ResearchGate vs. Google Scholar: Which finds more early citations?[1]

Mike Thelwall, Kayvan Kousha
Statistical Cybermetrics Research Group, University of Wolverhampton, UK.

ResearchGate has launched its own citation index by extracting citations from documents uploaded to the site and reporting citation counts on article profile pages. Since authors may upload preprints to ResearchGate, it may use these to provide early impact evidence for new papers. This article assesses the whether the number of citations found for recent articles is comparable to other citation indexes using 2,675 recently-published library and information science articles. The results show that in March 2017, ResearchGate found less citations than did Google Scholar but more than both Web of Science and Scopus. This held true for the dataset overall and for the six largest journals in it. ResearchGate correlated most strongly with Google Scholar citations, suggesting that ResearchGate is not predominantly tapping a fundamentally different source of data than Google Scholar. Nevertheless, preprint sharing in ResearchGate is substantial enough for authors to take seriously.

Keywords: ResearchGate, early impact, citation analysis, altmetrics, academic social network sites.

Introduction

Citation counts are frequently used to support research evaluations, for example to help compare the relative merits of individual researchers or research groups. An ongoing problem with traditional citation is that they take several years to appear in the Web of Science (WoS) and Scopus due to publication and publishing delays. This is a major drawback for research evaluators because the most recent research seems likely to be the most relevant for an evaluation. In response, several alternatives have been proposed for early impact data. These include social web citations, altmetrics (Priem, Taraborelli, Groth, & Neylon, 2010), and general web citations, webometrics (Vaughan & Shaw, 2003), as well as article download counts (Moed, 2005). Google Scholar is another logical alternative because its index can exploit public web documents, although its data can be time consuming to manually collect (Meho & Yang, 2007), when the Publish or Perish software (Harzing & Van Der Wal, 2009) is not suitable. Google Scholar seems to index more citations than Scopus (Moed, Bar-Ilan, & Halevi, 2016), which in turn has a bigger citation index than the Web of Science (WoS) (Mongeon & Paul-Hus, 2016). Another potential source is the citation data provided by ResearchGate since this is based upon an apparently large collection of publicly shared preprints, postprints and other documents. About half (51%) of the 78% user-uploaded articles (n=500) that are not open access violate publisher copyright agreements (Jamali, in press). This uploading may occur because authors believe that it will attract a greater audience for their work, and there is empirical evidence from Academia.adu that posting to an academic social network site helps to attract more citations than does posting to other parts of the public web (Niyazov, Vogel, Price, Lund, Judd, Akil, & Shron, 2016). More generally, some researchers use academic social network

[1] Thelwall, M., & Kousha, K. (in press). ResearchGate versus Google Scholar: Which finds more early citations? Scientometrics. 10.1007/s11192-017-2400-4

sites as the primary mechanism for document sharing (Laakso, Lindman, Shen, Nyman, Björk, 2017).

ResearchGate is part of a general rise in the importance of professional social network sites (Brandão & Moro, 2017). It is the most regularly used professional website for scientists, and the third most popular in the social sciences, arts and humanities, but Google Scholar is more popular in all cases (Van Noorden, 2014). Academic social networks like ResearchGate and Academia.edu seem to primarily replicate existing academic structures (Jordan, 2017), although they may give more space for younger researchers and women (e.g., Thelwall & Kousha, 2014). ResearchGate has allowed authors to upload their articles to the site since 2009 (ResearchGate, 2009). It added citation information to user profiles in 2013 (ResearchGate, 2013) and subsequently introduced the citation-related h-index (ResearchGate, 2016). Currently (April 2017), citation counts are displayed for individual articles in ResearchGate, along with the number of article reads and comments. The wide use of the site and the extensive uploading to it has apparently made it a competitor for Google Scholar in terms of a citation index derived from publicly-shared research papers.

ResearchGate provides an overall rating for each academic member, the RG Score, which reflects a combination of academic achievements and activities within the site (Orduña-Malea, Martín-Martín, & López-Cózar, 2016), although it correlates reasonably well with other indictors of academic prestige for individual researchers in at least one field (Yu, Wu, Alhalabi, Kao, & Wu, 2016). The number of times that an article has been viewed (now read) in ResearchGate has a positive correlation with its Scopus citation count, confirming that the site reflects scholarly-related activities and its indicators can be meaningful (Thelwall & Kousha, 2017). Despite this, the uptake of ResearchGate varies greatly on an international scale (Thelwall & Kousha, 2015) and so its data is likely to contain some systematic biases. Moreover, it can index low quality outputs, such as those from ghost journals (Memon, 2016) which may undermine its indicators.

