Same Question, Different World: Replicating an Open Access Research Impact Study

Julie Arendt, Bettina Peacemaker, and Hillary Miller

To examine changes in the open access landscape over time, this study partially replicated Kristin Antelman's 2004 study of open access citation advantage. Results indicated open access articles still have a citation advantage. For three of the four disciplines examined, the most common sites hosting freely available articles were independent sites, such as academic social networks or article-sharing sites. For the same three disciplines, more than 70 percent of the open access copies were publishers' PDFs. The major difference from Antelman's is the increase in the number of freely available articles that appear to be in violation of publisher policies.

Introduction

Open access advocates often promote an open access citation advantage: that is to say, the idea that open access research publications receive more citations than those only available through tolled access because their free online availability makes them more likely to be accessed, read, and cited. An open access citation advantage, unlike many other arguments for open access, appeals to authors' self-interest to increase their research impact. Research surrounding the existence, causes, and limits of an open access citation advantage has accumulated over nearly two decades. In that time, the open access environment has changed.

This study revisits a prominent early open access citation advantage study in the context of this changing environment. Kristin Antelman's 2004 "Do Open-Access Articles Have a Greater Research Impact?"[1] is a foundational work in the area of open access citation advantage, cited more than 200 times.[2] Several subsequent studies adapted its methods to examine the citation advantage in other subject areas or journals.[3] In the present study, nearly the same set of source documents and similar methods are used to examine free accessibility and the open access citation advantage over time.

Literature Review

Broadly speaking, most open access citation advantage studies have found there is an open access citation advantage.[4] One project tracking this research even abandoned its efforts be-

Julie Arendt is Science and Engineering Research Librarian, Bettina Peacemaker is Head, Academic Outreach, and Hillary Miller is Scholarly Communications Librarian, all in the Virginia Commonwealth University Libraries; email: jaarendt@vcu.edu, bjpeacemaker@vcu.edu, hmiller5@vcu.edu. *The authors would like to thank Jessica Nguyen for her assistance with the data collection and John Glover, Jimmy Ghaphery, and Marilyn Scott for helpful comments on drafts of this article.*

cause "…the citation advantage evidence has now become far more common knowledge to our authors."[5]

The earliest prominent investigation into a citation advantage was Steve Lawrence's report in 2001 documenting a correlation between the number of citations computer science papers received and the percentage of them available freely available on the web, controlling for year of publication.[6] Some journals were still in nascent stages of providing web access, so Lawrence's study did not fully distinguish an open access advantage from an online access advantage. Kristin Antelman's study, "Do Open Access Articles Have a Greater Research Impact?" followed in 2004, showing a higher average number of citations for freely accessible articles than for toll access articles from select journals in four subject areas.[7]

Much of the research on the open access citation advantage subsequent to Antelman has taken a similar approach: select a body of published research, method to establish open access status, and source for citation counts, and then compare the number of citations the open access documents received to the number of citations received by toll access documents. The details of these observational studies vary. The documents selected may be from one subject area or small set of subject areas, a set of journals, or a large set of publications across multiple subject areas.[8] The studies vary in the databases used to count citations[9] and document sample size, with some smaller studies using manual searches[10] and larger studies using computer programs to automate the process.[11]

These studies also vary in how they define or operationalize "open access." Both the Berlin Declaration and the Bethesda Statement define open access as a system with free access granted by the authors and other (copy)right holders.[12] These definitions stipulate that all users are granted the license to distribute works and to make derivative works, subject to proper citation of the original work, and the work be deposited in a repository suitable for long-term archiving.[13] Citation advantage studies rarely use these full definitions of open access. Often they simply include materials made freely available by a specific journal, publisher, or repository[14] without discussing rights for redistribution or derivative works. Other studies use even looser definitions of open access, such as findable and freely available on the web via search engines like Google.[15] Some documents accessed this way may not have permission from the author or copyright holder or be in a system suitable for long-term archiving.

Despite methodological differences, the bulk of the observational studies reproduced the finding that an open access citation advantage exists. A few found that it diminishes or disappears if other variables are controlled for.[16] Other observational studies that have not found a citation advantage compared open access journals to toll access journals.[17] Average quality of articles varies by journal, so this finding has been interpreted in different ways. It could be evidence for a lack of an open access citation advantage or that underlying "citability" of the articles was lower in the open access journals.

Some studies have taken an experimental approach. For example, Philip Davis randomly assigned articles freely available in select journals and found no open access citation advantage, even though downloads were higher for open access articles.[18]

In summary, the evidence that free availability increases readership,[19] increasing the pool of people who could cite a document, combined with the repeated finding of an open access citation advantage, bolsters proponents' claims that open access increases a work's citations.[20] Detractors, on the other hand, emphasize the similarities in the observational studies. Many studies' failure to control for relevant variables leaves open the question of whether open access

causes a citation advantage or whether alternative explanations, such as authors selectively making their best works freely available, fully explain the advantage seen in observational studies.[21]

Revisiting an older study such as Antelman's offers a lens to view changes over time. Ideally, data from the same body of research, using the same methods, allow for comparisons that could not be made by a new study or looking at broad trends. This lens, however, cannot clarify everything, both because of limitations inherent in the study design (which cannot address causality) and because of limitations in what methods can be reproduced in the changing environment.

In the years since Antelman's 2004 article, the open access environment has changed substantially, generally toward greater access.[22] Some changes affect the open access milieu but only indirectly affect the Antelman articles. For example, institutions and funders have established open access mandates, but these mandates came almost entirely after 2004.[23] Similarly, initiatives like the SPARC author addendum, introduced in 2004, that help authors retain rights[24] when signing publishing agreements do not directly affect the Antelman articles. Other changes more directly affect the articles in Antelman's data set.

Since 2004, some journals, including some included in Antelman's study, have instituted delayed open access, thereby increasing the number of articles that are freely available. For example, the American Mathematical Society began making all articles freely available five years after publication.[25] Similarly, another mathematics journal, this time from a for-profit publisher, *Computational Geometry—Theory and Applications*, now does so four years after publication.[26] Assuming the open access citation advantage found for this "delayed open access" holds for these journals,[27] older articles that have more recently been made open access could develop a citation increase after being made freely available. However, because citations received per year generally increase initially and then decrease and level off, adding few additional citations per year,[28] the number of citations added when an article becomes freely available many years after publication is likely to be small.

