An Update on Time Lag in Mathematical References, Preprint Relevance, and Subject Specifics[1]

Adam Bannister and Olaf Teschke (both FIZ Karlsruhe, Berlin, Germany)

Almost five years ago, we reported in this column on what turned out to be the most extensive study at the time of citation delay of mathematics publications [1]. Back in 2012, we were quite satisfied to have reference data available for about 50,000 EuDML articles and 170,000 zbMATH articles; this already accounted for a larger proportion of the mathematical literature when compared to commercial citation databases, which tend to be less comprehensive for the field of mathematics. Today, however, the situation has changed considerably. Further digitisation efforts and improved availability of reference data now allow the interrogation of about 20 million references for more than 900,000 mathematical articles in zbMATH. Furthermore, linking them to available zbMATH entries and arXiv submissions also facilitates an analysis of subject specifics and preprint effects. While a detailed investigation is beyond the limits of this column, we take the opportunity to outline some aspects that become visible when taking extended data into account.

Growing longevity confirmed

With much more data available, the trends discovered in [1] have been confirmed. The graph in Figure 1, covering the extended period from 1880 to 2017, for which reference data are now available, looks very similar to the one from [1]. However, some additional effects have become evident: the extremal points are slightly smoothed and, in particular, war effects, which were extensively discussed in [1], are less emphasised. This can be explained by the current database being broader, the restriction to publication years originating from matched references (in 2012, we had to extract plausible publication years from the reference strings since the matched data were still too sparse) and the bias originating from a dominance in 2012 of Springer data (which tended to be more affected by the World Wars). Still, the conclusion from 2012 holds that since World War II there has been a steady increase in the average citation longevity. Indeed, the prediction of further growth we made back then has become a reality.

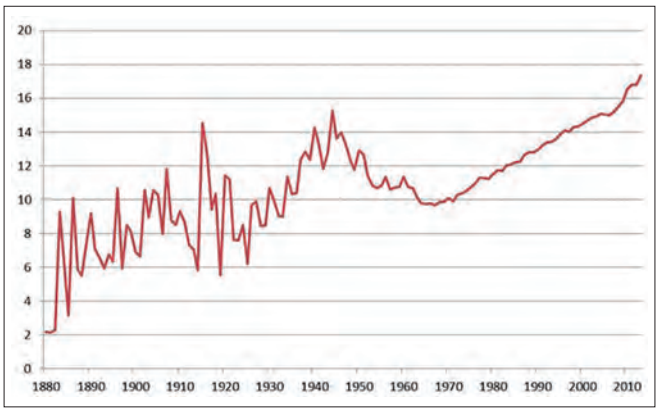

Figure 1. Average reference lag per year based on about 11 million matched zbMATH data.

Half-life: Infinity?

An interesting effect resulting from both the increased number of publications (and hence references) and growing longevity is that the notion of (absolute) citation half-life doesn't seem to be reasonable for mathematical publications. This is due to the fact that the distribution is heavily skewed toward a thick, long tail, without any indication of a convergence to zero. Hence, formal computation of half-lives actually leads to numbers much larger than half the period since publication, which still grow by about eight months every year. Figure 2, showing the distribution of references to publications for some fixed publication years, is typical. One should note, however, that this is mainly influenced by publication

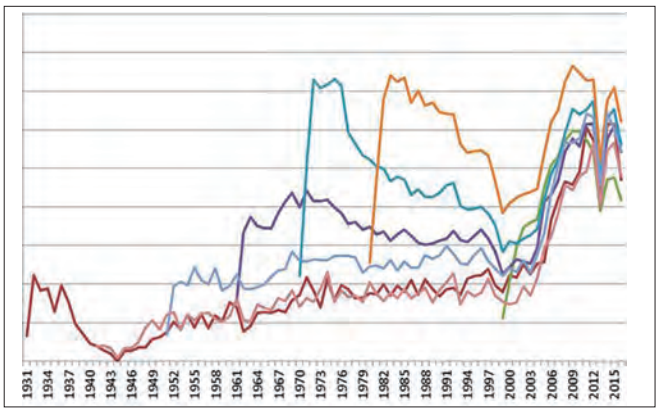

Figure 2. Long-term distribution of absolute citation numbers for fixed publication years.

[1] Several aspects of this report were discussed in the framework of a panel discussion on bibliometrics on Sep 12th, 2017, at the 19th ÖMG Congress and Annual DMV Meeting held in Salzburg. At this panel, representatives of the DMV, EMS, ÖMG, UMI, and zbMATH discussed the various challenges connected with the increased use of bibliometrical measures throughout Europe, and outlined practical approaches. Results of this discussion will also be published in a forthcoming article at the Mitteilungen der Deutschen Mathematiker-Vereinigung.

and corresponding citation growth; the relative citation frequency (adjusted for the available number of references) is indeed declining in the long term. Some aspects of this will be addressed in the last section.

