RESEARCH ARTICLE

Looking into Pandora's Box: The Content of *Sci-Hub* and its Usage [version 1; referees: 2 approved, 2 approved with reservations]

Bastian Greshake

Institute of Cell Biology and Neuroscience, Goethe University Frankfurt, Frankfurt, Germany

v1 **First published:** 21 Apr 2017, **6**:541 (doi: 10.12688/f1000research.11366.1)
Latest published: 21 Apr 2017, **6**:541 (doi: 10.12688/f1000research.11366.1)

Abstract

Despite the growth of Open Access, potentially illegally circumventing paywalls to access scholarly publications is becoming a more mainstream phenomenon. The web service Sci-Hub is amongst the biggest facilitators of this, offering free access to around 62 million publications. So far it is not well studied how and why its users are accessing publications through Sci-Hub. By utilizing the recently released corpus of Sci-Hub and comparing it to the data of ~28 million downloads done through the service, this study tries to address some of these questions. The comparative analysis shows that both the usage and complete corpus is largely made up of recently published articles, with users disproportionately favoring newer articles and 35% of downloaded articles being published after 2013. These results hint that embargo periods before publications become Open Access are frequently circumnavigated using Guerilla Open Access approaches like Sci-Hub. On a journal level, the downloads show a bias towards some scholarly disciplines, especially Chemistry, suggesting increased barriers to access for these. Comparing the use and corpus on a publisher level, it becomes clear that only 11% of publishers are highly requested in comparison to the baseline frequency, while 45% of all publishers are significantly less accessed than expected. Despite this, the oligopoly of publishers is even more remarkable on the level of content consumption, with 80% of all downloads being published through only 9 publishers. All of this suggests that Sci-Hub is used by different populations and for a number of different reasons, and that there is still a lack of access to the published scientific record. A further analysis of these openly available data resources will undoubtedly be valuable for the investigation of academic publishing.

Open Peer Review

Referee Status: ✓ ? ✓ ?

	Invited Referees			
	1	**2**	**3**	**4**
version 1 published 21 Apr 2017	✓ report	? report	✓ report	? report

1 **April Hathcock**, New York University USA

2 **Gabriel Gardner** , California State University USA, **Stephen McLaughlin**, University of Texas at Austin USA

3 **Jill Emery** , Portland State University USA

4 **Balázs Bodó**, University of Amsterdam Netherlands

Discuss this article

Comments (2)

Corresponding author: Bastian Greshake (bgreshake@googlemail.com)

How to cite this article: Greshake B. **Looking into Pandora's Box: The Content of *Sci-Hub* and its Usage [version 1; referees: 2 approved, 2 approved with reservations]** *F1000Research* 2017, **6**:541 (doi: 10.12688/f1000research.11366.1)

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing interests: The author uses *SciHub* regularly in his own research. Otherwise the author declares no competing financial, personal, or professional interests.

First published: 21 Apr 2017, **6**:541 (doi: 10.12688/f1000research.11366.1)

Introduction

Through the course of the 20th century, the academic publishing market has radically transformed. What used to be a small, decentralized marketplace, occupied by university presses and educational publishers, is now a global, highly profitable enterprise, dominated by commercial publishers[1]. This development is seen as the outcome of a multifactorial process, with the inability of libraries to resist price increases, the passivity of researchers who are not directly bearing the costs and the merging of publishing companies, leading to an oligopoly[2].

In response to these developments and rising subscription costs, the Open Access movement started out to reclaim the process of academic publishing[3]. Besides the academic and economic impact, the potential societal impact of Open Access publishing is getting more attention[4,5], and large funding bodies seem to agree with this opinion, as more and more are adopting Open Access policies[6–8]. These efforts seem to have an impact, as a 2014 study of scholarly publishing in the English language found that, while the adoption of Open Access varies between scholarly disciplines, an average of around 24 % of scholarly documents are freely accessible on the web[9].

