Does Dirty Data Affect Google Scholar Citations?

Güleda Doğan
Hacettepe University
Turkey
gduzyol@hacettepe.edu.tr

İpek Şencan
Hacettepe University
Turkey
ipeksencan@hacettepe.edu.tr

Yaşar Tonta
Hacettepe University
Turkey
tonta@hacettepe.edu.tr

ABSTRACT

Google Scholar (GS) is a database that enables researchers to create their scholarly profiles and keeps track of, among others, their citation counts, and h- and i10-index values. GS is now increasingly being used for research evaluation purposes. Although rich in bibliometric data, GS indexes some duplicate publications and citations, and therefore tends to inflate the citation counts to some extent. Based on a small sample of GS profiles of researchers, this paper aims to study the extent by which duplicates change the citation counts and metrics based thereupon. Findings show that duplicates in GS database somewhat inflates the citation metrics. The scale of the problem as well as the effect of dirty data on performance evaluations based on GS citations data need to be studied further using larger samples.

Keywords

Google Scholar, Google Scholar citations, Google Scholar citation metrics, dirty data.

INTRODUCTION

Google Scholar (GS) emerged in 2004 as a freely available database providing large scale searching of scholarly literature (Butler, 2004; Google Scholar, 2016a). It offers multifarious resource types (dissertations, articles, papers, reports, books, etc.) harvested by Google crawlers from online resources such as personal as well as university and publisher web pages and institutional archives/repositories.

Google introduced Google Scholar Citations (GSC) in 2011 that enabled researchers to track their citations and citation metrics (Connor, 2011). GSC also allowed users to create their own scholarly profiles and add publications thereto. GS runs similarity-matching algorithms on harvested data and identifies citations, counts them and calculates citation metrics such as h- and i10-indexes for all publications in a researcher's profile. Proposed by Hirsch (2005), h-index is "defined as the number of papers with citation number higher or equal to *h*". Google introduced i10-index ("the number of articles with at least ten citations") as part of GSC in 2011 (Connor, 2011). GSC also contains some features such as sorting results by publication year, title and the number of citations received (Google Scholar, 2016b; Jacsó, 2012).

Several studies compare GS with other citation databases like Scopus and Web of Science (Bartol and Machiewicz-Talarczyk, 2015; Harzing and Alakangas, 2016; Harzing, 2016b; Harzing and Wal, 2008; Levay, Ainsworth, Kettle and Morgan, 2016; Moed, Bar-Ilan and Halevi, 2016). Although GS offers a more comprehensive picture of a researcher's publications, citations and citation metrics than that of Web of Science and Scopus (Harzing, 2016b; Harzing and Wal, 2008, p. 62), it has been criticized for containing different types of errors (Jacsó, 2005; Jacsó, 2006a; Jacsó, 2006b) "including dirty data" (Konkiel, 2014). GS uses citations harvested from both scholarly and non-scholarly journals for citation analysis; lacks older publications; it does not represent different disciplines equally; has some problems with automatic matching algorithms and is not regularly updated (Harzing and Wal (2008, p. 65-66). Moreover, GS citations could easily be "gamed": it was shown that six documents written by a fictitious author citing papers of a research group and loaded on an institutional repository were enough to manipulate GS citation counts (López-Cózar, Robinson-García and Torres-Salinas, 2014).

Despite its shortcomings, GS is increasingly being used for research evaluation, particularly in recent years. Anne-Wil Harzing's Publish or Perish software uses GS as data source to analyze academic citations (Harzing, 2016a). Altmetrics refers to GS data as well as Mendeley, CiteULike, Twitter and Facebook (Hassan and Gillani, 2016; Martín-Martín, Orduna-Malea, Ayllón and López-Cózar, 2016). GS is used as data source for university ranking systems, too (e.g., Ranking Web of Universities, 2016a; URAP, 2016). Ranking Web of Universities (Webometrics), an international ranking system, ranks 2000 universities in the world according to GSC citation counts and scientists in 54

ASIST 2016, October 14-18, 2016, Copenhagen, Denmark.

countries according to GSC h-index values (Ranking Web of Universities, 2015, 2016b).

Rankings and research evaluation practices based on GS data should be interpreted carefully, as GS pulls citations from everywhere on the web and has other shortcomings, as already pointed out earlier (Konkiel, 2014; López-Cózar, Robinson-García and Torres-Salinas, 2014). The main goal of this study is to find out if GS citation metrics fluctuate on the basis of presence of duplicate publications and citations in the database. We addressed the following research question: Does GS database include duplicate publications and citations in researchers' profiles? If yes, what is the impact of this practice on citation counts and GSC metrics such as h- and i10-index values? Answering this question will shed some light on the size of the problem and help us better interpret the rankings and metrics based on GS data.

