The narrowing of literature use and the restricted mobility of papers in the sciences

Attila Varga

This study aims to answer the question of whether utilization of the literature narrows, and if so, how this tendency affects the circulation of ideas. If literature consumption focuses more on top papers, it may lead to inflexible scholarly communication, which, in turn, may hinder competition and may also affect research efficiency. This narrowed reception of papers implies that idea circulation is restricted and potentially useful findings are neglected. Scientists typically follow signals of reputation when they decide to consider new findings (1–3). Balancing out reputation-driven decisions with inclusivity is the key assurance that potential solutions, often discovered serendipitously, are not overlooked.

Aspects of scientific work are showing signs of resource concentration. The importance of collaborative work has increased (4) (SI Appendix, Fig. S1B) in parallel with an increasing share of publications going to the most productive scientists (5), which results in a higher proportion of citations received by the most highly cited authors (6). These influential scientists have a growing importance in the changing practices and knowledge flow at their institutions (5, 7, 8). Research teams deploy new organizational practices by employing an atypical, often temporary, academic workforce (9) and forming hierarchical leadership structures (10). These trends were foreseen some 50 y ago as a direct response to the growing complexity of research (11).

These developments raise the main question of this article: whether the attention to research papers is skewed toward the top of the citation distribution over time. If the research inputs (i.e., new publications) feed on a smaller proportion of research outputs (i.e., past papers), then this implies that the overall research productivity is, in this sense, declining. To put it simply, if the consumption of papers is skewed toward the top of the citation distribution, then this implies that the overall research productivity is, in this sense, declining. This is a long-standing trend, present since 1970. Citation concentration may lead to a less flexible scholarly communication system, whereby novel findings have a harder time resurfacing. In line with this argument, the popularity of a paper is more predictable today based on its previous citation impact. I present a mechanism of cumulative advantage that links increased concentration to the restricted mobility of papers. Citations are more dispersed from the 1990s, in the sense that papers are more likely to be cited at least a few times. Information technologies to disseminate papers are likely behind this latter trend.

Significance

A narrower range of literature is being read as citations are concentrating more on top papers. This is a long-standing trend, present since 1970. Citation concentration may lead to a less flexible scholarly communication system, whereby novel findings have a harder time resurfacing. In line with this argument, the popularity of a paper is more predictable today based on its previous citation impact. I present a mechanism of cumulative advantage that links increased concentration to the restricted mobility of papers. Citations are more dispersed from the 1990s, in the sense that papers are more likely to be cited at least a few times. Information technologies to disseminate papers are likely behind this latter trend.

Author affiliations: *Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47408

Author contributions: A.V. designed research, performed research, analyzed data, and wrote the paper.

The author declares no competing interest.

Author statement: A.V. designed research, performed research, analyzed data, and wrote the paper.

This article is a PNAS Direct Submission.

Copyright © 2022 the Author(s). Published by PNAS.

This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

1Email: atvarga@iu.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2117488119/-/DCSupplemental. Published April 18, 2022.
assumed that references concentrate on easily available online publications over time (19, 24). Larivière et al. (25) argued that the trend is actually less pronounced and, in fact, the literature has become more inclusive because the proportion of uncited papers is decreasing. In the past decades, several other papers have reported partial results on the topic, with no definitive conclusions (21, 26, 27).

In light of the aforementioned previous work, literature usage might be influenced by dual forces. Hypothetically, the continual development of information technology and bibliographic databases should counteract the postulated centralization tendency by providing easy access to a wider proportion of the literature over time. In this article, therefore, I also seek to answer questions about how citation dispersal potentially manifests itself.

**Results**

Citation concentration was measured for the references of all papers in a focal year made to papers published in the past 31 y, including the focal year. The data were retrieved from Web of Science’s Science Citation Index (SCI; Materials and Methods). As one concentration index, the Gini coefficient of the SCI has been calculated, which is arguably the most popular inequality index. If the coefficient is high, it means that the distribution is very unequal, or in other words, more concentrated at the top. Referenced papers having only one citation are omitted when the Gini coefficient is calculated. The reason is that, here, I focus on important papers with some scientific impact or, in other words, the tail of the citation distribution. In fact, non-cited papers do not follow the concentration trend, and their changing prevalence is presented later in this article in the context of citation dispersion.

The second approach to appraise concentration is to model specifically the tail of the citation distribution. This tail often follows the power law (28) and, in the current case, the function fits well (Materials and Methods) (SI Appendix, Fig. S2). The absolute value of the slope $\alpha$ of the power law informs us about the citation concentration: the smaller the value, the more the citation network is dominated by exceptionally high-impact papers.

