Understanding Citation Mobility in the Knowledge Space*

Shuang Zhang[1], Feifan Liu[1,2] and Haoxiang Xia[1,2,*]

[1] *Institute of Systems Engineering, Dalian University of Technology, No. 2 Linggong Road, Dalian, 116024, P.R. China*

[2] *Institute for Advanced Intelligence, Dalian University of Technology, No. 2 Linggong Road, Dalian, 116024, P.R. China*

Abstract

Despite persistent efforts to reveal the temporal patterns of citation dynamics, little is known about its spatial patterns in knowledge space, owing to the unquantifiability of citation diffusion in the virtual high-dimensional space. Here, drawing on millions of papers in the Physics field, we consider individual papers' citation sequences as a mobility process and track trajectories with embedding methods learning the semantic proximity. We first quantify the spatial scale of citation mobility and find Gaussian-distributed citation scope and exponentially-distributed citing embedding distance, indicating the constrained mobility of citations. Simulations with the Gravity model and Radiation model further confirm that epistemic distance and popularity are key push-and-pull factors, respectively, in citation mobility. It is then found that compared with high-cited papers, disruptive papers are more likely to receive distant recognition. As science evolves, papers nowadays make narrower citation mobility than those in earlier decades. These findings provide insights into understanding the diversified knowledge diffusion and scientific innovation efficiency.

Keywords

citation dynamics, spatial patterns, knowledge diffusion, disruptive innovation

1. Introduction

Citations encapsulate the dynamics of ideas circulation, unfolding both in temporal and spatial dimensions in the abstract knowledge space[1]. Extensive research has delved into citation patterns at levels from the paper[2], author[3], discipline[4], to nation[5]. For individual papers, despite the diversity of citation profiles[6], researchers attempt to quantify[2], model[7], and predict[8] citation dynamics. Key drivers of citation dynamics, including preferential attachment, aging, and fitness[2] have been identified. Universal patterns, such as scale laws in citation distributions[9], first mover effect[10], citation probability decreasing with papers' age[11], and "jump-decay" patterns[12] have been quantitatively revealed. Moreover, "sleeping beauties" whose atypical citation dynamics have been explored in terms of identification and awakening mechanism[13]. However, despite the fruitful efforts on the temporal aspects of the citation dynamics, our understanding of the spatial dimension remains limited.

On collective level, citations signify collective attention. Albeit with the explosion of papers and citation inflation[14], we find that citations are increasingly concentrated on elite scientists[15] and top papers[16], leaving new publications less likely to be recognized[17]. Growing citation inequality indicates a narrowing and decaying scientific attention, exacerbating the stratification of the scientific system and entrenching science trapped in

Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), April 23~24, 2024, Changchun, China and Online

* Corresponding author.

hxxia@dlut.edu.cn (H. Xia)
shuang94@mail.dlut.edu.cn (S. Zhang)
liufeifan@dlut.edu.cn (F. Liu)

existing norms[12,18]. This narrowing attention phenomenon warrants detailed investigation through the lens of a holistic knowledge landscape.

Papers receive citations spanning different epistemic distances. On citation dynamics of individual papers in the knowledge space, similar studies focus on mapping structure and evolution of disciplines with citation flows[19], associations between interdisciplinary citations and novelty[20], and measuring the breadth and depth of impact by examining textual proximity between citing papers[21]. However, exsiting studies remains inadequate for quantifying the knowledge aspect of citation trajectories due to their abstract nature and high dimensionality.

Major obstacles in large-scale quantitative investigations on individual papers' citation dynamics in knowledge space are the inability to track trajectories and the lack of an appropriate quantitative metric for this dynamical progress. It is unclear how papers diffuse impact and ideas in the knowledge space over lifecycles.

Here, we regard the sequential citation process of papers as mobility on a quantifiable epistemic landscape and use machine-learning techniques to trace the trajectories. In this manner, we introduce the theoretical and methodological framework of geospatial human mobility to characterize citation mobility. Some key research questions are quantitatively analyzed. First, we explore the spatial scale characteristics and collective-level mechanisms of citation mobility. Second, we probe whether different types of novel papers exhibit diversified spatial patterns. Third, evolutionary patterns of citation mobility over decades are checked.