Despite the apparent promise of ResearchGate citation counts, especially for recent papers, there is no research that compares their magnitudes with current citation indexes. The main research goal of this paper is therefore to assess the relative magnitude of the ResearchGate and Google Scholar citation counts. For completeness, these are also compared against WoS and Scopus. Since the ability of ResearchGate to index articles depends on journal copyright policies, it is possible that the relative magnitude of the citation counts may vary by journal, assuming a moderate amount of journal self-citation. Thus, the second research question assesses journal differences. Finally, if ResearchGate citations were to be used as an impact indicator then it is important to assess the extent to which they agree with the other sources.

- Which out of ResearchGate, Google Scholar, WoS and Scopus gives the most citations for recently published library and information science journal articles?
- Does the answer to the above question vary by journal?
- How similar are the rank orders of articles produced by the different sources?

Methods

English language research or review articles published in 86 Information Science & Library Science (IS&LS) journals during January 2016 to March 2017 were selected from the Thomson Reuters Web of Science (WoS). The list of IS&LS journals was extracted from Thomson Reuters Journal Citation Reports (JCR) Social Science 2015 edition.

DOIs of articles were searched through the syntax below using automatic Bing searches in Webometric Analyst (http://lexiurl.wlv.ac.uk) to locate article pages in ResearchGate by combining "DOI:" and the *site:researchgate.net/publication* command. Most ResearchGate publication pages contain DOIs of articles with "Reads, "Recommendations" and "Citations". The publication pages identified by the Bing searches were downloaded with SocSciBot (http://socscibot.wlv.ac.uk) and a program was written to extract the main bibliographic information and citation counts (if any) from the downloaded pages. ResearchGate citations were extracted from a crawl of the ResearchGate website in March 2017 at the maximum speed permitted (three pages per hour). Although ResearchGate appeared to allow unrestricted web crawling according to its robots.txt file in March 2017 (https://www.researchgate.net/robots.txt), in practice a speed of more than three pages per hour resulted in the additional requests returning blank pages. The titles of article from ResearchGate were matched with WoS records, giving 2,675 corresponding articles in both sources.

"DOI: 10.1007/s11192-016-2095-y" site:researchgate.net/publication

In order to save Scopus citations for further analysis, DOI of articles were searched in Scopus advance search option through OR operators (e.g., DOI (10.1108/ajim-03-2016-0036) OR DOI(10.1080/00048623.2016.1165645) OR …). The bibliographic and citation information of the records identified in Scopus were saved and matched with ResearchGate and WoS data through their DOIs. The *Publish or Perish* software (www.harzing.com/resources/publish-or-perish) was used to automatically extract Google Scholar citations to articles from each journal. Either ISSNs or journal names were searched in the *Google Scholar Query* option and publication years were limited to 2016-2017. Search results were saved and article titles were matched with the main data from ResearchGate, WoS and Scopus. From 2,675 records in the study, 244 had no matches from the Google Scholar automatic searches and were instead manually extracted from Google Scholar in March 2017 by article title searches.

Citation counts are highly skewed (de Solla Price, 1976) and so comparing mean citation counts could give a misleading impression of which source of citation data tends to give higher values. This problem can be remedied either by taking the geometric mean (Thelwall & Fairclough, 2015; Zitt, 2012) or by log-transforming the citation data with the formula ln(1+citations) to reduce skewing (Thelwall, 2017). In fact, since sets of citation counts tend to approximately follow a discretised lognormal distribution, whether for individual journals (Thelwall, 2016b) or entire fields (Thelwall, 2016a), it is reasonable to use normal distribution formula to calculate confidence intervals for the log-transformed data (Thelwall & Fairclough, in press; Thelwall, 2016c). Hence, log-transformed citation counts were used and the normal distribution formula, 1.96 +/- standard error, was used for 95% confidence intervals.

For the second question, average log-transformed citation counts were calculated for the journals with the most articles in the dataset, using 100 articles as a convenient cut-off. The choice of larger journals is pragmatic because smaller journals are less likely to produce statistically significant findings but will clutter the analysis.