More universities have developed institutional repositories where authors can post copies of their work, as well as institutional mandates for depositing articles in these repositories. To the extent that these repositories encourage and facilitate posting, faculty may be more inclined to post their publications. Although the number of institutional repositories and the growth rate for content is increasing, this activity "does not yet pose a challenge to traditional models of scholarly publication."[29] Even where institutions mandate deposit, faculty often ignore it.[30] Nevertheless, institutional repositories provide an avenue for researchers to make both their recent and past publications freely accessible, even years after publication.

Academic social networks, like Academia.edu and ResearchGate, did not exist in 2004 and have produced bigger changes to the open access environment and possibly the open access citation advantage.[31] Both platforms appeared in 2008 and have millions of registered users and uploaded papers.[32] At least two studies have found connections between posting in these networks and higher citations.[33] To the extent these sites encourage and facilitate posting, they would be expected to increase the number of articles authors make freely available.

Another potentially disruptive force is Sci-Hub, a large-scale article piracy site that makes toll access articles freely available. Sci-Hub includes tens of millions of publications.[34] Although Sci-Hub's activities, and those of related sites such as LibGen, have been found to violate US copyright law, the nature of its operations will make it difficult to shutter.[35] Sci-Hub currently

operates as a deep website, meaning articles can be found by using the search tool on the site, but its content typically does not surface in search engine results. It is possible for others to download from Sci-Hub and upload onto sites visible to search engines. Although Sci-Hub would not appear in results and count toward open access in studies relying on Google, Sci-Hub could nevertheless indirectly affect the availability of free articles found via Google.

The combined effect of these changes in article availability on the citation counts of freely available articles relative to toll access articles is complicated. It is even more complicated for works that may be made freely available years or even decades after the original publication. Given the amount of time that has passed, almost all of the articles in Antelman's study should have more citations. If all the articles' status as free access or toll access had remained unchanged since Antelman's observation of a citation advantage, it would be reasonable to expect the citation advantage to grow over time because getting cited in the past is associated with getting cited in the future.[36] It is likely, however, the access status of some of the articles has changed over time, complicating the relationship between access status and number of citations.

To examine how these changes intertwine with the open access citation advantage, the current study partially replicates Antelman's, using articles from the same journals and published in the same years as those in the original study, using similar methods. Because the methods used in the Antelman study were relatively straightforward and clearly laid out in the original article, they lend themselves to replication to investigate whether and how much the citation advantage has been sustained through these changes over more years for a group of articles.

One goal of replication is to determine if results are repeatable and represent a consistent pattern across multiple studies.[37] Follow-up studies or conceptual replications, using different populations or methods to study the same phenomenon, can provide a means of confirmation or disconfirmation, but publication bias[38] and other social factors surrounding the research and publication process[39] can build a line of research on an unstable foundation of preliminary findings. Direct replication or repetition of previous studies provides a means to establish that the foundational research is repeatable. However, replications may replicate flaws from original studies and therefore cannot wholly guarantee that repeatable results and the interpretations they support are true, only that the results are repeatable.

Recent large replication initiatives have had mixed results.[40] Enough findings from these initiatives have been at variance from the original studies for it to be called a crisis[41] and for questions of reproducibility to be raised in disciplines beyond those with large replication initiatives.[42] Aside from spurring discussion about replication, including whether there truly is a crisis,[43] these initiatives spurred discussion about appropriate replication approaches. Is it better to collaborate with the original authors to match the original study's conditions or to work independently to determine if results are robust, even with subtle methodological differences?[44]

The purpose of the current study was threefold: 1) to examine changes in the levels of free access and citation advantage more than a decade after Antelman's study; 2) to examine the changes in the sources and versions of documents available; and 3) to examine the replicability of the study based only on the content in the published article. Although this paper touches on replication challenges, the emphasis is on the first and second goals. Changes in the open access milieu could affect not just citation advantage but also source and locations of documents and types of documents available.

Methodology

Replication Choices

We preregistered our analysis plan with the Open Science Framework and partially replicated Antelman's methods, aiming to use the body of research Antelman used: articles from forty journals, with ten journals in each from four subject areas (mathematics, electrical and electronic engineering, political science, and philosophy), published in two selected years.[45] Antelman performed manual searches in Google to establish open access status and obtained citation counts from Web of Science.[46] Antelman's research also documented the types of websites and article versions (preprint or postprint) available.[47] In broad terms, we did the same, but some details of our study deviated from Antelman's.

We intended to closely replicate Antelman's methods, based only on the published article, to provide a longitudinal comparison to that study. However, we were not able to conduct as close a replication as intended. Prior to collecting the data reported here, we performed a pretest using the first 2004 issue of each journal used in the Antelman study and noticed challenges caused by the changing landscape. For example, we uncovered different types of document hosting sites than Antelman's categories. We also faced challenges making the distinction between preprints and postprints made in the original study. We probably also inadvertently deviated from Antelman's methods in some of our interpretations article. The most salient modifications and interpretations, as well as the reasons for them, are described below.

Sample

We attempted to use the same articles from the same forty journals, ten per discipline, as the original study.[48] This included all articles in the selected journals in mathematics from 2001 to 2002 and philosophy from 1999 to 2000. For political science and for electrical and electronic engineering, the sample included articles from 2001 and 2002 closest to the first 2002 issue until the desired sample size was reached.[49]

For all four subject disciplines, we searched Web of Science for articles from the appropriate journals and publication years, then limited results to those with the "article" document type. For electrical and electronic engineering and for political science, we sorted the results by date and used the 2001/2002 border of the sorted list to compile a sample in which roughly half the articles came from 2001 and half from 2002. The number was approximate because we included all items from a particular issue of a particular journal that fell near the cutoff to get the appropriate sample size. Metadata, including citation counts for the articles, were exported to a spreadsheet for data entry.

Despite our best efforts to match Antelman's sample selection procedure, we did not have the same sample size. Antelman used 2,017 articles total: 602 in philosophy, 299 in political science, 506 in electrical and electronic engineering, and 610 in mathematics.[50] We used 2,052 articles total: 575 in philosophy, 300 in political science, 508 in electrical and electronic engineering, and 669 in mathematics.