Another response to these shifts in the academic publishing world is what has been termed *Guerilla Open Access*[1], *Bibliogifts*[10] or *Black Open Access*[11]. Or in short, the usage of semi-legal or outright illegal ways of accessing scholarly publications, like peer2peer file sharing, for example the use of *#icanhazpdf* on Twitter[10], or centralized web services like *Sci-Hub*/*LibGen*[12].

Especially *Sci-Hub*, which started in 2011, has moved into the spotlight in the recent years. According to founder Alexandra Elbakyan, the website uses donated library credentials of contributors to circumvent publishers' paywalls and thus downloads large parts of their collections[13]. This clear violation of copyright not only lead to a lawsuit by Elsevier against Elbakyan[14], but also to her being called *"the Robin Hood of Science"*[15], with both sparking further interest in *Sci-Hub*.

Despite this, there has been little research into how *Sci-Hub* is used and what kind of materials are being accessed through it. A 2014 study has looked at content provided through *LibGen*[10]. In 2016 *Sci-Hub* released data on ~28 million downloads done through the service[16]. This data was subsequently analyzed to see in which countries the website is being used, which publishers are most frequent[13] and how downloading publications through *Sci-Hub* relates to socio-economic factors, such as being based in a research institution[17] and how it impacts interlibrary loans[12].

In March 2017 *Sci-Hub* released the list of ~ 62 million Digital Object Identifiers (DOIs) of the content they have stored. This study is the first to utilize both the data on which publications are downloaded through *Sci-Hub*, as well as the complete corpus available through them. This allows a data-driven approach to evaluate what is stored in the *Sci-Hub* universe, how the actual use of the service differs from that, and what different use cases people might have for *Sci-Hub*.

Methods

Data sources

The data on the around 62 million DOIs indexed by *Sci-Hub* was taken from the dataset released on 2017-03-19[18]. In addition, the data on the 28 million downloads done through *Sci-Hub* between September 2015 and February 2016[16] was matched to the complete corpus of DOIs. This made it possible to quantify how often each object listed in *Sci-Hub* was actually requested from its user base.

Resolving DOIs

The corresponding information for the publisher, the year of publication, as well as the journal in which it was published was gotten from doi.org, using the *RubyGem Terrier* (v1.0.2, https://github.com/Authorea/terrier). Acquiring the metadata for each of the 62 million DOIs in Sci-Hub was done between 2017-03-20 and 2017-03-31. In order to save time, the DOIs of the 28 million downloads were then matched to the superset of the already resolved DOI of the complete *Sci-Hub* catalog. In both cases, DOIs that could not be resolved were excluded from further analysis, but they are included in the dataset released with this article.

Tests for over- & under-representation

For each publisher, the number of papers downloaded was compared to the expected number of downloads, given the publishers' presence in the whole *Sci-Hub* database. For this the relative contribution to the database was calculated for each publisher, excluding all missing data. The number of actual downloads was then compared to the expected number of downloads using a binomial test. All p-values were corrected for multiple testing with False Discovery Rate[19] and post-correction $p<0.05$ were accepted.

Results

Resolving the *Sci-Hub* DOIs

For the 61,940,926 DOIs listed in the Sci-Hub data dump, a total of 46,931,934 DOIs could be resolved (75.77%). Manual inspection of the unresolvable 25% shows that nearly all of these could not be resolved as they are not available via doi.org, and are not a technical error in the procedure to resolve them (i.e. lack of internet connection). For the data on the downloads done through Sci-Hub, 21,515,195 downloads could be resolved out of 27,819,965 total downloads (77.34%).

The age of publications in *Sci-Hub*

To estimate the age distribution of the publications listed in *Sci-Hub*, and which fraction of these publications is actually requested by the people using *Sci-Hub*, the respective datasets were tabulated according to the year of publication, see Figure 1. While over 95% of the publications listed in *Sci-Hub* were published after 1950, there is nevertheless a long tail, reaching back to the 1619 edition of *Descriptio cometæ*[20].

