Limited Diffusion of Scientific Knowledge Forecasts Collapse

Donghyun Kang[1,2]
Robert S. Danziger[3,4,5]
Jalees Rehman[3,6]
James A. Evans[1,2,7*]

[1]Department of Sociology, University of Chicago, Chicago, IL, USA.
[2]Knowledge Lab, University of Chicago, Chicago, IL, USA.
[3]Department of Medicine, University of Illinois at Chicago, Chicago, IL, USA.
[4]Department of Pharmacology, University of Illinois at Chicago, Chicago, IL, USA.
[5]Department of Physiology and Biophysics, University of Illinois at Chicago, Chicago, IL, USA.
[6]Department of Biochemistry and Molecular Genetics, University of Illinois, College of Medicine, Chicago, IL, USA.
[7]Santa Fe Institute, Santa Fe, NM, USA.

*Corresponding author. Email: jevans@uchicago.edu

Abstract: Market bubbles emerge when asset prices are driven higher than the ultimate value without sustainability, as dramatic events and pivots of belief can burst them. We demonstrate the same phenomenon for biomedical knowledge when promising research receives inflated attention. We predict deflationary events by developing a diffusion index that captures whether research areas have been amplified within social and scientific bubbles or have diffused and become evaluated more broadly. By applying our diffusion approach, we first showcase two contrasting trajectories of cardiac stem cell research and cancer immunotherapy. We also trace the diffusion of 28,504 subfields in biomedicine comprising nearly 1.9M papers and more than 80M citations and demonstrate that limited diffusion of biomedical knowledge anticipates abrupt decreases in popularity. Our analysis emphasizes that this restricted diffusion, or a socio-epistemic bubble, leads to a dramatic collapse in relevance and attention accorded to scientific knowledge.

Market bubbles emerge when popular opinions about an asset, such as housing or securities, produce self-reinforcing information that drives its price much higher than its value to society[1]. These bubbles are characterized by a rapid increase in popularity, fueled by beliefs that the value may continue to rise and persist, leading to speculation. Bubbles burst when shifts in opinion "correct" the market, often triggered by new data or events, leading to radical discounts in pricing[2]. Here we extend this concept of bubbles and their collapse to the realm of scientific research and knowledge dissemination, where ideas or findings may experience a rise and fall in popularity and credibility.

In the system of biomedical knowledge, citation counts have come to function as an operational currency[3–5] and assessment of the importance and impact of scientific work. Citation metric manifests some distortion, however, as from the inflation of citation counts with historical growth in articles[6] and the unequal size of fields[7]. Drawing parallels with epidemiological modeling of diffusion[8,9], here we forecast substantial and dramatic declines in the popularity of research ideas—bursting epistemic bubbles—as the degree to which those ideas remain concentrated within the same collection of authors, institutions, and biomedical subfields, failing to diffuse across social and scientific space despite concentrated popularity. We argue that this lack of diffusion may indicate inflated attention to particular ideas or bubbles, which may not generalize, withstand scrutiny when examined more broadly, and ultimately lead to disappointment and disillusionment within the scientific community.

Consider the extreme case of cardiac regeneration in biomedicine. Dr. Piero Anversa and collaborators pioneered research in cardiac regeneration at the turn of the 21st Century that received tremendous attention and support—ballooning to receive tens of thousands of citations and more than $387M in NIH funding to top investigators—by asserting the possibility of heart

regeneration after myocardial infarction[10]. During the period of Anversa and collaborators' peak productivity, they also exercised powerful influence over the research narrative in how stem cells and progenitor cells from the bone marrow or within the heart could regenerate damaged heart muscle tissue, sitting on editorial boards of high-profile American Heart Association journals like *Circulation Research* (e.g., Anversa personally reviewed hundreds of papers for *Circulation Research* alone, more than any other researcher in this period). Anversa was a member of the NIH National Institute on Aging's Board of Scientific Counselors (2008-2013), and he and collaborators presided over an interlocking matrix of NIH grant review panels. Findings from the early work not only failed to generalize, however, but their experiments could not be replicated by other researchers[11]. This resulted in a dramatic breach of trust, the retraction of more than 30 related papers from leading journals, a marked discount in citations to the subfield and diminished confidence in the field of cardiac regeneration, and Anversa's forced departure from Harvard. This, in turn, adversely impacted even those researchers who had been studying cardiac regeneration using more rigorous scientific approaches and had identified reproducible mechanisms of the phenomenon[12]. Our approach generalizes well beyond research misconduct, however. Honestly reported and accurate medical findings can fail to generalize beyond the specific context of their initial investigation, despite optimism and hype regarding their transformative potential for medicine.

In this study, we demonstrate that fragile and overhyped biomedical findings could have been anticipated by analyzing how their work diffused through the system of science. Utilizing large-scale bibliographical databases of metadata, our work provides a framework that considers distances between publications and their citing papers within the "social space" inscribed by collaborating scientists and the "scientific space" constituted by co-investigated scientific entities

in the context of biomedicine. Specifically, we develop a diffusion index to capture whether ideas have been amplified by social and scientific bubbles[13], or diffused to become consumed more widely and tested for robustness across diverse research communities[14]. This approach allows us to gain insights into the diffusion of research ideas and their impact measured beyond citation counts, ultimately helping us to better assess the value and potential of scientific findings.

We then show how a lack of diffusion measured by this framework—the existence of a scientific bubble—can anticipate a decline of popularity as bubbles of confidence burst. Fig. 1A schematically illustrates this scenario as research gains rapid citations early on, but quickly loses attention as it fails to successfully spread. This pattern is indicative of fragile, overhyped ideas that may not withstand broader scrutiny or application in the scientific community, ultimately leading to a decline in interest and support. This contrasts with Fig. 1B, where research gradually diffuses to distant research groups and topics before garnering outsized attention. By applying conceptual tools and measurement, we first trace two contrasting trajectories of cardiac stem cell research and cancer immunotherapy. We then measure the diffusion of 28,504 subfields in biomedicine[15] comprising nearly 1.9M papers and more than 80M citations. We demonstrate that limited diffusion of biomedical knowledge systematically anticipates abrupt decreases in popularity. Our analysis emphasizes that restricted diffusion, indicative of a socio-epistemic bubble, leads to a dramatic collapse in relevance and attention accorded to scientific knowledge.

Results

Contrasting Trajectories of Cardiac Stem Research and Cancer Immuno-Therapy

Applying neural embedding models to MEDLINE data enables us to project all biomedical research articles onto social and scientific manifolds. This indicates their relative positions within the landscape of collaborating scientists and co-occurring biomedical entities (Methods). The cosine or angular distances between citing and cited research measured over social and scientific spaces aggregate into a straightforward, continuous metric of diffusion. To illustrate our diffusion index constructed across scientific and social spaces, we first contrast trajectories of two highly cited publications at the individual paper level, drawn from Cardiac Stem Cell and Cancer Immunotherapy research, respectively.

Our first example is drawn from a research article published (PMID: *11777997*) in *the New England Journal of Medicine* in 2002[16]. The research was conducted by a team led by Dr. Piero Anversa, who argued and supported the existence of substantial numbers of endogenous myocardial stem and progenitor cells thought to regenerate infarcted heart muscles. This line of research received outsized attention because it highlighted the possibility of regenerating the heart after severe myocardial infarctions with massive loss of heart tissue. However, several researchers outside of this author network raised questions about the purported regeneration potential of the heart. This resulted in the retraction of more than 30 papers from the Anversa team published in many of the leading biomedical research journals due to data fabrication and scientific malpractice by 2018[17].

The second example refers to an article published (PMID: 11015443) in *the Journal of Experimental Medicine* in 2000[18]. The research was conducted by a group of researchers who pioneered the field of cancer immunotherapy, with discoveries on the inhibition of negative immune regulation and its role in treating cancer. This publication and related work gave rise to a wide range of cancer immunity and cancer immunotherapy research programs across many

research groups and countries, laying the foundation for what has become one of the most impactful innovations in cancer treatment.

The bottom panels (C and D) of Fig. 1 visualize the 3D kernel density estimation based on contrasting trajectories of the two publications in size of attention and diffusion with the same conceptual frame as the upper panels (A and B). The annual diffusion indices are computed by using two rolling years and averaging the cosine distances between focal and forward citing papers across social and scientific space (see Methods).

