$ to the equation:

$$G(t) = \frac{1}{\beta} e^{\frac{-(t-\alpha)}{\beta}} e^{-e^{\frac{-(t-\alpha)}{\beta}}} \quad (6)$$

using a least-squares optimization fitting procedure. After obtaining the location α and scale β parameters from the fitting procedure, we computed for each field the re-scaled

time:

$$t' = \frac{t - \alpha}{\beta}. \quad (7)$$

We finally computed for each researcher the minimum RT value achieved across their articles, defining the “Field age” feature.

Second, we used another independent method to assess the innovative potential of the articles. This method leverages how disruptive an article is by comparing the attention it receives compared to the articles it cites. More precisely, the Disruptive Index (DI) measures how much the citing articles following up a certain work will cite the given work and not its sources, informing on the fact that this work has made the previous works “obsolete”. It has been used as a measure of innovation using patent data [73] and used in the context of academic journals [58]. Citation data was obtained from [47]. The disruptive index was then computed using the method from [74] for each article. Given a paper p followed by N_s subsequent works, if N_i papers out of the N_s cite only p , N_j out of N_s cite both p and its references and N_k out of N_s cite only its references, then the Disruptive Index (DI) of the paper p is defined by:

$$DI_p = \frac{N_i - N_j}{N_s}, \quad (8)$$

where $N_s = N_i + N_j + N_k$ [73, 74]. For each author, we computed the maximum DI across their articles. Finally, we computed the percentile of the obtained value across articles to compute the disruptiveness of an author.

To measure the impact of a work, we used two methods. First, for each author i , we computed the maximum number of citations \hat{c}_i received by any of their articles across their career. Since citation counts are distributed with a heavy-tailed function, we used the transformation $\log(\hat{c}_i + 1)$ to quantify the impact. In addition, we further quantified the importance of works citing another work by using a more global network measure: the PageRank algorithm. It quantifies not just the local importance of a work, but its ability to integrate the core of the citation network. We used in our case the logarithm of the maximum Pagerank value across an author’s articles.

4.10 Cognitive distance

We observe in Fig. 5 that compared to exploiters, explorers use a larger number of field tags per article, as well as a larger number of unique tags across their articles. However some tags might be more closely related than others in terms of research area, which the simple measure for linear estimate of tags used does not differentiate. To account for this effect, we use the network based cognitive distance measure from [48], where the cognitive distance C_{ij} between field tags i and j is the weighted distance along the shortest path between tags i and j in the tag co-occurrence network.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1140/epjds/s13688-024-00451-8>.

Additional file 1. Supplementary figures (PDF 8.8 MB)

Acknowledgements

We thank Emma Barme for early discussions on this topic, as well as Ariel Lindner and Robert Ward for advice on relevant literature.

Funding

Thanks to the Bettencourt Schueller Foundation long term partnership, this work was partly supported by the LPI Research Fellowship to Marc Santolini. This work was also partly supported by the French Agence Nationale de la Recherche (ANR), under grant agreement ANR-21-CE38-0002.

Data availability

The datasets used and analysed during the current study are available on Zenodo [75].

Declarations

Competing interests

The authors declare that they have no competing interests.

Author contributions

MSa, LT conceptualized the study; CS, LT, MSt, MSa designed research; CS and MSa analysed the data; CS, MSa, FL visualized the data; CS, MSa, LT, FL, MSt wrote the paper. All authors read and approved the final manuscript.

Author details

¹Université Paris Cité, Inserm, System Engineering and Evolution Dynamics, F-75004, Paris, France. ²Learning Planet Institute, F-75004, Paris, France. ³Nokia Bell Labs, Nozay, France. ⁴Sorbonne Université, CNRS, LIP6, F-75005, Paris, France. ⁵CENTA Institute, Turin, Italy. ⁶Departament de Física, Universitat Politècnica de Catalunya, Campus Nord, 08034, Spain.

