Can Microsoft Academic be used for Citation Analysis of Preprint Archives? The Case of the Social Science Research Network[1]

Mike Thelwall, University of Wolverhampton, UK.

Abstract
Preprint archives play an important scholarly communication role within some fields. The impact of archives and individual preprints are difficult to analyse because online repositories are not indexed by the Web of Science or Scopus. In response, this article assesses whether the new Microsoft Academic can be used for citation analysis of preprint archives, focusing on the Social Science Research Network (SSRN). Although Microsoft Academic seems to index SSRN comprehensively, it groups a small fraction of SSRN papers into an easily retrievable set that has variations in character over time, making any field normalisation or citation comparisons untrustworthy. A brief parallel analysis of arXiv suggests that similar results would occur for other online repositories. Systematic analyses of preprint archives are nevertheless possible with Microsoft Academic when complete lists of archive publications are available from other sources because of its promising coverage and citation results.
Keywords: Microsoft Academic; SSRN; arXiv; Digital repositories; Preprint archives

Introduction

Citation analysis is sometimes used in research evaluations and to analyse scholarly communication but tends to deal exclusively with journal articles. Citation analyses of other types of document (e.g., patents: Jaffe, Trajtenberg, & Henderson, 1993; Karki, 1997) are difficult to conduct systematically because journal articles are the primary document type indexed by the Web of Science (WoS) and Scopus. Despite this, other types of output are essential to the smooth functioning of many fields. For example, working papers are routinely shared in economics and physics (Di Cesare, Luzi, Ricci, Ruggieri, della Ricerche, & della Repubblica, 2011; Luce, 2001), either as a stepping stone to formal journal publication or recording other research-related information. If large preprint repositories could be analysed with scientometric methods, then their role could better be understood and methods could be developed to help assess the impact of individual papers or groups of papers to help reward the creators of successful content.

In the past, scientometric analyses of repositories have been difficult because they are not covered by the major citation indexes, Scopus or Web of Science. For example, one study of arXiv used citation counts to mathematics articles published in journals and available in arXiv (Davis & Fromerth, 2007), ignoring arXiv deposits that did not subsequently appear in journals. Another used download records in SSRN but not citations from WoS or Scopus (Eisenberg, 2006). Research Papers in Economics (RePEc) is unusual in providing data on download counts and citations from other RePEc papers and other online papers via CitEc (Zimmermann, 2013) but also does not report WoS or Scopus citations. Although it is possible to identify citations to individual papers in Scopus through its advanced reference search function, it is not possible to download from it a systematic

[1] Thelwall, M. (in press). Can Microsoft Academic be used for citation analysis of preprint archives? The case of the Social Science Research Network. *Scientometrics*.

collection of SSRN articles irrespective of whether they have been cited. In theory, Google Scholar and Microsoft Academic can fill this gap because they can index any scholarly document found online, which includes the contents of digital repositories (Halevi, Moed, & Bar-Ilan, 2017). This enables them to report much higher citation counts for recent documents than Scopus, for example (Harzing, & Alakangas, 2017b; Thelwall, 2017b). Of these, Microsoft Academic is the most promising for citation analysis because it allows automatic data harvesting (Harzing, 2016), whereas Google Scholar support for this is limited to author centred analyses with the Publish or Perish software (Harzing, 2007). It therefore has the potential to support new types of citation analysis for collections of documents that are not indexed by Scopus and the Web of Science.

This paper focuses on one important preprint archive, the Social Science Research Network, as an extended case study to analyse in detail whether Microsoft Academic could be used for effective citation analyses of online repositories. A previous study has found Microsoft Academic to find more citations to in press articles from journals (Kousha, Thelwall, & Abdoli, 2018), but repositories have not previously been investigated. SSRN was created by financial economists but includes papers from the wider social sciences as well as the humanities and some natural sciences[2]. Its papers are in series published by academic departments, journals, non-academic organisations (e.g., the World Bank) or separate submissions. It is important enough that researchers have attempted to create bibliometrics from its data (Brown & Laksmana, 2004; Brown, 2003). SSRN publishes "SSRN eJournals", which are subject-based collections of articles that meet a minimum content requirement but have not been peer reviewed (SSRN, 2017). For example, *International Administrative Law eJournal*[3] is a simple list of qualifying articles, without volumes or issues. This is a dissemination device rather than a type of formal publishing and papers in these may be published in traditional journals.

Background: Microsoft Academic

Microsoft Academic is the replacement for the former Microsoft Academic Search (Sinha, Shen, Song, Ma, Eide, Hsu, & Wang, 2015). It was released in 2016 in a trial version and formally in July 2017. It has similar functionality to Google Scholar in terms of indexing both publisher databases and open web content (Falagas, Pitsouni, Malietzis, & Pappas, 2008; Harzing & Van der Wal, 2008), and providing author level (Orduña-Malea, Martín-Martín, & Delgado-López-Cózar, 2016) and journal-level (Delgado López-Cózar, & Cabezas-Clavijo, 2012) information. It has two additional important features.

First, Microsoft Academic attempts to automatically classify documents into fields. Field classifications are important for many scientometric analyses, such as field normalisation (e.g., van Leeuwen, & Calero Medina, 2012; Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). Nevertheless, the Microsoft Academic scheme does not seem to be coherent enough to be useful yet (Hug, Ochsner, & Brändle, 2017).

Second, Microsoft Academic allows automatic data harvesting through its Applications Programming Interface (API). This makes it a practical data source for large scale analyses. This is its biggest advantage in comparison to Google Scholar (Harzing & Alakangas, 2017b).

[2] https://www.ssrn.com/en/index.cfm/rps/

[3] https://papers.ssrn.com/sol3/JELJOUR_Results.cfm?form_name=journalBrowse&journal_id=2127729

Microsoft Academic's coverage of the academic literature has been tested through the works of individual scholars in multiple disciplines (Harzing & Alakangas, 2017ab), the contents of journal articles in multiple disciplines (Thelwall, in press, 2017b), and the documents in the digital repository of an institution (Hug & Brändle, 2017). Taken together, these studies suggest that the coverage of Microsoft Academic, in terms of the number of papers indexed and the average citation counts, is like Google Scholar and usually greater than Scopus and WoS. Its average citation counts are especially high relative to Scopus and WoS for recently published articles, giving it an early citation advantage. There may be broad disciplinary differences in the advantage of Microsoft Academic and there are differences between individual journals for its early citation advantage.

A major practical drawback of Microsoft Academic is that, like Google Scholar, it does not have a formal quality control mechanism and therefore cannot be used for formal evaluations where the participants are aware of its use in advance. This is because it is straightforward to manipulate its citation counts by uploading low quality citing documents into places that it indexes (for Google Scholar, see: Delgado López-Cózar, Robinson-García, & Torres-Salinas, 2014).