Citation Counts and Free Availability

Antelman obtained citation counts from Web of Science, which has since introduced additional databases, such as the Book Citation Index. For this study, we recorded a count from the databases we thought would have been used in Antelman's study: Science Citation Index

Expanded, Social Science Citation Index, and Arts and Humanities Citation Index. In the Antelman study, self-citations by any of the coauthors or from the same journal issue as the article were excluded from the citation count,[51] so we did likewise.

Antelman also excluded citations from 2004 from the citation counts.[52] In this study, citations were included regardless of the year received. From March 3, 2017 to May 4, 2017, we gathered citation counts and removed self-citations.

Google searches were conducted from May 2, 2017 to July 31, 2017. The first two pages of search results were examined for links to free access copies. To ensure that institutional subscriptions did not affect the results, searches were conducted away from university networks.

Antelman's work occurred before Google Scholar was available. To deal with Google Scholar links that sometimes appeared at the top of search results, we explored those links only if the regular results did not include a free access copy in the first two pages.

Antelman searched for article titles "as a phrase" in Google and removed parenthetical additions to the title and nontext or encoded characters that may not have been indexed by Google.[53] In our searches, we entered the title of the article into the Google search box, without quotation marks.

Some articles, particularly in philosophy, had titles such as "posthumous harm" that did not have any version, free access or toll access, appear in the first two pages of Google results. Rather than indicating no free version was available for these 111 articles, we used an escalation procedure. If the results failed to include any version of the article (even those that were toll access) within the first five Google results, the title was searched again with nontext encoded characters and parenthetical comments (if any) removed, as Antelman had. If this search failed, the title was searched as a phrase in quotes. Finally, if the search still failed, we searched with the title as a phrase in quotes and the surname of the first author.

Another twenty-four articles failed to have any version appear in the first two pages of Google results, even though they had unusual titles. These articles all came from two of the mathematics journals that included articles published in French. Web of Science provided the article titles in English, probably causing the Google searches to fail. For these twenty-four articles, we conducted another Google search using the French titles of the articles.

For the purposes of this study, we followed Antelman's operational definition that free availability, located via Google, was considered open access.[54] However, to accurately reflect what we examined, relative to other definitions of open access, this paper uses "free access" or "freely available" rather than "open access" to describe articles in our data set. Some clarifications to Antelman's operational definition were necessary due to changes in the landscape since Antelman's study. Articles available with read-only access and without requiring registration were counted, but freely available articles requiring registration or login to view were not. The rise of academic social network sites not in existence at the time of Antelman's study led to an increase in the latter situation.

Articles were counted as free if they were accessible directly by clicking on the link from Google's results page, or if the result led to a page that contained article metadata, and clicking on a link, such as one labeled "PDF" or "Full text," led to a free access copy. Like Antelman, PDF and PostScript files were included; zipped and dvi files were not.[55]

Sources and Versions of Free Access Copies

Antelman subsampled fifty free access articles from each discipline and categorized the type

of site where the article was found and the article version (preprint or postprint). In this replication attempt, we categorized the full sample.

To accurately and reliably categorize the sites and documents, we modified Antelman's original categories, which were author's site, discipline repository, other repository, departmental/company site, conference/association/project site, working paper series, another person's site, or course archive. To represent the categories uncovered during the pretest, we used these categories: Author/departmental, Institutional repository, Discipline repository, Publisher/JSTOR, and Independent/external.

Antelman made distinctions between author and departmental/company sites based on what pages linked to them.[56] Because we sometimes reached PDFs directly from Google or Google Scholar, we collapsed these into one category. Author/departmental sites included the authors' pages on their employers' sites, sites maintained by the author's department for publications produced in that department, and sites that appeared to belong to the author (such as author's name as the URL, name, and photographs of the author on the site).

Institutional repositories, a category created for this study, differed from departmental or author websites in that they were centralized repositories for scholarly work across the entire organization. The institutional repositories category included sites that described themselves as institutional repositories and sites indicating that their purpose was to distribute works produced across an entire university or business, even if they did not label themselves as repositories.

Disciplinary repositories included sites that accepted papers from authors for sharing within a discipline. Established repositories were included, as were smaller disciplinary repositories, such as those specialized for subfields of mathematics. A site such as PubMed Central or CiteSeerX was included in this category even if its area of specialization was not one of the four disciplines in this study.

The Publisher/JSTOR category included publisher sites and sites that have arrangements with publishers to distribute their content—primarily JSTOR with a few instances of the Philosophy Documentation Center.

Independent/External sites included sites where articles may not have been posted by the author or in accordance with publisher policies. This category included academic social network sites and sites for which a connection to the author could not be found. Publishers generally allow authors to post preprints and postprints on their personal or institutional sites, often after an embargo period, but not on another website.[57] Academic social networks, such as Academia.edu or ResearchGate, where the author may have posted their own article, were included in this category. It could be argued that author pages on academic social networks are author websites, with the network simply supplying a platform. In some cases, however, articles on ResearchGate have not been posted by the authors and would not align with publisher policies.[58] The independent/external category also included sites with less of a connection to the author or publisher, like Semantic Scholar, that harvest content from other sites, as well as sites where the article was posted by someone other than the author for course instruction or other purposes. Sites where it was unclear how the document arrived there, such as docslide.net, also were included in this category.

In a footnote, Antelman states, "If there was a repository copy, the article was coded repository even when a copy was also on the author's site."[59] We deviated from Antelman on this. If multiple freely available copies of an article appeared in our results, rather than

attempting to sort through multiple copies, with repository copies given deference, we coded the location based on whichever copy appeared first in the Google results.

In addition to indicating the type of hosting site for the subsample, Antelman categorized the posted articles as "preprint" or "postprint." In practice, freely accessible access articles lacking publisher imprints rarely were labeled as being copies created before peer review (preprints) or after peer review (postprints). As we had no simple, reliable way of determining their statuses, we categorized articles by whether they appeared to be the publisher's imprint—with signals such as formatting, a header with the name of the journal, copyright notice, and page numbering—or not.