2. Data and methods

2.1. Data

This study focuses on the discipline of Physics. The dataset used is SciSciNet[22], a large-scale scientific dataset built on MAG[23], covering over 134 million scientific publications up to the year 2021.

Using the "fields of study" classification, we extract 3,263,546 papers labeled "Physics". Then we select focal papers satisfying: (*i*) number of citations no less than 10, to ensure sufficient trajectory points for quantification; (*ii*) citation history spanning at least 10 years, to ensure sufficient timespans to capture spatiotemporal patterns; (*iii*) receiving at least one citation every five years, to exclude noisy data. Finally, we obtain 214,867 focal papers.

2.2. Construction of citation trajectories on the epistemic landscape

We develop a framework, which combines representative learning algorithms and manifold learning algorithms, for the construction of the quantifiable disciplinary knowledge landscape based on semantics association. Unlike citation networks merely representing the topological connections of elements, this landscape provides a continuous distance scale, allowing for the tracking and quantifying of citation trajectories of individual papers.

Here, we employ the Doc2Vec algorithm[24], capturing the semantics of content, and the popular UMAP algorithm[25] preserving the global and local topology in dimension reduction. The majority of architectures and hyperparameters we utilized were set to their default values throughout the model training process.

Figure 1 illustrates the proposed framework for constructing the knowledge landscape. After building the corpus with the title and abstract, we train the Doc2vec model to obtain semantic vectors of papers. The UMAP algorithm is subsequently applied to project the semantic vectors into a two-dimensional space based on their cosine distance. Finally, we obtain the coordinates of each paper and the epistemic landscape. Thus, the citation trajectories of individual papers are traced by mapping their citation sequences onto this landscape.

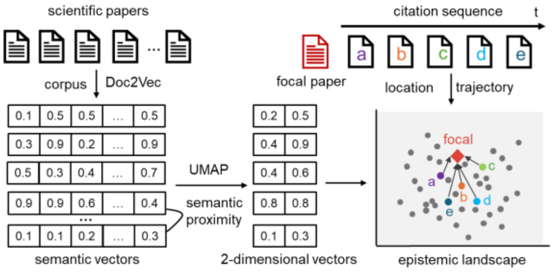


Figure 1: Illustration for constructing the epistemic landscape and citation trajectories based on the semantic proximity embedded in the textual content of papers

2.3. Radius of gyration and jump lengths

Two indicators are applied to characterize the spatial scale of citation mobility[26,27]. The radius of gyration (r_g) refers to the typical distance from individual trajectories from their centroid of mass. The jump length (Δr) measures the epistemic distance between a citing-cited pair.

In the context of citation mobility, r_g is applied to measure the degree to which one's citations are concentrated or dispersed. *Δr* quantifies the research proximity of the focal paper to its citing papers.

$$r_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{r}_i - \boldsymbol{r}_{cm})^2}, r_{cm} = \sum_{i=1}^{N}\boldsymbol{r}_i/N \qquad (1)$$

$$\Delta r = \boldsymbol{r}_i - \boldsymbol{r}_0 \qquad (2)$$

In formulas (1-2), r_0 is the coordinates of the focal paper; r_i and r_{i-1} are the coordinates of its *i*th and *(i-1)*th citing paper; r_{cm} is the centroid of the *N* citing papers.

2.4. Gravity model and Radiation model

The distance-based Gravity model, and the opportunity-based Radiation model, are introduced to characterize aggregated citation flows on the epistemic landscape. These two classical population-level models depict distinct flow generation mechanisms and could reveal key drivers of citation flows in terms of research popularity, knowledge distance, and opportunities.