It is not possible to check the coverage of Microsoft Academic because it merges records for article preprints with the final published article versions (Thelwall, in press, 2017b). It may therefore not report a version of a paper that it has indexed because a different version is its primary copy.

Research questions

The first research question targets the comprehensiveness of Microsoft Academic's coverage of SSRN because any gaps will undermine citation analyses. If the gaps are systematic then they also risk biasing the results of evaluations.

If Microsoft Academic citations are to be used for analyses of SSRN papers then it is important to know whether they reflect scholarly impact. The standard first way to do this is to assess the strength of correlation between them an alternative recognised source of citation impact data, such as WoS or Scopus (Sud & Thelwall, 2014) but this is not available for SSRN. It is possible to use heuristics to estimate the total number of Scopus documents citing SSRN (Li, Thelwall, & Kousha, 2015) but not to obtain accurate citation counts for each document on a large scale. The full text of SSRN papers can be parsed to extract citations between them (West, Jensen, Dandrea, Gordon, & Bergstrom, 2013) but this does not include citations from papers outside SSRN. Mendeley reader counts are therefore used instead because they are known to have a significant positive correlation with citation counts for journal articles in many contexts (e.g., Thelwall & Sud, 2016; Thelwall & Wilson, 2016) and to positively correlate with peer review scores for journal articles in many fields (HEFCE, 2015). They are more suitable than other altmetrics because they correlate more strongly with citation counts (Haustein, Larivière, Thelwall, Amyot, & Peters, 2014; Thelwall, Haustein, Larivière, & Sugimoto, 2013; Zahedi, Costas, & Wouters, 2014). Mendeley reader counts are also better than citation counts for identifying early impact evidence (Maflahi & Thelwall, 2018; Thelwall, 2017a), which may be important for preprints. A Mendeley reader here is anyone that has added a paper to their Mendeley library, signalling interest in it (Gunn, 2013). Mendeley readers have unusually read, or intend to read, a paper (Mohammadi, Thelwall, & Kousha, 2016) and so the "reader" terminology is appropriate here, even though roughly 90% of researchers do not use Mendeley (Van Noorden, 2014) so the data reflects the activities of a minority of scholars.

- RQ1: How comprehensive is Microsoft Academic's indexing of SSRN?
- RQ2: When are average Microsoft Academic citation counts higher than Mendeley reader counts (and therefore statistically more powerful)?
- RQ3: Do Microsoft Academic citation counts reflect traditional citation impact?

Methods

The research design was to download from Microsoft Academic records for all articles identifiable as being within SSRN and to explore the coverage and citation count characteristics of this data set. Mendeley reader counts were used to help check the meaningfulness of the citation counts. As mentioned above, Mendeley was chosen because both Scopus and Web of Science do not index SSRN. Mendeley is also a good choice because it belongs to the same company as SSRN (Elsevier), it has a strong correlation with citation counts for journal articles in almost all fields (Thelwall, 2017c) and it is straightforward to check reader counts in Mendeley for papers with DOIs.

Data

SSRN's eJournals are not separately indexed by Microsoft Academic. Instead, it groups together some SSRN papers into a "journal" called *SSRN Electronic Journal*. This term is not used by SSRN but is used by some citation indexes, such as scilit[4] as well as some individual authors[5]. It was also used by Mendeley for some of the publications analysed here[6]. To extract SSRN papers from Microsoft Academic, the following query was submitted to the Microsoft Academic API. Webometric Analyst was used as the interface to the API and the query was submitted on 13 August 2017.

```
Composite(J.JN=='ssrn electronic journal')
```

Some of the records returned from the above query had a SSRN DOI and these were searched for in Mendeley on 13 August 2017. Records without a DOI were ignored since many SSRN papers are preprints of papers published elsewhere, so the standard practice of finding extra matching papers in Mendeley by searching for author, title and publication year (Zahedi, Haustein, & Bowman, 2014) may give many false matches. Similarly, papers with non-SSRN DOIs were ignored as these DOIs pointed to other versions of the papers. Exact numbers are given in the results section.

Five similar datasets were created to check whether the results depended on the decisions made constructing each dataset. First, both Mendeley and Microsoft Academic publication years can include errors so either, both or neither could be correct. Second, Microsoft Academic documents without a matching Mendeley record could be missing their DOI in Mendeley. Thus, it is not clear whether it is better to interpret missing Mendeley records as indicating zero readers or as missing data.

- *All papers*: All papers returned by Microsoft Academic for the query Composite(J.JN=='ssrn electronic journal'), and using the Microsoft Academic publication year.
- *All papers with a SSRN DOI, using the Microsoft Academic publication year, treating missing Mendeley records as having 0 readers*. As above except that papers without

[4] https://www.scilit.net/journals/28185
[5] http://orcid.org/0000-0003-4416-0164
[6] http://www.mendeley.com/research/exchangerate-passthrough-g7-countries-1

a DOI were deleted and papers without a Mendeley record were given a Mendeley reader count of 0.

- *All papers with a SSRN DOI, using the Microsoft Academic publication year, treating missing Mendeley records as missing data*: As above except that publications without a Mendeley record were deleted.
- *All papers with a SSRN DOI, using the Mendeley publication year, treating missing Mendeley records as missing*: As above except that the Mendeley publication year was used instead of the Microsoft Academic publication year.
- *All papers with a SSRN DOI, a Mendeley record and the Mendeley and Microsoft Academic publication years agreeing*. As above except deleting records for which the Microsoft Academic and Mendeley publication years do not match.

Analysis

Average citation counts were calculated using geometric means instead of arithmetic means because these are more suitable for highly skewed data (Zitt, 2012). The standard technique of adding 1 to all data before starting and subtracting 1 from the result was used because of the presence of zeros in the data (uncited articles) (Thelwall & Fairclough, 2015).

Spearman correlations were used to compare Mendeley reader counts and Microsoft Academic citation counts separately for each year. Spearman is preferable to Pearson because the data is skewed. Correlations need to be calculated separately for different years because citation counts and reader counts increase over time, generating spuriously high correlations due to the common influence of time. Correlations should also be calculated separately for fields because there are different citation rates between fields but the data does not have a natural classification scheme. Although Microsoft Academic records incorporate multiple classifications, these have been shown to be not useful for scientometric purposes (Hug, Ochsner, & Brändle, 2017). The lack of subject classification is therefore a limitation of the analysis.