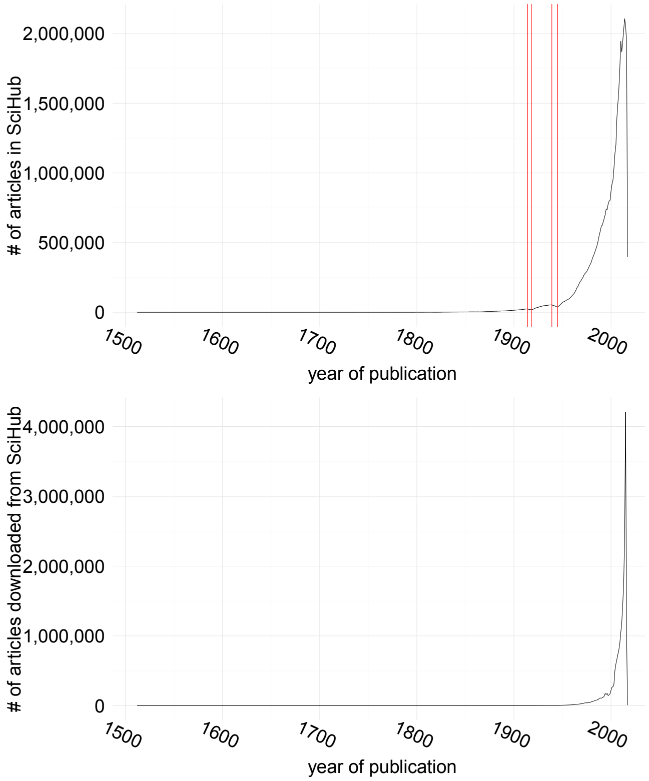

Figure 1. Top: Number of Publications in *Sci-Hub* by year of publication. Red bars denote the years 1914, 1918, 1939 and 1945. Bottom: Number of publications downloaded by year of publication.

As a general trend the number of publications listed in *Sci-Hub* increases from year to year. Two notable exceptions are the time periods of the two World Wars, at which ends the number of publications dropped to pre-1906 and pre-1926 levels, respectively (red bars in Figure 1).

When it comes to the publications downloaded by *Sci-Hub* users, the skew towards recent publications is even more extreme. Over 95% of all downloads fall into publications done after 1982, with ~35% of the downloaded publications being less than 2 years old at the time they are being accessed (i.e. published after 2013). Despite this, there is also a long tail of publications being accessed, with articles published even in the 1600s being amongst the downloads, and 0.04% of all downloads being made for publications released prior to 1900.

Which journals are being read?

The complete released database contains ~177,000 journals, with ~60% of these having at least a single paper downloaded. The number of articles per journal likely follows an exponential function, for both the total number of publications listed on *Sci-Hub* as well as the number of downloaded articles (see Supplementary Figure S1), with <10% of the journals being responsible for >50% of the total content in *Sci-Hub*. The skew for the downloaded content is even more extreme, with <1% of all journals getting over 50% of all downloads.

Contrasting the 20 most frequent journals in the complete database with the 20 most downloaded ones (Figure 2), one observes a clear shift not only in the distribution but also in the ranking, with the most abundant journal of the whole corpus not appearing in the 20 most downloaded journals. In addition, chemical journals appear to be overrepresented in the downloads (12 journals), compared to the complete corpus (7 journals), with no other discipline showing an increase amongst the 20 most frequent journals.

Are publishers created equal?

Looking at the data on a publisher level, there are ~1,700 different publishers, with ~1,000 having at least a single paper downloaded. Both corpus and downloaded publications are heavily skewed towards a set of few publishers, with the 9 most abundant publishers having published ~70% of the complete corpus and ~80% of all downloads respectively (see Supplementary Figure S2).

Given the background frequency in the complete corpus, the download numbers were compared to the expected numbers using a binomial test. After false discovery rate correction for multiple testing, 982 publishers differed significantly from the expected download numbers, with 201 publishers having more downloads than expected and 781 being underrepresented. Interestingly, while some big publishers like Elsevier and Springer Nature come in amongst the overly downloaded publishers, many of the large publishers, like Wiley-Blackwell and the Institute of Electrical and Electronics Engineers (IEEE) are being downloaded less than expected given their portfolio (Figure 3).