Results

Citation Advantage and Free Availability

Free access versions of articles were found for 37 percent of the articles in philosophy, 47 percent of the articles in political science, 59 percent of the articles in electrical and electronic engineering, and 86 percent of the articles in mathematics (see table 1). As with Antelman's study, mathematics had the highest percentage of free access articles, followed by electrical and electronic engineering, political science, and philosophy, in that order. All four disciplines had a free access article percentage at least seventeen points higher than in Antelman's study.

TABLE 1
Sample (Number of Articles) and Frequency of Free Access

Discipline	Articles Total	Articles Free	Articles Not Free	% of Total Free Access
Philosophy	575	210	365	36.5%
Political Science	300	142	158	47.3%
Electrical and Electronic Engineering	508	299	209	58.9%
Mathematics	669	576	93	86.1%

At the time of Antelman's study, the articles had few years to accumulate citations, and average citation counts were under 2.5. The articles now have had more than a decade and a half to accumulate citations, so citation counts in this study were understandably higher. The distributions of citations were skewed (see figure 1), so the median, rather than the mean, was used to measure the average citation count. For each of the four disciplines, the median citation counts were higher for the free access articles than for toll access articles (see table 2). The differences in the medians were statistically significant for each of the disciplines (see independent samples median test on table 2). From figure 1, it appears that there is a wider range of citation counts for the free access articles, especially in the 50th to 75th percentile. Although this specific observation was not statistically tested, a Mann-Whitney U test for equal distributions did suggest the distributions were not equal in any of the disciplines for the free access articles compared to toll access articles (see table 2).

Sources and Versions of Free Access Copies

In Antelman's subsample of free access article hosting sites, the majority of articles were on the author's site for all subject disciplines but mathematics, which had the majority on

FIGURE 1
Comparison of Citation Rates across Disciplines between Free and Not Free (outliers >155 excluded)

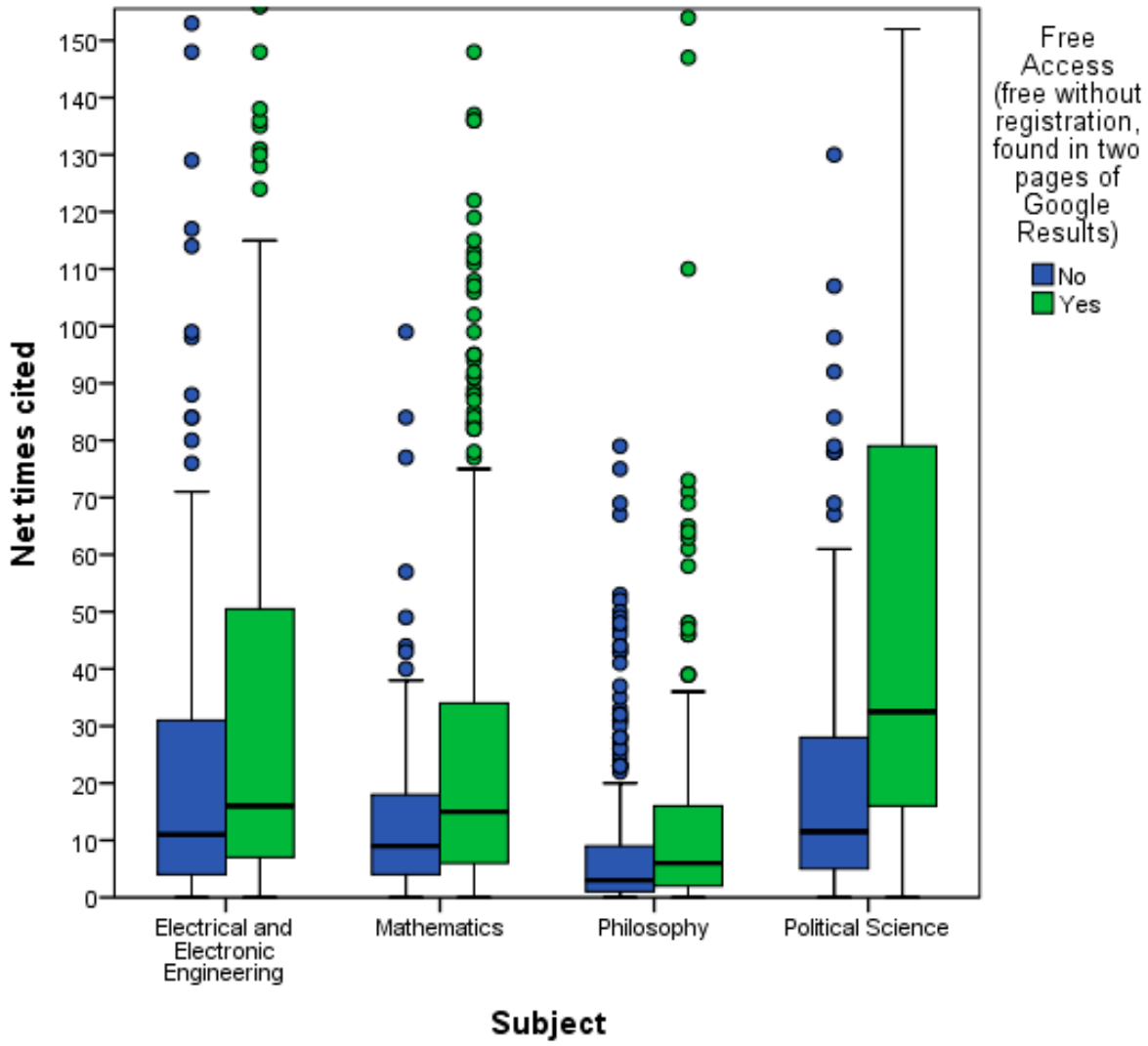

TABLE 2
Comparison of Median Citation Rates between Freely Available Articles and Those That Are Not Freely Available

Discipline	Median (free)	Median (not free)	Difference in Median	Percent Difference in Medians	Independent Samples Median Test Two-tailed P Value	Mann-Whitney U (equal distributions)
Philosophy	6	3	3	100%	<.001	<.001
Political Science	32.5	11.5	21	182%	<.001	<.001
Electrical and Electronic Engineering	16	11	5	45%	0.032	0.001
Mathematics	15	9	6	67%	0.005	<.001

a disciplinary repository, followed by the author's site. In the current study, mathematics continued to have disciplinary repositories as the most common hosting site and an author or department site as the second most common (see table 3). For the other three disciplines, the most common hosting sites were independent/external sites. These sites hosted the majority of free access articles in political science and in electrical and electronic engineering and more than 48 percent of the free access articles in philosophy (see table 3). For mathematics, more than a seventh of the free access articles were available on either the official site of the publisher or another official site such as JSTOR. In Antelman's original study, article hosting sites and journal publishers were so unusual that they were not mentioned.