Dr. Anversa's publication experienced a meteoric rise in total citations during the first 5 years following debut but with limited diffusion across the scientific space of distinct subfields and social space of author teams citing the paper (Fig. 1C). Put differently, the Cardiac Stem Cell paper remained highly concentrated in its influence. By contrast, the article from Cancer immunotherapy, which demonstrated the potential to inhibit negative immune regulation in treating cancer, experienced a much slower pace in gaining early attention (Fig. 1D). Nevertheless, the ideas diffused much more broadly before becoming one of the most influential innovations in recent cancer treatment and research, resulting in the 2018 Physiology and Medicine Nobel Prize to Drs. Tasuku Honjo and James P. Allison. These contrasting examples illustrate how accounting for epistemic bubbles with our diffusion metric can complement traditional citation counts by capturing latent diffusion dynamics through social and scientific space.

Knowledge Concentration Anticipates Collapse

We elevate our analysis to the level of scientific subfields in order to test the generalizability of our approach beyond high-profile papers, allowing us to capture the typical

dynamics of diffusion and shifting attention. If work from a focal subfield is cited by abundant work concentrated in nearby social and scientific space, that subfield's insights have not diffused in demonstrated influence and may maintain an inflated value based on local reinforcement. In other words, we anticipate that substantial and dramatic declines in the popularity of research ideas, conceptualized as knowledge "bubbles bursting," may be predicted by the degree to which ideas, despite their apparent popularity, fail to diffuse across the social and scientific space via citations.

We apply our framework to ~28.5K biomedical subfields curated by Azoulay et al.[15], each spanning a subfield in biomedical science, with approximately 1.9M unique research articles published before 2020. Our primary outcome of interest is "bubble bursting," an abrupt decline in the relevance of a given subfield, measured in the drop of relative citations as illustrated in the upper left panel of Fig. 1. We specifically time a bubble burst as when the standardized citation difference falls below a cutoff (i.e., 0.5%, 0.25%, 0.1% of the distribution; see Method and Figs. S1 and S2 in Supplementary Information). We compute our main predictors—knowledge diffusion indices for each subfield—by separating within-subfield and outside-subfield citations to distinguish the degree to which knowledge pools within or transfers beyond a subfield (Fig. 2 and Method).

Using a nonparametric Cox survival model to predict the probability of bubble bursting, our estimation reveals the knowledge diffusion index as a strong leading signal that precedes a sudden collapse in attention. By splitting our observations into three groups with diffusion percentiles ranked by calendar year and subfield age—the bottom 10th percentile, the top 10th percentile, and the middle between them— Fig. 3 visualizes diffusion beyond subfields in social space that forecasts the bursting of attention bubbles. Results indicate that low-outside subfield

diffusion rates in social space is a leading indicator of poor long-term subfield survival. By contrast, high-outside subfield diffusion is related to subfield survival in the long-term, avoiding extreme subfield-level deflationary events and ultimate extinction.

We confirm this pattern with discrete-time event history models that allows us to consider temporal covariates, including field size and growth rate, total cumulative citations, citation concentration across papers, paper retractions, and unexpected deaths of elite scientists (see Method). The analysis consistently shows that the lower a paper's diffusion of influence, especially beyond the origin subfields in social space, the greater the hazard that subfield will experience an abrupt collapse of attention (Table 1 and Tables S1 and S2 in Supplementary Information). Specifically, as diffusion in social space reduces outside the original subfield from one standard deviation above to one below the mean, translates into a 39.1% (=exp[-.165 *-2] - 1) increase in the odds of experiencing a major deflationary event, accounting for calendar years and subfield ages. This affirms the significant association of social concentration on the likelihood bubbles will burst in the economy of attention and its impact on future advances in scientific knowledge.

Discussion

Current metrics of scientific attention and confidence pay scant attention to patterns of research consumption and diffusion across diverse people, institutions, disciplines, regions, and beyond. This lack of consideration can lead to an incomplete understanding of a research field's true impact and potential. Our knowledge diffusion index contrasts with and complements citation counts, the conventional unit of scientific credit, which remains blind to who, where, and how far

across the landscape of science those building on research reside, providing a more comprehensive view.

A constriction in diffusion identifies an epistemic bubble or social echo chamber that forms a strong leading indicator of future collapse in relevance and attention accorded to scientific and biomedical knowledge. Researchers can anticipate the collapse of biomedical approaches years before their occurrence by systematically tracking the diffusion of their ideas across scientists and biomedical areas. Moreover, science and biomedical policy that analyzes knowledge diffusion patterns can anticipate such collapses and reduce their occurrence by incentivizing and accounting for diverse, disconnected support for robust scientific and medical claims[14]. In this way, we demonstrate the importance of idea diffusion for advancing scientific knowledge, its ability to transfer across broad science communities, and the relevance of these signals for forecasting robust ideas on which to build novel and important scientific and biomedical knowledge.

Like other efforts to quantitatively evaluate research impact, our framework for measuring diffusion and its implementation cannot and should not replace holistic judgment of research quality. For example, small, dense research networks may be necessary to launch risky research projects with a high probability of failure in their early stages. Nevertheless, we believe that our finding holds strong implications for biomedical researchers, science-based industries, and science policymakers. Accounting for diffusion and diversity, funding agencies could spot echo chambers and adjust resource allocation by diversifying groups of researchers sponsored to work on a research topic. Research information systems and platforms could also add a strong, leading signal from which analysts can anticipate the future relevance of current research by indicating the robustness of trending insights versus their fragility when reinforced within an

echo chamber. Regular self-assessments of knowledge diffusion could enable individual researchers and research groups to better gauge the robustness and impact of their work. Further, documenting associations between scientific knowledge diffusion and its rate of application such as clinical translation from basic biomedical research can better inform science policy.

Our result draws on subfields identified in academic science using a particular delineation of research subfields. Nevertheless, our analysis demonstrates clear evidence for the wisdom of diverse crowds in science and technology to sustain advance. It suggests the importance of both social and scientific diversity for robust evaluation of an idea's relevance to science as a whole. Moreover, our proposed framework for measuring diffusion extends to other domains of knowledge, such as the spread of misinformation, by allowing us to measure diversity in information consumption. In social media, algorithmic metrics that account for diversity in diffusion would be far less susceptible to strategic, concentrated efforts seeking to misclassify information as a legitimate, widespread trend (e.g., on Facebook's Newsfeed), just as they would decrease the intentional or unintentional illusion of scientific support.

Ultimately, our analysis affirms that scientific fields are social enterprises, which advance with cultivated diversity. This manifests the relative importance for identifying the path of an idea's consumption over its point of production for predicting lasting, far-reaching impact. Accounting for this will allow the design of wise and diverse research, development, and clinical crowds leading to better scientific and biomedical research policy, greater reproducibility and more sustained impact on scientific knowledge.

Methods

Manifold Representations of Social and Scientific Space

To assess the diffusion of ideas in science from biomedicine, we train two high-dimensional vector representations using neural embedding models[19] for publications cataloged in the PubMed Knowledge Graph (PKG)[20]. It provides 15,530,165 disambiguated author IDs and 481,497 unique combinations of Medical Subject Headings (MeSH) from 29,339 MeSH descriptors and 76 qualifiers, each assigned to 28,329,992 and 26,666,615 MEDLINE-indexed publications by the end of 2019. Each document from the PubMed database is assigned a PMID, document identifier from PubMed. The database also includes the publications to the publication reference records, which integrates PubMed's citation data, NIH's open citation collection, OpenCitations, and the Web of Science.

We specifically adapt the Doc2vec model[19], a variant of the Word2vec model[21], initially developed to produce dense vector representations for documents or paragraphs from the words that compose them. This approach has previously been extended to generate high-dimensional representational vectors geometrically proximate to the degree that entities frequently share neighbors, contexts[21–23], or are connected via social ties[24,25].

We treat that a biomedical research article can be characterized by 1) a list of MeSH terms and 2) researchers authoring it. Considering this, we build two separate representational vector spaces - "scientific space" and "social space". We use the Python Gensim package[26] to train our vector representations. We detail the training and validation procedure in the Supplementary Information (S2).