Discussion

Earlier investigations into the data provided through *Sci-Hub* and *LibGen* focused large on either on the material being accessed[13] or on the data stored in these resources[10]. This study is the first to make use of both the whole corpus of *Sci-Hub* as well as data on how this corpus is being accessed by its users.

Why *Sci-Hub?*

Comparing actual usage with the background set of articles shows that articles from recent history are highly sought for, giving some evidence that embargoes prior to making publications Open Access seem to become less effective. These findings are in line with prior research into the motivations for crowd-sourced, peer2peer academic file sharing[21]. While embargoes have impact on the use of those publications[22], these hurdles are being surpassed more and more by *Black Open Access*[11], as provided by *Sci-Hub.*

While a good part of the literature available through *Sci-Hub* seems to be rarely accessed, the long tail of, publications, especially older ones, seems to be put to use - albeit at a lower frequency. With DOIs that are unresolvable due to issues on publishers' sides[23], and with Open Access publications that disappear behind accidental

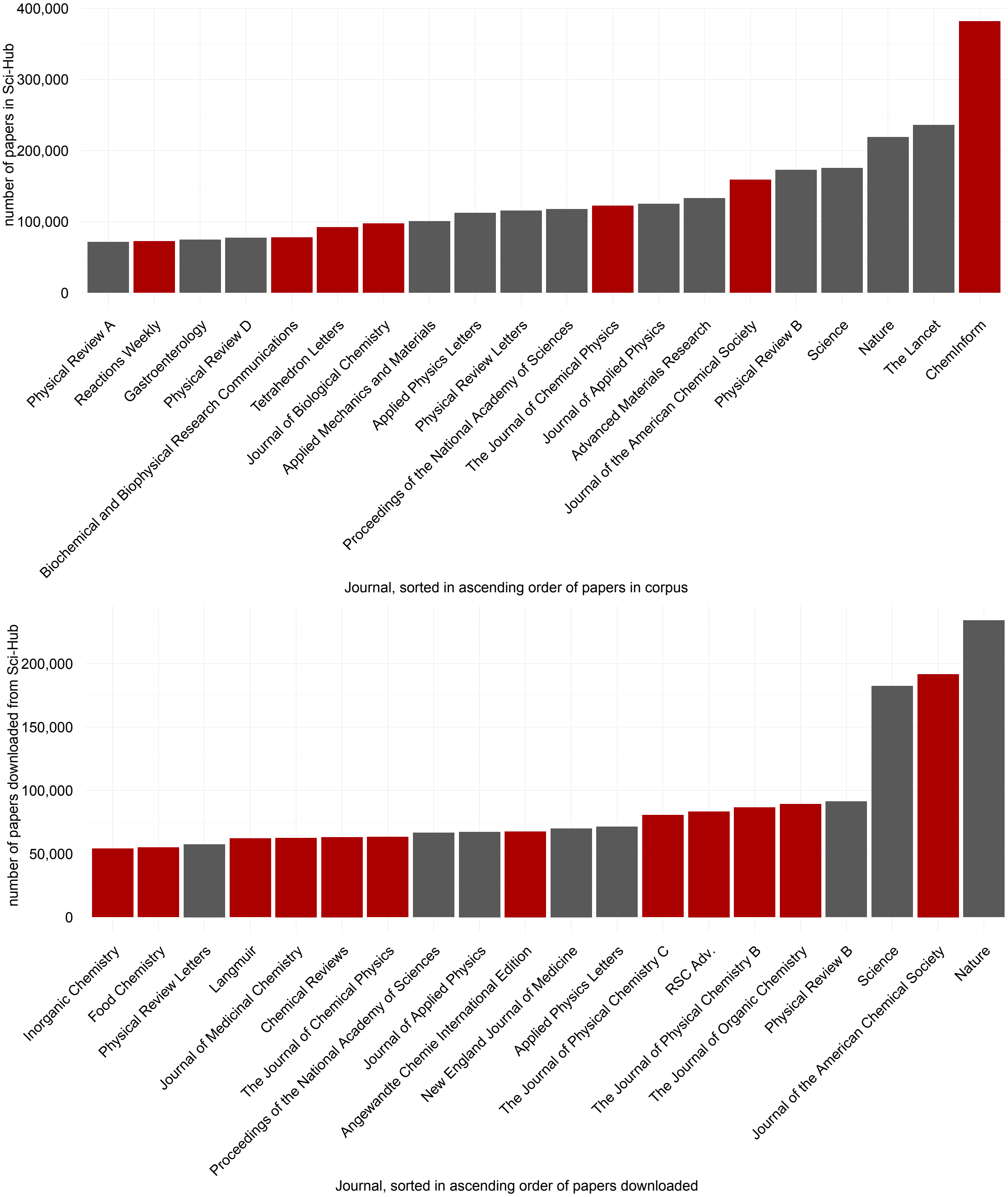

Figure 2. Top: The 20 most frequent journals in all of *Sci-Hub.* Bottom: The 20 journals with the most downloads. In both panels Chemistry journals are highlighted in red.