It might be worth noting that this is contrary to the general development in the sciences: a study [2] considering the fields of medicine, molecular biology, chemistry and physics found that there has recently been a quicker decay of citation numbers, which, according to the authors, is due to a general increase in the number of published papers that supersede earlier results.[2] There is no indication of a similar effect in mathematics.

Preprint citations

While the long-term pattern seems to be very stable, it goes without saying that the publication landscape has changed significantly over the last few decades. The diagrams above, showing an average citation lag larger than 15 years for traditional publications, indicate that it may yet be too early to see the effects of changing publication behaviour at this level. It is natural to ask about these patterns when arXiv preprints are included. As shown earlier in this column [3], the arXiv has established itself as the standard preprint repository for many areas in mathematics, often preceding the actual publication by several years. Taking arXiv submission years into account, one might be able to get rid of backlog effects affecting the publication year. Since the arXiv version is matched to the zbMATH entry and it is easy to identify the arXiv submission year, one might wonder about the results when taking preprints into account.

The comparison, however, shows no significant difference in long-term citation behaviour. Of course, the average citation lag is initially much smaller for references to the arXiv (which, by definition, has no submission years before 1991) but it very closely resembles the behaviour of references to traditional publications when publications before 1991 are omitted (see Figure 3). Any local differences can be linked to the fact that the arXiv corpus has grown more quickly and shows a different subject pattern (as explained in [3]). In particular, the distribution of references to arXiv submissions for a fixed year shows the same right-skewed pattern related to the "immortality" of mathematical research (Figure 4). As a conclusion, this would support handling citations to the arXiv on an equal footing with those of traditional publications, taking advantage of avoiding the publication gap associated to journal backlog.[3]

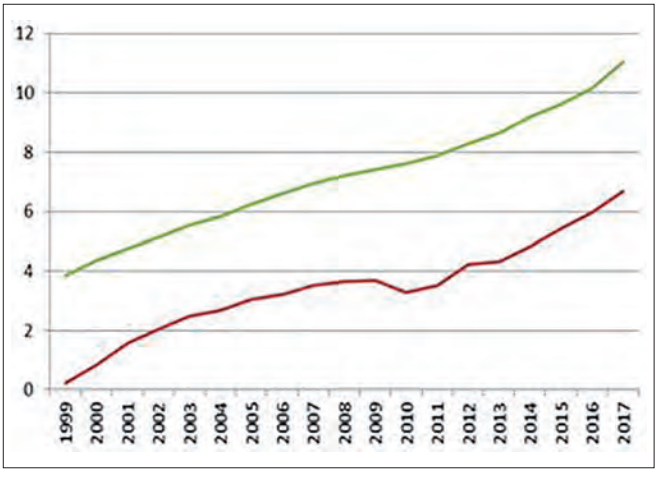

Figure 3. Time lag for references to arXiv submissions compared to publications after 1991.

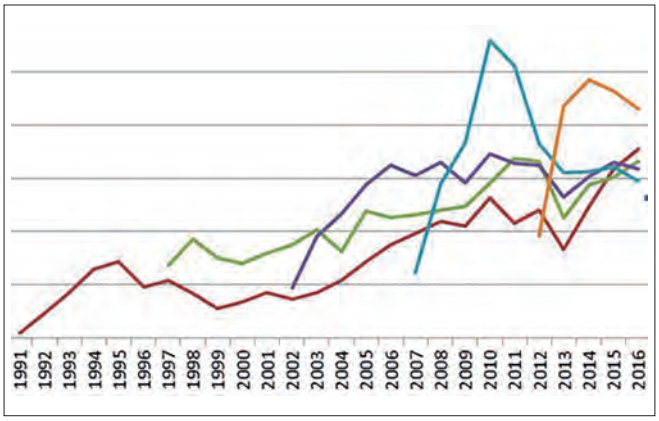

Figure 4. Absolute citation numbers for fixed arXiv submission years.

One step further: subject specifics

One might wonder whether it is possible to differentiate this general picture further by taking mathematical subjects into account. Matching citations to zbMATH provides MSC information and perhaps the first natural question is whether the topic is reflected by the citation network. Figure 5 shows that there is indeed a strong concentration in the diagonal (which means that the bulk of references go to papers with the same MSC), although there obviously exist further cross-references that should