Delineating Biomedical Subfields

Biomedical knowledge obtains influence when others recognize and build on it[27,28]. Here we seek to understand the dynamics of diffusion and shifting attention at the level of biomedical

subfields, which we define here as a group of biomedical publications tightly related to a medically relevant research topic. This approach has been previously adopted in the context of studying the impact of publication retraction[29] and the consequences of premature death of elite life scientists[15] on subfields.

We specifically repurpose the 28,504 biomedical subfields captured by Azoulay et al.,[15] spanning over 1.9 million unique articles, constructed through the PubMed Related Algorithm (PMRA)[30] and ~86.8 million paper-to-paper citations identified within PKG. More comprehensive illustration of the original data source and our extension can be found from S3 in the Supplementary Information.

Model

Using a nonparametric Cox model and discrete-time event history model, we relate the annual diffusion indices for each subfield calculated across social and scientific spaces with an abrupt decline in the relevance of a given subfield, or "bubble burst," as illustrated in Fig. 1A. Formally, the discrete-time event history analysis model can be written as:

$$log(\frac{p_{ti}}{1-p_{ti}}) = \alpha D_{ti} + \beta x_{(t-1)i} \quad \text{···Eq. (1).}$$

p_{ti} denotes the probability of event happens at t for i,
D_{ti} is a vector for the function of the time duration by interval t with coefficients α,
x_{ti} is a vector for covariates (time varying and constant over time) with coefficients β.

Outcome Event: Bubble bursting

Our primary outcome of interest is the event of socio-epistemic bubbles bursting, an abrupt decline of the popularity of a given subfield, measured in the decline of citation counts as

illustrated in Fig. 1. Specifically, we time bubble bursts as when the standardized citation count difference of a given year from a subfield falls below extreme cutoffs within the life cycle of each subfield. This requires distinguishing subfields that experienced the deflationary 'bursting' event from those that did not. We do this by the following steps.

We first compute $\Delta_{i,t} = c_i(t) - c_i(t-2)$, where $c_i(t)$ is the citations that a subfield i garnered during year t across 1970 to 2019. Unlike Azoulay et al. 's (2019) work that uses publications indexed both in Web of Science and MEDLINE, we use all PMID to PMID citation links identified in PKG 2020 data to compute citation counts. Moreover, we include all MEDLINE indexed publications, even when MeSH terms or author disambiguated IDs are not assigned to them. Then, we standardize $\Delta_{i,t}$ within the life cycle of each subfield to make the $\Delta_{i,t}$ values comparable across 28,504 subfields. That is, we transform $\Delta_{i,t}$ to $z_{i,t}$ by subtracting the mean of $\Delta_{i,t}$, $\overline{\Delta_i} = \frac{1}{N}\sum \Delta_{i,t}$, from $\Delta_{i,t}$ and dividing it by the standard deviation of $\Delta_{i,t}$ computed within a subfield. By doing so, we obtain the distribution of the standardized two-year citation difference, $z_{i,t}$, across 28,504 subfields. The distribution of $z_{i,t}$ is presented in Fig. S1 in Supplementary Information.

We operationalize bubble bursts as when the standardized citation count difference of a given year from a subfield falls below given cutoffs such as 0.5%. 0.25, 0.1% of the distribution. Using $k = 2$ and the 0.5% cutoff (i.e., $z_{i,t} < -2.64$) identifies 5,699 subfields (20.0% of 28,504 subfields) that experienced a drastic collapse in collective scientific attention relative to other subfields. Applying the 0.25% ($z_{i,t} < -2.91$) and 0.1% ($z_{i,t} < -3.26$) cutoffs return 2,943 and 1,200 subfields with the bubble bursting events, respectively. When a subfield experiences a bubble burst more than once, we treat the earliest one within a subfield as the event of our interest. Fig. S2 in Supplementary Information shows an example of a subfield with a substantial

decline of attention based on a subfield captured by PMID 8392054. In this case, we denote that a deflationary event took place in 2002.

Key Indicator: Knowledge Diffusion

The key leading indicator for our analysis is *subfield level knowledge diffusion.* We measure the knowledge diffusion by calculating mean values of cosine distance (or 1 - cosine similarity) between focal papers included in a subfield and citing papers with two years windows in our scientific and social embedding space. We specifically distinguish temporal dynamics by separating citations observed within a subfield (Fig. 2) and that come beyond a given subfield when we compute time-varying variables to characterize the evolution of subfields. This consideration leads us to measure four diffusion indices: *1) Within Diffusion across Scientific Space (MeSH); 2) Outside Diffusion across Scientific Space (MeSH); 3) Within Diffusion across Social Space; 4) Outside Diffusion across Social Space.*

Further Characterization of Subfield Dynamics

A. Time Effect

- ***Subfield Age*:** The difference between calendar years and the year focal articles were published. We include subfield age dummies within each subfield. It is the same treatment from Azoulay et al. [15] but again the coverage is extended to the end of 2019.

- ***Calendar Year Fixed-Effect*:** Potential confounding effects of the calendar year from 1970 and 2019 are controlled using calendar year dummies.

B. Subfield Growth Pattern

- ***Cumulative Subfield Size*:** Number of articles captured in a subfield up to the given year, t. Formally, $CS_i(t) = \sum_{k=1}^{t} n_i(k)$, where $n_i(k)$ is the number of articles published in year k and captured in a subfield i. We take the logarithmic to address skewness of the variable.

- ***Two Rolling-Years Marginal Growth*:** The proportion of articles published in a given year and the past year divided by cumulative subfield size, which is $\frac{n_i(t)+n_i(t-1)}{CS_i(t)}$. This dynamically measures how actively a subfield grows.

C. Citation Dynamics

- ***Total Cumulative Citations***: Aggregated citation counts that publications capture in a subfield received until year t, $\sum_{k=1}^{t} c_i(k)$ where $c_i(k)$ is the citations a subfield receives during year k as defined earlier. The natural logarithm is taken to account for skewness of its distribution in the following analysis.

- ***Two-year Rolling Citation Counts, Within and Outside***: Citations a subfield accumulates during the given year t and the past year, $t-1$, in order to control for fluctuations in the size of attention paid a subfield. We keep the within-outside distinction parallel to knowledge diffusion indices described above and take the natural logarithm of the raw counts.

- ***Gini Coefficient of Citation Counts***: We include the Gini coefficient of citation counts to consider the degree of centralization on citation counts. The coefficient ranges from 0

(every article in a subfield receives the same number of citations) to 1 (a single article receives all citation attention). We compute Gini coefficients for 1) ***Total Cumulative Citations*** and 2) ***Two-year Rolling Citations*** for a subfield annually.

D. Other Controls

- ***Article Retraction Notification:*** Indicator that turns from 0 to 1 once a retraction notification is observed in a subfield to control the potential impact of experiencing a retraction event at the subfield level on overall attention to the subfield.

- ***Strata IDs (from Coarsened Exact Matching)***: This denotes 3,076 strata IDs based on subfields identified from publications by research superstars who prematurely died. The Strata IDs are assigned to matched subfields based on coarsened exacting matching.

- ***After Death (of Superstar Scientists)***: Indicator variable that switches from to 1 with the death of superstar scientists. The times of counterfactual death of elite scientists for subfields that did not lose superstars prematurely are assigned based on the initial coarsened exact matching procedure.

- ***After Death (of Superstar Scientists) * Subfields Associated with Premature Death of Superstar Scientists:*** The first term is as described above. The latter denotes an indicator variable to distinguish subfields associated with premature deaths of elite scientists from those not.

Acknowledgement: We express gratitude for funding from the Fetzer Franklin Fund in association with the 2019 MetaScience Symposium (RD, JE, DK), the Air Force Office of Scientific Research (AFOSR: FA9550-19-1-0354 and FA9550-15-1-0162) (JE, DK), and the National Science Foundation (NSF: 1829366 and 1800956) (JE, DK). The funders have/had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank the members of the Knowledge Lab at the University of Chicago for helpful discussions. This work was completed in part with resources provided by the University of Chicago's Research Computing Center. We also appreciate the support from Jian Xu and Ying Ding in facilitating access to PubMed Knowledge Graph.