The goal of the correlation test is to provide evidence whether Mendeley readers and Microsoft Academic citations could reflect the same underlying factors, such as academic impact (Sud & Thelwall, 2014). In some cases, Mendeley readers will directly lead to Microsoft Academic citations because a Mendeley user has had an article published. This would probably account for a maximum of 10% of cases since most scientists do not use Mendeley (Van Noorden, 2014). A more complex mathematical model might be able to take into account the likely time delay between Mendeley readers and Microsoft Academic citations when estimating the underlying strength of association. This is less relevant than the correlation coefficient for research evaluators deciding which source to use, however.

Results

RQ1: Microsoft Academic's coverage of SSRN

The Microsoft Academic 'ssrn electronic journal' queries returned 41923 papers, of which 25290 (60.3%) included a SSRN DOI (e.g., 10.2139/ssrn.270780) and a further 580 contained a non-SSRN DOI (e.g., 10.6092/unibo/amsacta/5426). SSRN's homepage[7] on 14 August 2017 stated, "SSRN´s eLibrary provides 751,159 research papers". The API figure closely matches

[7] https://www.ssrn.com/en/

the figure reported on the Microsoft Academic homepage for SSRN[8] on the download date (41,926), suggesting that the API returns all results known to Microsoft Academic. Microsoft Academic therefore indexes 5.6% of SSRN papers within its *SSRN Electronic Journal*. The following contribute to this.

1. SSRN includes duplicates, such as updates and revisions, and/or non-academic documents within its reported number of research papers. For example, the final version of paper 2171622 is version 7, and SSRN supports "version groups" to include multiple versions of the same paper[9].
2. SSRN has added papers since the last Bing crawl processed by Microsoft Academic. This must be true since it is a large site and articles can be added at any time (one every 8 minutes, on average in 2017[10]).
3. Bing/Microsoft Academic is unable to crawl SSRN comprehensively. This seems unlikely to be a significant cause because, in a trial, all 25 SSRN papers matching the query 'Scientometrics' in SSRN were all found in by a Bing search for them in SSRN (e.g., site:papers.ssrn.com "Monitoring Global Supply Chains").
4. Bing/Microsoft Academic cannot read all SSRN file formats. This may occur if the text is stored in image format within a PDF file.
5. Microsoft Academic does not recognise all SSRN papers as being academic, or some are not academic. For example, 2817493 is "Sample Student Project Assignments for Public Health Law Seminar".
6. Microsoft Academic associates SSRN records with subsequently published journal articles, conference papers or book chapters and reports them as part of these instead of SSRN. For example, 897063 is in Microsoft Academic only as a journal article. It lists papers.ssrn.com as a "source", proving that Microsoft Academic knows it to be in SSRN. This is probably the most important factor, as discussed below.
7. Microsoft Academic categorises SSRN records as individual preprints or in other repositories rather than regarding them as part of *SSRN Electronic Journal*. For example, 2257540 has two records in Microsoft Academic – one for an arXiv copy and one for a RePEc copy. This article is indexed by Bing in SSRN but there is no mention of SSRN in the two matching Microsoft Academic records.

Microsoft Academic's coverage of SSRN is clearly incomplete. For example, both the author's SSRN papers were not returned by Microsoft Academic. One was subsequently published as a journal article (1734850) and is indexed as such by Microsoft Academic[11] (number 6 in the list above) and the other (2587962) was also indexed by Microsoft Academic with an SSRN DOI but was not recorded as part of the SSRN eJournal[12] (number 7 above). This paper contains no metadata that would allow a SSRN-related Microsoft Academic query to be constructed to download it because the information that can be queried is author name, year, authors, journal, paper id, title[13]. Thus, the set of papers classified by Microsoft Academic as *SSRN Electronic Journal* are a subset of the SSRN papers indexed by Microsoft Academic.

[8] https://academic.microsoft.com/#/detail/60730585
[9] https://www.ssrn.com/en/index.cfm/ssrn-faq/#version_group
[10] https://papers.ssrn.com/sol3/displayabstractsearch.cfm
[11] https://academic.microsoft.com/#/detail/2133840132
[12] https://academic.microsoft.com/#/detail/2199326065
[13] https://docs.microsoft.com/en-gb/azure/cognitive-services/academic-knowledge/entityattributes

SSRN does not provide a complete list of its articles, does not have a public API and does not allow fast web crawling so it is not possible to generate a random sample of SSRN articles to check. Instead, random numbers were generated on 22 December 2017 up to the number of the article most recently posted (3088032, from browsing the archive by date). These random numbers were tested by adding them to the URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id= to generate an SSRN URL. In 76% of cases an error message was returned by SSRN but the first 100 URLs without an error messages were checked for presence in Microsoft Archive. This is a random sample of 100 SSRN records. Each article was searched for in Microsoft Academic using its web interface. Of these, 89 (i.e., 89% of the sample) had a Microsoft Academic page and 84 of these pages included a link to SSRN. Five articles were recorded as published in the "Social Science Research Network" journal, 29 in other journals and 55 were not recorded as being part of any journal. Microsoft Academic's coverage of SSRN is extensive but it assigns only about 5% of articles that it has found in SSRN to SSRN as a journal. The records not found in Microsoft academic had short names (e.g., "Insurance"), were book reviews, book chapters, conference papers or panels, and one was recent posted 9 days before the check. Thus, Microsoft Academic might not index SSRN records that it is not able to identify as scholarly contributions.

It is not clear how Microsoft Academic classifies papers found in SSRN as being part of *SSRN Electronic Journal* (later renamed *Social Science Research Network*). For example, "The Principle of Effective Legal Protection in Administrative Law" (SSRN ID: 2839823; Microsoft Academic ID: 2523456852) is in *International Administrative Law eJournal* from SSRN and in Microsoft Academic but is not associated with the Microsoft Academic SSRN eJournal. Thus, Microsoft Academic's *SSRN Electronic Journal* is not just the combination of all SSRN's eJournals.

To investigate articles classified by Microsoft Academic as *SSRN Electronic Journal* papers further, SSRN identification numbers were plotted against date for papers with a DOI (which contains the SSRN number) to explore the discrepancy between the SSRN reported number of papers and the MA count of papers (Figure 1). The highest SSRN number of any paper with a DOI was 2,993,152, suggesting that 25% of SSRN numbers are in use. The missing 75% of IDs may be for revised or withdrawn papers or used for other purposes. The graph shows several characteristics that illuminate the SSRN system, irrespective of Microsoft Academic's coverage of it.