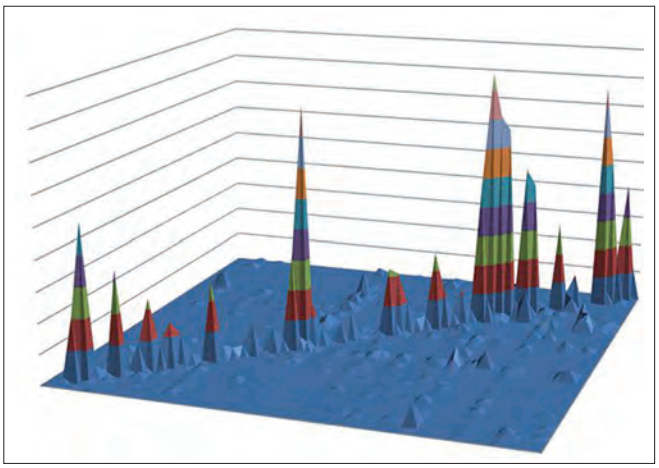

Figure 5. Cross-MSC citation map.

[2] It is not fully clear whether this study might actually miss a large number of references due to not comprehensively covering commercial citation databases. Computed formally, the "half-lives" of mathematical publications also seem to be declining rapidly for recent work, if only due to the fact that the bulk of citations are likely to happen in the future.

[3] One caveat to be aware of is that arXiv submissions are usually not yet peer-reviewed, whereas only citations of arXiv submissions by peer-reviewed papers were taken into account.

not be neglected in detailed studies. For a first impression, however, it might be justified to restrict to MSC-preserving citations.

Subject specifics

Since the number of publications and references is very unevenly distributed for different mathematical subject classes, it makes sense to study long-term referencing behaviour within the main MSC classes subject to availability of citation data. This adjustment also aims to eliminate the growth effects mentioned in Section 2. Figure 6 shows the relative distribution of references for several mathematical subjects in relation to the gap years (from 0 to 24).

It should be noticed that, at least for the relative citation frequencies, there is a long-term decay and also a clearly visible long tail. The only clearly different distribution belongs to quantum mechanics, where the initial relative citation rate is much higher before descending much more quickly. For the remaining subjects (with such diverse areas as number theory, algebraic geometry, partial differential equations, functional analysis, mathematical statistics and mathematical programming), the long-term behaviour is surprisingly similar, although there exist significant initial differences for the relative citation frequency. Therefore, a computation of relative citation half-lives on this basis yields somewhat different results for the mathematical subjects (mostly between 7 und 10 years, with the exception of mathematical physics, as shown in Figure 7). Even in this setting, it once more becomes obvious that the most widely used citation metrics (like impact factors, which usually consider a span of at most five years) miss the bulk of relevant information.

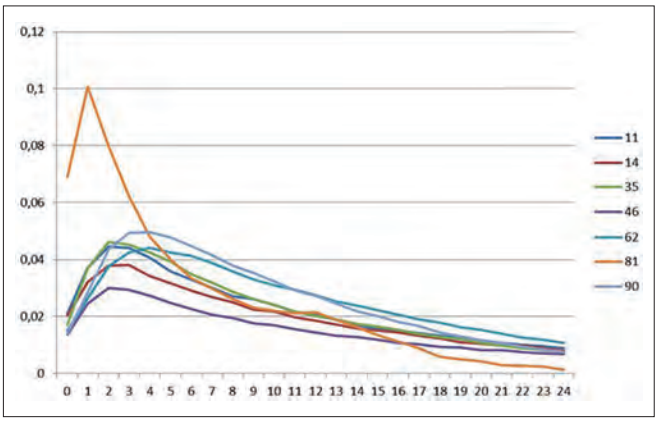

Figure 6. Relative time lags for MSC-preserving citations.

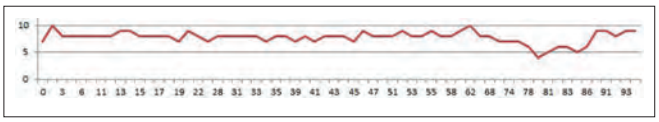

Figure 7. Relative half-lives of MSC-preserving citations.

References

1 Th. Bouche, O. Teschke, K. Wojciechowski: Time lag in mathematical references. *Eur. Math. Soc. Newsl.* 86, 54–55 (2012)

2 P. Della Briotta Paroloa et al.: Attention decay in science. *J. Informetrics* 9, No. 4, 734–745 (2015)

3 F. Müller, O. Teschke: Will all mathematics be on the arXiv (soon)? *Eur. Math. Soc. Newsl.* 99, 55–57 (2016).

Adam Bannister [adam.bannister@fiz-karlsruhe.de] has a postgraduate diploma in Geographic Information Systems and currently works on the Scalable Author Disambiguation for Bibliographic Databases at zbMath in cooperation with Schloss Dagstuhl and Heidelberg Institute for Theoretical Studies.

Olaf Teschke [olaf.teschke@fiz-karlsruhe.de] is a member of the Editorial Board of the EMS Newsletter, responsible for the zbMATH Column.