Looking under the COUNTER for overcounted downloads

Ted Bergstrom
UCSB Economics Dept

Richard Uhrig
UCSB Economics Dept

Kristin Antelman
UCSB Library

February 8, 2019[1]

[1]The authors are grateful to Doug Steigerwald and Alex Wood-Doughty for advice and stimulating conversation.

Abstract

Libraries frequently use journal download statistics to estimate the value of subscriptions. Download statistics are made available by journal publishers in summary form approved by the COUNTER organization. These summary statistics conceal much information that is essential for evaluating reported downloads. We have obtained copies of server log files from four publishers that specify for each reported download, the article downloaded, the time of download, the format of the download and the IP address from which the download occurred. This enables us to estimate the frequency of multiple reported downloads of the same article by the same user, as well as the frequency of bulk downloads. We suggest that libraries using download statistics to evaluate bundled subscriptions from large publishers would benefit from requesting access to the publisher's server log files that record download details.

1 Introduction

Libraries often employ "usage statistics" to evaluate journal subscriptions. Publishers are aware of the importance that libraries place on usage (download) reports in making subscription decisions. In 2004, Sir Crispin Davis, then the CEO of Reid-Elsevier testified to the British House of Commons that:

> "The biggest single factor is usage. That is what librarians look at more than anything else and it is what they [use to] determine whether they renew, do not renew and so on... I would say that [usage] is the single biggest factor." [1]

Download statistics are not collected by the libraries themselves, but are supplied to them by journal publishers in summarized form. Additionally, many content licenses signed by libraries include a non-disclosure clause that places restrictions on sharing these download reports.[2] In 2006, Philip Davis and Jason Price [3] expressed discomfort with this state of affairs, saying that: "publishers, who control the raw data on downloads have a strong incentive to release statistics that may overstate the number of users." Davis and Price suggest that publishers may inflate the number of downloads by inducing users to access html versions of papers before downloading pdf versions, thus counting two downloads for a single usage.[3]

One might expect that even if publisher-reported downloads exaggerate usage, the exaggeration would be fairly uniform across publishers, so that download counts could be used to compare the relative usage of journals with different publishers. But a recent study [5] suggests that this may not

[1] From testimony in 2004 to the British House of Commons, quoted in Davis and Price [3]

[2] For example, Section 2.4 of the standard Elsevier contract states that

> "Elsevier will make usage data reports on the subscriber's usage available to the librarians/administrators employed by the Subscriber for internal use only. Such reports may be accessed by vendors or other third parties only with permission of Elsevier and for the purpose of usage analysis of the subscriber."

In contrast, the University of California contract with Elsevier states that

> "The Subscriber reserves the right to collect, analyze, and make results of such analysis available to both internal and external constituencies of usage data compiled by Elsevier and made available to the Subscriber."

[3] Similar concerns are expressed by Chan Li and Jacqueline Wilson in a paper delivered at the American Library Association in 2016. [4]

be the case. The authors of this study obtained copies of publisher-supplied download statistics for all libraries in the University of California system. They fitted an equation that predicts annual downloads from each journal published by seven major publishers as a function of the journal's number of recent citations, number of articles, academic discipline, the year in which the download takes place, and the journal's publisher. They found that after controlling for other variables, there remains a significant "publisher effect." The number of downloads reported for journals published by Elsevier, Nature, and the American Chemical Society is about twice as large as the number of downloads reported for journals in the same discipline, with similar citation rates, that are published by Springer, Taylor & Francis, Wiley, or IEEE.