In citation scenarios, Gravity models assume flows between two locations are proportional to research hotness and decay with knowledge distance[28]. Radiation models assume movement probability of citations is proportional to destination opportunities and inversely proportional to intervening opportunities[29].

$$T_{ij} \propto m_i m_j f(r_{ij}) \qquad (3)$$

$$T_{ij} = O_i \frac{m_i m_j}{(m_i+s_{ij})(m_i+m_j+s_{ij})} \qquad (4)$$

where T_{ij} is citation flows from tile *i* of the citing paper to tile *j* of the focal paper. m_i and m_j are the paper density in tile *i* and *j*; $f(r_{ij})$ is the distance function modeled with power-law form. O_i represents flows from tile *i*; s_{ij} is the number of intervening opportunities (paper density) between tile *i* to *j*. Model performance is assessed with metrics: R^2, RMSE, Spearman, and Pearson correlations.

3. Results

3.1. The spatial characteristics of trajectories of citation mobility

We start by visualizing the individual papers' citation trajectories on the epistemic landscape. In Fig. 2c, paper points are clustered and semantically distributed, depicting the knowledge structure. After mapping citation dynamics (Fig. 2a) of papers onto the epistemic landscape, we find citations are not homogeneous, as they span different knowledge distances (Fig. 2b). However, the visualization in Fig. 2d intuitively shows one's trajectory is locally distributed.

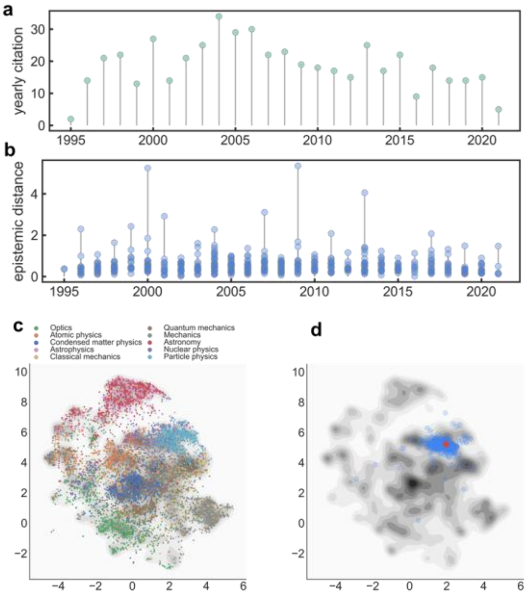


Figure 2: Visualization of individual papers' citation mobility on the epistemic landscape

We quantify spatiotemporal characteristics with two indicators. The citing epistemic distance *Δr* is more approximated by an exponential function, than power-law (Fig. 3a). It indicates that papers are likely to receive massive short-distanced citations and a few longer-distanced ones. Then, the radius of gyration r_g approximates lognormal distribution, suggesting the narrower impact of most papers and the broader impact of a few papers (Fig. 3b). These findings indicate that both citing distance and overall impact scope follow the typical scale variation in citation mobility, in contrast to the fat-tailed spatial scale displayed by human mobility in the biological world[26,27,30].

Furthermore, we note the more citations papers receive, the wider their impact scope (Fig. 3d). However, exponentially distributed citing distance and lognormal-distributed citation concentration are independent of the number of citations (Fig. 3c&d). In a word, we observe constrained mobility of citations in the knowledge space.

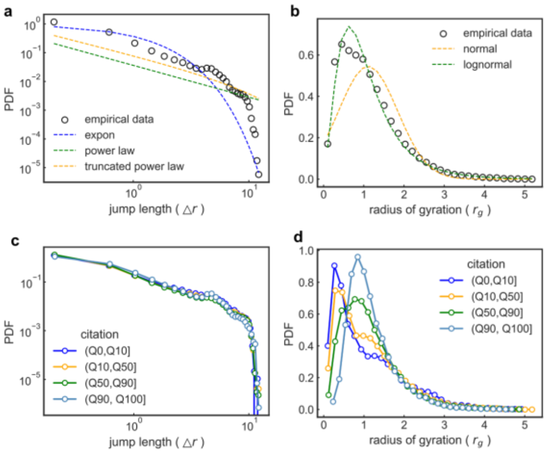
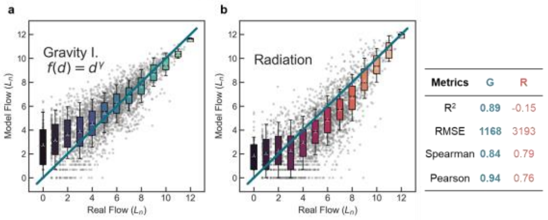