Journals within each subject area varied in the percentage of free access articles. For philosophy journals, free access percentages ranged from 13 percent for *British Journal of Aesthetics* to 54 percent for *Philosophical Review*. Political science ranged from 21 percent for *Public Opinion Quarterly* to 68 percent for *American Political Science Review*. In electrical and electronic engineering, *IEEE Electron Device Letters* was the lowest with 44 percent, and the highest was 80 percent for *IEEE Network*. In mathematics, the lowest was 70 percent for *Communications on Pure and Applied Mathematics*. Four mathematics journals had 100 percent free access: *Annales Scientifiques de l'école Normale Supérieure*, *Bulletin of the American Mathematical Society*, *Computational Geometry—Theory and Applications*, and *Journal of the American Mathematical Society*.

TABLE 3
Where Freely Available Articles Are Found

Discipline	Author/Department/ Company	Discipline Repository	Institutional Repository	Independent Site	Publisher/ JSTOR
Philosophy (n = 210)	77 (36.7%)	16 (7.6%)	9 (4.3%)	102 (48.6%)	6 (2.9%)
Political Science (n = 142)	40 (28.2%)	20 (14.1%)	7 (4.9%)	74 (52.1%)	1 (0.7%)
Electrical and Electronic Engineering (n = 299)	78 (26.1%)	16 (5.4%)	26 (8.7%)	177 (59.2%)	2 (0.7%)
Mathematics (n = 576)	131 (22.7%)	250 (43.4%)	12 (2.1%)	82 (14.2%)	101 (17.5%)

In the original study, Antelman tabulated whether a subsample of free access articles were preprints or postprints and found that a majority for three of the subject disciplines—philosophy, political science, and mathematics—were preprints. Because we instead tabulated whether the articles were publishers' imprints or not, direct comparison cannot be made. In our study, however, more than 70 percent of the free access articles in philosophy, electrical and electronic engineering, and political science were publishers' imprints (see table 4): that is to say, not preprints. For mathematics, in contrast, fewer than half of the free access copies were publishers' imprints. For all four disciplines, the majority of documents on independent sites bore the publishers' imprint (see table 4).

TABLE 4
Fraction (and Percentage) of Free Articles with the Publisher's Imprint

Discipline	Total	Author/ Department/ Company	Discipline Repository	Institutional Repository	Independent Site	Publisher/ JSTOR
Philosophy	151/210 (71.9%)	44/77 (57.1%)	9/16 (56.3%)	3/9 (33.3%)	89/102 (87.3%)	6/6 (100%)
Political Science	110/142 (77.5%)	30/40 (75.0%)	10/20 (50.0%)	5/7 (71.4%)	64/74 (86.5%)	1/1 (100%)
Electrical and Electronic Engineering	233/299 (77.9%)	53/78 (67.9%)	12/16 (75.0%)	19/26 (73.1%)	147/177 (83.1%)	2/2 (100%)
Mathematics	268/576 (46.5%)	36/131 (27.5%)	86/250 (34.4%)	2/12 (16.7%)	43/82 (52.4%)	101/101 (100%)

Discussion

This study observed similar patterns as those found by Antelman: free access articles receiving more citations than toll access articles and similar disciplinary differences in the depths of penetration of free access. Even with more articles available via free access overall, the order of the four subject areas remained the same.

However, compared to the time of Antelman's paper, the types of hosting sites have changed. Other differences observed in this study, relative to Antelman's, are that more articles are freely available across all four disciplines, and more of the free access articles are publishers' imprints.

The magnitude of the change in the number of citations to free access articles compared to toll access articles found in this study should be approached with caution. Because this study had a different sample size and found more free access articles, it is clear that the sample of free access articles found here is not exactly the same as Antelman's. Some of the additional free access documents we found were likely not freely available for the entire period since the Antelman study. The results here do not supply information about the rate at which a citation advantage accumulates, because we did not know how long the documents had been freely available and did not track when the citations were received.

Antelman treated open access as an authorial choice. Today, the choice to make the article freely available is sometimes made by publishers, given their increasing involvement in open access. For example, some journals in this study have instituted delayed open access since 2004. Unsurprisingly, we found especially high free access percentages for these journals. At the opposite end of the journals' percentages, there may be systematic differences among publishers' antipiracy efforts, influencing free article availability. We did not investigate this directly, but, for both philosophy and political science, Oxford University Press published the journals with the lowest rates of freely available articles.

These journal practices pose a problem for attributing the higher citations of open access articles to open access. If journals' open access and antipiracy practices were unrelated to their citation rates, they would not be an issue, but it is plausible the citation advantage could be biased by journal characteristics rather than the articles' open access statuses. For example, one study has shown that journals that institute delayed open access have average citation rates much higher than journals that do not.[60]

The meaning of the change in hosting sites should be interpreted with caution, as this study only examined the version appearing first in Google results and did not preference repository copies, unlike Antelman. Multiple free access copies may exist on more than one type of site, and the independent sites may have been higher in the Google results. We did not check how often free access copies appeared on multiple sites, but, for example, one study has found that 66.5 percent of a sample of articles in Academia.edu also had free access articles posted elsewhere.[61]

Because the independent/external sites included both author-posted and independently posted articles, the results of this study cannot support claims about changes in author behavior.

Authors could have been responsible for posting articles to these sites, or the articles may have been harvested from wherever the author posted them. At the opposite extreme, additional articles could have been posted without any author involvement. Taken to this extreme, assuming authors posted none of the free access articles on independent or publishers' sites, the percentage of articles made freely available by authors could even have decreased since Antelman's study.

