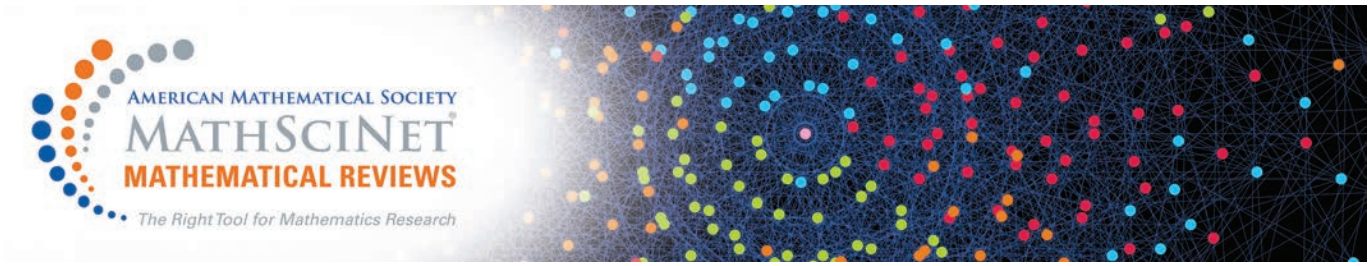


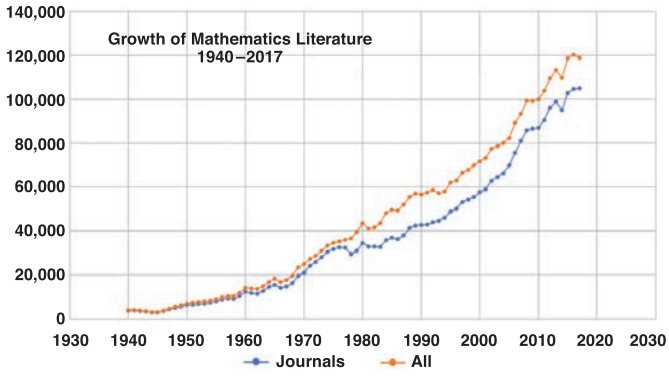

# Looking at the Mathematics Literature

*Edward Dunne*

If you think your job is getting harder, you are correct. The mathematics literature is growing relentlessly, and becoming harder to figure out along the way. The vehicles for publishing are more varied than ever. Meanwhile, bibliometrics are tempting to administrators for their apparent objectivity, which then tempts researchers to respond accordingly. This is a look at these three issues, using information contained in the Mathematical Reviews Database (MRDB), which is what powers MathSciNet®. Mathematical Reviews has been indexing and reviewing the research literature in mathematics since 1940. We have collected a considerable amount of information about this corpus over the years. As of this writing, the database contains roughly 3.6 million items and profiles for over 900,000 authors.

## 1. Growth of the Literature

Counting the number of items indexed by Mathematical Reviews per year from 1985 to 2017, the number of new articles per year is well modelled by exponential growth at a rate of about 3% percent per year. Counting just journal articles, the rate is about 3.6%. That rate has a doubling time of just over 19 years. So far, we have counted 104,953 journal articles published in 2017. When I finished my PhD, in 1984, just over 35,000 mathematical articles in journals were published that year. For graduate students finishing their PhDs last year, that number had essentially tripled. Moreover, this model says they should expect more than 400,000 journal articles to be published per year by the time they are thinking about retirement.




This trend is not unique to mathematics, of course. Looking at the data in Web of Science from 1985 to today, the scientific literature overall is growing at about 3.9% per year.

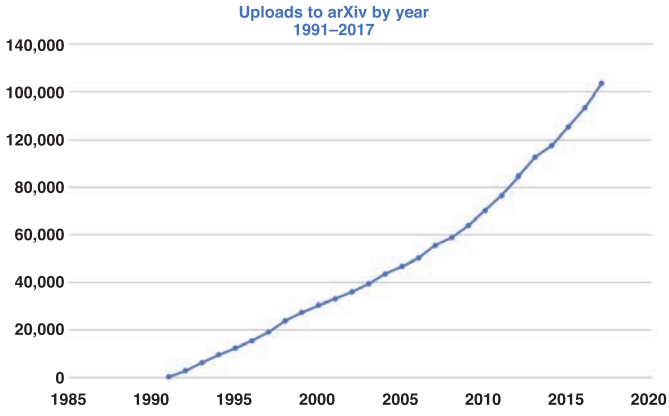
The data from the arXiv are harder to compare with publication data, as there are two phenomena occurring simultaneously. The arXiv started in 1991, with almost all the submissions being in certain areas of physics, such as high-energy physics. One way that the arXiv has grown is by attracting papers in different subjects and by the increasing participation of researchers in those subjects. As a result, the increasing number of uploads per year to the arXiv reflects

*Edward Dunne is the Executive Editor of* Mathematical Reviews *at the American Mathematical Society. His email address is* egd@ams.org.