Figure 3: Empirical distribution of spatiotemporal characteristics of citation trajectories

3.2. The Gravity and Radiation modeling in citation mobility

To further delineate the observed narrow movements, we use the classic Gravity model and Radiation model to fit the aggregated flow of citation mobility. After discretizing the Physics epistemic landscape to a spatial tessellation, we aggregate individual trajectories into origin-destination citation flows. Most citation flows are intra-flows and only inter-flows between two different grids are used to employ parameter fitting and flow generation.



Metrics	G	R
R²	0.89	-0.15
RMSE	1168	3193
Spearman	0.84	0.79
Pearson	0.94	0.76

Figure 4: Actual and simulated citation flows generated by Gravity and Radiation models

Fig. 4 shows the simulated results. It could be seen that the gravity model outperforms the radiation model, especially for long-distance flows. This suggests that epistemic distance and popularity are key factors in citation behavior, whereas the research gap representing potential research intersection area, is not significant in attracting citations.

3.3. Comparisons of high-cited, sleeping beauties, and disruptive papers

The further question is how citation mobility differs across papers with various types of novelty. We focus on three attributes of papers: popularity, delayed recognition, and disruptiveness, and measure them with the number of citations, sleeping beauty coefficient[13], and disruption index[31], respectively. The top 10% of papers by each metric are identified as highly cited, sleeping beauties, and disruptive papers (Fig. 5a).

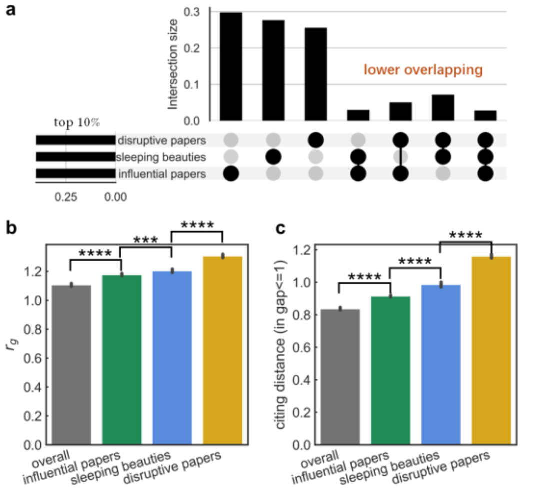


Figure 5: The r_g and Δr of citation trajectories of high-cited papers, sleeping beauties, and disruptive papers. *****p≤.0001, ***: p≤ .001, ns: p ≥ 0.05*

We observe that these three representative novel papers with a low degree of overlap (Fig. 5a), have above-average impact scopes, with disruptive papers standing out in particular (Fig. 5b). The finding that sleeping beauties with a broader impact is in line with their interdisciplinary nature [13].

We further examine the citing distance in the first year post-publication. The consistent patterns observed in Fig. 5c reinforce our previous findings. It suggests that compared with the influential high-cited papers, sleeping beauties and high-disruptive papers promptly attract attention from more distant knowledge communities once published.

3.4. Evolution of citation mobility

Finally, we group focal papers into different decades according to their publication year to investigate how citation mobility evolved over decades.

The first finding is that papers nowadays make more restricted mobility than those in the early years, as shown in Fig. 6a. To rule out the possibility that this result is due to semantic differences between papers from different decades, we analyze the citing distance of citing pairs with one year gap. In Fig. 6b, the observed decrease in the trend of citing distance over publication years indicates the narrowing of literature use. These two results suggest a possible shortersightedness for scientists' information foraging nowadays.