At the outset of the study, we decided to categorize academic social network sites that enable article sharing in the same category with other independent sites. We did this because even for articles posted by authors, the legality of the article sharing on these networks is in dispute. For example, in November 2017, ResearchGate moved 1.7 million articles from free access to access only by requesting a copy from the author because of such disputes.[62] However, these sites are not as clearly piracy sites as sites fully independent of authors. In retrospect, because so many articles were in our category of independent/external sites, it would have been informative to have distinguished between social network sites and sites where it was clearer that the posting was a likely copyright violation.

Although not tracked in this study, some of the articles found on independent sites were found on sites with no clear connection to the authors. Many of these articles were on docslide.net or similar sites. Docslide sites provided minimal information about who was responsible for the site or how the articles got there. The Docslide sites did not require a login, but they typically required the user to enter a reCAPTCHA response and to wait up to a minute for the article's download to begin or for an error message to appear. These barriers would make it difficult for an automated system to detect these free articles. Both researchers attempting to map the open access landscape and publishers attempting to stop piracy would be challenged to access these full-text articles by automated means.

For articles made freely available by external operators, rather than by the authors or publishers, uncertainty about the causality of the open access citation advantage persists. Some piracy sites use crowdsourcing, with articles added based on demand.[63] Preliminary results show that highly cited journals have a higher proportion of articles in Sci-Hub than journals with fewer citations.[64] To the extent that interest generated by previous citations leads to requests and posts of bootleg articles, articles could become freely accessible because of being highly cited, rather than the reverse.

In the current study, in three of the four disciplines, more than two-thirds of the freely available articles were publishers' PDFs on sites other than the publisher's site. In 2004, Antelman did not even include publishers' PDFs as a category; but, within a couple of years, Antelman documented that publisher versions were approaching half of the free access copies in some social science disciplines.[65] Although the frequency with which we found free copies

in the form of publishers' PDFs is consistent with other recent research,[66] that fraction is high relative to how often publishers permit authors to post the published version. Estimates have put that percentage in the range of 11[67] to 29 percent.[68] Although we did not assess whether the publisher PDFs were permitted for our sample, other studies have found that posted PDFs generally were not permitted.[69]

Conclusion: Open Access, Free Access, and Piracy

Although we originally set out to closely replicate Antelman's study, we found this was not possible and the more interesting questions were around changes in how articles are shared online. Throughout this article, we followed Antelman's definition, "If any freely available full-text version (including drafts, preprints, and postprints) was available, the article was considered to be open access."[70] We sidestepped questions of whether versions were made available legally and in accordance with publication agreements and ignored questions of additional rights granted under some definitions of open access. With so many publisher PDFs and articles posted on independent/external sites, the search in Google for free articles likely uncovered copies in violation of copyright or publication agreements. Because the fraction of faculty who use a general purpose search engine as a starting point for locating research literature has grown to more than a third,[71] scholars are finding these copies too.

For open access as a system, the legal permissibility of the free distribution matters. Many free access articles are not made available in accordance with publisher policies, and sometimes their posting may not even be authorized by the author. A free access system built on unauthorized copies is vulnerable to legal action taken by copyright holders. Conversely, a tolled access system—with authors and other parties benefiting from sharing unauthorized copies for free—is vulnerable to multiple forces working against publishers' goal of receiving payment for their work.

Notes

1. Kristin Antelman, "Do Open-Access Articles Have a Greater Research Impact?" *College & Research Libraries* 65, no. 5 (2004): 372–82, doi:10.5860/crl.65.5.372.

2. Clarivate Analytics, "Web of Science" (2017), available online at https://clarivate.com/products/web-of-science/ [accessed 16 October 2017].

3. Here are two examples: Michael Norris, Charles Oppenheim, and Fytton Rowland, "The Citation Advantage of Open-Access Articles," *Journal of the Association for Information Science and Technology* 59, no. 12 (2008): 1963–72, doi:10.1002/asi.20898; Amy Atchison and Jonathan Bull, "Will Open Access Get Me Cited? An Analysis of the Efficacy of Open Access Publishing in Political Science," *PS: Political Science & Politics* 48, no. 1 (2015): 129–37, doi:10.1017/S1049096514001668.

4. Alma Swan, "The Open Access Citation Advantage: Studies and Results to Date" (2010), available online at https://eprints.soton.ac.uk/268516 [accessed 27 September 2017]; SPARC Europe, "The Open Access Citation Advantage Service (OACA)" (2015), available online at http://sparceurope.org/what-we-do/open-access/sparc-europe-open-access-resources/open-access-citation-advantage-service-oaca/ [accessed 27 September 2017]; A. Ben Wagner, "Citation Impact Advantage of Open Access (OA) Articles over Non-OA Articles: An Annotated Bibliography Version 5" (2016), available online at https://ubir.buffalo.edu/xmlui/handle/10477/75108 [accessed 27 October 2017].

5. SPARC Europe, "The Open Access Citation Advantage Service (OACA)."

6. Steve Lawrence, "Free Online Availability Substantially Increases a Paper's Impact," *Nature* 411, no. 521 (2001): 521, doi:10.1038/35079151.

7. Antelman, "Do Open-Access Articles Have a Greater Research Impact?"

8. Philip M. Davis and William H. Walters, "The Impact of Free Access to the Scientific Literature: A Review of Recent Research," *Journal of the Medical Library Association* 99, no. 3 (2011): Table 2, doi:10.3163/1536-5050.99.3.008; Wagner, "Citation Impact Advantage of Open Access (OA) Articles over Non-OA Articles."

9. Davis and Walters, "The Impact of Free Access to the Scientific Literature," Table 2.

10. Here are two examples: Norris, Oppenheim, and Rowland, "The Citation Advantage of Open-Access Articles"; Atchison and Bull, "Will Open Access Get Me Cited?"

11. For example: Chawiki Hajjem, Stevan Harnad, and Yves Gingras, "Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How It Increases Research Citation Impact," *Bulletin of the Technical Committee on Data Engineering, IEEE Computer Society* 28, no. 4 (2005): 39–46, available online at http://sites.computer.org/debull/A05dec/hajjem.pdf [accessed 23 March 2018].