Authors Contributions:

Conceptualization: DK, RD, JR, JE

Methodology: DK, JE

Visualization: DK

Funding acquisition: RD, JE

Writing – original draft: DK, JE

Writing – review & editing: DK, RD, JR, JE

Competing Interests: The authors declare no competing interests.

Data availability: This work uses PubMed Knowledge Graph *(13)* and the replication data from Azoulay and colleagues *(14)*. They are available publicly.

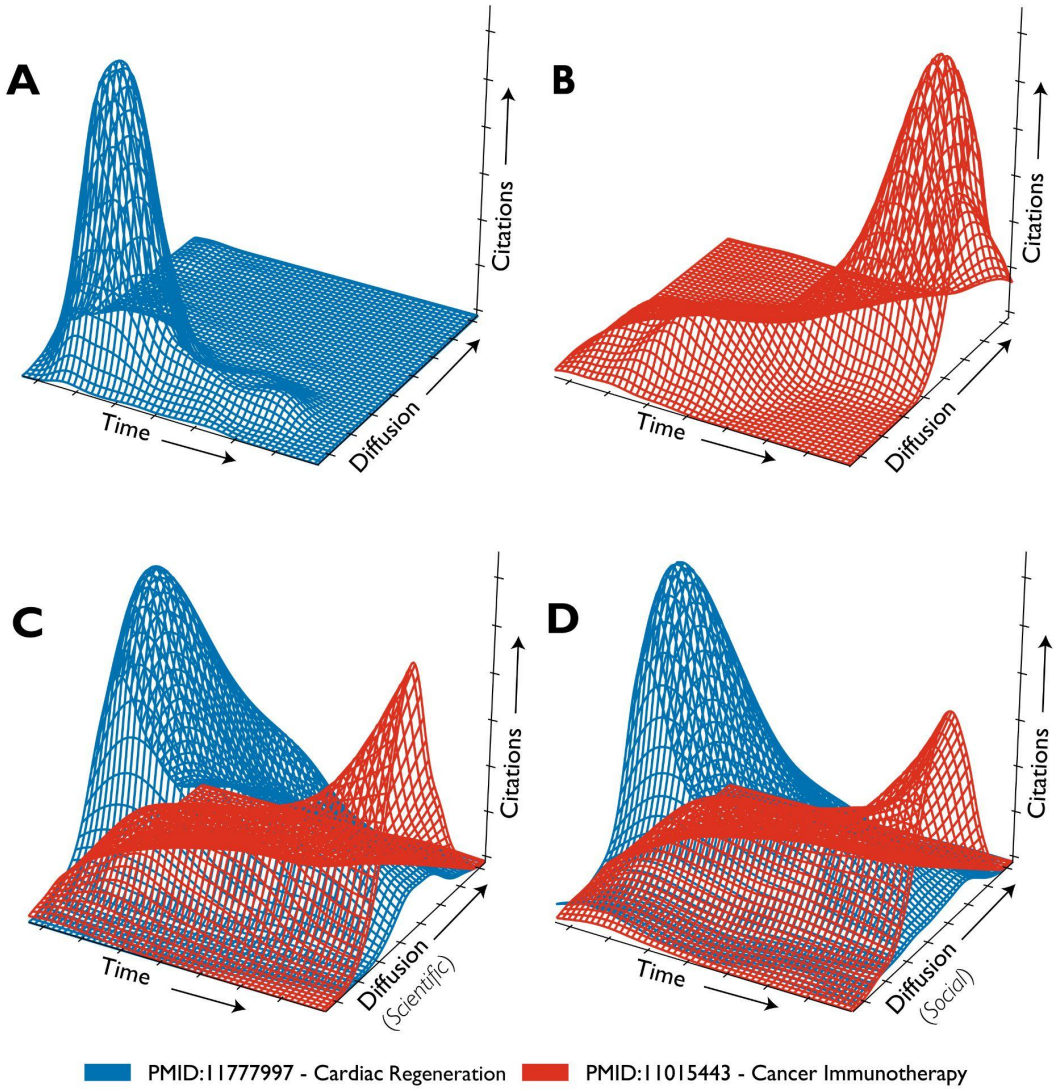

Fig. 1. Conceptual representation of field-level citations with different diffusion levels and examples of contrasting diffusion trajectories. (A) visualizes a hypothetical case where research attracts massive popularity early on but quickly loses attention by failing to diffuse broadly: subsequent studies are either concentrated in similar topics or conducted by research groups overlapping with or close to the original research team. (B) displays the inverted case where research does not receive immediate attention but steadily diffuses to distant research groups and topics and later garners sustained, outsized attention. (C) and **(D)** show 3D kernel density plots of diffusion indices and citations from PMID 11777997 (Cardiac Regeneration) and PMID 11015443 (Cancer Immunotherapy) across scientific and social space, respectively. The blue mesh refers to the 3D kernel density estimation based on the diffusion and citation count trajectories for PMID 11777997 (Cardiac Regeneration). The red mesh charts the 3D kernel density estimation from diffusion and citation count trajectories for PMID 11015443 (Cancer Immunotherapy). Starting years are aligned to zero for publication years associated with each article for comparison. The annual diffusion indices and citation counts are computed using a two-year rolling average.

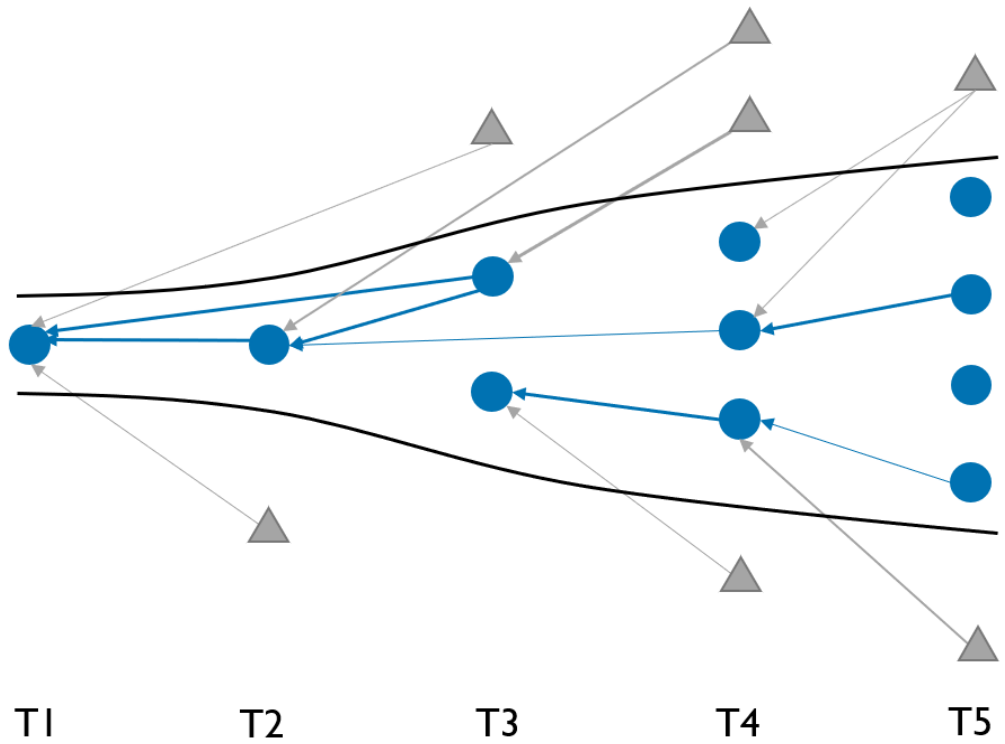

Fig. 2. Schematic representation of the within-outside subfield citation distinction. Blue dots represent publications captured as a similar article from a seed article thus constituting a subfield and gray triangles are publications citing at least one paper in the subfield. The thickness of the edges reflects the weights measured as the cosine similarity between publications, based on either the scientific or social embedding spaces.