For the third research question, Spearman correlations were calculated to assess the similarity in the rank orders produced by the different citation sources. Spearman is more appropriate than Pearson because it directly assesses rank order similarity. The results are likely to be misleadingly high because recently published articles have longer to attract citations than older articles, an unfair advantage. Hence, in the unlikely event that there is

no underlying (i.e., long term) correlation between the data sources, there is still likely to be a positive correlation between all of them. Thus, the correlations should not be interpreted as statistical evidence of a relationship between the citation sources, but it is nevertheless reasonable to compare the relative magnitudes of the correlations between different pairs of citation sources since the time lag is the same for all of them.

Results

ResearchGate found statistically significantly fewer citations than did Google Scholar, but more than both Scopus and Web of Science. Scopus found more citations than did WoS, although this excludes the results for 155 articles not indexed in Scopus (the *All articles* bar in Figure 1).

As a simple heuristic for interpreting the confidence limits in Figure 1, if the confidence intervals for two bars do not overlap then the difference is statistically significant at the 95% level. The converse is not necessarily true, however, because a small overlap is still consistent with statistical significance (Austin & Hux, 2002; Julious, 2004). Taking this into account, for all six large journals, the results are consistent with Google Scholar always tending to find more citations for each individual journal than ResearchGate, and with ResearchGate tending to find more than both WoS and Scopus, although the difference is smallest for *Scientometrics*.