- From the dominant linear trend in the bottom right of Figure 1, SSRN IDs up to about 2,700,000 were usually given out in approximate publication date order.
- Outlier dots to the right of the main diagonal line are probably for papers that were updated or published in a journal unusually long after initial deposit.
- Outlier dots to the left of the main diagonal line are probably for papers that were deposited long after the paper had been published elsewhere.
- The two lines to the left of the main diagonal line are probably for groups of papers that were deposited long after they had been published elsewhere – such as a for a journal or working paper series.
- On the main diagonal, there is an initial thick low slope from about 1998 to 2005, then a thinner steeper slope to about 2016 then a thick horizontal bar in the top right-hand corner.
    - The steeper slope from about 2005 shows that at this time there was a big increase in the allocation of SSRN IDs, and presumably also of the depositing

of papers. The thinness of this slope suggests that fewer of these IDs were used for documents, many more of the documents were eventually withdrawn from SSRN or Microsoft Academic found fewer of them.

- The thick bar about SSRN ID 2,700,000 shows that from 2016 there was a systematic attempt (from Elsevier?) to import large numbers of previously published documents into SSRN. If Microsoft Academic's coverage is systematic, then this import focused on papers from about 2010 or is ongoing and will eventually reach earlier years.

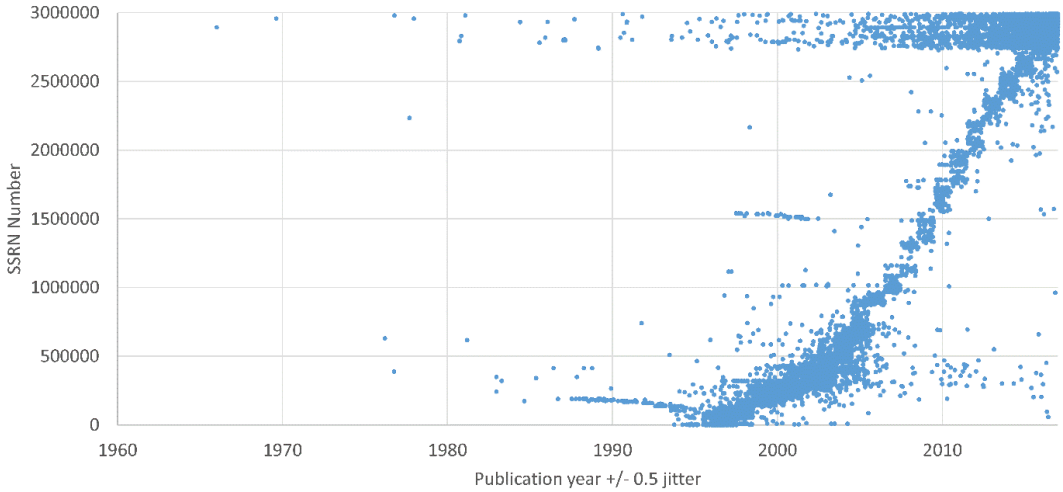

Figure 1. SSRN number against Microsoft Academic publication date with +/- 0.5 year jitter added; for *SSRN Electronic Journal* papers in Microsoft Academic with a DOI.

The gaps between the IDs used (Figure 2) are large in the middle range of SSRN IDs. Although they are typically single digit for low or high numbers, they average above 1000 at three points in time. This confirms large gaps in SSRN or in Microsoft Academic's indexing of it. Manual checks of some of the gaps confirmed that some of the IDs were no longer in use.

Random checks of papers in SSRN were made to find examples that were not indexed in any form by Microsoft Academic but none were found, although the most recent SSRN papers would presumably be missing, due to indexing delays. It seems, therefore, that Microsoft Academic indexes SSRN quite comprehensively but does not support query syntax that would return a complete set of SSRN papers (i.e., numbers 6 and 7 in the list above are most important).

In summary, there are gaps in the SSRN IDs used by SSRN, with no obvious reason for them; the depositing pattern in SSRN has changed substantially twice; some papers in SSRN are indexed by Microsoft Academic but not returned as part of *SSRN Electronic Journal;* Microsoft Academic's coverage of SSRN seems to be close to comprehensive.

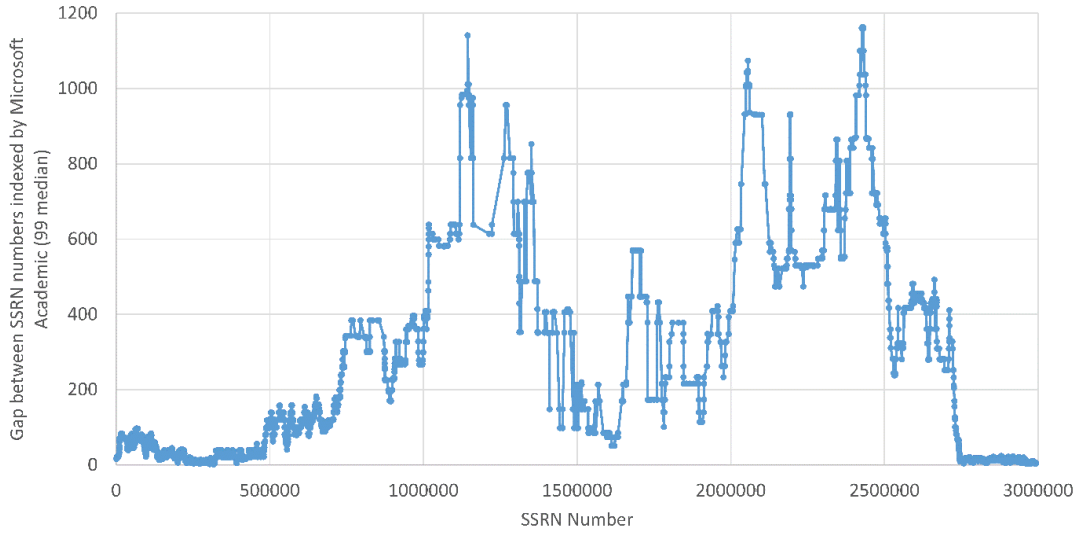

Figure 2. Average (median over 99 consecutive MA-indexed papers) SSRN number gap between consecutive MA-indexed *SSRN Electronic Journal* papers with DOIs.

The number of publications found for each year varied broadly in line with the above findings (Figure 3). The peak years for deposits returned by Microsoft Academic were 2003 and 2016, with a substantial proportion of the articles before 2006 missing a DOI. Using Mendeley or Microsoft Academic dates makes a small difference to the number of articles in each year.