Publishers have attempted to impose some regularity in reporting of download statistics through an organization called COUNTER. According to its own website,

> "COUNTER is a non-profit organization supported by a global community of library, publisher and vendor members... [which] provides the Code of Practice that enables publishers and vendors to report usage of their electronic resources in a consistent way. This enables libraries to compare data received from different publishers and vendors."[1]

Despite this statement of good intentions, the COUNTER-compliant reports that are released to libraries arrive in a summary form that conceals much information that would be relevant to libraries in evaluating journal subscriptions. The two COUNTER formats that are most commonly used by libraries are the JR1 and JR5 reports. The JR1 format reports the "number of successful full-text article requests by month and journal." The JR5 format reports the annual number of downloads by year of publication and journal. The JR5 report is the more useful metric for assessing the prospective value of subscribing to the next subscription year, but its utility is limited by the fact that html and pdf downloads are not recorded separately, as they are in the JR1 report specification. Neither of these reports offers information about the frequency of multiple downloads of the same article by the same user, nor do they report on the number of downloads that are the result of bulk downloading.

2 Peeking into the black box

We have obtained copies of server log files from publishers Springer, Nature, and Wiley that record detailed information about each download made from the University of California Santa Barbara during the time interval from Jan 1-June 30, 2018. We also obtained a copy of the server log file from Elsevier for the California Institute of Technology (Caltech) for the time interval from Jan 1-June 30, 2015. These records include the IP address[4] from which each download was requested, along with the date and time (to the nearest second), and a unique identification number of the downloaded article. For Wiley and Elsevier, the log files also specify the format (pdf or html) of each download.[5] These files can be used to assess the frequency with which a single user's access to a single article is counted as multiple downloads.

2.1 Time t duplicates

Often a user will look at an html version of an article to have a quick look at it, or simply because an article link automatically "downloads" an html copy, and then will download a pdf copy for reading, saving, or printing. The download reports that libraries receive will count this single use as two downloads. Users who have previously downloaded and examined an article will sometimes want to take a second, third, or fourth look at this article. Some users find it convenient to access an article repeatedly from the publisher's server rather than to cache it locally.

To account for repeated downloads of the same article by the same user, we define a download as a *time t duplicate* if within a time interval of t, the same article has been previously downloaded from the same private[6] IP

[4]There may be privacy or data regulation compliance concerns with the release of log files that show the IP address from which each download occurred. These concerns could be addressed by anonymizing the IP addresses. For the purposes described here, it is useful to know when multiple downloads come from the same IP address, but there is no need to know *which* addresses are doing the downloading. Assigning a randomly chosen alias for each address should serve the purpose of protecting user privacy while not degrading the utility of the server log files for understanding user behavior.

[5]The log files for UCSB were obtained by direct request from the publishers. Our Elsevier files were provided by the Caltech Library, which was provided with the log files for Caltech downloads from Elsevier.

[6]Not every IP address is associated with a single user. Some addresses correspond to publicly accessible computers that could be used by many individuals, one corresponds to the library proxy server, and some to VPN addresses used to access the campus network from off campus. The UCSB and Caltech libraries supplied us with the IP number ranges

address in either pdf or html form. The number of *t-unique downloads* is then the total number of logged downloads minus the number of time t duplicates. We have written a python program that allows us to use publishers' server log files to estimate the fractions of all reported downloads from each private IP address that are t duplicates and t unique for various values of t. Table 1 reports the results.

Table 1: Fraction of reported downloads from private IPs that are time t duplicates: by publisher

IP type	Number of Downloads	Fraction that are time t duplicates 15 min	1 hour	1 day	1 week	6 weeks
Elsevier (Caltech)	130,109	0.32	0.34	0.38	0.43	0.46
NPG (UCSB)	95,058	0.26	0.28	0.33	0.37	0.39
Springer (UCSB)	62,245	0.25	0.26	0.29	0.32	0.33
Wiley (UCSB)	123,978	0.22	0.24	0.27	0.30	0.31

We see from Table 1 that the fraction of duplicates found from the Elsevier log file is substantially higher than for the other publishers. Some of this difference may be due to the fact that the Elsevier data is for Caltech in Jan-June, 2015 and the data for the other publishers is for UCSB in Jan-June, 2018. But these results are also consistent with results reported in [5], which found that the reported number of downloads for Elsevier journals were significantly higher relative to those for Springer and Wiley than would be expected, given the citation rates, impact factors, and subject areas of these journals.