Received: 29 April 2023 Accepted: 5 February 2024 Published online: 20 February 2024

References

1. Belikov AV, Rzhetsky A, Evans J (2022) Prediction of robust scientific facts from literature. *Nat Mach Intell* 4(5):445–454
2. Iacopini I, Milojević S, Latora V (2018) Network dynamics of innovation processes. *Phys Rev Lett* 120(4):048301. <https://doi.org/10.1103/PhysRevLett.120.048301>
3. Ferreira MR, Reisz N, Schueller W, Servedio VDP, Thurner S, Loreto V (2020) Quantifying exaptation in scientific evolution. *arXiv preprint*. [arXiv:2002.08144](https://arxiv.org/abs/2002.08144)
4. Bornmann L, Haunschild R, Mutz R (2021) Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanit Soc Sci Commun* 8(1):1–15
5. Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, Petersen AM, Radicchi F, Sinatra R, Uzzi B et al (2018) Science of science. *Science* 359(6379):0185
6. Shwed U, Bearman PS (2010) The temporal structure of scientific consensus formation. *Am Sociol Rev* 75(6):817–840. <https://doi.org/10.1177/0003122410388488>
7. Lin Y, Evans JA, Wu L (2022) New directions in science emerge from disconnection and discord. *J Informetr* 16(1):101234
8. Liu L, Wang Y, Sinatra R, Giles CL, Song C, Wang D (2018) Hot streaks in artistic, cultural, and scientific careers. *Nature* 559(7714):396–399
9. Sun Y, Latora V (2020) The evolution of knowledge within and across fields in modern physics. *Sci Rep* 10(1):1–9
10. Yin Y, Wang D (2017) The time dimension of science: connecting the past to the future. *J Informetr* 11(2):608–621. <https://doi.org/10.1016/j.joi.2017.04.002>
11. Pan RK, Petersen AM, Pammolli F, Fortunato S (2018) The memory of science: inflation, myopia, and the knowledge network. *J Informetr* 12(3):656–678
12. Wang D, Song C, Barabási A-L (2013) Quantifying long-term scientific impact. *Science* 342(6154):127–132
13. Sinatra R, Wang D, Deville P, Song C, Barabási A-L (2016) Quantifying the evolution of individual scientific impact. *Science* 354(6312):5239
14. Chavalarías D, Cointet J-P (2013) Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PLoS ONE* 8(2):54847
15. Battiston F, Musciotto F, Wang D, Barabási A-L, Szell M, Sinatra R (2019) Taking census of physics. *Nat Rev Phys* 1(1):89–97
16. Jia T, Wang D, Szymanski BK (2017) Quantifying patterns of research-interest evolution. *Nat Hum Behav* 1(4):1–7
17. Zeng A, Shen Z, Zhou J, Fan Y, Di Z, Wang Y, Stanley HE, Havlin S (2019) Increasing trend of scientists to switch between topics. *Nat Commun* 10(1):3439. <https://doi.org/10.1038/s41467-019-11401-8>
18. Aleta A, Meloni S, Perra N, Moreno Y (2019) Explore with caution: mapping the evolution of scientific interest in physics. *EPJ Data Sci* 8(1):27. <https://doi.org/10.1140/epjds/s13688-019-0205-9>
19. Tuninetti M, Aleta A, Paolotti D, Moreno Y, Starnini M (2021) Prediction of new scientific collaborations through multiplex networks. *EPJ Data Sci* 10(1):25. <https://doi.org/10.1140/epjds/s13688-021-00282-x>
20. Holden C (1974) Federation of scientists plans “great leap forward”. *Science* 185(4145):47. <https://doi.org/10.1126/science.185.4145.47.a>
21. Wu L, Hasan S, Chung Y, Kang JE (2021) Understanding the heterogeneity of human mobility patterns: user characteristics and modal preferences. *Sustainability* 13(24):13921. <https://doi.org/10.3390/su132413921>
22. Ubaldi E, Monechi B, Chiappetta C, Loreto V (2021) Heterogeneity and segregation of mobility patterns. In: *Handbook on entropy, complexity and spatial dynamics*. Edward Elgar, Cheltenham Glos, pp 486–509. ISBN: 9781839100598
23. Barbosa H, Barthelemy M, Ghoshal G, James CR, Lenormand M, Louail T, Menezes R, Ramasco JJ, Simini F, Tomasini M (2018) Human mobility: models and applications. *Phys Rep* 734:1–74. <https://doi.org/10.1016/j.physrep.2018.01.001>

