Too Many Papers? Slowed Canonical Progress in Large Fields of Science

Johan S. G. Chu (johan.chu@chicagobooth.edu)

James A. Evans (jevans@uchicago.edu)

University of Chicago

For SocArxiv. March 1, 2018

Abstract

We argue that paradigmatic progress may be slowed as scientific fields grow large. This assertion is supported by evidence from citation patterns across 251 fields—over 1 billion citations among 57 million papers over 54 years—covered by the Web of Science dataset. A deluge of papers in a scientific field does not lead to quick turnover of central ideas, but rather to the ossification of canon.

A straightforward view of scientific progress would suggest that more is better. The more papers published in a field, the greater the rate of scientific progress; the more researchers, the more ground covered. Even if not every article is earth-shaking in its impact, each can contribute a metaphorical grain of sand to the sandpile, increasing the probability of an avalanche, wherein the scientific landscape is reconfigured and new paradigms arise to structure inquiry (1, 2). Also, with more papers, the probability at least one of them contains an important innovation increases. A disruptive new idea can destabilize the status quo, siphoning attention from previous work and garnering the lion's share of new citations (3).

This more-is-better view is reflected in policy. Scholars are evaluated and rewarded based on productivity; publishing a large number of articles within a set period of time is the surest path to tenure and promotion. Quantity is the measuring stick at the university (4) and the national levels (5) also, where comparisons focus on the total number of publications, patents, and scientists, and the amount of spending.

When assessed in addition to quantity, quality is predominantly judged by number of citations. Citation counts are used to measure the importance of individuals (6), teams (7), and journals (8) within a field. At the paper level, the assumption is that the best and most valuable papers will attract more attention, shaping the research trajectory of the field (9). While some papers will garner early recognition, others may take longer to become widely cited (10). Whether immediate or delayed, the underlying process of citation accumulation is one of preferential attachment, where the number of previous citations to a paper is a good predictor of future citations (11).

When the quantity of papers grows very large, however, the sheer size of the body of

knowledge in the field can limit the number of citations for less-established papers—even those with novel and useful ideas. A deluge of new publications, rather than causing faster turnover of field paradigms, may entrench top-cited papers, precluding less-cited papers from rising into the most-cited, commonly-known canon of the field.

There are two ways this could happen (12). First, when many papers are published within a short period of time, scholars are forced to resort to heuristics to make continued sense of the field. Rather than encountering and considering intriguing new ideas each on their own merits, cognitively-overloaded reviewers and readers process new work only in relationship to existing exemplars (13–15). A novel idea that does not fit within extant schemas will be less likely to be published, read, or cited. Faced with this dynamic, authors are pushed to frame their work firmly in relationship to well-known papers, which serve as “intellectual badges” (16) identifying how the new work is to be understood, and discouraged from working on too-novel ideas that cannot be easily related to existing canon. The probabilities of a novel idea being produced, published, and widely-read all decline, and indeed, the publication of each new paper adds to the citations for the already most-cited papers.

Second, if the arrival rate of new ideas is too fast, competition between new ideas may prevent any of the new ideas from becoming known and accepted field-wide. To see why this is so, consider a sandpile model of idea spread in a field. When sand is dropped on a sandpile slowly, one grain at a time, waiting for movement on the sandpile to stop before dropping the next grain of sand, the sandpile over time reaches a scale-free critical state wherein one dropped grain of sand can trigger an avalanche over the whole area of the sandpile (2). But when sand is dropped at a rapid rate, neighboring mini-avalanches interfere with each other, and no individual grain of sand can trigger pile-wide shifts. The faster the rate of sand dropping the smaller the

domain each new grain of sand can affect (17). If the publication rate of novel papers is too fast, no new paper can rise into canon through localized processes of diffusion.

The arguments above yield four empirical predictions, each of which are borne out in citation patterns from the Web of Science: Compared to when a field has few publications each year, when a field has many new publications in a year, 1) the list of most-cited papers will change little year to year, 2) new papers will be more likely to cite the most-cited papers rather than less-cited papers, 3) the probability a new paper eventually becomes canon will be small, and 4) localized diffusion and preferential attachment will not explain the rise of a new paper into the ranks of the most-cited.