Both the Gini coefficient and $\alpha$ measures follow the same trend (Fig. 1A); the concentration grows throughout the period, except in the mid-1990s, when the concentration falls back to the level of the mid-1980s, and then, by the mid-2000s, it resets to the initial trend. The field-level comparison of citation concentration reveals that this S shape predominantly affects the biological sciences, and the rest of the fields are relatively unaffected (SI Appendix, Fig. S3). Within the biological sciences, it is prevalent especially among fields that are strongly connected to biochemistry and genetics. Removing them substantially reduces this trend change (SI Appendix, Fig. S4).

Annualized citation distributions provide evidence for a more nuanced interpretation of the unfolding dynamics. These distributions offer a test of the question of whether citation concentration can be the result of a narrowed attention to papers of a certain age, as some authors have suggested (21). Each plot in Fig. 1B pertains to an age, which is the age of the references at the year indicated on the x-axis.

If the concentration could be explained only by a narrowing focus on younger papers (myopia) or old ones (e.g., because of slower forgetting), the plots in Fig. 1B should be flat. Generally speaking, entering cohorts from 1970 onward received a narrower focus on their top papers in most years. Then the S-shaped pattern appears in the trend from the mid-1990s until the papers become about 10 y old. Even for these younger cohorts, however, it is gradually fading out and becoming less noticeable. This finding shows that this pattern does not relate to some inherent quality of the cited cohort; instead, the S-shaped pattern is the result of the changing citation behavior of the year in which the references were made. Finally, I should note that old papers (14 to 16 y old) that still show up in the citation distribution are not affected by the concentration trend until 2000.

Considering the concentration trend, and especially that it increasingly affects new papers (Fig. 1B), how does this influence the flexibility of idea flow? Intuitively, the steeper the slope to the top, the harder it should be to climb for a currently less popular but novel knowledge claim. In the present framework, the flexibility of idea circulation is conceptualized as the predictability of papers’ future impact ranking based on their past ranking. If the importance of papers can change as the research program unfolds, the field is flexible, and if the ranking is more stable over time, the field is less flexible. In the latter case, the field is more focused on exploiting established findings. I tested the association between paper rankings at two time intervals (Fig. 2A). Papers typically reach their citation peak 2 y after publication. The two comparisons contrast the early citation impact with the second-year citation frequency peak and the second-year performance to the long-term impact. Both measures of ranking similarity show the same increasing trend (Fig. 2B). Fields exhibit similar trends to the aggregated results (SI Appendix, Fig. S5). The characteristic S-shaped trend change is clearly noticeable on the early year’s comparison, and it is tied again to the biological sciences.
The explanation I suggest herein relates this restricted mobility to the trend of concentrated citation activity in the tail. In short, the higher concentration of citations at a given moment reduces the likelihood that the paper ranking would change in the future. First, note that while the “rich-get-richer” process is responsible for the general evolution of inequality in networks, what needs to be explained is why the mobility of papers in the citation network, induced by the progress of research, is constrained. Consider the growth of the citation network. The citations on a set of papers published in the same year grow by the entering of new publications that distribute their citations based on a preferential attachment process. The latter means that the papers receive new citations proportional to their already accumulated citations: this is the rich-get-richer process (1, 29). If the accumulated citation distribution is more concentrated at the top at a certain moment, it is less likely that two papers will switch ranking when they receive new citations in the future. This is because their impact is farther away in a more unequal setting.

The detailed argument in terms of random network growth can be found in SI Appendix, and it is summarized as follows. In the random network, a paper cohort has cumulated an initial citation distribution \( p_k \) and, after preferential attachment, they received new citations distributed as \( p_k \). While in a real citation network, mobility would happen due to new evidence in the field, in the random model it emerges as random variation around the expected number of new citations based on preferential attachment. When the inducing past impact \( p_k \) is distributed as a power law or follows a lognormal function, the fatter the tail of \( p_k \) the more the variation in \( p_k \) is explained by \( p_k \), which determines the preferential attachment process, and it is less likely to be the result of random fluctuations. See Materials and Methods for the details of a simulation study of this process; Fig. 2C presents the results of the simulations.

In the final section, I investigate the dual aspect of citation concentration, citation dispersion. Here, I demonstrate that while citations concentrate in the tail of the citation distribution, the proportion of papers with at least a few citations also increase. Both of these tendencies fit into the more general trend of citation inflation, which means that the papers send out and receive more and more citations over time (21, 30). The average number of references made by papers in a given year grows exponentially in this dataset, with a 2.9% growth rate (SI Appendix, Fig. S1C). Note, however, that only references to journals in the SCI are counted, as reported in Materials and Methods.