2.2 Double-clicking and html-pdf duplicates

The COUNTER Code of Practice [2], states that COUNTER reports are screened to eliminate double-clicking by impatient users. The Code states that if a user clicks a link to an html copy twice within 10 seconds, or requests a pdf copy twice within 30 seconds, the two clicks count as only one access. We find that about 2% of the downloads recorded in the Elsevier log file meet COUNTER's definition of double-click and about 0.5% of those in the Wiley log file.

for addresses of each of these types.

Inadvertent double-clicks are not the only source of multiple counting of a single download event. Users typically open an html copy of an article to view it on screen and then download a pdf copy of this article, but occasionally they first view a pdf copy and then view the html version online. The COUNTER reports count either of these events as two separate downloads. The log files for Elsevier and for Wiley specify whether each download is of an html file or a pdf file. This enables us to look at specific types of double-counting. We find that about 21% of non-bulk downloads were pdf files that were downloaded within 5 minutes of a previous html download of the same article at the same IP address, while about 2% were html downloads downloaded within 5 minutes of a previous pdf download.

Table 2: Quick Duplicates for Elsevier and Wiley

Type of Duplication	Elsevier Fraction	Wiley Fraction
Counter Double-clicks	0.02	0.005
html-pdf within 1 minute	0.18	0.10
html-pdf within 15 minutes	0.22	0.13
pdf-html within 1 minute	0.01	0.02
pdf-html within 15 minutes	0.03	0.03

2.3 Bulk downloads and comparison with JR1 reports

The server log files from Springer, Nature, and Wiley show no evidence of bulk downloads. For these publishers, the number of downloads reported in the log files are within 1% of the number that appears in the JR1 reports.

The Elsevier log files show two instances of massive bulk downloading from a single Caltech IP address. The JR1 reports evidently include the bulk downloads. Table 3 shows that for the month of June 2015, when the bulk downloads occurred, the number of downloads reported by JR1 is about 25% higher than the average for the previous three months. This table also shows that for the month of June, the number of downloads reported in Elsevier's JR1 document is very close to the number of downloads found in the log files, which are known to include bulk downloads. Table 3 also shows that for the months of January-March 2015, the number of downloads reported in the JR1 document exceeds the number reported in the log files by 15% or more.

Table 3: Monthly comparison of reported downloads in Elsevier JR1 report with those in log file

Month	Reported Downloads JR1	Reported Downloads Log file	COUNTER double-clicks	Ratio JR1/Log file
Jan	26,342	22,608	438	1.17
Feb	27,231	23,588	766	1.15
March	29,602	25,624	561	1.16
Apr	28,664	28,937	503	0.99
May	27,621	27,955	591	0.99
June	34,555	35,290	503	0.98

* COUNTER double-clicks are defined in section 2.2

Since bulk downloads occur only sporadically, our limited sample is not adequate to predict the overall proportion of a publisher's reported downloads that come from bulk downloading. However it is important to recognize that when bulk downloads occur, COUNTER reports can significantly overestimate the number of legitimate downloads. Unless a library examined its COUNTER report by looking at month-by-month data, it would not be aware that bulk downloads may have been included in their downloads total.

3 The distribution of html and pdf downloads

Log files from Elsevier and Wiley (but not from Springer and Nature) identify each reported download as being in either html or pdf format. Looking at the distribution of each type separately allows a glimpse into the downloading habits of users. For example, some users never download a pdf file, but download (or view) an html copy several times. Others download an html copy on just one occasion and never download a pdf file. Some who download html files also download a pdf file at least once.