DOI: http://dx.doi.org/10.1090/noti1799

both a growth in the number of papers being written and a growth in the number of researchers choosing to post their papers on the arXiv. A rough analysis of arXiv data indicates that it is growing according to a power law that is in between linear and quadratic.




A note about the arXiv: There is a belief that most of what is published is available on the arXiv. However, that is not so true for the items in MathSciNet. In particular, for all items in the Math Reviews Database with publication dates in the five-year period from 2013 to 2017 (inclusive), just 23% were also in the arXiv. This corresponds reasonably well with the findings of Larivière, et al., who reported that 21% of all mathematics papers in journals covered in Web of Science were also on the arXiv, using data with publication year 2010. Mathematics had the highest proportion of published papers on the arXiv. Physics was second, at 20%, and the global percentage for all disciplines was 3.6%. Going the other direction, Larivière, et al. found that in their data set, 64% of all arXiv papers were published in a journal indexed by Web of Science. The percentage was highest in condensed-matter physics (80%), and was about 45% for mathematics papers. Of course, there are some interesting examples of papers posted on the arXiv that were never published in a journal, such as Perelman's papers on the Poincaré Conjecture.

## 2. New Models

Publishing models for journals run from traditional journals that are a hundred years old to new overlay journals. Journals offer several versions of open access: green, gold, diamond. Many journals are now hybrid journals, where you can pay an APC to make your article open access, but some of the other articles in the journal will be behind a paywall. When Mathematical Reviews started in the 1940s, it was difficult to start a journal.[1] At a minimum, you needed resources to print and to distribute the journal. You also needed to establish a subscription base. Now, those obstacles have fallen away, making it much easier to start a journal. The effect has been to broaden the types of journals that are started.

There have been some interesting new journals: The *Cambridge Journal of Mathematics* is a traditional journal published by International Press that started publishing in 2013 and quickly attracted good authors who submitted good papers. The now-hybrid journal *Research in Mathematical Sciences* from Springer, which began as an open access journal in 2014, also quickly attracted good papers from recognized researchers. The overlay journal *Discrete Analysis*, started in 2016, hosts the papers on the arXiv, rather than developing their own infrastructure. There has also been a surge in journals of questionable value. Mathematical Reviews tries to identify those that meet the standards in our Editorial Statement `www.ams.org/publications/math-reviews/mr-edit`, in particular that are publishing refereed research in the mathematical sciences.

How do you know if a journal is "good"? For a few journals, this is easy to answer. They have been around for some years and have consistently published good papers—papers that you know have had an impact in the field. However, for most journals, this is very difficult to answer. Ideally, the measure would be the quality of the papers published in the journal. But is this "quality" the depth of the ideas, the correctness of the paper, its influence, or something else? Since most of these characteristics are hard to quantify, we end up adopting what we can count.

## 3. The Mismeasure of Math

Bibliometrics are designed to provide a statistical analysis of published research. There are two basic quantities that are used: citations and publication counts. Citations are meant to be a proxy for "influence," with the assumption being that highly cited papers are influential, hence important and deep. Counts are meant to be a measure of productivity. A researcher who writes many papers is presumed to be more productive than a researcher who writes few.

Counting citations is harder than it appears. People often ask why their citation counts are lower in MathSciNet than in Google Scholar. It is because they count different things. Google does not give a precise definition for their source of citations, but it is clearly broad. Mathematical Reviews provides a list on MathSciNet of the sources used in our citation counts. In looking at a particular paper from number theory from 2006, Google Scholar finds 48 citations, while Mathematical Reviews finds 19. The Google Scholar list includes: duplicates—a preprint on the arXiv and the published version of the paper; papers posted on a web site (not the arXiv) but not published; papers from conference proceedings; a research plan by one of the authors of the paper; the syllabus to a course at MIT posted on MIT OpenCourseWare plus a separate set of lecture notes for that course. These non-journal references indicate the reach of the work but are different from what Mathematical

[1] *Duke University Press has posted a nice history of the founding of the* Duke Mathematical Journal, *which is well worth a read.*

Reviews citations are trying to indicate: adoption in the research literature.

Matching citations can be tricky. References are matched to the original works algorithmically. It is hard for the algorithms to predict all the ways a citation might vary. An author contacted Mathematical Reviews to say that some citations were missing from their book. They provided specifics. In looking at the papers that were citing the book, we found that the citations were careless. The title was missing or was wrong. In at least one case, the year of publication given was actually the volume number. Meanwhile, the formats for references in many physics journals omit the title, give just initials for the authors' names, and sometimes give only the starting page number instead of the full page range. While these formats may be traditional, they provide fewer hooks for finding a match. Perhaps someday physicists will adopt a less telegraphic format for their references. If they do, they may find that their citation counts go up immediately.