METHODOLOGY

To address the research question, we selected the 11 researchers based at Hacettepe University's Department of Information Management with public GS profiles (January 27, 2016), collected and cleaned data between January 27-March 18, 2016. Obviously, publication and citation counts were updated during the course of the study. Therefore, we used the citation counts and GSC metrics of h- and i10-index values as of January 27, 2016, for our analysis to minimize the effect of updates.

GS profiles may include more than one records for the same publication. Therefore, we used the term *record* for what GS uses for publication. We used the term *publication* for the cleaned up singular records.

We checked GS profiles of 11 researchers to identify duplicate records for the same publications. Next, we identified the number of different records for each publication and citations thereto as well as singular publication counts for each researcher and combined citation counts for each publication. We then re-calculated the h- and i10-indexes for each researcher using their new publication and combined citation counts and compared them with GSC metrics to see if there was any discrepancy between the two using the *Wilcoxon Signed Rank Test*.

FINDINGS

In total, 11 researchers had 617 records and 3,144 citations listed in their GS profiles. Citation counts for five researchers were over 100 and their h-indexes ranged between 6 and 18.

The effect of duplicate records on GSC metrics

Eight out of 11 researchers (73%) had one or more duplicate records listed in their profiles. Table 1 shows the then existing GSC metrics along with re-calculated ones for these eight researchers based on 591 records listed in their profiles. The total publication count was 499, indicating that 14% (n=69) of publications were represented with more than one records (mostly 2, max. 5) in the GS database. However, excluding duplicate records did not reduce the number of citations, as citation counts of only 4 out of 69 publications got affected. As a result, none of the researchers' re-calculated h-index was changed and only one researcher's i10-index has increased by 1. This finding suggests that sometimes GS's matching algorithm identifies different copies of the same publication that exist in personal and institutional web sites incorrectly, which does not seem to change the existing GSC metrics much.

Researcher	# of		# of Citations		h-index		i10-index	
	Rec	Pub	P	RC	P	RC	P	RC
R1	243	200	1244	1239	18	18	44	44
R2	88	75	831	830	16	16	18	18
R3	50	40	244	244	10	10	11	12
R4	78	68	485	485	12	12	18	18
R5	29	27	19	19	3	3	0	0
R6	53	47	76	76	5	5	2	2
R7	32	27	34	31	4	4	0	0
R8	18	15	173	173	6	6	5	5

*Rec: Record, Pub: Publication (re-calculated), P: Present, RC: Re-calculated

Table 1. Present GSC metrics along with re-calculated ones excluding duplicate publications

The effect of duplicate citations on GSC metrics

Ten researchers had at least one publication in their profiles that was cited twice or more (Table 2). (One researcher's profile did not meet this criterion and therefore was excluded from further analysis.) The total number of such publications was 245 and they were cited 3,079 times in total. Of 245, 135 publications (55%) received a total 364 duplicate citations (12% of all citations). When duplicate citations removed, citation counts of half of 135 publications decreased by at least two citations. The effect of the removal of duplicate citations was even more pronounced for researchers: citation counts of almost all researchers decreased, some as much as by 20%. We re-calculated h- and i10-indexes for all 10 researchers accordingly. Consequently, h-indexes of more than half the researchers decreased by at least 1. Similarly, i10-indexes of four researchers decreased by 2 and 4, although one researcher's i10-index increased by 1. The *Wilcoxon Signed Rank Test* results showed that researchers' existing GSC metrics and the re-calculated ones differ significantly (Z=-10.219; $p<0.001$), suggesting that GS's citation matching algorithm fails to identify the correct citations in some cases, thereby generating somewhat inflated GSC metrics. The rate of inflation is expected to be much higher for researchers with higher h- and i10-indexes.

Researcher	# of Citations		h-index		i10-index	
	P	RC	P	RC	P	RC
R1	1214	1080	18	17	44	40
R2	826	734	16	13	18	16
R3	242	209	10	9	11	7
R4	474	402	12	12	18	15
R5	18	17	3	3	0	0
R6	67	59	5	4	2	2
R7	35	32	5	4	0	0
R8	171	156	6	6	5	6
R9	30	24	4	3	0	0
R10	2	2	1	1	0	0

*P: Present, RC: Re-calculated

Table 2. Present GSC metrics along with re-calculated ones excluding duplicate citations

CONCLUSION

GS is used as an alternative data source for the evaluation of researchers' scholarly performance. As GS collects (sometimes duplicate) data from all types of sources that are readily available on the web, it generates somewhat higher citation metrics than those of proprietary databases of Thomson Reuters' Web of Science and Elsevier's Scopus. This paper investigated the extent of duplicate publications and citations in GS database and their impact on GS citation metrics. Findings indicate that 16% of publications and 12% of citations identified by GS's matching algorithms and included in the GS database were duplicates. Even though the sample was small and the h- and i10-indexes of researchers in the sample were relatively low, duplicate citations increased the values of GS citation metrics. The difference between h- and i10-index values provided by GS and the ones we re-calculated after removing duplicates was statistically significant.