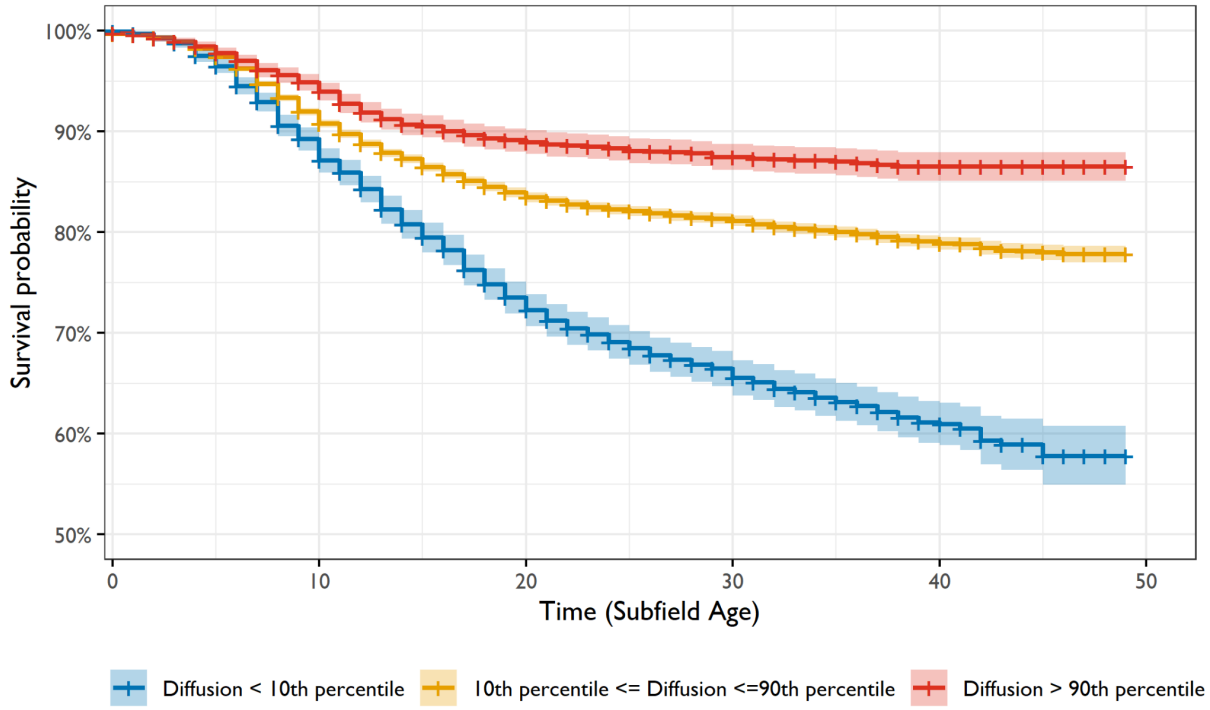

Fig. 3. Survival probability against bubble bursting as a function of knowledge diffusion outside-subfields in social space. Events are defined as a drastic decline of 2 year citation counts at the subfield level with 0.5% cutoff (Method). Survival refers to the converse, i.e., not experiencing a subfield-level extreme deflationary event. Subfield ages are set to 0 in the year when the focal seed article spanning a subfield was published. Diffusion percentile is ranked within calendar years and subfield ages. Bands depict 95% confidence intervals.

Table 1. Discrete-Time Event History Model Estimates

Dependent Variable	*Substantial Decline of Citations*		
	Estimate	**Std Error**	***z*-value**
Knowledge Diffusion			
Within, Scientific Space	-.011	.024	.445
Within, Social Space	-.065	.027	2.410
Outside, Scientific Space	-.118	.034	3.509
Outside, Social Space	-.165	.042	3.892
Subfield Growth Pattern			
Cumulative Subfield Size (logged)	.127	.084	1.509
Two-year Subfield Marginal Growth	-.225	.009	-26.452
Citation Dynamics			
Total Cumulative Citations (logged)	-2.113	.194	-10.915
Gini Coef. of Total Cumulative Citations	.009	.006	1.604
Two-year Citations, Within (logged)	.536	.033	16.373
Two-year Citations, Outside (logged)	2.129	.176	12.095
Gini Coef. of Two-year Citations	-.012	.005	-2.346
Other Controls			
Retraction Notice Published	.057	.174	.329
After Death	.105	.132	.796
After Death * Superstar Death Treatment	-.088	.065	-1.346
Total Observations	1,358,615		

Table 1. Model estimates with the bottom 0.5% cutoff for citation differences in the two-year rolling period. Coefficients for fixed effects of field age, calendar year, and strata ID dummies are omitted. Variables under *Knowledge Concentration*, *Subfield Growth Pattern*, *and Citation Dynamics* are all one-year lagged. The concentration indices are standardized within field ages and calendar years across 28,504 subfields. Standard errors are clustered with strata ID and calendar years.

References

1. Arthur, W. B. Complexity in economic and financial markets. *Complexity* **1**, 20–25 (1995).
2. Harras, G. & Sornette, D. How to grow a bubble: A model of myopic adapting agents. *J. Econ. Behav. Organ.* **80**, 137–152 (2011).
3. Fortunato, S. *et al.* Science of science. *Science* **359**, (2018).
4. Weis, J. W. & Jacobson, J. M. Learning on knowledge graph dynamics provides an early warning of impactful research. *Nat. Biotechnol.* **39**, 1300–1307 (2021).
5. Partha, D. & David, P. A. Toward a new economics of science. *Res. Policy* **23**, 487–521 (1994).
6. Petersen, A. M., Pan, R. K., Pammolli, F. & Fortunato, S. Methods to account for citation inflation in research evaluation. *Res. Policy* **48**, 1855–1865 (2019).
7. Hutchins, B. I., Yuan, X., Anderson, J. M. & Santangelo, G. M. Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. *PLoS Biol.* **14**, e1002541 (2016).
8. Kiss, I. Z., Broom, M., Craze, P. G. & Rafols, I. Can epidemic models describe the diffusion of topics across disciplines? *J. Informetr.* **4**, 74–82 (2010).
9. Bettencourt, L. M. A., Cintrón-Arias, A., Kaiser, D. I. & Castillo-Chávez, C. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. *Physica A: Statistical Mechanics and its Applications* **364**, 513–536 (2006).
10. Taylor, M. & Heath, B. Years after Brigham-Harvard scandal, U.S. pours millions into tainted stem-cell field. *Reuters* (2022).
11. Murry, C. E. *et al.* Haematopoietic stem cells do not transdifferentiate into cardiac myocytes in myocardial infarcts. *Nature* **428**, 664–668 (2004).

12. Osafune, K. *et al.* Marked differences in differentiation propensity among human embryonic stem cell lines. *Nat. Biotechnol.* **26**, 313–315 (2008).

13. Teplitskiy, M., Acuna, D., Elamrani-Raoult, A., Körding, K. & Evans, J. The sociology of scientific validity: How professional networks shape judgement in peer review. *Res. Policy* **47**, 1825–1841 (2018).

14. Belikov, A. V., Rzhetsky, A. & Evans, J. Prediction of robust scientific facts from literature. *Nat. Mach. Intell* **4**, 445–454 (2022).

15. Azoulay, P., Fons-Rosen, C. & Zivin, J. S. G. Does Science Advance One Funeral at a Time? *Am. Econ. Rev.* **109**, 2889–2920 (2019).

16. Quaini, F. *et al.* Chimerism of the Transplanted Heart. *N. Engl. J. Med* **346**, 5–15 (2002).

17. Chien, K. R. *et al.* Regenerating the field of cardiovascular cell therapy. *Nat. Biotechnol.* **37**, 232–237 (2019).

18. Freeman, G. J. *et al.* Engagement of the PD-1 immunoinhibitory receptor by a novel B7 family member leads to negative regulation of lymphocyte activation. *J. Exp. Med.* **192**, 1027–1034 (2000).

19. Le, Q. & Mikolov, T. Distributed Representations of Sentences and Documents. in *Proceedings of the 31st International Conference on Machine Learning* (eds. Xing, E. P. & Jebara, T.) vol. 32 1188–1196 (PMLR, 2014).

20. Xu, J. *et al.* Building a PubMed knowledge graph. *Sci Data* **7**, 205 (2020).

21. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv [cs.CL]* (2013).

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

23. Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E3635–E3644 (2018).

24. 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* 701–710 (Association for Computing Machinery, 2014).

25. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. *KDD* **2016**, 855–864 (2016).

26. Radim Rehurek, P. S. Software Framework for Topic Modelling with Large Corpora. in *IN PROCEEDINGS OF THE LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS* (2010).