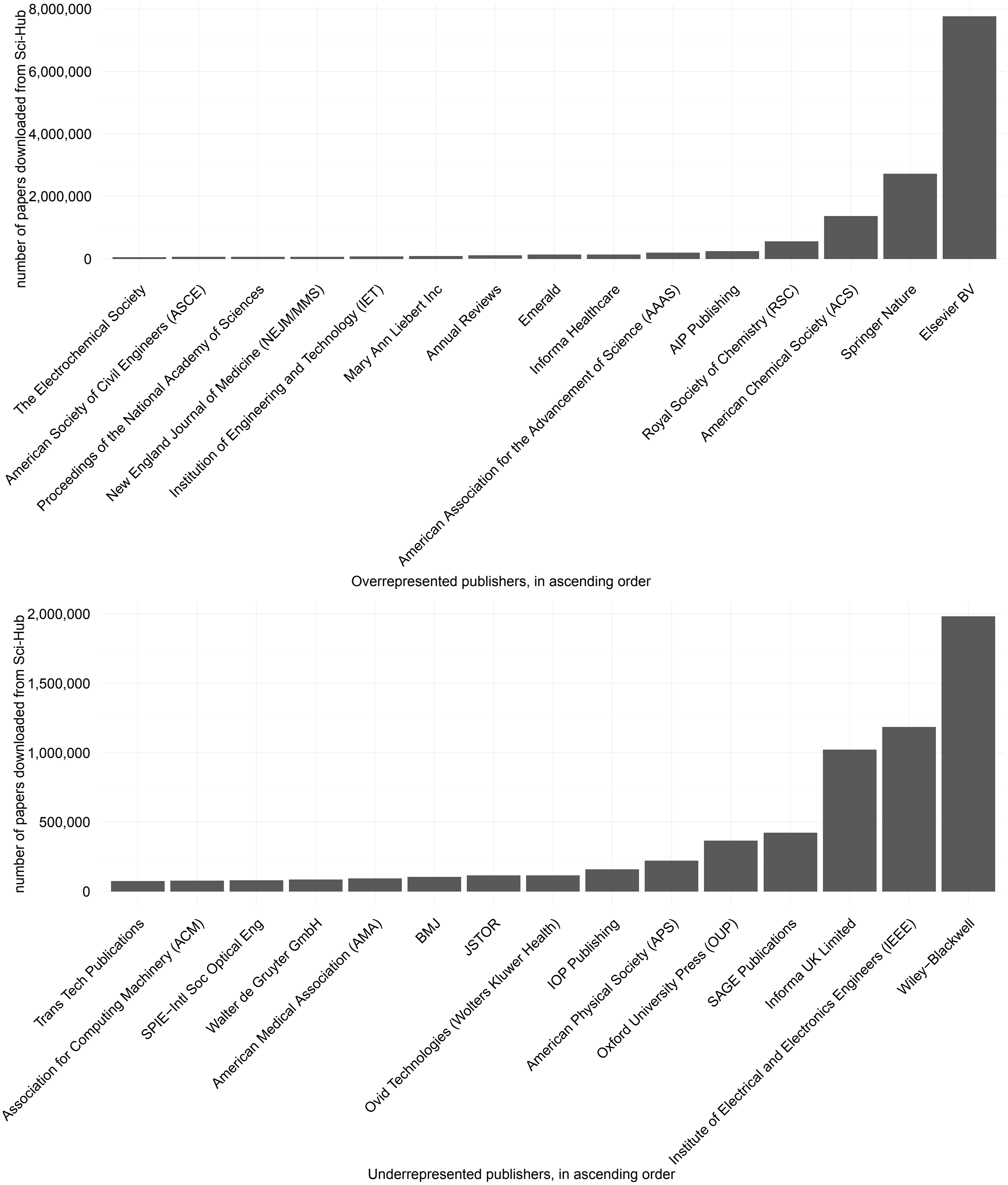

Figure 3. The most downloaded publishers that are either overrepresented (top) or underrepresented (bottom).

paywalls[24], this use for *Black Open Access* might play an important role and needs to be investigated more closely. It is worth noting that all analyses related to the number of downloads are limited to the six month period between September 2015 and February 2016, and do not necessarily reflect the complete use of *Sci-Hub*.

Who's reading?

Looking at the disproportionately frequented journals, one finds that 12 of the 20 most downloaded journals can broadly be classified as being within the subject area of chemistry. This is an effect that has also been seen in a prior study looking at the downloads done from *Sci-Hub* in the United States[12]. In addition, publishers with a focus on chemistry and engineering are also amongst the most highly accessed and overrepresented. While it is unclear whether this imbalance comes due to lack of access by university libraries, it's noteworthy that both disciplines have a traditionally high number of graduates who go into industry. The 2013 *Survey of Doctorate Recipients* of the National Center for Science and Engineering Statistics (NCSES) of the United States finds that 50% of chemistry graduates and 58% of engineering graduates move to private, for-profit industry while only 32% and 27% respectively stay at educational institutions[25]. In comparison, in the life sciences these numbers are nearly switched, with 52% of graduates staying at educational institutions, which presumably offer more access to the scientific literature.

Non solus. Or at least not completely