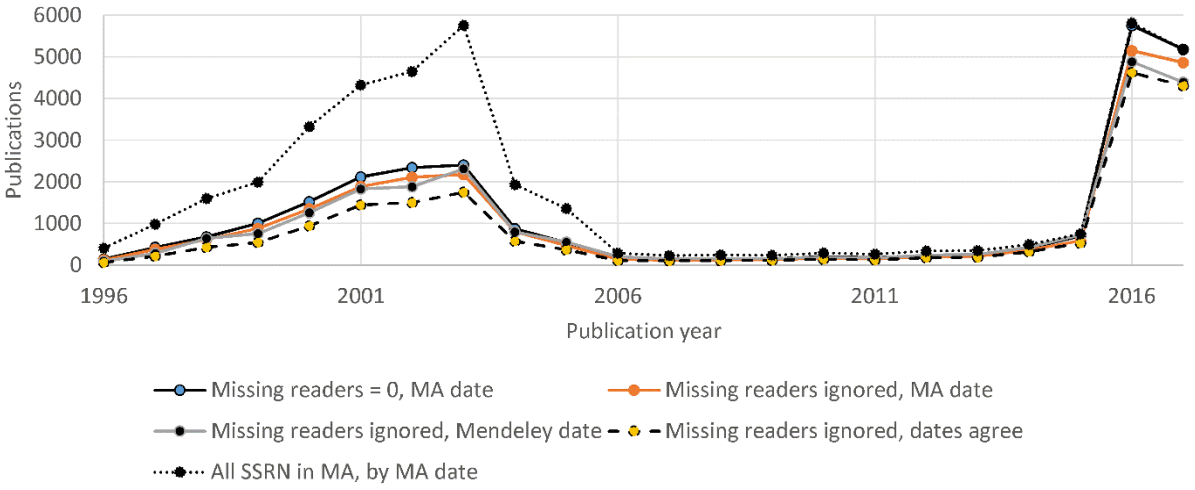

Figure 3. Number of *SSRN Electronic Journal* publications indexed by Microsoft Academic (MA) 1996-2017 and four subsets.

RQ2: Microsoft Academic citation counts for SSRN

The average number of citations for *SSRN Electronic Journal* papers indexed by Microsoft Academic is higher for older articles, reflecting citations taking time to accrue. (Figure 4). Whilst the average varies by dataset, the trend is the same for all. There is an anomalous increase in the average for 2006 and 2007. The steepness of the slope from 1999 to 2002 is also surprising given that these papers are all over 15 years old and so their citation counts should mostly have stabilised. Both anomalies point to non-regularity of the indexing of *SSRN Electronic Journal* by Microsoft Academic or the contents of SSRN. These anomalies cast doubt on the validity of using field normalised indicators on this data set. For example,

it seems that unusually high impact papers were indexed in 2006 or 2007 and so a field normalised indictor for a more typical paper in these years would be unfairly low.

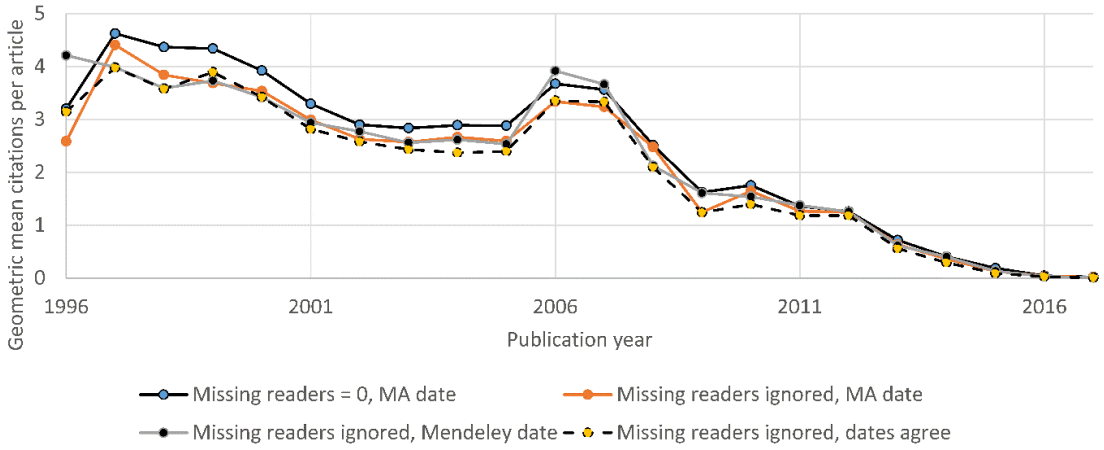

Figure 4. Average (geometric mean) number of Microsoft Academic citations per paper for three subsets (one twice, with different publication dates) of Microsoft Academic *SSRN Electronic Journal* papers.

The Mendeley reader counts gradually increase from 1996 to 2011 and then sharply decrease (Figure 5). This broad pattern is reasonable for Mendeley given that it is a relatively new tool and researchers tend to rely upon recent research more than upon older papers. The citation anomaly in Figure 4 for 2006-7 is not echoed in the Mendeley data except for the *Missing readers ignored, Mendeley date* set. This discrepancy is possible because of the relatively small number of articles published in 2006 and the high proportion of discrepancies. For example, 196 articles had a Microsoft Academic year of 2006 but only 109 also had a Mendeley year of 2006. Ignoring this exception, it looks like a substantial number of articles in SSRN 2006-7 have too many citations. The most likely explanation seems to be a degree of group self-citation for a set of articles from this period. Almost a quarter of the papers from 2006 with publication years agreeing (24 out of 109) had U.S. Federal Reserve System authors, for example, so a degree of self-citation amongst these would explain the result.

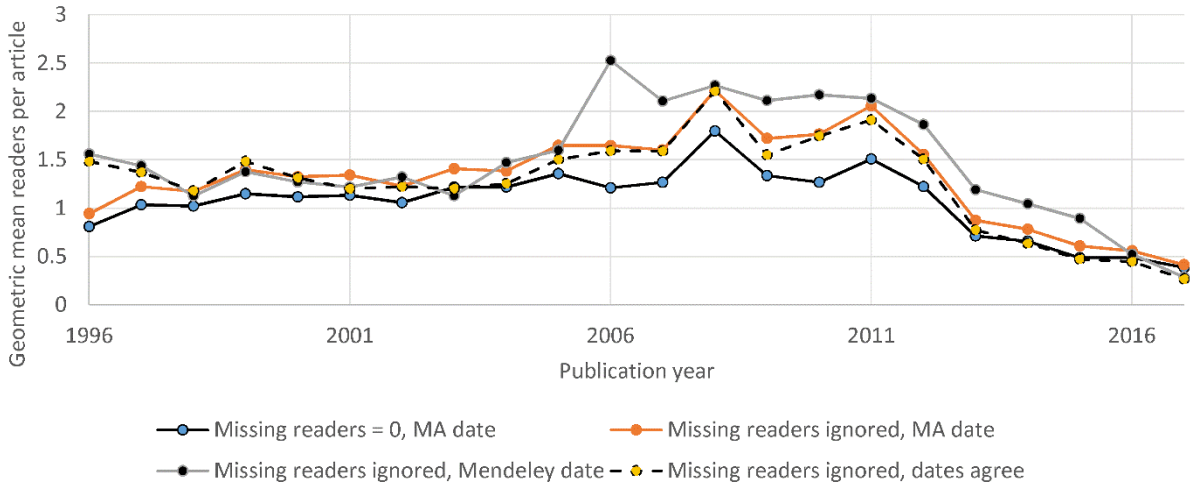

Figure 5. Average (geometric mean) number of Mendeley readers per paper for three subsets (one twice) of Microsoft Academic *SSRN Electronic Journal* papers.