27. Bourdieu, P. The specificity of the scientific field and the social conditions of the progress of reason. *Soc. Sci. Inf.* **14**, 19–47 (1975).

28. Foster, J. G., Rzhetsky, A. & Evans, J. A. Tradition and innovation in scientists' research strategies. *Am. Sociol. Rev.* **80**, 875–908 (2015).

29. Azoulay, P., Furman, J. L. & Murray, F. Retractions. *Rev. Econ. Stat.* **97**, 1118–1136 (2015).

30. Lin, J. & Wilbur, W. J. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinform.* **8**, 423 (2007).

SUPPLEMENTARY INFORMATION FOR

Limited Diffusion of Scientific Knowledge Forecasts Collapse

S1. Supplementary Figures and Tables

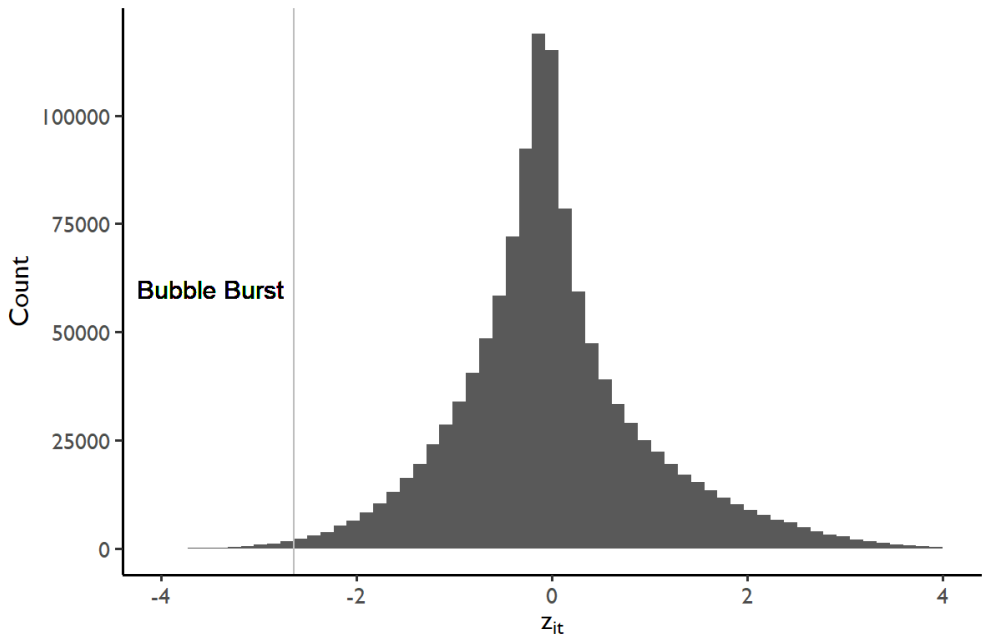

Fig. S1. Distribution of $z_{i,t}$ from 28,504 subfields. The cutoff value for bubble burst here is set to -2.64, the bottom 0.5% percentile.

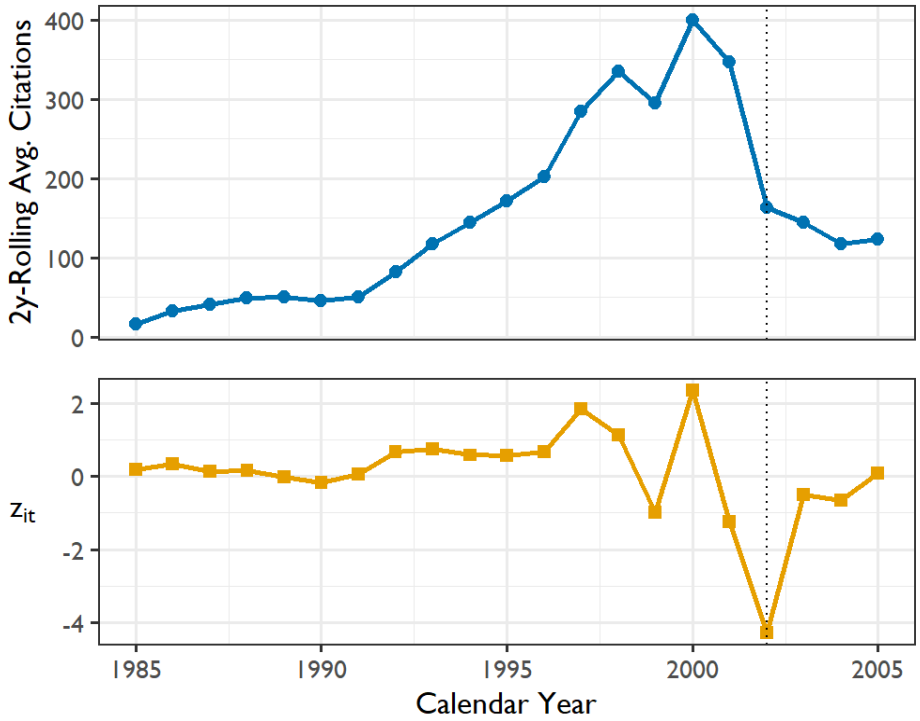

Fig. S2. Example of bubble bursting. Annual citation counts are aggregated at the subfield level, using forward citations to related publications captured based on PMID 8392054 published in 1993.

Dependent Variable	*Substantial Decline of Citations*		
	Estimate	**Std Error**	***z*-value**
Knowledge Diffusion			
Within, Scientific Space	-.034	.039	.857
Within, Social Space	-.061	.043	1.421
Outside, Scientific Space	-.099	.039	2.560
Outside, Social Space	-.216	.054	4.020
Subfield Growth Pattern			
Cumulative Subfield Size (logged)	.111	.123	0.902
Two-year Subfield Marginal Growth	-.251	.012	-21.624
Citation Dynamics			
Total Cumulative Citations (logged)	-2.256	.231	-9.782
Gini Coef. of Total Cumulative Citations	.009	.008	1.105
Two-year Citations, Within (logged)	.617	.046	13.527
Two-year Citations, Outside (logged)	2.261	.210	10.767
Gini Coef. of Two-year Citations	-.013	.007	-1.782
Other Controls			
Retraction Notice Published	-.138	.223	-.620
After Death	.227	.171	1.326
After Death * Superstar Death Treatment	-.105	.087	-1.208
Total Observations	1,420,986		

Table S1. Model estimates with the bottom 0.25% cutoff for citation differences in the two-year rolling period. Coefficients for fixed effects of field age, calendar year, and strata ID dummies are omitted. Variables under *Knowledge Concentration*, *Subfield Growth Pattern*, *and Citation Dynamics* are all one-year lagged. The concentration indices are standardized within field ages and calendar years across 28,504 subfields. Standard errors are clustered with strata ID and calendar years.

Dependent Variable	*Substantial Decline of Citations*		
	Estimate	**Std Error**	***z*-value**
Knowledge Diffusion			
Within, Scientific Space	-.089	.051	1.739
Within, Social Space	-.038	.085	.452
Outside, Scientific Space	-.058	.049	1.184
Outside, Social Space	-.279	.091	3.069
Subfield Growth Pattern			
Cumulative Subfield Size (logged)	.326	.167	1.949
Two-year Subfield Marginal Growth	-.288	.016	-18.091
Citation Dynamics			
Total Cumulative Citations (logged)	-2.594	.277	-9.37
Gini Coef. of Total Cumulative Citations	.012	.015	.830
Two-year Citations, Within (logged)	.671	.073	9.156
Two-year Citations, Outside (logged)	2.387	.263	9.074
Gini Coef. of Two-year Citations	-.009	.014	-.644
Other Controls			
Retraction Notice Published	.152	.340	.449
After Death	.597	.245	2.440
After Death * Superstar Death Treatment	-.046	.171	-.271
Total Observations	1,459,992		

Table S2. Model estimates with the bottom 0.1% cutoff for citation differences in the two-year rolling period. Coefficients for fixed effects of field age, calendar year, and strata ID dummies are omitted. Variables under *Knowledge Concentration*, *Subfield Growth Pattern*, *and Citation Dynamics* are all one-year lagged. The concentration indices are standardized within field ages and calendar years across 28,504 subfields. Standard errors are clustered with strata ID and calendar years.