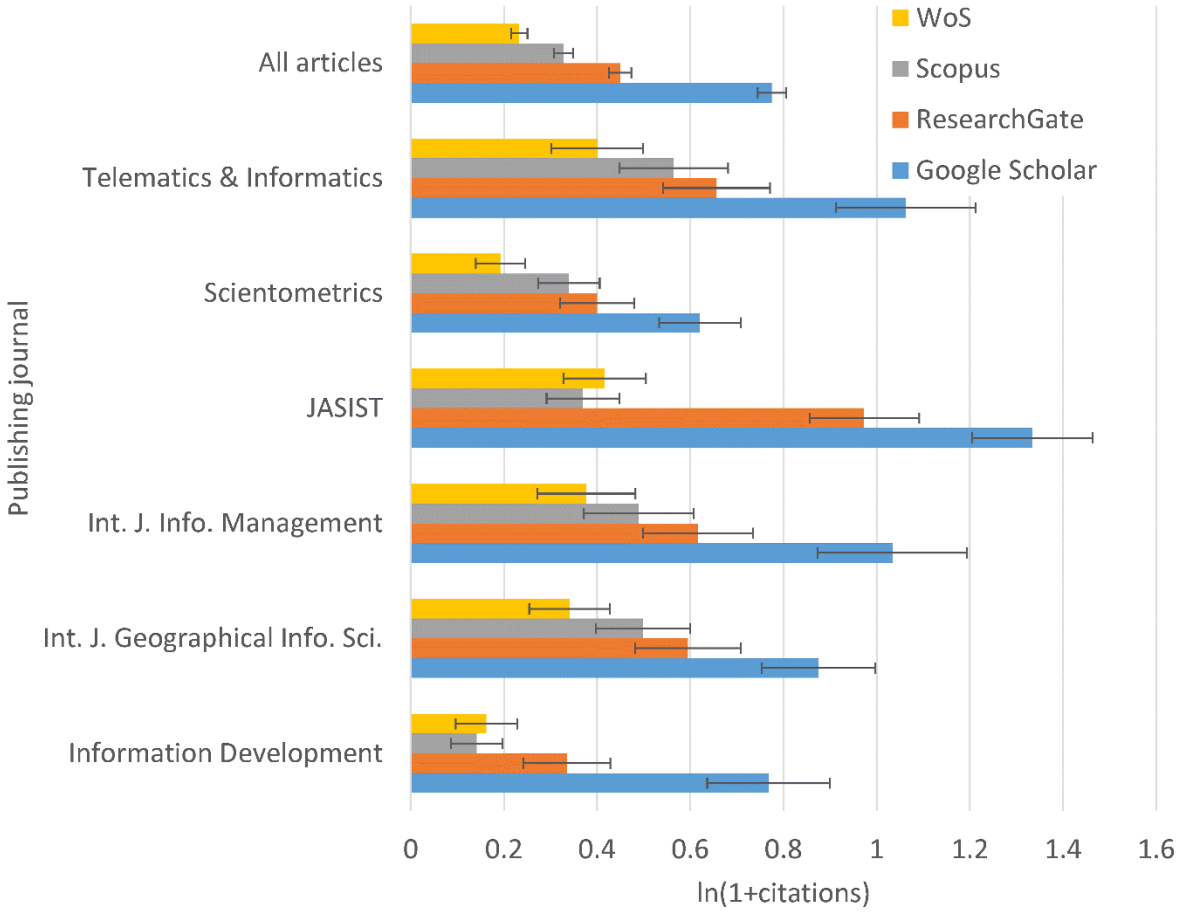

Figure 1. Log-transformed citation counts and 95% confidence intervals for the six journals with over 100 articles in the sample, as well as for all articles in the sample (n=2675 for all except n=2520 for Scopus, excluding non-indexed articles).

Out of all the pairs of data sources, the most similar article ranks are given by Google Scholar and ResearchGate (Table 1). It is perhaps surprising that this correlation is higher than that between WoS and Scopus, which presumably rely upon similar publisher data sources, but the reason may be the higher numbers of uncited articles in the latter case.

Table 1. Spearman correlations between citation counts from the four sources for all articles in the sample (n=2,675 for all correlations except those involving Scopus, otherwise n=2,520). All correlations are statistically significant at the 0.001 level, but this is misleading due to the shared influence of publication delays.

Citation source	Research Gate	WoS	Scopus	Google Scholar
Research Gate	1	0.609	0.587	0.732
WoS		1	0.635	0.582
Scopus			1	0.624
Google Scholar				1

Despite the overall results, there were individual articles for which there were many more Google Scholar citations than ResearchGate citations and some articles for which there were more ResearchGate citations. For example, "FEDS: a framework for evaluation in design science research" in the European Journal of Information Systems had 53 Google Scholar citations but only 6 ResearchGate citations. This was due to Google Scholar indexing documents from publishers (e.g., Springer) that were not available on the open web. At the other extreme, the paper "Evaluating the academic trend of RFID technology based on SCI and SSCI publications from 2001 to 2014" in Scientometrics had 30 ResearchGate citations but only 12 Google Scholar citations. All 30 citing documents in ResearchGate and all 12 Google Scholar citations were from PDF presentations uploaded by one of the authors (Nader Ale Ebrahim) and so in this case the results include no peer reviewed citations. Thus, there can be problems at the level of individual articles despite the overall positive correlations.

Limitations and conclusions

This study is limited by the focus on a single field and the results may not apply to other fields, particularly those that use ResearchGate less or upload preprints to ResearchGate less. The findings may also change over time if publishers enforce their copyright on ResearchGate more actively, if the popularity of ResearchGate changes, or if the indexing practices of Google Scholar change.

The results are primarily negative because they suggest that ResearchGate cannot yet challenge Google Scholar for early citation impact indicators. Moreover, although ResearchGate in theory allows automated data collection, unlike Google Scholar (except for *Publish or Perish*), its current maximum crawling speed is a major practical limitation on its use for large scale data gathering.

More generally, the results show that ResearchGate has indexed impressively many citations for a single website and has become a major source of academic papers, perhaps even starting to challenge Google Scholar in this regard. Combined with the apparent citation advantage of uploading to academic social network sites (Niyazov et al., 2016),

scholars should take ResearchGate seriously as a venue for disseminating their research. Nevertheless, like many web extracted indicators, such as Google Scholar citations (Delgado Lopéz-Cózar et al. 2014), ResearchGate citations can potentially be manipulated by uploading non-peer reviewed or fake documents and hence should be used cautiously for research evaluation.

References

Austin, P. C., & Hux, J. E. (2002). A brief note on overlapping confidence intervals. Journal of Vascular Surgery, 36(1), 194-195.

Brandão, M. A., & Moro, M. M. (2017). Social professional networks: A survey and taxonomy. Computer Communications, 100(1), 20–31.

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.

de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. Journal of the American society for Information science, 27(5), 292-306.

Halevi, G., & Moed, H. F. (2014). Usage patterns of scientific journals and their relationship with citations. Proceedings of the Science and Technology Indicators Conference 2014 (STI 2014), Leiden, Netherlands (pp. 241-251).