24. Schläpfer M, Dong L, O'Keeffe K, Santi P, Szell M, Salat H, Anklesaria S, Vazifeh M, Ratti C, West GB (2021) The universal visitation law of human mobility. *Nature* 593(7860):522–527
25. Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási A-L (2015) Returners and explorers dichotomy in human mobility. *Nat Commun* 6(1):1–8
26. Simini F, Barlacchi G, Luca M, Pappalardo L (2021) A deep gravity model for mobility flows generation. *Nat Commun* 12(1):1–13
27. Alessandretti L, Aslak U, Lehmann S (2020) The scales of human mobility. *Nature* 587(7834):402–407
28. Schneider CM, Belik V, Couronné T, Smoreda Z, González MC (2013) Unravelling daily human mobility motifs. *J R Soc Interface* 10(84):20130246
29. Simini F, González MC, Maritan A, Barabási A-L (2012) A universal model for mobility and migration patterns. *Nature* 484(7392):96–100
30. Wilson AG (1967) A statistical theory of spatial distribution models. *Transp Res* 1(3):253–269. [https://doi.org/10.1016/0041-1647\(67\)90035-4](https://doi.org/10.1016/0041-1647(67)90035-4)
31. Hills TT, Todd PM, Lazer D, Redish AD, Couzin ID, Group CSR et al (2015) Exploration versus exploitation in space, mind, and society. *Trends Cogn Sci* 19(1):46–54
32. Hills TT (2006) Animal foraging and the evolution of goal-directed cognition. *Cogn Sci* 30(1):3–41
33. Bonabeau E, Dorigo M, Theraulaz G, Theraulaz G (1999) Swarm intelligence: from natural to artificial systems, vol 1
34. Hills TT, Jones MN, Todd PM (2012) Optimal foraging in semantic memory. *Psychol Rev* 119(2):431
35. March JG (1991) Exploration and exploitation in organizational learning. *Organ Sci* 2(1):71–87
36. Zhao Z-D, Huang Z-G, Huang L, Liu H, Lai Y-C (2014) Scaling and correlation of human movements in cyberspace and physical space. *Phys Rev E* 90(5):050802
37. Hu T, Luo J, Liu W (2018) Life in the “matrix”: human mobility patterns in the cyber space. In: Twelfth international AAAI conference on web and social media
38. Barbosa HS, de Lima Neto FB, Evsukoff A, Menezes R (2016) Returners and explorers dichotomy in web browsing behavior—a human mobility approach. In: Cherifi H, Gonçalves B, Menezes R, Sinatra R (eds) *Complex networks VII: proceedings of the 7th workshop on complex networks CompleNet 2016*. Studies in computational intelligence. Springer, Cham, pp 173–184. https://doi.org/10.1007/978-3-319-30569-1_13
39. Baronchelli A, Radicchi F (2013) Lévy flights in human behavior and cognition. *Chaos Solitons Fractals* 56:101–105
40. Thompson WH, Wojtowicz Z, DeDeo S (2018) Lévy flights of the collective imagination. *arXiv preprint*. [arXiv:1812.04013](https://arxiv.org/abs/1812.04013)
41. Milojević S (2015) Quantifying the cognitive extent of science. *J Informetr* 9(4):962–973
42. Milojević S, Sugimoto CR, Yan E, Ding Y (2011) The cognitive structure of library and information science: analysis of article title words. *J Am Soc Inf Sci Technol* 62(10):1933–1953. <https://doi.org/10.1002/asi.21602>
43. Peng H, Ke Q, Budak C, Romero DM, Ahn Y-Y (2021) Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Sci Adv* 7(17):9004
44. Guevara MR, Hartmann D, Arístarán M, Mendoza M, Hidalgo CA (2016) The research space: using career paths to predict the evolution of the research output of individuals, institutions, and nations. *Scientometrics* 109:1695–1709
45. Ying QF, Venkatramanan S, Chiu DM (2015) Modeling and analysis of scholar mobility on scientific landscape. In: *Proceedings of the 24th international conference on world wide web. WWW '15 companion*. Association for Computing Machinery, New York, pp 609–614. <https://doi.org/10.1145/2740908.2741737>
46. Chinazzi M, Gonçalves B, Zhang Q, Vespignani A (2019) Mapping the physics research space: a machine learning approach. *EPJ Data Sci* 8(1):33
47. Clement CB, Bierbaum M, O'Keeffe KP, Alemi AA (2019) On the use of arXiv as a dataset. *arXiv preprint*. [arXiv:1905.00075](https://arxiv.org/abs/1905.00075)
48. Singh CK, Barne E, Ward R, Tupikina L, Santolini M (2022) Quantifying the rise and fall of scientific fields. *PLoS ONE* 17(6):0270131
49. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9:2579–2605
50. Van Der Maaten L (2009) Learning a parametric embedding by preserving local structure. In: *Artificial intelligence and statistics*, pp 384–391. PMLR
51. Wilson AG (1967) A statistical theory of spatial distribution models. *Transp Res* 1(3):253–269. [https://doi.org/10.1016/0041-1647\(67\)90035-4](https://doi.org/10.1016/0041-1647(67)90035-4)
52. Senior ML (1979) From gravity modelling to entropy maximizing: a pedagogic guide. *Prog Hum Geogr* 3(2):175–210
53. Wilson A (2013) *Entropy in urban and regional modelling*. Routledge revivals. Routledge, London
54. Ribeiro FL, Rybski D (2021) Mathematical models to explain the origin of urban scaling laws: a synthetic review. *arXiv preprint*. [arXiv:2111.08365](https://arxiv.org/abs/2111.08365)
55. Iovan C, Olteanu-Raimond A-M, Couronné T, Smoreda Z (2013) Moving and calling mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies. In: Vandenbroucke D, Bucher B, Crompvoets J (eds) *Geographic information science at the heart of Europe*. Lecture notes in geoinformation and cartography. Springer, Cham, pp 247–265. https://doi.org/10.1007/978-3-319-00615-4_14
56. Klafter J, Sokolov IM (2011) First steps in random walks: from tools to applications. <https://doi.org/10.1093/acprof:oso/9780199234868.001.0001>
57. Cui H, Wu L, Evans JA (2022) Aging scientists and slowed advance. *arXiv preprint*. [arXiv:2202.04044](https://arxiv.org/abs/2202.04044)
58. Park M, Leahey E, Funk RJ (2023) Papers and patents are becoming less disruptive over time. *Nature* 613(7942):138–144. <https://doi.org/10.1038/s41586-022-05543-x>
59. Kuhn TS (1979) *The essential tension: selected studies in scientific tradition and change*. University of Chicago Press, Chicago. <https://press.uchicago.edu/ucp/books/book/chicago/E/bo5970650.html>
60. Lanoiselée Y, Grebenkov DS (2017) Unraveling intermittent features in single-particle trajectories by a local convex hull method. *Phys Rev E* 96(2):022144. <https://doi.org/10.1103/PhysRevE.96.022144>
61. Kobak D, Linderman GC (2021) Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat Biotechnol* 39(2):156–157. <https://doi.org/10.1038/s41587-020-00809-z>
62. González-Márquez R, Schmidt L, Schmidt BM, Berens P, Kobak D (2023) The landscape of biomedical research. *Sci Commun Educ*. <https://doi.org/10.1101/2023.04.10.536208>