The prior analysis of the roughly 28 million downloads done through *Sci-Hub* showed a bleak picture when it came to the diversity of actors in the academic publishing space, with around 1/3 of all articles downloaded being published through Elsevier[13]. The analysis presented here puts this into perspective with the whole space of academic publishing available through *Sci-Hub*, in which Elsevier is also the dominant force with ~24% of the whole corpus. The general picture of a few publishers dominating the market, with around 50% of all publications being published through only 3 companies, is even more pronounced at the usage level compared to the complete corpus, perpetuating the trend of *the rich getting richer*. Only 11% of all publishers, amongst them already dominating companies, are downloaded more often than expected, while publications of 45% of all publishers are significantly less downloaded.

Conclusions

The analyses presented here suggest that *Sci-Hub* is used for a variety of reasons, by different populations. While most usage is biased towards getting access to recent publications, there is a subset of users interested in getting historical academic literature. Compared to the complete corpus, *Sci-Hub* seems to be a convenient resource, especially for engineers and chemists, as the overrepresentation shows. Lastly, when it comes to the representation of publishers, the *Sci-Hub* data shows that the academic publishing field is even more of an oligopoly in terms of actual usage when compared to the amount of literature published. Further analysis of how, by whom and where *Sci-Hub* is used will undoubtedly shed more light onto the practice of academic publishing around the globe.

Data availability

All the data used in this study, as well as the code to analyze the data and create the figures, is archived on Zenodo as *Data and Scripts for Looking into Pandora's Box: The Content of Sci-Hub and its Usage* (DOI, 10.5281/zenodo.472493)[26].

In addition the analysis code can also be found on GitHub at http://www.github.com/gedankenstuecke/scihub.