Notwithstanding the inherent shortcomings of such metrics (e.g., Rousseau, García-Zorita and Sanz-Casado, 2013), findings as such should be taken into account when using GS citation metrics for performance evaluation of researchers for funding and promotion as well as for ranking of universities. On the other hand, GS should monitor the performance of its matching algorithms and de-duplicate the records accordingly. In addition, GS can use stricter matching rules to verify publication and citation data using, say, DOI and ORCID numbers.

REFERENCES

Bartol, T. and Mackiewicz-Talarczyk, M. (2015). Bibliometric analysis of publishing trends in fiber crops in Google Scholar, Scopus, and Web of Science. *Journal of Natural Fibers*, 12(6), 531-541. DOI: 10.1080/15440478.2014.972000

Butler, D. (November 2004). Science searches shift up a gear as Google starts Scholar engine. *Nature*, 432, 423. DOI: 10.1038/432423a

Connor, J. (2011, November 16). Google Scholar citations open to all. Google Scholar blog. http://googlescholar.blogspot.com.tr/2011/11/google-scholar-citations-open-to-all.html

Google Scholar. (2016a). About. https://scholar.google.com.tr/intl/tr/scholar/about.html

Google Scholar. (2016b). Citations. https://scholar.google.com.tr/intl/tr/scholar/citations.html

Harzing, A-W. (2016a, February 6 - updated May 29). Publish or Perish. Harzing.com, Research in International Management. (Blog post). http://www.harzing.com/resources/publish-or-perish

Harzing, A-W. (2016b, February 6 - updated April 11). Google Scholar: A new data source for citation analysis. Harzing.com, Research in International Management. (Blog post). http://www.harzing.com/publications/white-papers/google-scholar-a-new-data-source-for-citation-analysis

Harzing, A-W. and Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics*, 106, 787–804. DOI: 10.1007/s11192-015-1798-9

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

Hassan, S. and Gillani, U. A. (2016). Altmetrics of "altmetrics" using Google Scholar, Twitter, Mendeley, Facebook, Google-plus, CiteULike, Blogs and Wiki. https://arxiv.org/abs/1603.07992

Hirsch, J.E. (15 November 2005). An index to quantify an individual's scientific research output. *PNAS*, 102(46), 16569–16572.

Jacsó, P. (2005) Google Scholar: the pros and the cons. *Online Information Review*, 29(2), 208-214.

Jacsó, P. (2006a) Dubious hit counts and cuckoo's eggs. *Online Information Review*, 30(2), 188-193.

Jacsó, P. (2006b) Deflated, inflated and phantom citation counts. *Online Information Review*, 30(3), 297-309.

Jacsó, P. (2012). Google Scholar Author Citation Tracker: Is it too little, too late? *Online Information Review*, 36(1), 126-141. DOI: http://dx.doi.org/10.1108/14684521211209581

Konkiel, S. (2014, July 23). 4 reasons why Google Scholar isn't as great as you think it is. Impactstory blog post. http://blog.impactstory.org/googe-scholar-profiles-fail/

López-Cózar, E.D., Robinson-García, N. and 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. DOI: 10.1002/asi.23056

Levay, P., Ainsworth, N., Kettle, R. and Morgan, A. (2015). Identifying evidence for public health guidance: a comparison of citation searching with Web of Science and Google Scholar. *Research Synthesis Methods*, 7, 34-45. DOI: 10.1002/jrsm.1158

Martín-Martín, A., Orduna-Malea, E., Ayllón, J.M. and López-Cózar, E.D. (2016). The counting house, measuring those who count: Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in the Google Scholar Citations, ResearcherID, ResearchGate, Mendeley & Twitter. *EC3 Working Papers*, 21. DOI: 10.13140/RG.2.1.4814.4402

Moed, H.F., Bar-Ilan, J. and Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. *Journal of Informetrics*, 10, 533–551. DOI: 10.1016/j.joi.2016.04.017

Ranking Web of Universities. (2015). Top 2000 universities by Google Scholar citations. http://www.webometrics.info/en/node/169

Ranking Web of Universities. (2016a). Objectives. http://www.webometrics.info/en/Objetives

Ranking Web of Universities. (2016b). Ranking of individuals by country (54 countries), Ranking Web of Researchers. http://www.webometrics.info/en/node/116

Rousseau, R., García-Zorita, C. and Sanz-Casado, E. (2013). The h-bubble. *Journal of Informetrics*, 7, 294-300.

URAP. (2016). Genel bilgi (General information). http://tr.urapcenter.org/2010/2010_10.php