The first row of Fig. 1 presents correlations between the most-cited papers in a year and in the previous year by field, with each dot representing a field-year. The y-value for blue dots is the Spearman rank correlation between the two field-years, while that for red dots is the proportion of top-50 most-cited papers from the previous year remaining in the top 50 in the current year. The x-axis is logged (base 10) number of papers published in the focal year in the field. The pattern is consistent when looking at data across all fields and at individual large fields separately: When the number of papers published is large, change in the list of most-cited papers shrinks.

The second row shows that the most-cited papers gain a larger advantage in number of citations over less-cited papers when the number of papers published in a year was large. The y-axis is the decay rate of number of citations from the previous year. A data point with decay rate of 0.5 indicates that, on average, a paper will receive half the number of citations this year as it did last year. A decay rate of 1 or above indicates a paper's citations will remain steady or increase from the previous year. In years where few papers are published, the decay rate for the

most-cited papers (the blue line represents papers within the top 1%, the red line papers within the top 1-5%) is significantly below 1 and not much different from less-cited papers. When the number of papers published is large, however, the decay rate of citations for the most-cited papers is close to 1 and is significantly higher than that of less-cited papers.

The probability of a paper rising into the top 0.1% most widely-cited in the field for any subsequent year in the dataset shrank when it was published in the same year as many others. This was true cross-sectionally across fields in the same year (All panel), and across years in individual fields (A-E panels).

When a paper did rise into the top 0.1%, it took longer when the field was small—suggesting a slow climb through local diffusion—and much less time when the field was large. In large fields, papers did not become widely-cited by preferential attachment accumulation of citations. They instead jumped into the top 0.1%.

These findings suggest troubling implications for the current direction of science. If too many papers are published in short order, new ideas cannot be carefully considered against old, and processes of cumulative advantage cannot work to select valuable innovations. The more-is-better, quantity metric-driven nature of today's scientific enterprise may ironically be retarding fundamental progress in the largest scientific fields. Proliferation of journals and the blurring of journal hierarchies due to online article-level access can exacerbate this problem.

The current study is at the level of fields and large subfields, and progress may now occur at lower sub-disciplinary levels. To examine lower levels requires more precise methods for classifying papers, perhaps using temporal network community detection, than are available to us at the moment. But it is worth noting that the fields and subfields identified in the Web of Science correspond closely to real-world self-classifications of journals and departments.

Established scholars transmit their cognitive view of the world to their students via reading lists and syllabi, and field boundaries are enforced through career considerations.

It may be that progress still occurs, even though the most-cited articles don't change. While the most-cited article in molecular biology (*18*; published in 1976) has not changed since 1982, one would be hard-pressed to say that the field has been stagnant. But recent evidence (*19*) suggests that much more research effort and money are now required to produce similar scientific gains—productivity is declining precipitously. Could we be missing fertile new paradigms because we are locked into over-worked areas of study?

References and Notes

1. T. S. Kuhn, *The Structure of Scientific Revolutions*, 2nd ed. (Univ. of Chicago Press, Chicago, 1970).
2. P. Bak, C. Tang, K. Wiesenfeld, *Phys. Rev. Lett.* **59**, 381–384 (1987).
3. R. J. Funk, J. Owen-Smith, *Manage. Sci.* **63**, 791–817 (2017).
4. C. Baden-Fuller, F. Ravazzolo, T. Schweizer, *Long Range Plann.* **33**, 621–650 (2000).
5. Scientific American, *The World's Best Countries in Science* (2017; <https://www.scientificamerican.com/article/the-worlds-best-countries-science/>).
6. S. Alonso, F. J. Cabrerizo, E. Herrera-Viedma, F. Herrera, *Scientometrics* **82**, 391–400 (2010).
7. B. F. Jones, S. Wuchty, B. Uzzi, *Science* **322**, 1259–1262 (2008).
8. G. F. Davis, *Admin. Sci. Quart.* **59**, 193–201 (2014).
9. J. G. Foster, A. Rzhetsky, J. A. Evans, *Am. Sociol. Rev.* **80**, 875–908 (2015).
10. Q. Ke, E. Ferrara, F. Radicchi, A. Flammini, *Proc. Nat. Acad. Sci. U.S.A.* **112**, 7426–7431 (2015).
11. D. Wang, C. Song, A.-L. Barabási, *Science* **342**, 127–132 (2013).
12. J. S. G. Chu, “A theory of durable dominance” (Univ. of Chicago working paper, 2017).
13. A. Tversky, D. Kahneman, *Science* **185**, 1124–1131 (1974).
14. B. Schwartz, *The Paradox of Choice: Why More is Less* (Harper Collins, New York, 2004).
15. E. W. Zuckerman, *Am. J. Sociol.* **104**, 1398–1438 (1999).
16. A. L. Stinchcombe, *Am. Sociol.* **17**, 2–11 (1982).
17. C. Adami, J. Chu, *Phys. Rev. E* **66**, 011907 (2002).
18. M. M. Bradford, *Anal. Biochem.* **72**, 248–254 (1976).
19. N. Bloom, C. I. Jones, J. Van Reenen, M. Webb, “Are ideas getting harder to find?” (Stanford Univ. working paper, 2017).