Fig. 3A shows that the increased reference activity does not affect the prevalence of low- and high-impact papers in the same way throughout the period. In the first half of the studied period, the proportion of papers receiving at least a few citations (≥1 to 5) slightly declines while the proportion of high-impact papers is increasing. From around the mid-1990s, the proportion of papers having at least a few citations starts to increase dramatically. The proportion of higher-impact papers increases at a higher rate from this date as well. In fact, if noncited papers are also included in the calculation of the Gini coefficient, by the mid-1990s, the value of the coefficient begins to steeply decrease (SI Appendix, Fig. S6). In conclusion, the usage of scholarly literature follows dual influences from the mid-1990s. On the one hand, citation behavior is more inclusive by referencing a larger proportion of papers. On the other hand, the bulk of attention increasingly falls on the top papers, which is the continuation of a trend started decades before.

Where do the extra citations fall over time in terms of age of the referenced papers? This question is motivated by the observation that the average age of the references is increasing (14, 16) (SI Appendix, Fig. S1D). Does all the growth distribute evenly to older papers? SI Appendix, Fig. S7 illustrates the computation and fit of the summary statistics for the age distributions of cited references in the given year. These distributions have two components. The first component is recency, which is the proportion of fresh papers among the references. This includes the first 3 y until the citation peak at age 2 y. The second component is decay, which is the rate of abandoning the older literature after the second-year peak. To quantify this decay, I use the exponential function \( p_t = p_0 e^{-\alpha t} \).

Overall, the popularity of older papers is increasing (indicated by the slowing rate of decay); however, later in the period, recently published papers are becoming more popular (Fig. 3B). From the 1990s, the relative citation rate to the freshest papers published in the same year in which the

PNAS 2022 Vol. 119 No. 17 e2117488119 3 of 5
https://doi.org/10.1073/pnas.2117488119
the SEs of the coefficients. Lines showing recency indicate the proportions of papers among the references with the corresponding ages. To ease comparison, the scales of these proportions are normalized with the percentages at the beginning of the period. Shadings around the line representing recency are the SEs of the coefficients.

Fig. 3. Citation (cit.) dispersion and aging. (A) Percentage of papers from the past 30 y (relative to the given year on the x-axis) that received at least a certain number of citations. The title of each figure shows the citation threshold. The shaded year is 1995. The dashed lines running from 1970 to 1995 are to guide the eye. They are the results of fitting ordinary least squares linear regression models for that period. (B) Recency and decay constants. Lines showing recency indicate the proportions of papers among the references with the corresponding ages. To ease comparison, the scales of these proportions are normalized with the percentages at the beginning of the period. Shadings around the line representing recency are the SEs of the coefficients.

reference was made (0th y) starts to substantially increase. Later in the 2010s, 1- and 2-y-old papers constitute a larger proportion of the total references as well. Some of these tendencies vary at the disciplinary level. Not all disciplines exhibit a continuously slowing decay, as many fields recover from that trend in the 1990s and 2000s (SI Appendix, Fig. S8). However, the latter observation about recency (i.e., increased proportion of 0th-y papers) is generalizable across disciplines.

In summary, from the start, the tail of the citation distribution inflates, which leads to higher citation inequality. However, from the 1990s, the bottom of the distribution also shows inflation. In terms of age, inflation favors older papers, but again, from around the 1990s, recently published material starts to increase its share. Given the nature and the timing of these trend changes, one could speculate that improved scholarly communication technologies are behind these developments. Perhaps dispersion began to occur because the search space for literature enlarged and it is now easier to learn about less popular, albeit personally interesting, publications. Regarding the recency trend, it is possible that easier access to the most up-to-date findings, coupled with a motivation to demonstrate the timeliness of one’s own work, boosted the prevalence of references to fresh findings. Citations, to some extent, are rhetorical devices used by researchers to support their argument and signal the importance of their findings (31–33). Citing fresh papers could signal the timeliness and, therefore, the importance of a paper.

Discussion

At the general level, the main finding of this paper—that citation concentration is increasing—is in line with arguments that, over time, the scientific workforce tends to concentrate its attention on fewer problems relative to its size. Rewards in the sciences are distributed very unevenly in general. If the trend is that more papers are delegated to one problem, it is expected that the resultant enlarged citation volume would fall exceedingly to the top papers. This profusion of research effort could be the consequence of a newly emerging production system in which a larger and more diverse group of scientists is delegated to deal with more demanding research tasks in a coordinated fashion (22). Several attributes of this system, such as interdisciplinarity, shrinking distances in the knowledge network, and more collaboration and concentration, are, indeed, characteristic developments of the current scientific landscape (5, 7, 8, 10, 23, 26). On the other hand, it is also possible that concentration is simply due to an increased pool of researchers, due to the expansion of research institutions globally or an increased temporary workforce boosting paper production (6, 9). I am unable to disentangle these two factors, which are probably related to field-specific funding and innovation cycles (34) and require more focused research.