Citation counts are higher than Mendeley reader counts for published before 2011, about the same for articles 2011-2013 and lower for subsequent articles (Figure 6). The difference is especially large for articles from the current year. This pattern (except for the 2015-7 anomaly) as in line with previous comparisons of the two data sources for journals (Thelwall, 2018, 2017b).

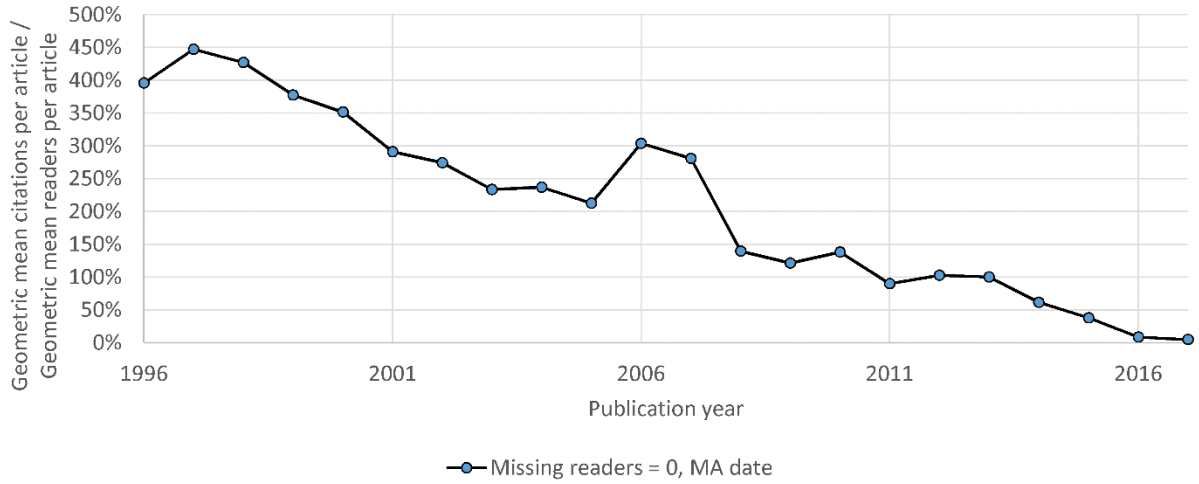

Figure 6. Microsoft Academic citations per Mendeley reader for *SSRN Electronic Journal* publications in Microsoft Academic with a DOI, against the publication year recorded in Microsoft Academic.

RQ3: Traditional citation impact

Spearman correlations between Microsoft Academic citation and Mendeley readers are moderate except for recent articles (Figure 7). The lower correlations 2006-7 reflect the anomaly discussed above. The low correlations for recent years are an expected side-effect of the very low numbers (overwhelmingly 0) from Microsoft Academic. These correlations are consistent with, but do not prove, the hypothesis that Microsoft Academic citations reflect traditional citation impact (RQ3). More information would be needed to fully support this claim because Mendeley readers have been used as a proxy for citation counts. The correlations may underestimate the degree of association between Mendeley and Microsoft Academic since Mendeley has multiple records for some articles, not all of which may have been found by Webometric Analyst and Microsoft Academic may have merged multiple versions of articles together, perhaps not always correctly.

Figure 7. Spearman correlations between Mendeley readers and Microsoft Academic citations for *SSRN Electronic Journal* publications calculated separately for each year 1996-2017. Error bars indicate 95% confidence intervals (Fisher transformation). Jitter has been added to years to reduce error bar overlaps.

Discussion

The results are limited by date since the coverage and features of Microsoft Academic may evolve over time. They are also limited by the methods to search for SSRN-indexed publications: it is possible that there is another method to find them that has not been found. From a broader perspective, the results are also limited by the choice of repository and so may not generalise to other types.

The correlation results are limited by the absence of an effective method to decide whether a paper not found in Mendeley is absent from Mendeley. It is therefore not possible to determine how many of the papers not found in Mendeley have no Mendeley readers. This issue was addressed through multiple data sets (e.g., Figure 7). Assuming that less correct data would generate lower correlations, the most plausible explanation is that most publications not found in Mendeley were uncited. If a method was available to separate out the articles with no Mendeley readers from the articles not found by the DOI searches then the correlations would presumably be higher.

The results are of a different type to previous studies of Microsoft Academic and so cannot be directly compared. They agree with previous findings that Mendeley reader counts are higher than Microsoft Academic citation counts for newer articles and lower for old articles (Thelwall, in press, 2017b). The Microsoft Academic citation counts are likely to be higher than Scopus citation counts for SSRN articles since a previous study found that 73% of SSRN papers that are cited in Scopus are only cited once (Li, Thelwall, & Kousha, 2015). Extrapolating and assuming that most SSRN papers are not cited by Scopus, the average Scopus citation count should be considerably lower than 1.

Another de-facto preprint archive is ResearchGate (Jamali, 2017), but Microsoft Academic does not have a named journal for it and so it is not possible to systematically harvest all papers from ResearchGate with Microsoft Academic queries. The same is true for the Astrophysics Data System, but there are Microsoft Academic journals for others, such as bioRxiv (*biorxiv*), and several for arXiv.

Comparison with arXiv

Although this paper focuses on SSRN, a comparison with another archive is useful to assess whether the treatment of SSRN might be untypical. The physics preprint repository arXiv was chosen for this as a popular simple format archive. It is indexed by Microsoft Academic into 148 journals, each corresponding to an arXiv category. For example, the arXiv category Cosmology and Nongalactic Astrophysics with arXiv name astro-ph.CO corresponds to the Microsoft Academic journal *arxiv astro ph co*. Unlike SSRN, ArXiv does not assign DOIs to any of its articles but authors can add DOIs from other sources. All records from the 148 arXiv journals were downloaded from Microsoft Academic on 13 August 2017 (Table 1). All arXiv records were also downloaded from arXiv.org via its API in August 2017 for comparison purposes.