Harzing, A. W., & Van Der Wal, R. (2009). A Google Scholar h-index for journals: An alternative metric to measure journal impact in economics and business. Journal of the American Society for Information Science and Technology, 60(1), 41-46.

Jamali, H. R. (in press). Copyright compliance and infringement in ResearchGate full-text journal articles. Scientometrics. doi:10.1007/s11192-017-2291-4

Jordan, K. (2017). Understanding the structure and role of academics' ego-networks on social networking sites. PhD thesis, The Open University. http://oro.open.ac.uk/48259/

Julious, S. A. (2004). Using confidence intervals around individual means to assess statistical significance between two means. Pharmaceutical Statistics, 3(3), 217-222.

Laakso, M., Lindman, J., Shen, C., Nyman, L., Björk, B-C. (2017). Research output availability on academic social networks: Implications for stakeholders in academic publishing. Electronic Markets. doi:10.1007/s12525-016-0242-1

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. Journal of the American Society for Information Science and Technology, 58(13), 2105-2125.

Memon, A. R. (2016). ResearchGate is no longer reliable: Leniency towards ghost journals may decrease its impact on the scientific community. Journal of the Pakistan Medical Association, 66(12), 1643-1647.

Moed, H. F., Bar-Ilan, J., & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. Journal of Informetrics, 10(2), 533-551.

Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. Journal of the Association for Information Science and Technology, 56(10), 1088-1097.

Orduña-Malea, E., Martín-Martín, A., & López-Cózar, E. D. (2016). ResearchGate como fuente de evaluación científica: desvelando sus aplicaciones bibliométricas. El Profesional de la Información (EPI), 25(2), 303-310.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics, 106(1), 213-228.

Niyazov, Y., Vogel, C., Price, R., Lund, B., Judd, D., Akil, A., & Shron, M. (2016). Open access meets discoverability: Citations to articles posted to Academia.edu. PloS ONE, 11(2), e0148257.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. http://altmetrics.org/manifesto/

ResearchGate (2009). Self-Archiving Repository goes online. https://www.researchgate.net/blog/post/self-archiving-repository-goes-online

ResearchGate (2013). Introducing citations on ResearchGate. ResearchGate blog (7 February 2013). https://www.researchgate.net/blog/post/introducing-citations-on-researchgate

ResearchGate (2016). Introducing the h-index on ResearchGate. ResearchGate blog (8 March 2016). https://www.researchgate.net/blog/post/introducing-the-h-index-on-researchgate

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

Thelwall, M. & Fairclough, R. (in press). The accuracy of confidence intervals for field normalised indicators. Journal of Informetrics. doi:10.1016/j.joi.2017.03.004

Thelwall, M., & Kousha, K. (2014). Academia.edu: social network or academic network? Journal of the Association for Information Science and Technology, 65(4), 721-731.

Thelwall, M. & Kousha, K. (2015). ResearchGate: Disseminating, communicating and measuring scholarship? Journal of the Association for Information Science and Technology, 66(5). 876–889. doi:10.1002/asi.23236

Thelwall, M., & Kousha, K. (2017). ResearchGate articles: Age, discipline, audience size and impact. Journal of the Association for Information Science and Technology, 68(2), 468-479.

Thelwall, M. (2016a). Are the discretised lognormal and hooked power law distributions plausible for citation data? Journal of Informetrics, 10(2), 454-470.

Thelwall, M. (2016b). Citation count distributions for large monodisciplinary journals. Journal of Informetrics, 10(3), 863-874. doi:10.1016/j.joi.2016.07.006

Thelwall, M. (2016c). The discretised lognormal and hooked power law distributions for complete citation data: Best options for modelling and regression. Journal of Informetrics, 10(2), 336-346.

Thelwall, M. (2017). Three practical field normalised alternative indicator formulae for research evaluation. Journal of Informetrics, 11(1), 128–151. 10.1016/j.joi.2016.12.002

Van Noorden, R. (2014). Scientists and the social network. Nature, 512(7513), 126.

Vaughan, L., & Shaw, D. (2003). Bibliographic and web citations: what is the difference? Journal of the American Society for Information Science and Technology, 54(14), 1313-1322.

Yu, M. C., Wu, Y. C. J., Alhalabi, W., Kao, H. Y., & Wu, W. H. (2016). ResearchGate: An effective altmetric indicator for active researchers? Computers in Human Behavior, 55(B), 1001-1006.

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