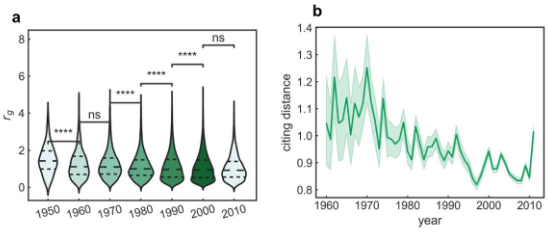


Figure 6: Spatial characteristics of citation trajectories in different decades

4. Conclusion and discussion

An empirically detailed investigation of the spatial pattern of papers' citation mobility in knowledge space is indispensable for understanding knowledge diffusion. In this study, we trace and quantify individual papers' citation sequences on the epistemic landscape based on semantic proximity.

We primarily examine two spatial scale characteristics and observe the overall conserved citation mobility independent of citation counts, which is distinct from the fat-tail characteristics displayed in human mobility. By applying the Gravity model, epistemic distance and popularity are identified as two key divers. Next, compared with high-cited papers, disruptive and sleeping beauties present wider citation mobile scopes. Finally, current papers have narrower mobility than earlier papers, reflecting more myopic information foraging in current scientific practice.

Several research extensions can be performed. Further with a whole picture of science, citation mobilities within and across disciplines could be explored, gaining more comprehensive insights. The framework could be applied to patents, open-source software, and online searching behavior.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 71871042 and 72371052).

References

[1] S. Fortunato, C.T. Bergstrom, K. Boerner, J.A. Evans, D. Helbing, S. Milojevic, A.M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, A. Barabasi, Science of science, Science 359 (2018). doi:10.1126/science.aao0185.

[2] D. Wang, C. Song, A.L.A. Barabási, Quantifying long-term scientific impact, Science 342 (6154) (2013) 127-133. doi:10.1126/science.1237825.

[3] R. Sinatra, D. Wang, P. Deville, C. Song, A.L. Barabasi, Quantifying the evolution of individual scientific impact, Science 354 (6312) (2016). doi:10.1126/science.aaf5239.

[4] R.K. Pan, S. Sinha, K. Kaski, J. Saramäki, The evolution of interdisciplinarity in physics research, Sci. Rep. 2 (1) (2012). doi:10.1038/srep00551.

[5] R.K. Pan, K. Kaski, S. Fortunato, World citation and collaboration networks: uncovering the role of geography in science, Sci. Rep. 2 (1) (2012). doi:10.1038/srep00902.

[6] A. Avramescu, Actuality and obsolescence of scientific literature, Journal of the American Society for Information Science 30 (5) (1979) 296-303. doi:10.1002/asi.4630300509

[7] Y.H. Eom, S. Fortunato, Characterizing and modeling citation dynamics, Plos One 6 (9) (2011) e24926. doi:10.1371/journal.pone.0024926.

[8] A. Abrishami, S. Aliakbary, Predicting citation counts based on deep neural network learning techniques, J. Informetr. 13 (2) (2019) 485-499. doi:10.1016/j.joi.2019.02.011.

[9] M. Golosovsky, S. Solomon, Runaway events dominate the heavy tail of citation distributions, The European Physical Journal Special Topics 205 (1) (2012) 303-311. doi:10.1140/epjst/e2012-01576-4.

[10] M.E.J. Newman, The first-mover advantage in scientific publication, Epl 86 (6) (2009). doi:10.1209/0295-5075/86/68001.

[11] M. Golosovsky, S. Solomon, Stochastic dynamical model of a growing citation network based on a self-exciting point process, Phys. Rev. Lett. 109 (2012) 98701. doi:10.1103/PhysRevLett.109.098701.

[12] P.D.B. Parolo, R.K. Pan, R. Ghosh, B.A. Huberman, K. Kaski, S. Fortunato, Attention decay in science, J. Informetr. 9 (4) (2015) 734-745. doi:10.1016/j.joi.2015.07.006.

[13] Q. Ke, E. Ferrara, F. Radicchi, A. Flammini, Defining and identifying sleeping beauties in science, Proc. Natl. Acad. Sci. U. S. A. 112 (24) (2015) 7426-7431. doi:10.1073/pnas.1424329112.