When it is costless for a user to download the same file many times, the download data show that users find it convenient to do so rather than to save a pdf copy of an article and use this copy for later viewing. Table 4 shows that slightly more than half of recorded html downloads are accompanied by a pdf download of the same article from the same private IP address. This table also shows that 6-7% of users download html versions more than once and never download a pdf copy. The needs of these users could also be satisfied by access to a single pdf copy of the article. For Wiley, about 32%

of those who download an html copy of the article do so only once and do not download a pdf copy. For many of these users it is likely that the article is of little interest, as the "download" of the html full text was a byproduct of looking at a citation or abstract.

We define *redundant* html downloads in the following way. If a private IP address downloads one or more html copies of an article and also downloads a pdf copy during the six month period for which we have data, then the html copies are said to be redundant. If a private IP address downloads multiple html copies, but no pdf copy, then we define all but one of these html downloads as redundant.

In searching the journal literature, users frequently take a cursory look at articles in which they might be interested. Most publishers provide access to the abstract of all of their papers without a subscription. If a scholar whose university subscribes to a journal accesses an article from the journal's website, clicks on the article name, she will download both the abstract and an html copy of the entire article. The websites of Wiley and Elsevier do make it possible for subscribers to access the abstract alone, but this takes a conscious effort and offers no advantage over downloading the abstract plus the entire article. If a user downloaded an html copy only once and never downloaded it again, either as html or pdf, it is likely that the author found the article to be of little interest for his or her research and may well have looked only at the abstract. We have no way of knowing whether the user's purposes would have been served equally well by looking at the freely available abstract of the paper. To get a sense of the magnitude of this effect, we assume that half of the instances in which users download an html copy of an article and never download it again are cases where the author's needs would be satisfied by a look at the abstract alone.

Table 4: Patterns of html downloads from Private IP addresses

	Wiley		Elsevier	
	Total	ratio	Total	ratio
html downloads reported	56,035	1.00	81,247	1.00
html downloads accompanied by pdf	24,267	0.43	42,009	0.52
Two or more htmls, no pdf	10,978	0.20	15,548	0.19
Distinct two+ html pairs	4,269	0.08	5,291	0.07
Single html download, no pdf	20,790	0.37	23,690	0.29
Redundant html downloads	30,976	0.55	52,266	0.64
Inessential html downloads	41,371	0.74	64,111	0.79

We see from Table 4 that about 55% of Wiley's and 64% of Elsevier's reported html downloads from private IP addresses are redundant. If we define *inessential* downloads to include half of the cases of a single html download, then the proportions of html downloads that are inessential are 74% for Wiley and 79% for Elsevier.

If the same private IP address downloads the same article more than once during our six-month window, we define all but one of these downloads as redundant. Table 5 shows that about 13% of Wiley's pdf downloads and 18% of Elsevier's pdf downloads are redundant.

Table 5: Redundant pdf downloads

	Wiley		Elsevier	
	Total	ratio	Total	ratio
pdf downloads reported	67,943	1.00	48,862	1.00
Single pdf downloads	53,823	0.79	33,680	0.69
First of multiple pdf downloads	5,482	0.08	6,557	0.13
Redundant pdf downloads	8,638	0.13	8,625	0.18

Table 6 shows the proportion of all downloads for Wiley and for Elsevier that are redundant and the proportion that are inessential by our definitions.

Table 6: Redundant and inessential downloads

	Wiley		Elsevier	
	Total	ratio	Total	ratio
Downloads reported	123,978	1.00	130,109	1.00
Redundant downloads	39,614	0.32	60,891	0.47
Inessential downloads	50,009	0.40	72,736	0.56

The results displayed in Tables 4-6 suggest a simple method for constructing crude estimates of the proportions of redundant and inessential downloads if one does not have log files, but does have JR1 download reports. The JR1 files for each publisher report the fractions of html and pdf downloads. A relatively high proportion of html downloads is an indicator of frequent double-counting. These proportions differ substantially between publishers and between journals. If we assume that the fractions of html and pdf downloads that are redundant and/or inessential are the same as those that we found for Wiley at UCSB, then, if according to the JR1 report the fraction x of all downloads are html downloads, then the

fraction of redundant downloads would be $.55x + .13(1 - x) = .13 + .42x$. If we assume that these fractions are the same as those that we found for Elsevier at Caltech, then the fraction of redundant downloads would be $.64x + .18(1 - x) = .18 + .46x$.