12. "Berlin 3 Open Access: Progress in Implementing the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities" (Oct. 22, 2003), available online at https://openaccess.mpg.de/Berlin-Declaration [accessed 27 October 2017]; "Bethesda Statement on Open Access Publishing" (June 20, 2003), available online at https://dash.harvard.edu/bitstream/handle/1/4725199/suber_bethesda.htm?sequence=1 [accessed 27 October 2017].

13. Ibid.

14. Here are a few examples: Gunther Eysenbach, "Citation Advantage of Open Access Articles," *PLoS Biology* 4, no. 5 (2006): 692–98, doi:10.1371/journal.pbio.0040157; Lifang Xu, Jinhong Liu, and Qing Fang, "Analysis on Open Access Citation Advantage: An Empirical Study Based on Oxford Open Journals," in *iConference 2011* (New York, 2011), 426–32, doi:10.1145/1940761.1940819; Jim Ottaviani, "The Post-Embargo Open Access Citation Advantage: It Exists (Probably), It's Modest (Usually), and the Rich Get Richer (Of Course)," *PLoSOne* (2016), doi:10.1371/journal.pone.0159614.

15. For example: Norris, Oppenheim, and Rowland, "The Citation Advantage of Open-Access Articles."

16. Davis and Walters, "The Impact of Free Access to the Scientific Literature," Table 2; Wagner, "Citation Impact Advantage of Open Access (OA) Articles over Non-OA Articles."

17. Here are two examples: Pablo Dorta-González, Sara González-Betancor, and María Dorta-González, "Reconsidering the Gold Open Access Citation Advantage Postulate in a Multidisciplinary Context: An Analysis of the Subject Categories in the Web of Science Database 2009–2014," *Scientometrics* 112, no. 2 (2017): 877–901, doi:10.1007/s11192-017-2422-y; Mikael Laakso and Bo-Christer Björk, "Delayed Open Access: An Overlooked High-Impact Category of Openly Available Scientific Literature," *Journal of the Association for Information Science and Technology* 64, no. 7 (2013): 1323–29, doi:10.1002/asi.22856.

18. Philip M. Davis, "Open Access, Readership, Citations: A Randomized Controlled Trial of Scientific Journal Publishing," *FASEB Journal* 25, no. 7 (2011): 2129–34, doi:10.1096/fj.11-183988.

19. Here are two examples: Xianwen Wang et al., "The Open Access Advantage Considering Citation, Article

Usage and Social Media Attention," *Scientometrics* 103, no. 2 (2015): 555–64, doi: 10.1007/s11192-015-1547-0; Davis, "Open Access, Readership, Citations," 2131.

20. Wagner, "Citation Impact Advantage of Open Access (OA) Articles over Non-OA Articles."

21. For example: Iain D. Craig et al., "Do Open Access Articles Have Greater Citation Impact? A Critical Review of the Literature," *Journal of Informetrics* 1, no. 3 (2007): 239–48, doi:10.1016/j.joi.2007.04.001.

22. Eric Archambault et al., "Proportion of Open Access Peer-Reviewed Papers at the European and World Levels—2004–2011" (2013), available online at www.science-metrix.com/pdf/SM_EC_OA_Availability_2004-2011.pdf [accessed 27 October 2017].

23. Jingfeng Xia et al., "A Review of Open Access Self-Archiving Mandate Policies," *portal: Libraries and the Academy* 12, no. 1 (2012): 85–102, doi:10.1353/pla.2012.0000.

24. Michael Seadle, "Copyright in the Networked World: Author's Rights," *Library Hi Tech* 23, no. 1 (2005): 130–36, doi:10.1108/07378830510586766.

25. Eric Friedlander and Donald McClure, "AMS Views on Journal Publishing," American Mathematical Society (Feb. 27, 2012), available online at www.ams.org/publications/journals/AMS-Views-on-Journal-Publishing [accessed 16 October 2017].

26. Elsevier, "Free Access to Archived Articles of Primary Mathematics Journals" (2017), available online at https://www.elsevier.com/physical-sciences/mathematics/free-access-to-archived-articles-of-primary-mathematics-journals [accessed 16 October 2017].

27. Ottaviani, "The Post-Embargo Open Access Citation Advantage."

28. Wolfgang Glänzel and Urs Schoepflin, "A Bibliometric Study on Ageing and Reception Processes of Scientific Literature," *Journal of Information Science* 21, no. 1 (1995): 37–53, doi:10.1177/016555159502100104.

29. Ellen Dubinsky, "A Current Snapshot of Institutional Repositories: Growth Rate, Disciplinary Content and Faculty Contributions," *Journal of Librarianship and Scholarly Communication* 2, no. 3 (2014): eP1167, doi:10.7710/2162-3309.1167.

30. Julia A. Lovett, Andrée J. Rathemacher, Divana Boukari, and Corey Lang, "Institutional Repositories and Academic Social Networks: Competition or Complement? A Study of Open Access Policy Compliance vs. ResearchGate Participation," *Journal of Librarianship and Scholarly Communication* 5, no. 1 (2017): p.eP2183, doi:10.7710/2162-3309.2183.

31. Mikael Laakso et al., "Research Output Availability on Academic Social Networks: Implications for Stakeholders in Academic Publishing," *Electronic Markets* 27, no. 2 (2017): 125–33, doi:10.1007/s12525-016-0242-1; Mike Thelwall and Kayvan Kousha, "ResearchGate Articles: Age, Discipline, Audience Size, and Impact," *Journal of the Association for Information Science and Technology* 68, no. 2 (2017): 468–79, doi:10.1002/asi.23675.

32. Richard Van Noorden, "Online Collaboration: Scientists and the Social Network," *Nature* 512, no. 7513 (Aug. 13, 2014): 126–29, doi:10.1038/512126a.

33. Yuri Niyazov et al., "Open Access Meets Discoverability: Citations to Articles Posted to Academia.edu," *PLoSOne* (2016): e0148257, doi:10.1371/journal.pone.0148257; Mohammad Sababi et al., "How Accessibility Influences Citation Counts: The Case of Citation to the Full Text Articles Available from ResearchGate," *RT. A Journal on Research Policy & Evaluation* 5, no. 1 (2017), doi:10.13130/2282-5398/7997.

34. Bastian Greshake, "Looking into Pandora's Box: The Content of Sci-Hub and Its Usage" [version 1; referees: 2 approved, 2 approved with reservations], *F1000Research* 6:541 017, doi:10.12688/f1000research.11366.1.