The main two differences with SSRN are that (a) because arXiv does not assign its own DOIs, it is difficult to make accurate comparisons with Mendeley readers or other citation sources for the articles that lack these, and (b) Microsoft Academic separates arXiv articles into multiple field-based journals rather than a single journal. For example, the article with Microsoft Academic ID 2529095144 is registered as published in the Microsoft Academic journal *arXiv: High Energy Physics – Phenomenology* only, but in arXiv it is in both High Energy Physics – Phenomenology (hep-ph) and High Energy Physics - Experiment (hep-ex).

Microsoft Academic's records from arXiv journals accounted for 28% of arXiv's 1294141 records: 7% of arXiv's 709632 records with DOIs and 54% of its 584509 records without DOIs (comparisons cannot be conducted at the category level because many articles have multiple arXiv categories). The discrepancy is probably due to arXiv articles with DOIs usually being assigned to their publishing journal rather than arXiv (number 6 in the RQ1 list above). Some arXiv records are indexed by Microsoft Academic but attributed to non-journal sources (e.g., 1624351243) such as the Astrophysics Data System (adsabs.harvard.edu), which incorporates all arXiv papers but does not seem to have its own Microsoft Academic journal name (number 7 in the RQ1 list above). Ad-hoc investigations within the site failed to find articles that were not indexed by Microsoft Academic in any form.

In summary, since arXiv records without DOIs may also be journal articles that lack DOIs or for which the authors have not added DOIs, the results are consistent with Microsoft Academic having close to comprehensive coverage of arXiv but tending to assign articles that are published in journals to those journals rather than to arXiv journals.

Combining the arXiv and SSRN results, Microsoft Academic is able to index open archives quite comprehensively, but its "journals" cannot be relied upon for finding the indexed papers.

Table 1. Numbers of articles extracted from 148 Microsoft Academic arXiv Journals on 13 August 2017, organised into broad categories.

Category	MA root name	MA records	MA records with DOI
High Energy Physics	arxiv hep	55099	11688 (21%)
Astrophysics	arxiv astro ph	17060	3451 (20%)
Condensed Matter	arxiv cond mat	34447	5270 (15%)
Computer Science	arxiv cs	56668	6454 (11%)
Gen. Relativity & Quantum Cosmology	arxiv gr qc	10563	1798 (17%)
Mathematics	arxiv math	117122	11047 (9%)
Nonlinear Sciences	arxiv nlin	4113	528 (13%)
Nuclear Physics	arxiv nucl	8378	2592 (31%)
Quantitative Biology	arxiv q bio	4867	483 (10%)
Physics	arxiv physics	29433	3564 (12%)
Quantitative finance	arxiv q fin	3065	885 (29%)
Quantum Physics	arxiv quant ph	16368	2141 (13%)
Statistics	arxiv stat	7580	668 (9%)
Total	arxiv	**364763**	**50569 (14%)**

Conclusions

The results suggest that it is impossible to design a query to identify all SSRN papers that it indexes and the same is true for arXiv and probably all other digital repositories. The main root cause is Microsoft Academic assigning a small percentage of articles to repositories out of the total amount that it indexes from them. It would be possible to identify the records in Microsoft Academic through metadata searches (e.g., title, authors, publication years) with an expected 90% success rate (Hug & Brändle, 2017; Thelwall, 2018) but this process would require an initial comprehensive set of repository papers to check. This is in addition to the previously-identified field categorisation problems (Hug, Ochsner, & Brändle, 2017). Thus, Microsoft Academic indexes enough papers in preprint archives and finds enough citations to them to be used for comprehensive analyses of the role of preprint archives in scholarly communication when complete sets of papers are available from other sources. Metadata searches would consume one query per paper and would therefore be slower and more expensive than journal-based queries. They may also introduce retrieval biases.

For the case of SSRN, the content indexed by Microsoft Academic within the *Social Science Research Network* "journal", or published by SSRN, has changed character over time, as reflected by changes in papers per year uploaded, which is likely to bias any citation analysis based upon normalised indicators. This further undermines the use of Microsoft Academic's SSRN journal set for evaluative citation analysis purposes.

If Microsoft Academic could provide a simple mechanism to identify all documents in a repository, even if subsequently published in a traditional journal or available elsewhere, then this would greatly facilitate future evaluations of digital repositories. An effective field categorisation scheme would also help to support effective field normalisation for citation-based comparisons between fields or years.

References

Brown, L. D., & Laksmana, I. (2004). Ranking accounting Ph. D. programs and faculties using social science research network downloads. Review of Quantitative Finance and Accounting, 22(3), 249-266.

Brown, L. D. (2003). Ranking journals using social science research network downloads. Review of Quantitative Finance and Accounting, 20(3), 291-307.

Davis, P., & Fromerth, M. (2007). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? Scientometrics, 71(2), 203-215.

Delgado López-Cózar, E., & Cabezas-Clavijo, Á. (2012). Google Scholar Metrics: An unreliable tool for assessing scientific journals. El Profesional de la Información, 21(4), http://www.elprofesionaldelainformacion.com/contenidos/2012/julio/15_eng.pdf

Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. Journal of the Association for Information Science and Technology, 65(3), 446-454.

Di Cesare, R., Luzi, D., Ricci, M., Ruggieri, R., della Ricerche, C. N., & della Repubblica, S. (2011). A profile of Italian Working papers in RePEc. In Proceedings of the Twelfth International Conference on Grey Literature. Amsterdam: TextRelease (pp. 1-12).

Eisenberg, T. (2006). Assessing the SSRN-Based Law School Rankings. Indiana Law Journal, 81(1), 285-291.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. The FASEB Journal, 22(2), 338-342.

Gunn, W. (2013). Social signals reflect academic impact: What it means when a scholar adds a paper to Mendeley. Information Standards Quarterly, 25(2), 33-39.

Halevi, G., Moed, H., & Bar-Ilan, J. (2017). Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation - Review of the Literature. Journal of Informetrics, 11(3), 823-834.

Harzing, A. W., & Alakangas, S. (2017a). Microsoft Academic: Is the phoenix getting wings? Scientometrics, 110(1), 371-383.

Harzing, A. W., & Alakangas, S. (2017b). Microsoft Academic is one year old: The Phoenix is ready to leave the nest. Scientometrics, 112(3), 1887-1894.

Harzing, A. W. K., & Van der Wal, R. (2008). Google Scholar as a new source for citation analysis. Ethics in Science and Environmental Politics, 8(1), 61-73.

