Editorial: The Dark Citations of TODS Papers and What to Do about It—or: Cite the Journal Paper

When assessing the excellence of a scientific article, for example, in a review, important aspects include the novelty and significance of its contribution, its scientific depth, and its mastery of the pertinent apparatus of computer science. The *excellence* of a researcher can be measured by his or her ability to publish in the scientific outlets with the highest reputation.

In contrast, the academic impact of the content of an article can be measured by the number of citations to the article. In some areas, it is easier to get citations than in other areas. However, when comparing two articles from the same area, one article with many citations and one article with few, the former can generally be considered as the more interesting, relevant, important, and/or impactful one. The academic *impact* of a researcher can then be measured by the number of citations to their articles.

However, although impact as measured by citations then differs from excellence, citations are still used for the rating of journals. Notably, journals are rated according to their citation-based impact factors, and a number of publishers advertise these statistics of their journals. Further, in some countries, the impact factors of a journal play an important role when different institutions assess the excellence of the journal. If a journal is not rated highly by funding agencies then researchers who rely on funding from those agencies are effectively encouraged to publish in other journals. Likewise, if a journal is not rated highly by hiring or promotion committees then candidates are effectively encouraged to publish in other journals. Because of reasons such as these, I find that it is not advisable to simply ignore citations.

A journal's two-year impact factor for a particular year n is calculated as the sum of the number of citations given during year n to each article published in the journal during years $n-1$ and $n-2$, divided by the count of articles published during years $n-1$ and $n-2$. Thus, an impact factor of 2.5 for year 2015 means that articles published in that journal during 2013 and 2014 received an average of 2.5 citations during 2015. This definition does not state explicitly which citations are counted. When considering the two-year impact factor computed by Thomson Reuters, it is not entirely clear which citations are counted. Thomson Reuters maintains a master journal list. If citations are counted only from articles published in journals on this list then only some journals are counted, and no conferences are counted. The latter is not good for computer science, where conference publication is important.

Having argued that citations are important, I will argue next that many citations to results published in *Transactions on Database Systems* (TODS) are not counted and that TODS articles should really have many more citations. This would substantially increase the citation statistics of TODS, including its two-year impact factor, and it would thus better reflect the externally perceived excellence of the journal and its articles.

A concrete example illustrates the issue. In June 2011, I and three coauthors published an article in TODS entitled “*Design and Analysis of a Ranking Approach to Private Location-Based Services*.” This article is an extension of a conference article entitled “*SpaceTwist: Managing the Trade-Offs Among Location Privacy, Query Performance, and Query Accuracy in Mobile Services*” that we published in *International*

2016 Copyright is held by the owner/author(s).

0362-5915/2016/06-ART8e

DOI: <http://dx.doi.org/10.1145/2946798>

Fig. 1. Citations to a 2008 ICDE article and its TODS 2011 extended version (Source: Google Scholar as of May 14, 2016).

Conference on Data Engineering (ICDE) in 2008. We chose to extend this article into a journal article because we felt that its approach was quite novel. Also, the article received encouraging reviews and was considered for the best article award at ICDE. The journal article offers more comprehensive coverage; for example, we involved a statistics professor in order to be able to analyze better the article's ranking approach. Thus, the journal article contains everything that the conference article contains and significantly more.

Figure 1 shows the citations to the two articles. The journal article received 3, 4, 6, 4, and 4 citations in the years 2011 to 2015, respectively. If these citations are all counted then the article contributes 4 citations to the TODS 2012 impact factor and 6 citations to the 2013 impact factor. The conference article received 5, 41, 26, 47, 31, 30, 39, and 60 citations in the years 2008 to 2015. If these citations are all counted, the article contributes 41 citations to the ICDE 2009 impact factor and 26 citations to the 2010 impact factor.

In this example, a total of 21 citations are counted for the results published in the journal article from 2011 to 2015, but considering also the citations to the conference article, the citations to the results are 228 from 2011 to 2015. The 207 concurrent citations to the conference article are the *dark citations* that are not counted. The difference between the counted citations and the uncounted dark citations is an order of magnitude. Imagine the difference it would make if these citations were counted.

It is common practice in the database area and other areas of computer science to first publish articles in conferences and only then publish extended versions in journals. Indeed, database and other journals accept extended conference articles, and they publish many articles that are extensions of conference articles. TODS requires that extended versions include at least 30% of new content material (see

<http://tods.acm.org/ThirtyPercentRulePolicy.cfm>), and I estimate that around three quarters of the articles published each year are extensions of conference articles.

So far, I have argued that we cannot simply ignore citations and that results published in TODS receive many more citations than are actually counted. Why does the problem occur and how can we fix the problem?

The example shows that other articles continue to cite the conference article, even when it has been superseded by an extended journal article. This practice may occur because the conference article is cited initially, as only it exists. (This was true for 2008 through 2010.) Then the authors of subsequent articles just keep citing the conference article. They may not have noticed that an extended journal version had become available, as they already have something to cite. That said, in my view, this practice is generally not one that makes the most sense from an academic perspective.

One possible action that addresses the problem is for TODS to publish a higher fraction of articles that do not extend a conference article. Such articles have no dark citations. TODS has already started to encourage more submissions of such original articles by making them eligible for presentation at the Special Interest Group on Management of Data (SIGMOD; see the editorial *The Best of Two Worlds—Present Your TODS Paper at SIGMOD* in the June 2015 issue of TODS). Other journals have established fast-track publication schemes for original articles. TODS could do something similar. However, this action can only partially fix the problem.

Another possible action is to develop a citation metric and system that takes the dark citations into account when assessing the citation performance for the results published in journals. While I think that such a metric and system makes sense, the result is yet another metric that may not be adopted where it counts. Specifically, it is going to be a long, tedious, and up-hill battle to get publishers to use yet another metric, and it may be even harder to get institutions to adopt the new metric.

I propose a very practical action that authors can start taking right now and that I think is good for science. Specifically, *I propose to address the problem of dark citations by always citing the extended journal version of an article whenever it is available.* The journal version is the definitive and most recent account of the research. The journal version has gone through an additional and more formal review process. The journal version extends and, likely, consolidates the conference version's results. And the journal version is likely to offer a better and more up-to-date coverage of related work. These are all good arguments for citing the journal version.

There can be reasons for also citing the conference version. One is that it may be important to establish the order of invention. It may have taken several years for an extended version to appear in a journal because it takes time to develop the new results, because the review process and revisions take time, and because there may be a delay from acceptance to actual publication in an issue. A possible reason for citing only the conference version occurs if one wants to make reference to content in the conference version that is not present in the journal version. However, in my experience as an editor and an author, this situation occurs rarely.

In summary, it is important for the database community to have journals that are not only excellent but also highly cited. Results published in TODS have many more citations than are counted. *You can help by citing the extended journal article when one exists.*

My colleague Rick Snodgrass, a former TODS Editor-in-Chief, provided valuable comments that helped improve the presentation.

Christian S. Jensen
Editor-in-Chief