Even if the matches are perfect, the question of whether they mean anything remains. There is a story that people like to tell when discussing consultants:

> A mathematical prodigy named Jedediah Buxton was taken to see David Garrick perform in Shakespeare's *Richard III* at the Drury Lane theatre. When asked whether he had enjoyed the play, his reply was that it contained 12,445 words. His analysis was correct, but did seem to miss some significant aspects.

The Jedediah Buxton tale is often a prelude to a discussion of the maxim: You get what you measure. That is to say, the system responds to the measuring, which leads to the problem of people gaming the system. For impact factors of journals, gaming includes excessive self-citations and citation stacking. The first is self-explanatory. The latter, also known as a citation cartel, is an arrangement whereby a group of journals have a policy of inflating citations to other journals in the group.

Individual researchers can try to game the system, aiming to maximize publication counts. The best-known tool is "salami slicing." The goal is to find the least publishable unit (LPU) or publon, which is the minimum amount of content that can be used to get a paper accepted in a journal. We have seen more than one series of papers where a result was first proved for second-order equations, followed by another paper proving the result for third-order equations, possibly even a paper on fourth-order equations.

**Prolific authors and award-winning mathematicians**

In mathematics, having a significant impact need not be correlated to having published lots of papers. The following is a list of prolific authors combined with a list of some award-winning authors. The lists of award-winning mathematicians and prolific authors have overlaps, but the symmetric difference is interesting.

| Author | Total pubs in MRDB | # Citations in MRDB | # of Citing Authors in MRDB | h-index in MRDB |
|---|---|---|---|---|
| Ravi Agarwal | 1,549 | 13,807 | 5,211 | 51 |
| Paul Erdős | 1,445 | 17,097 | 9,563 | 58 |
| Donal O'Regan | 1,295 | 10,220 | 3,716 | 42 |
| Israel Gel'fand | 486 | 9,075 | 7,171 | 45 |
| Terence Tao | 305 | 13,186 | 6,199 | 56 |
| Stanley Osher | 272 | 14,802 | 8,210 | 54 |
| Jean-Pierre Serre | 270 | 14,650 | 7,701 | 62 |
| Michael Atiyah | 262 | 9,284 | 5,471 | 47 |
| George Lusztig | 244 | 8,277 | 2,011 | 47 |
| Masaki Kashiwara | 240 | 6,044 | 2,110 | 42 |
| Peter Lax | 211 | 6,649 | 5,612 | 37 |
| Sylvia Serfaty | 84 | 1,524 | 582 | 22 |
| Cathleen Morawetz | 82 | 1,045 | 603 | 19 |
| Andrew Wiles | 28 | 1,824 | 789 | 15 |
| Maryam Mirzakhani | 20 | 324 | 267 | 9 |
| Peter Scholze | 19 | 233 | 143 | 8 |

Data current as of October 3, 2018.

## 4. The Human Factor

At Mathematical Reviews, we rely on computers and data, but we rely even more on experts. We have two departments that function like departments in a research library: maintaining relations with publishers, keeping track of what we have received and what we are still waiting for, and carefully cataloging material—especially author identification. Many of them have graduate degrees in library science or information science. They are valuable internal resources on publishing standards.

Our database tools help our staff do their work more quickly and more accurately. For instance, the program that helps with author identification tries to match the author of a newly received paper to an author already in our database, using subject area, coauthors, institution, and more. Many times, there is an obvious match. Plenty of other times, there are multiple possible matches. At this point, the catalogers need to rely on their training, experience, and guile to find the correct match.

We also have 18 PhD mathematicians who make editorial decisions every day. All of them have research programs, which help them identify what journals, proceedings, and books meet our editorial standards. They also rely on their years of experience in working with the mathematical literature, which gives them a powerful perspective.

Finally, we have 22,000 research mathematicians who serve as reviewers. We rely on their training and expertise in specific areas of mathematics to comment on the published literature. The reviewers' familiarity with subfields can be particularly helpful in pointing out overlaps between papers or duplicate publications.

The mathematics literature is complex. It is useful to count it and to measure it in various different ways. But it is also subtle, and we only truly understand it by reading it and engaging with it.

## References

Larivière, Vincent; Sugimoto, Cassidy R.; Macaluso, Benoit; Milojevic, Stasa; Cronin, Blaise; Thelwall, Mike. arXiv E-Prints and the Journal of Record: An Analysis of Roles and Relationships. *Journal of The Association for Information Science* and Technology **65** (2014), no. 6, 1157-1169. DOI: 10.1002/asi.23044. See also: https://arxiv.org/abs/1306.3261.

Edward Dunne

## Credits

All images are courtesy of the author.