S2. Measuring Knowledge Diffusion Through Document Embedding Spaces

We train vector representation models for biomedical science publications from PKG 2020 to locate positions of scientific publications based on their contents and measure the similarity/distance between papers linked through citation. We specifically adapt the Doc2vec model[1], a variant of the Word2vec model[2,3], initially developed to produce dense vector representations for documents or paragraphs from the words that compose them. Word embedding models generate a high-dimensional vector space in which geometrically proximate word vectors correspond to words that frequently share local linguistic contexts in the training data[2–4]. This approach has previously been extended to generate representational vectors for entities connected in networks by substituting connections among entities as shared contexts[5,6].

We treat that a research article can be characterized by 1) a list of MeSH terms and 2) researchers authoring it. Considering this, we build two separate representational vector spaces - "scientific space" and "social space". We use the Python Gensim package (version 4.0)[7] to train our vector representations. We choose to use the Distributed Bag of Words (DBOW) model, analogous to the skip-gram model from the Word2vec framework, simultaneously training the document vectors and constituting elements (MeSH terms and author IDs). Detailed implementation procedures are as follows.

S2.1 Scientific Space from MeSH Descriptors

We treat the MeSH terms as constituting words to build a "scientific space" for the biomedical literature. Because nominal terminologies are subject to change, we use MeSH terms' unique IDs from the National Library of Medicine. For instance, a MeSH descriptor, *Mesenchymal Stem Cells* (Descriptor ID: *D059630*), was indexed as *Mesenchymal Stromal Cells* from 2012 to 2018. However, it began to be reindexed as *Mesenchymal stem cells* in 2019, while its uniquely assigned descriptor ID, *D059630*, remains the same.

MeSH terms

- Antibodies, Monoclonal / therapeutic use
- Antineoplastic Agents / metabolism
- Antineoplastic Agents / pharmacology*
- Humans
- Immunologic Factors / therapeutic use*
- Immunotherapy*
- Neoplasms / therapy*
- Programmed Cell Death 1 Receptor / therapeutic use*
- Signal Transduction

Fig. S3. MeSH terms assigned to PMID 28376884

When a MeSH qualifier is attached to a MeSH descriptor, we consider both a descriptor with a qualifier and without it. Note Fig. S3. that displays MeSH terms assigned to "Cancer immunotherapies targeting the PD-1 signaling pathway" (PMID 28376884**),** published in *Journal of biomedical science* in 2017, authored by Iwai, Hamanishi, Chamoto, and Honjo[8]. The second term *Antineoplastic Agents / metabolism* can be broken down into the primary MeSH

descriptor, *Antineoplastic Agents,* and the qualifier, *metabolism*, narrowing down the scope. The third term *Antineoplastic Agents/pharmacology**, also has a qualifier, *pharmacology*. (The asterisk denotes that the given term is a major topic of the publication.) For this case, we include 1) *Antineoplastic Agents, 2) Antineoplastic Agents / metabolism,* and 3) *Antineoplastic Agents / pharmacology* for our model training. We do this to reflect that PubMed search queries using only MeSH terms (without qualifiers), *Antineoplastic Agents* for this case*,* capture publications like PMID 28376884. We exclude the asterisks for the same reason, taking into consideration co-searchability. As a result, the final list of MeSH terms fed into the training process for PMID 28376884 is *Antibodies, Monoclonal, Antibodies, Monoclonal / therapeutic use, Antineoplastic Agents, Antineoplastic Agents / metabolism, Antineoplastic Agents / pharmacology, Humans, Immunologic Factors, Immunologic Factors / therapeutic use, Immunotherapy, Neoplasms, Neoplasms / therapy, Programmed Cell Death 1 Receptor, Programmed Cell Death 1 Receptor / therapeutic use, Signal Transduction*.

With these MeSH combinations, we train 100-dimensional vectors for 26,666,615 PMIDs and 303,492 MeSH combinations that appear at least ten times with 100 training epochs. The mean number of MeSH terms (after the procedure detailed above) per PMID from our dataset is 16.34 (std=9.04). However, we set the sliding window size that defines the boundary of training context as 110, the maximum number from the data, to ensure that each training instance includes all the other MeSH combinations on a given article without splitting them up by imposing arbitrary contexts.

We validate the resulting vector representations by attempting to retrieve resulting publication vectors using MeSH combination vectors across 20 random samples, each containing 1,000 publications. We first take the vectors of MeSH terms assigned to each publication, infer the position of a document combining the MeSH terms, and check its proximity to the original vector representation of the article containing those MeSH terms. It is, for instance, a test if we can retrieve PMID 28376884 in Fig. S3. by inferring the position of a document combining the vectors of MeSH terms assigned to it. Because it is impossible to differentiate publications with the same set of MeSH terms with this model, we consider the 1, 5, and 10 most similar documents from the inferred vector, using cosine similarity. We find that it is possible to retrieve the target PMIDs with the rate of 92.48% (sd= .81), 96.14% (sd=.59), 97.18% (sd=.52) from the top 1, 5, 10 most similar documents, respectively, which suggests documents sharing MeSH terms are located close together in the 100-dimensional embedding space.

An advantage of using this Doc2vec model is that it reflects the high order proximity of constituting words beyond their direct co-occurrence in a context. Consider two documents, PMID 23142641, a review article titled “Challenges measuring cardiomyocyte renewal” published in 2013 [9] and PMID 11287958, an original research article “Bone marrow cells regenerate infarcted myocardium” published in 2001 [10]. The former review article cited the latter article. A simple but popular similarity metric would be the Jaccard coefficient ranges from 0 to 1, computed by dividing the number of MeSH terms that two articles share by the size of the union set of all MeSH terms assigned to the two publications.

The MeSH terms assigned to PMID 23142641 are *Animals; Bromodeoxyuridine; Cell Differentiation; Cell Nucleus / metaboli`m; Cell Nucleus / ultrastructure; Cell Proliferation; Cell Tracking; Genes, Reporter; Integrases; Mice; Mice, Transgenic; Myocardium / cytology*; Myocardium / metabolism; Myocytes, Cardiac / cytology*; Myocytes, Cardiac / metabolism; Regeneration; Stem Cells / cytology*; Stem Cells / metabolism; beta-Galactosidase*.

The MeSH terms assigned to PMID 11287958 are as follows: *Animals; Bone Marrow Transplantation*; Cell Differentiation; Connexin 43 / metabolism; DNA-Binding Proteins / metabolism; Female; Green Fluorescent Proteins; Ki-67 Antigen / metabolism; Luminescent Proteins / metabolism; MEF2 Transcription Factors; Male; Mice; Mice, Inbred C57BL; Mice, Transgenic; Myocardial Infarction / therapy*; Myocardium / cytology; Myocardium / pathology*; Myogenic Regulatory Factors; Proto-Oncogene Proteins c-kit / metabolism; Transcription Factors / metabolism*.

The Jaccard coefficient of the two publications based on the MeSH terms is .133 despite the close relationship between the two articles. However, the cosine similarity between the two documents based on our trained model is .844, much better reflecting the overall topic similarity between the two publications.

S2.2 Social Space with Disambiguated Author IDs

Analogous to the content embedding space from MeSH terms, we also build a 100-dimensional social embedding space using Doc2vec, anchored by 8,359,189 disambiguated biomedical authors within which we locate the vector space position of 28,329,992 PMIDs published by the end of 2019. In other words, we consider the author IDs as constituting document units. To inscribe the co-author information per publication, we included only authors that appeared more than once. The mean number of authors per publication from 28,329,992 PMIDs is 3.97 (std=5.01) with a median of 3. However, we set the window size for training context as 2000 – arbitrarily larger than the maximum number of authors in the dataset – to include all author IDs in the training process for a given publication. We do this to ensure that the resulting article embedding model assigns similar vectors to articles co-authored by the same groups of overlapping co-authors who are directly or indirectly close in the social space of biomedical research collaboration. We trained our social embedding space using 100 epochs (or training iterations).