4 Conclusion

The server log files that we have obtained from Elsevier, Springer, Nature, and Wiley reveal much useful information that cannot be found in the JR1 and JR5 download reports. The log files show that the fraction of downloads that repeat previous downloads of the same article from the same private IP address in the same week ranges from 27% for Wiley to 38% for Elsevier, while the fraction of recorded downloads that repeat previous downloads from the previous 6 weeks ranges from 31% for Wiley to 46% for Elsevier.

Log files for Wiley and Elsevier specify whether each download is in html form or pdf form. We found frequent instances of double-counting downloads in which it appears that the user first downloaded an html version of a paper to view it onscreen and then within the next 15 minutes downloaded a pdf copy of the same paper for reading, saving, or printing. About 20% of reported downloads for Elsevier and 13% of reported downloads for Wiley are of this form. While multiple accesses of an article indicate genuine interest, it is also true that a single cached copy of the article would serve the same purpose. Thus if the library did not subscribe to this journal, a single copy obtained by interlibrary loan, or some other means, would serve this user's needs.

A recent paper, [5], finds that, controlling for journals' citation rates and disciplinary specialization, reported downloads for Elsevier and Nature journals are significantly higher than those of Springer and Wiley. This paper's finding that duplication rates for Elsevier are substantially higher, and those for Nature are somewhat higher than those for Wiley and Springer, is consistent with these results.

Examination of the Elsevier log file showed two instances of a large number of bulk downloads. These downloads were included in the count of JR1 downloads, although such downloads will be of far less value to a library community than ordinary downloads of a single article.

Not only do the COUNTER JR1 and JR5 reports conceal information on duplicate downloads that is crucial for evaluating subscriptions, they are aggregated at the journal level rather than the article level. The log files, which show downloads by article as well as by journal could be used to study

such things as the frequency distribution of downloads across articles in the same journal and the time distribution of downloads for single articles.

Much as the impact factor of a journal in which an article is published is an unreliable indicator of the article's own citation rate, the average number of downloads per article of the journal in which an article appeared is not an accurate indicator of the number of times the article itself is downloaded. Because downloads appear sooner in the life an article than citations, access to article-level downloads would be especially valuable for evaluating the research impact of young scholars.

To us it seems remarkable that libraries have allowed publishers to selectively release statistics of their customers' usage of these licensed resources, and then to restrict use of even these statistics. This does not appear to be a procedure well-suited to delivering credible and reliable information. We believe that there is a compelling case for subscribing libraries to insist on receiving publishers' server logs for their institutions rather than relying solely on data found in COUNTER reports.

References

[1] COUNTER. About counter. https://www.projectcounter.org/about/, 2017. accessed 2-23-2018.

[2] Counter Project Release 4. Code of practice, release 4. https://www.projectcounter.org/code-of-practice-sections/data-processing/, 2017. Discussion in Data Processing section. Accessed 2-17-2018.

[3] Philip M. Davis and Jason S. Price. ejournal interface can influence usage statistics: Implications for libraries, publishers, and projectcounter. *Journal of the American Society for Information Science and Technology*, 57(9), July 2006.

[4] Gabriella Wiersma. Report of the ALCTS CMS collection evaluation and assessment interest group meeting. american library association conference, san francisco, june 2015. *Technical Services Quarterly*, 33(2):183–192, 2016.

[5] Alex Wood-Doughty, Ted Bergstrom, and Douglas Steigerwald. Do download reports reliably measure journal usage? trusting the fox to count your hens. UCSB working paper, 2018.