Some argue that the average scientific work has only a negligible contribution to scientific progress (35). Concentration, from this perspective, could be considered beneficial, because it sorts out important findings faster. The faster adoption of new literature in the past 20 y or so and the faster selection of top papers can then alternatively be interpreted as signs of accelerating information flow and consensus formation within the sciences, which eventually leads to faster progress. Here, I present evidence that these changes could lead to a more deterministic research-agenda evolution in scientific fields by demonstrating that citation concentration and the restricted mobility of papers work hand in hand. It is open to interpretation whether these empirical findings and model suggest that communication is more effective, or whether this is an indicator of a waning capability of the scientific community to remain open to novel findings.

It is worth noting that the trend switch in concentration from the mid-1990s coincided with the accelerated dispersion of citation activity. This accelerated dispersal of citations in the 1990s reversed the tendency of concentration only briefly. Because it emerged within fields related to biochemistry and genetics, it is also perhaps associated with the Human Genome Project, which began to report results in the mid-1990s, leading to cheaper instrumentation for gene sequencing to widen the spectrum of biomedical research (36).

Given the periodization of the studied trends, information technology has likely influenced citation dispersal. Electronic publishing became an important factor in the 1990s (37), and that is when recency and the widening of literature use really start to show themselves as robust trends. It is also possible that the per capita publishing-productivity increase (driven by more coauthorships) (38) might have boosted citation activity in general and contributed to the decrease in noncited papers: as scientists contribute to more papers, there is more opportunity for self-citation. Finally, I would like to point out that while citation concentration predates electronic publishing, most likely information technology influenced this trend as well: in the past two decades, recommender systems utilizing citation counts have possibly exacerbated concentration.

Materials and Methods

Data. The bibliometric data used for the analysis contain journal articles from the Web of Science’s current SCI collection. The analyzed data contain only the citations between articles that were published by the journals indexed by SCI.
The first year of the analysis is 1970. The reason for choosing 1970 as the start date is that peer review, which fundamentally affects citation behavior, became a dominant practice during the 1960s (16). SI Appendix, Fig. S1 shows the annual number of articles and several other descriptive statistics of the data. The number of papers grows exponentially by a rate of 4%.

The primary focus of the analysis is the annual references to research articles. I delimited these references by taking a 31-y window: only those papers that are no older than 30 y old at the time they are referenced are involved in the analysis. The papers that are published in the focal year are in the 0th y. Altogether, the window comprises 31 y. Imposing such a limit has the benefit that papers that were published centuries ago do not act as statistical outliers. It also ameliorates the problem that, over time, the chance of citations for older papers may grow disproportionately (39).

Papers are classified into disciplines. This classification is based on Web of Science’s subject categories. Presently, there are 179 such categories in the SCI. For the analysis, I compiled the categories into larger fields or disciplines, based on the Integrated Postsecondary Education Data System Completions Survey (IPEDS) conducted by the National Center for Education Statistics. It contains 18 relevant categories; therefore, I made changes as follows to reduce that number: 1) All the engineering fields are under Engineering; 2) astronomy and astrophysics are categorized under Physics; and 3) I omitted mathematics from the field-level analysis, because it is a very small field and quite distinct from the rest of the disciplines. These changes resulted in a more manageable eight categories. Furthermore, I relabeled the IPEDS category “Earth, atmospheric, and ocean sciences” to the shorter label “Environmental sciences.”

**Fitting Power Laws.** To fit the power laws to the citation distributions, I used the Python package powerlaw (40) (SI Appendix, Fig. S2). The minimum number of citations for the fit in every focal year was set to 10 for consistent measurement. As an alternative candidate distribution, I compared the fit of the power law with lognormal distribution. The likelihood ratio tests, comparing the fit of the two candidate distributions, were not statistically significant in any cases (P ≥ 0.05). However, in 80% of the years, the likelihood ratio statistic indicated that the power law fits better.

**Simulation Study of Restricted Mobility.** This model simulates the attachment of a new cohort of papers to a set of earlier published papers. The model has the same characteristics that are presented in SI Appendix. The old cohort has 100,000 nodes and the new cohort has 104,000, which corresponds to a 4% growth rate estimated from data (SI Appendix, Fig. S1). The distribution of the references of the new cohort is fixed across all the experiments, and it is a lognormal distribution (μ = 1.2, σ = 0.4) with a mean of approximately 3.6 references.