35. Quirin Schiermeier, "US Court Grants Elsevier Millions in Damages from Sci-Hub," *Nature* (June 22, 2017), doi:10.1038/nature.2017.22196.

36. Derek De Solla Price, "A General Theory of Bibliometric and Other Cumulative Advantage Processes," *Journal of the American Society for Information Science* 27, no. 5 (1976): 292–306, doi:10.1002/asi.4630270505.

37. Stefan Schmidt, "Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences," *Review of General Psychology* 13, no. 2 (2009): 90–100, doi:10.1037/a0015108.

38. Harold Pashler and Christine R. Harris, "Is the Replicability Crisis Overblown? Three Arguments Examined," *Perspectives on Psychological Science* 7, no. 6 (Nov. 7, 2012): 531–36, doi:10.1177/1745691612463401.

39. John P.A. Ioannidis, "Why Most Published Research Findings Are False," *PLoS Medicine* 2, no. 8 (Aug. 30, 2005): e124, doi:10.1371/journal.pmed.0020124.

40. Open Science Collaboration, "Estimating the Reproducibility of Psychological Science," *Science* 349, no. 6251 (Aug. 28, 2015): aac4716, doi:10.1126/science.aac4716; Jocelyn Kaiser, "Rigorous Replication Effort Succeeds for Just Two of Five Cancer Papers," *Science* (Jan. 18, 2017), doi:10.1126/science.aal0628.

41. One of the earliest uses of "crisis" was in Pashler and Harris, "Is the Replicability Crisis Overblown?"

42. Monya Baker, "1,500 Scientists Lift the Lid on Reproducibility," *Nature* 533, no. 7604 (May 25, 2016): 452–54, doi:10.1038/533452a.

43. Daniel T. Gilbert et al., "Comment on Estimating the Reproducibility of Psychological Science," *Science*

351, no. 6277 (March 4, 2016): 1037, doi:10.1126/science.aad7243.

44. Chris Chambers, "Physics Envy: Do 'Hard' Sciences Hold the Solution to the Replication Crisis in Psychology?" *The Guardian* (June 10, 2014), available online at https://www.theguardian.com/science/head-quarters/2014/jun/10/physics-envy-do-hard-sciences-hold-the-solution-to-the-replication-crisis-in-psychology [accessed 15 November 2017]; Gordon J. Lithgow, Monica Driscoll, and Patrick Phillips, "A Long Journey to Reproducible Results," *Nature* 548, no. 7668 (2017): 387–88, doi:10.1038/548387a.

45. Antelman, "Do Open-Access Articles Have a Greater Research Impact?" 374–76. Our study was preregistered with an analysis plan at https://osf.io/2ae99.

46. Ibid.

47. Ibid.

48. Ibid., 381, note 17.

49. Ibid., 375.

50. Ibid., 375, table 1.

51. Ibid., 375.

52. Ibid.

53. Ibid., 375–76.

54. Ibid., 375.

55. Ibid.

56. Ibid., 381, note 23.

57. Mikael Laakso, "Green Open Access Policies of Scholarly Journal Publishers: A Study of What, When, and Where Self-Archiving Is Allowed," *Scientometrics* 99 (2014): 475–94, doi:10.1007/s11192-013-1205-3.

58. Van Noorden, "Online Collaboration," 127.

59. Antelman, "Do Open-Access Articles Have a Greater Research Impact?" 381.

60. Laakso and Björk, "Delayed Open Access."

61. Niyazov et al., "Open Access Meets Discoverability."

62. David Matthews, "Publishers Push ResearchGate Harder in Copyright Battle," *Times Higher Education* (Nov. 9, 2017), available online at https://www.timeshighereducation.com/news/publishers-push-researchgate-harder-copyright-battle [accessed 17 November 2017].

63. Guillaume Cabanac, "Bibliogifts in LibGen? A Study of a Text-Sharing Platform Driven by Biblioleaks and Crowdsourcing," *Journal of the Association for Information Science and Technology* 67, no. 4 (2016): 874–84, doi:10.1002/asi.23445.

64. Daniel S. Himmelstein et al., "Sci-Hub Provides Access to Nearly All Scholarly Literature," *PeerJ Preprints* (Oct. 12, 2017), 5: e3100v1, doi:10.7287/peerj.preprints.3100v1.

65. Kristin Antelman, "Self-Archiving Practice and the Influence of Publisher Policies in the Social Sciences," *Learned Publishing* 19 (2006): 85–95, doi:10.1087/095315106776387011.

66. Hamid R. Jamali and Majid Nabavi, "Open Access and Sources of Full-Text Articles in Google Scholar in Different Subject Fields," *Scientometrics* 105 (2015): 1635–51, doi:10.1007/s11192-015-1642-2; Hamid R. Jamali, "Copyright Compliance and Infringement in ResearchGate Full-Text Journal Articles," *Scientometrics* 112, no. 1 (2017): 241–54, doi:10.1007/s11192-017-2291-4.

67. Laakso, "Green Open Access Policies of Scholarly Journal Publishers," 475–94.

68. David Hansen, "Understanding and Making Use of Academic Authors' Open Access Rights," *Journal of Librarianship and Scholarly Communication* 1, no. 2 (2012): eP1050, doi:10.7710/2162-3309.1050.

69. Antelman, "Self-Archiving Practice and the Influence of Publisher Policies in the Social Sciences," 85–95; Denise Troll Covey, "Self-Archiving Journal Articles: A Case Study of Faculty Practice and Missed Opportunity," *portal: Libraries and the Academy* 9, no. 2 (2009): 223–51, doi:10.1353/pla.0.0042; Jamali, "Copyright Compliance and Infringement in ResearchGate Full-Text Journal Articles," 241–54.

70. Antelman, "Do Open-Access Articles Have a Greater Research Impact?" 375.

71. Christine Wolff, Alisa B. Rod, and Roger C. Schonfeld, "Ithaka S+R US Faculty Survey 2015" (2016), available online at www.sr.ithaka.org/wp-content/uploads/2016/03/SR_Report_US_Faculty_Survey_2015040416.pdf [accessed 27 October 2017].