63. Beltagy I, Lo K, Cohan A (2019) SciBERT: a pretrained language model for scientific text. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, pp 3615–3620. <https://doi.org/10.18653/v1/D19-1371>. <https://aclanthology.org/D19-1371>
64. Duval A (2019) Explainable artificial intelligence (XAI). <https://doi.org/10.13140/RG.2.2.24722.09929>
65. Larivière V, Sugimoto CR, Macaluso B, Milojević S, Cronin B, Thelwall M (2014) ArXiv E-prints and the journal of record: an analysis of roles and relationships. *J Assoc Inf Sci Technol* 65(6):1157–1169. <https://doi.org/10.1002/asi.23044>
66. Yu X, Szymanski BK, Jia T (2021) Become a better you: correlation between the change of research direction and the change of scientific performance. *J Informetr* 15(3):101193
67. Sun Y, Livan G, Ma A, Latora V (2021) Interdisciplinary researchers attain better long-term funding performance. *Commun Phys* 4(1):263
68. Wuchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. *Science* 316(5827):1036–1039
69. Sekara V, Deville P, Ahnert SE, Barabási A-L, Sinatra R, Lehmann S (2018) The chaperone effect in scientific publishing. *Proc Natl Acad Sci* 115(50):12603–12607
70. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
71. Wang Y, Huang H, Rudin C, Shaposhnik Y (2021) Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *J Mach Learn Res* 22(201):1–73
72. McInnes L, Healy J, Melville J (2018) Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint. [arXiv:1802.03426](https://arxiv.org/abs/1802.03426)
73. Funk RJ, Owen-Smith J (2017) A dynamic network measure of technological change. *Manag Sci* 63(3):791–817. <https://doi.org/10.1287/mnsc.2015.2366>
74. Wu L, Wang D, Evans JA (2019) Large teams develop and small teams disrupt science and technology. *Nature* 566(7744):378–382. <https://doi.org/10.1038/s41586-019-0941-9>
75. Singh CK, Tupikina L, Ward R, Barme E, Santolini M (2022) arXiv articles metadata and citation network for the “Quantifying the rise and fall of Scientific fields” 2022 article. Zenodo. <https://doi.org/10.5281/zenodo.6598737>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)