When these new nodes choose references, they follow linear preferential attachment. This preferential attachment is based on the initial citation impact distribution of the old, or focal, cohort, and it is denoted as ρi. ρi represents the past cumulative citations received by the old cohort up to this point when the new cohort enters. When the reference choices are made in the model, the new nodes select old papers only once. In other words, multiple lines are forbidden in this network. This is achieved by repeating the random choice if multiple lines occur. ρi varies across the experiments. In one set of simulations, it is a lognormal distribution, and in another set, it is a power-law distribution. The tails of these distributions are varied by manipulating the appropriate parameters of these distributions. In case of the lognormal distribution, it is the shape parameter σ, and in case of the power law, it is the slope α. See Fig. 3C in the main text for the actual values of these parameters and the results. The simulation was repeated 100 times for each parameter value.

**Data Availability.** The data used in the study are partially or fully available with a subscription to Web of Science. All other data are included in the manuscript and/or SI Appendix.

**ACKNOWLEDGMENTS.** This article is based upon work supported by the Air Force Office of Scientific Research under Award FA9550-19-1-0391. I thank Yong-Yeol Ahn, Staša Milojević, and Sadamori Kojaku for their help.

---

1. R. K. Merton, *The Matthew effect in science*. The reward and communication systems of science are dominant practice during the 1960s (16).
2. Y. Xie, *Sociology of science: “Undomesticated”. Inequalities in science.* Science 344, 809–810 (2014).
3. A. M. Petersen et al., *Reputation and impact in academic careers*. Proc. Natl. Acad. Sci. U.S.A. 111, 15316–15321 (2014).
4. S. Wuchty, B. J. Jones, B. Uzzi, The increasing dominance of teams in production of knowledge. Science 316, 1036–1039 (2007).
5. A. A. Gargawal, J. Michalec, A. Detti, “Collaboration, stars, and the changing organization of science evidence from evolutionary biology” in The Changing Frontier: Rethinking Science and Innovation Policy, A. B. Jaffe, B. F. Jones, Eds. (The University of Chicago Press, 2013), pp. 75–102.
6. M. W. Nielsen, J. P. Andersen. Global citation inequality is on the rise. Proc. Natl. Acad. Sci. U.S.A. 118, e2020128118 (2021).
7. C. M. Rawlings, D. A. McFarland, L. Dahlander, D. Wang. Streams of thought: Knowledge flows and intellectual cohesion in a multidisciplinary era. Soc. Forces 93, 1687–1722 (2015).
8. T. Stark, J. Rambaran, D. McFarland, The meeting of minds: Forging social and intellectual networks within universities. Social. Sci. 7, 433–446 (2020).
9. S. Milojčić, F. Radjčić, J. P. Walsh, Changing demographics of scientific careers: The rise of the temporary workforce. Proc. Natl. Acad. Sci. U.S.A. 115, 12616–12623 (2018).
10. J. P. Walsh, Y. N. Lee, The bureaucratization of science. Res. Policy 44, 1584–1600 (2015).
11. D. J. de Solla Price, *Little Science, Big Science* (Columbia University Press, 1963).
12. S. Milojčić, Quantifying the cognitive extent of science. J. Informetrics 9, 962–973 (2015).
13. N. Bloom, C. J. Jones, J. Van Reenen, M. Webb, Are ideas getting harder to find? Am. Econ. Rev. 110, 1104–1144 (2020).
14. V. Lavière, É. Archambault, Y. Gingras, Long-term variations in the aging of scientific literature. From exponential growth to steady-state science (1900-2004). J. Am. Soc. Inf. Technol. 59, 288-296 (2008).
15. S. Fortunato, Prizes: Growing time lag threatens Nobels. Nature 508, 186 (2014).
16. R. Sinatra, P. Deville, M. Szell, D. Wang, A.-L. Barabási, A century of physics. Nat. Phys. 11, 791–796 (2015).
17. P. O. Larsen, M. von Ins, The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 84, 575–603 (2010).
18. L. Bornmann, R. Mutz, Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. J. Assoc. Inf. Sci. Technol. 66, 2215–2222 (2015).
19. J. A. Evans, Electronic publication and the narrowing of science and scholarship. Science 321, 395–399 (2008).
20. P. D. B. Parolo et al., Attention decay in science. J. Informetrics 9, 734–745 (2015).
21. R. K. Pan, A. M. Petersen, F. Pammolli, S. Fortunato, The memory of science: Inflation, myopia, and the knowledge network. J. Informetrics 12, 655–678 (2018).