FIGURE 1

A. Engineering, Electrical & Electronic B. Biochemistry & Molecular Biology C. Oncology D. Computer Science, Theory & Methods E. Mathematics

Fig. 1. Changes in citation dynamics by size of field.

(1st row) Two types of top-50 rank correlations between adjacent years. For (All) panel, each blue dot corresponds to a subject-year (the Web of Science classifies academic fields, or in some cases, large subfields, into what it terms *subjects*) in the dataset, with the y-position indicating Spearman rank correlation (S) of the top-50 most-cited list of the previous year to their rank in the focal year, and the x-position indicating the logged number of articles (N) published in the subject in the focal year. The blue shaded region is the 95% confidence interval from a linear regression of S on $\log N$. The red-shaded region is the 95% confidence interval of a linear regression of the retention rate (R)—the number of papers from the previous year’s top-50 remaining in the top-50 in the focal year—regressed on $\log N$. Blue dots in panels (A) through (E) indicate S and N for each year for one field each, red dots indicate R and N . The pattern is consistent across panels; churn in the most-cited articles decreases as the number of articles published per year increases. Panel regressions with fixed effects for subject and year confirm the positive relationship between number of papers and adjacent-year correlations.

(2nd row) Coefficient of next year number of citations for an article (n_{t+1}) regressed on current year number of citations (n_t). Blue, red, green, purple, and cyan lines indicate coefficients for top 0-1%, 1%-5%, 5%-10%, 10%-25%, and 25-50% most-cited bins respectively. The (All) panel shows results from a sample of 100 papers from each subject-bin-year taken from the top of the bin—e.g., the 100 most-cited papers in Mathematics’ 1%-5% most-cited papers in 1998. Panels (A) to (E) display regressions over all binned papers in subject. The x-axis indicates the logged number of articles (N) published in the subject in the focal year. When few articles are published in a subject each year, the number of citations for higher-cited and lower-cited articles decays at similar rates over time. When many articles are published, the number of citations received by a

highly-cited article decays slowly year-to-year compared to less-cited articles. Panel regressions with paper and fixed effects and controls for paper age, age-squared, and cumulative number of citations show the same pattern of results.

(3rd row) Probability (p) of a paper reaching the top 0.1% of most-cited articles. The x-axis indicates the number of articles published in the same year as the paper (N_p). Blue dots are subject-year observations and the blue line is a linear fit. The (All) panel displays data across subjects for papers published in 1980. Panels (A) to (E) present data for years up to and including 1984 in the respective subjects. Papers published in the same year as many others have a lower probability of reaching the top 0.1% of most-cited articles in any year.

(4th row) Median number of years (τ) for a paper to reach the top 0.1% of most-cited articles. The x-axis indicates the number of articles published in the same year as the paper (N_p). Blue dots are subject-year observations and the blue line is a linear fit. The (All) panel displays data across subjects for papers published in 1980. Panels (A) to (E) present data for years up to and including 1984 in the respective subjects. For papers that do become widely-cited, the time for the paper to reach the top 0.1% is shorter for papers published in the same year as many others in the subject.

Note: The analyses excluded all papers that were never cited.