Harzing, A.W. (2007). Publish or Perish. http://www.harzing.com/pop.htm

Harzing, A. W. (2016). Microsoft Academic (Search): A phoenix arisen from the ashes? Scientometrics, 108(3), 1637-1647.

Haustein, S., Larivière, V., Thelwall, M., Amyot, D., & Peters, I. (2014). Tweets vs. Mendeley readers: How do these two social media metrics differ? IT-Information Technology, 56(5), 207-215.

HEFCE (2015). The Metric Tide: Correlation analysis of REF2014 scores and metrics (Supplementary Report II to the Independent Review of the Role of Metrics in Research Assessment and Management). http://www.hefce.ac.uk/pubs/rereports/Year/2015/metrictide/Title,104463,en.html

Hug, S. E., & Brändle, M. P. (2017). The coverage of Microsoft Academic: Analyzing the publication output of a university. Scientometrics. doi:10.1007/s11192-017-2535-3

Hug, S. E., Ochsner, M., & Brändle, M. P. (2017). Citation analysis with Microsoft Academic. Scientometrics, 111(1), 371-378. doi:10.1007/s11192-017-2247-8

Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. the Quarterly journal of Economics, 108(3), 577-598.

Jamali, H. R. (2017). Copyright compliance and infringement in ResearchGate full-text journal articles. *Scientometrics*, 112(1), 241-254.

Karki, M. M. S. (1997). Patent citation analysis: A policy analysis tool. World Patent Information, 19(4), 269-272.

Kousha, K, Thelwall, M. & Abdoli, M. (2018). Can Microsoft Academic assess the early citation impact of in-press articles? A multi-discipline exploratory analysis. Journal of Informetrics, 12(1), 287-298.

Li, X., Thelwall, M., & Kousha, K. (2015). The role of arXiv, RePEc, SSRN and PMC in formal scholarly communication. Aslib journal of information management, 67(6), 614-635.

Luce, R. E. (2001). E-prints intersect the digital library: inside the Los Alamos arXiv. Issues in science and technology librarianship, 29(Winter). http://webdoc.sub.gwdg.de/edoc/aw/ucsb/istl/01-winter/article3.html

Maflahi, N, & Thelwall, M. (2018). How quickly do publications get read? The evolution of Mendeley reader counts for new articles. Journal of the Association for Information Science and Technology, 69(1), 158–167.

Mohammadi, E., Thelwall, M. & Kousha, K. (2016). Can Mendeley bookmarks reflect readership? A survey of user motivations. Journal of the Association for Information Science and Technology, 67(5), 1198-1209. doi:10.1002/asi.23477

Orduña-Malea, E., Martín-Martín, A., & Delgado-López-Cózar, E. (2016). The next bibliometrics: ALMetrics (Author Level Metrics) and the multiple faces of author impact. El Profesional de la Información, 25(3), 485-496.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B. J. P., & Wang, K. (2015). An overview of Microsoft Academic service (mas) and applications. In Proceedings of the 24th international conference on world wide web (pp. 243-246). New York, NY: ACM Press.

Sud, P. & Thelwall, M. (2014). Evaluating altmetrics. Scientometrics, 98(2), 1131-1143. 10.1007/s11192-013-1117-2

SSRN (2017). Is my paper eligible for distribution in a SSRN eJournal? https://www.ssrn.com/en/index.cfm/ssrn-faq/#distribution_eligibility

Thelwall, M. & Fairclough, R. (2015). Geometric journal impact factors correcting for individual highly cited articles. Journal of Informetrics, 9(2),263–272.

Thelwall, M., Haustein, S., Larivière, V. & Sugimoto, C. (2013). Do altmetrics work? Twitter and ten other candidates. PLOS ONE, 8(5), e64841. doi:10.1371/journal.pone.0064841

Thelwall, M. & Sud, P. (2016). Mendeley readership counts: An investigation of temporal and disciplinary differences. *Journal of the Association for Information Science and Technology*, 57(6), 3036-3050. doi:10.1002/asi.2355

Thelwall, M. & Wilson, P. (2016). Mendeley readership altmetrics for medical articles: An analysis of 45 fields, *Journal of the Association for Information Science and Technology*, 67(8), 1962-1972. doi:10.1002/asi.23501

Thelwall, M. (2017a). Are Mendeley reader counts high enough for research evaluations when articles are published? *Aslib Journal of Information Management*, 69(2), 174-183. doi:10.1108/AJIM-01-2017-0028

Thelwall, M. (2017b). Microsoft Academic: A multidisciplinary comparison of citation counts with Scopus and Mendeley for 29 journals. Journal of Informetrics, 11(4), 1201-1212.

Thelwall, M. (2017c). Are Mendeley reader counts useful impact indicators in all fields? Scientometrics, 113(3), 1721–173.

Thelwall, M. (in press). Does Microsoft Academic find early citations? *Scientometrics*. doi:10.1007/s11192-017-2558-9 http://wlv.openrepository.com/wlv/bitstream/2436/620806/1/DoesMicrosoftAcademicFindEarlyCitations_Preprint.pdf

Thelwall, M. (2018). Microsoft Academic automatic document searches: Accuracy for journal articles and suitability for citation analysis. Journal of Informetrics, 12(1), 1-9.

van Leeuwen, T. N., & Calero Medina, C. (2012). Redefining the field of economics: Improving field normalization for the application of bibliometric techniques in the field of economics. *Research Evaluation*, 21(1), 61-70.

Van Noorden, R. (2014). Online collaboration: Scientists and the social network. Nature, 512(7513), 126-129.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: An empirical analysis. *Scientometrics*, 87(3), 467-481.

West, J. D., Jensen, M. C., Dandrea, R. J., Gordon, G. J., & Bergstrom, C. T. (2013). Author-level Eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. *Journal of the Association for Information Science and Technology*, 64(4), 787-801.

Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A crossdisciplinary analysis of the presence of 'alternative metrics' in scientific publications. Scientometrics, 101(2), 1491-1513.

Zahedi, Z., Haustein, S. & Bowman, T. (2014). Exploring data quality and retrieval strategies for Mendeley reader counts. Presentation at SIGMET Metrics 2014 workshop, 5 November 2014. Available: http://www.slideshare.net/StefanieHaustein/sigmetworkshop-asist2014

Zimmermann, C. (2013). Academic rankings with RePEc. Econometrics, 1(3), 249-280.

Zitt, M. (2012). The journal impact factor: Angel, devil, or scapegoat? A comment on JK Vanclay's article 2011. *Scientometrics*, 92(2), 485-503.