We validate the quality of vector representations in the same manner we did for the MeSH content space across 20 random samples of 1,000 publications each. We take the author vectors for each publication, infer the position of a hypothetical publication those authors could have written within the 100-dimensional embedding space, and check its proximity to the vector representation written by the same author(s). Considering the impossibility to distinguish publications written by the same author(s), we also assess the 1, 5, 10, and 20 most similar PMIDs from the inferred vector using cosine similarity. The target PMIDs could be retrieved with the rate of 65.26% (sd= 1.73), 86.16% (sd=1.06), 90.27% (sd=0.74), 92.9% (sd=0.77) from the top 1, 5, 10, 20 most similar documents, respectively. The sharp increase in self-retrieval for relaxed conditions demonstrates that documents written by the same author(s) are contiguous in the resulting 100-dimensional social embedding space.

S2.3 Aggregated Temporal Pattern of Diffusion from Highly Cited Articles Published in 1980, 1990, 2000, 2010.

Here we provide an aggregate-level description of how our diffusion indices temporally evolve using highly cited papers (top 5% percent in citation counts by the end of 2019) from four cohorts of research articles published in 1980, 1990, 2000, and 2010. We first make subsets of publications that the raw citation obtained by the end of 2019 fall over the 5% percentile in each

cohort year (10,967 of 219,358 in 1970; 14,031 of 280,622 in 1980; 20,527 of 410,555 in 1990; 26,513 of 530,271 in 2000; 41,156 /823,129 in 2010), also accordingly extract cosine distances between the focal papers and citing papers measured in social and scientific space. With data from two rolling years, medians of cosine distances each calendar year from two spaces are computed. For example, the median cosine distance assigned to 1991 for the 1990 cohort is computed using all the citations observed in 1990 and 1991.

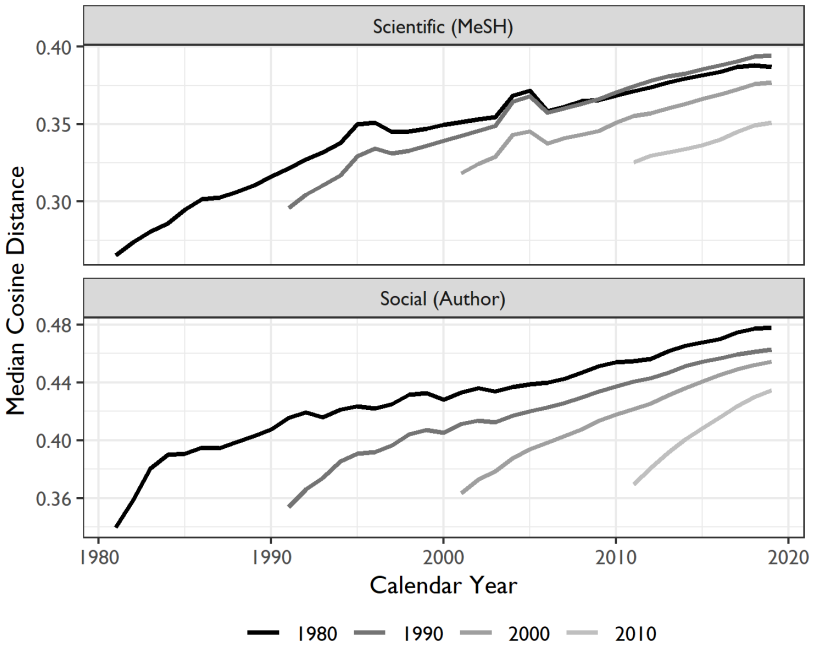

Fig. S4. Temporal pattern of diffusion from highly cited articles (Top 5%) published in 1980, 1990, 2000, 2010. Median cosine distances for each year (t) are computed based on a two-year rolling (t and t -1) period.

Fig. S4 shows the temporal evolution of diffusion metrics from scientific and social space. As the universe of biomedical entities and scientists expands, distances between focal papers and citing papers tend to increase in both scientific and social spaces by 2019. Then, the pattern, especially from the scientific space, indicates that our 100-dimensional representational spaces may allocate publications in some years (e.g., 2004 and 2005) in relatively distant locations within a trained manifold in the training process. Together, these suggest a necessity to consider the calendar year effect when a research article was published for the following analysis.

S3. *Delineating Biomedical Subfields* Measuring Knowledge Diffusion

Science is a social enterprise: like any other intellectual product, biomedical science obtains its meaning when others in the field recognize it and build upon each other[11–13]. In this sense, we attempt to understand the dynamics of diffusion and shifting attention beyond individual publications, at the subfield level defined as a group of biomedical science publications tightly related in terms of their research topics. This approach has been previously adopted in the context of studying the impact of publication retraction[14] and the consequences of premature death of elite life scientists[15] on subfields.

We harness the "Similar Article" (or "Related Articles") function from PubMed[16]. It allows us to capture a set of intellectually neighboring articles from a seed article using words in the abstracts, titles, and the MeSH terms, to identify research subfields in biomedical sciences. Our subfields identification hinges on 28,504 unique seed articles meticulously curated by Azoulay and their colleagues[15], consisting of research papers from U.S. elite life scientists published between 1970 and 2002 (inclusive). As the study attempts to estimate the effects of premature death of elite biomedical researchers, they first identified 3,076 seed articles authored by 452 researchers who died prematurely. Then, they collected articles published in the same year and the journals as the seed articles and conducted "coarsened exact matching" by keeping only publications from elite scientists who did not experience premature death and considering other factors such as the team sizes, ages of elites scientists, citation counts for seed articles[15]. We repurpose the existing dataset to study a different question from the original study.

Consistent with the original approach, we treat each seed article to represent distinct subfields in biomedicine, but we extended the period to the end of 2019 as their subfield panel data stopped in 2006. In total, we identify 1,941,680 unique publications (including the seed articles) constituting 28,504 subfields. By the end of 2019, the mean and median size of subfields is 122.52 and 102, respectively. We extract research articles that have cited any of 1,941,680 publications from the PKG 2020 citation database, which returns 11,421,194 publications and 86,804,637 paper-to-paper citations. Not all publications are associated with MeSH or Author IDs from PKG. We identify 10,894,779 publications that PKG assigns author IDs constituting 84,389,548 citations from social space; and 10,454,104 publications associated with MeSH terms linked through 82,228,828 citations.

Reference for Supplementary Information

1. Le, Q. & Mikolov, T. Distributed Representations of Sentences and Documents. in *Proceedings of the 31st International Conference on Machine Learning* (eds. Xing, E. P. & Jebara, T.) vol. 32 1188–1196 (PMLR, 2014).
2. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv [cs.CL]* (2013).
3. Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E3635–E3644 (2018).
4. Kozlowski, A. C., Taddy, M. & Evans, J. A. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *Am. Sociol. Rev.* **84**, 905–949 (2019).
5. 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* 701–710 (Association for Computing Machinery, 2014).
6. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. *KDD* **2016**, 855–864 (2016).
7. Radim Rehurek, P. S. Software Framework for Topic Modelling with Large Corpora. in *IN PROCEEDINGS OF THE LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS* (2010).
8. Iwai, Y., Hamanishi, J., Chamoto, K. & Honjo, T. Cancer immunotherapies targeting the PD-1 signaling pathway. *J. Biomed. Sci.* **24**, 26 (2017).
9. Soonpaa, M. H., Rubart, M. & Field, L. J. Challenges measuring cardiomyocyte renewal. *Biochim. Biophys. Acta* **1833**, 799–803 (2013).
10. Orlic, D. *et al.* Bone marrow cells regenerate infarcted myocardium. *Nature* **410**, 701–705 (2001).
11. Bourdieu, P. The specificity of the scientific field and the social conditions of the progress of reason. *Soc. Sci. Inf.* **14**, 19–47 (1975).
12. McMahan, P. & McFarland, D. A. Creative Destruction: The Structural Consequences of Scientific Curation. *Am. Sociol. Rev.* **86**, 341–376 (2021).
13. Crane, D. *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. (University of Chicago Press, 1972).
14. Azoulay, P., Furman, J. L. & Murray, F. Retractions. *Rev. Econ. Stat.* **97**, 1118–1136 (2015).
15. Azoulay, P., Fons-Rosen, C. & Zivin, J. S. G. Does Science Advance One Funeral at a Time? *Am. Econ. Rev.* **109**, 2889–2920 (2019).
16. Lin, J. & Wilbur, W. J. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinform.* **8**, 423 (2007).