[14] A.M. Petersen, R.K. Pan, F. Pammolli, S. Fortunato, Methods to account for citation inflation in research evaluation, Res. Policy 48 (7) (2019) 1855-1865. doi:10.1016/j.respol.2019.04.009.

[15] M.W. Nielsen, J.P. Andersen, Global citation inequality is on the rise, Proceedings of the National Academy of Sciences 118 (7) (2021). doi:10.1073/pnas.2012208118.

[16] A. Varga, The narrowing of literature use and the restricted mobility of papers in the sciences,

Proceedings of the National Academy of Sciences 119 (17) (2022). doi:10.1073/pnas.2117488119.

[17] J.S.G. Chu, J.A. Evans, Slowed canonical progress in large fields of science, Proceedings of the National Academy of Sciences 118 (41) (2021). doi:10.1073/pnas.2021636118.

[18] R.K. Pan, A.M. Petersen, F. Pammolli, S. Fortunato, The memory of science: inflation, myopia, and the knowledge network, J. Informetr. 12 (3) (2018) 656-678. doi:10.1016/j.joi.2018.06.005.

[19] R. Sinatra, P. Deville, M. Szell, D. Wang, A. Barabsi, A century of physics, Nat. Phys. 11 (10) (2015) 791-796. doi:10.1038/nphys3494.

[20] Y. Bu, L. Waltman, Y. Huang, A multidimensional framework for characterizing the citation impact of scientific publications, Quant. Sci. Stud. 2 (1) (2021) 155-183. doi:10.1162/qss_a_00109.

[21] V. Larivière, S. Haustein, K. Börner, Long-distance interdisciplinarity leads to higher scientific impact, Plos One 10 (3) (2015) e122565, . doi:10.1371/journal.pone.0122565.

[22] Z. Lin, Y. Yin, L. Liu, D. Wang, Sciscinet: a large-scale open data lake for the science of science research, Sci. Data 10 (1) (2023). doi:10.1038/s41597-023-02198-9.

[23] Z. Shen, H. Ma, K. Wang, A web-scale system for scientific knowledge exploration, Melbourne, Australia, 2018, pp. 87-92.

[24] Q. Le, T. Mikolov, Distributed representations of sentences and documents, Proceedings of Machine Learning Research, Bejing, China, 2014, pp. 1188-1196.

[25] L. Mcinnes, J. Healy, N. Saul, L. Großberger, Umap: uniform manifold approximation and projection, Journal of Open Source Software 3 (29) (2018) 861. doi:10.21105/joss.00861.

[26] M.C. González, C.A. Hidalgo, A.L. Barabási, Understanding individual human mobility patterns, Nature 453 (2008) 779-782. doi:10.1038/nature.

[27] C. Song, T. Koren, P. Wang, A. Barabási, Modelling the scaling properties of human mobility, Nat. Phys. 6 (10) (2010) 818-823. doi:10.1038/nphys1760.

[28] M. Lenormand, A. Bassolas, J.J. Ramasco, Systematic comparison of trip distribution laws and models, J. Transp. Geogr. 51 (2016) 158-169. doi:10.1016/j.jtrangeo.2015.12.008.

[29] F. Simini, M.C. González, A. Maritan, A. Barabási, A universal model for mobility and migration patterns, Nature 484 (7392) (2012) 96-100. doi:10.1038/nature10856.

[30] D.W. Sims, E.J. Southall, N.E. Humphries, G.C. Hays, C.J.A. Bradshaw, J.W. Pitchford, A. James, M.Z. Ahmed, A.S. Brierley, M.A. Hindell, D. Morritt, M.K. Musyl, D. Righton, E.L.C. Shepard, V.J. Wearmouth, R.P. Wilson, M.J. Witt, J.D. Metcalfe, Scaling laws of marine predator search behaviour, Nature 451 (7182) (2008) 1098-1102. doi:10.1038/nature06518.

[31] L. Wu, D. Wang, J.A. Evans, Large teams develop and small teams disrupt science and technology, Nature (2019). doi:10.1038/s41586-019-0941-9.