Regular article

# Can Microsoft Academic assess the early citation impact of in-press articles? A multi-discipline exploratory analysis

Kayvan Kousha*, Mike Thelwall, Mahshid Abdoli

*Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK*

## ARTICLE INFO

## ABSTRACT

Many journals post accepted articles online before they are formally published in an issue. Early citation impact evidence for these articles could be helpful for timely research evaluation and to identify potentially important articles that quickly attract many citations. This article investigates whether Microsoft Academic can help with this task. For over 65,000 Scopus in-press articles from 2016 and 2017 across 26 fields, Microsoft Academic found 2–5 times as many citations as Scopus, depending on year and field. From manual checks of 1122 Microsoft Academic citations not found in Scopus, Microsoft Academic's citation indexing was faster but not much wider than Scopus for journals. It achieved this by associating citations to preprints with their subsequent in-press versions and by extracting citations from in-press articles. In some fields its coverage of scholarly digital libraries, such as arXiv.org, was also an advantage. Thus, Microsoft Academic seems to be a more comprehensive automatic source of citation counts for in-press articles than Scopus.

© 2018 Elsevier Ltd. All rights reserved.

## 1. Introduction

Citation indicators derived from conventional scholarly databases, such as Scopus and the Web of Science (WoS), are often used for the impact assessment of published articles. They are rarely useful for recently-published and in-press articles, however, since their citation counts tend to be zero. The overall publication delay (the time between submission or acceptance and publication) also negatively influences citation indicators (Luwel & Moed, 1998; Yu, Wang, & Yu, 2005; Tort, Targino, & Amaral, 2012; Shi, Rousseau, Yang, & Li, 2017). Traditional citation indexes seem to wait for articles to be formally published by journals before processing their references. For instance, on 15 October 2017 Scopus had indexed over 277,000 "In-Press" articles that had been published as "Online First" or similar in journals. Nevertheless, Scopus does not index or display the cited references of in-press articles until their final version is published in a journal issue (as of 20 October 2017[1]). Hence, it seems likely that millions of citations from in-press articles are not included in any Scopus citation counts. WoS seems to wait for in-press articles to be published in an issue before reporting them. For instance, although on 15 October 2017 Scopus found 52 and 46 in-press articles in 2017 from *Scientometrics* and *Journal of the Association for Information Science and Technology,* respectively, none had been indexed in WoS.

---

Advance online publication increases citations to articles, Journal Impact Factors and Immediacy Index values (Alves-Silva et al., 2016; Al & Soydal, 2017; Echeverría, Stuart, & Cordón-García, 2017; Todorov & Glänzel, 1988). Many academic publishers provide early online access to their journal articles to minimize publication delays and perhaps to increase citation rates such as *Springer* (Online First), *Wiley* (Early View), *Taylor & Francis* (Latest Articles), and *Nature Publishing Group* (Advance Online Publication). Some authors deposit preprints or postprints (final drafts after peer review) of their articles to open access repositories, such as arXiv.org, or share them via academic social websites, such as ResearchGate or Academia.edu before they are available as online first or final versions in publishers' websites.[2] These strategies reduce article publication delays and presumably make it more likely that an article is cited before it is formally published, especially for journals with long publication backlogs. For example, the article "*Stationary graph processes and spectral estimation*" was published online first in *IEEE Transactions on Signal Processing* on 11 August 2017 and cited 42 times in Google Scholar (16 October 2017) but had not been cited in Scopus or WoS. All 42 citations were to a preprint version of the article that had been deposited in arXiv.org on 14 March 2016. In general, citation rates are influenced by online availability, publication date, and indexing date (Haustein, Bowman, & Costas, 2015).

From the above discussion, advance online publication is an increasingly important phenomenon that needs to be investigated by scientometricians. This paper assesses the ability of Microsoft Academic to find citations to recently published research by comparing Microsoft Academic citations to over 65,000 in-press articles from 2016 and 2017 with Scopus citations across 26 fields.

## 2. Early citation impact

Early citation impact evidence could help to identify cutting-edge research that quickly attracts citations, differentiating it from typical articles that need longer to be cited (Moed, 2005). Early impact evidence for recent research can also be useful to predict the long-term citation impact of articles. This can support timely research evaluation exercises, academic promotion, the employment of early-career researchers, and the evaluation of research funding programs (Levitt & Thelwall, 2011; Bornmann, 2013; Bruns & Stern, 2016). In justification of these applications, early citation impact (1–2 years after publication) for scientific articles positively correlates with citation indicators calculated in subsequent years (Adams, 2005). Nevertheless, predicting the future citation impact of articles based on early citation counts is challenging due to factors like field differences in citation behaviour (Wang, 2013).

Some alternative sources of evidence have been proposed to identify the early intellectual impact of research, including article downloads (Kurtz et al., 2005; Brody, Harnad, & Carr, 2006; Bollen, Sompel, Smith, & Luce, 2005), Mendeley reader counts (Thelwall & Sud, 2016; Maflahi & Thelwall, 2018) and social web mentions (Thelwall, Haustein, Larivière, & Sugimoto, 2013). These all reflect types of use that are likely to appear before citations. Some, such as download counts, may reflect different degrees of interest or uses of academic research compared to citations (Kurtz & Bollen, 2010). Similarly, Mendeley reader counts partly reflect professional, teaching and educational uses (Mohammadi, Thelwall, & Kousha, 2016).

### 2.1. Free citation indexes for early citation analysis

Several free scholarly websites, including Google Scholar, ResearchGate and Microsoft Academic, index or host preprint versions of articles that could be used for early citation impact assessment.

#### 2.1.1. Google scholar

Google Scholar may be the largest index for the early citation impact of research because it generates higher citation counts than traditional citation databases. It is helpful in this regard by indexing different publishers and wider online sources, such as open access publications (Bar-Ilan, 2008; Khabsa & Giles, 2014; Halevi, Moed, & Bar-Ilan, 2017). For instance, the in-press article "*Stochastic multicriteria decision-making approach based on SMAA-ELECTRE with extended gray numbers*" published online first in the journal *International Transactions in Operational Research* on 7 February 2017 (DOI: 10.1111/itor.12380), had no Scopus citations by 25 September 2017 but had received eight Google Scholar citations from other recently-published journal articles (mostly online first). All eight citing journals found in Google Scholar were covered by Scopus and so indexing delays in Scopus were the reason for its missing citations. On 18 October, Scopus found two of the missing citations to the above article, confirming that indexing delays were the cause. These delays may be for technical (delays in accessing or processing publications) or quality control (waiting for the version of record) reasons. Despite the substantial Google Scholar coverage of scholarly publications and citations, it cannot be used for most research evaluations because it does not allow automatic data collection, which is a practical necessity for large scale analyses. The Publish or Perish software can extract Google Scholar citations and other citation impact indicators for individual papers, academics or journals, however (Harzing, 2007).

---

[2] http://www.sherpa.ac.uk/romeo/statistics.php?colour=green