Competing interests

The author uses *SciHub* regularly in his own research. Otherwise the author declares no competing financial, personal, or professional interests.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgements

The author wants to acknowledge Alexandra Elbakyan, for releasing both data sets used in this study. Further thanks go to John Bohannon, who analyzed and helped release the initial data on downloads from *Sci-Hub*, and to Athina Tzovara and Philipp Bayer, for fruitful discussion of this manuscript as well as the statistics and analyses involved.

Supplementary materials

Supplementary Figure S1: Top: The distribution of publications per journal in the whole corpus, sorted in ascending order of articles. Bottom: The distribution of downloads per journals, sorted in ascending order of downloads.

Click here to access the data

Supplementary Figure S2: The proportion of the whole content as aggregated by publisher, both for the corpus (top) and downloads (bottom). Sorted by number of publications in the respective dataset. Only the 9 most frequent publishers are listed, smaller ones are grouped as *other*.

Click here to access the data

References

1. Balázs B: **Pirates in the library – an inquiry into the guerilla open access movement.** *Paper prepared for the 8th Annual Workshop of the International Society for the History and Theory of Intellectual Property, CREATe, University of Glasgow, UK July 6–8, 2016.* 2016.
Reference Source
2. Larivière V, Haustein S, Mongeon P: **The Oligopoly of Academic Publishers in the Digital Era.** *PLoS One.* 2015; **10**(6): e0127502.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Royster P: **A brief history of open access.** (accessed 4th of april, 2017), 2016.
Reference Source
4. Tennant JP, Waldner F, Jacques DC, *et al.*: **The academic, economic and societal impacts of Open Access: an evidence-based review [version 3; referees: 3 approved, 2 approved with reservations].** *F1000Res.* 2016; **5**: 632.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Wildschut D: **The need for citizen science in the transition to a sustainable peer-to-peer-society.** *Futures.* 2017.
Publisher Full Text
6. Butler D: **Gates Foundation announces open-access publishing venture.** *Nature.* 2017; **543**(7647): 599.
PubMed Abstract | Publisher Full Text
7. Jahn N, Tullney M: **A study of institutional spending on open access publication fees in Germany.** *Peer J.* 2016; **4**: e2323.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Giles J: **Trust gives warm welcome to open access.** *Nature.* 2004; **432**(7014): 134.
PubMed Abstract | Publisher Full Text
9. Khabsa M, Giles CL: **The number of scholarly documents on the public web.** *PLoS One.* 2014; **9**(5): e93949.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Cabanac G: **Bibliogifts in libgen? a study of a text-sharing platform driven by biblioleaks and crowdsourcing.** *J Assoc Inf Sci Technol.* 2015; **67**(4): 874–884.
Publisher Full Text
11. Björk BC: **Gold, green, and black open access.** *Learned Publishing.* 2017; **30**(2): 173–175.
Publisher Full Text
12. Gardner GJ, McLaughlin SR, Asher AD: **Shadow libraries and you: Sci-hub usage and the future of ill.** *In ACRL 2017, Baltimore, Maryland, March 22–25, 2017.* 2017.
Reference Source
13. Bohannon J: **Who's downloading pirated papers? Everyone.** *Science.* 2016; **352**(6285): 508–12.
PubMed Abstract | Publisher Full Text
14. **Elsevier inc. *et al.* v. sci-hub *et al.* case no. 1:15-cv-04282.** 2015.
Reference Source
15. Oxenham S: **Meet the robin hood of science.** (accessed 4th of april, 2017). 2016.
Reference Source
16. Bohannon J, Elbakyan A: **Data from: Who's downloading pirated papers? everyone.** 2016.
Publisher Full Text
17. Greshake B: **Correlating the sci-hub data with world bank indicators and identifying academic use.** *The Winnower.* 2016.
Publisher Full Text
18. Hahnel M: **List of dois of papers collected by scihub.** *figshare.* 2017.
Publisher Full Text
19. Benjamini Y, Hochberg y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society. Series B (Methodological).* 1995; **57**(1): 289–300.
Reference Source
20. Snell W: **De cometarum materia, qui in solis vicinia non exarserunt.** In *Descriptio Cometæ.* Elsevier BV, 1619. 53–57.
Publisher Full Text
21. Gardner CC, Gardner GJ: **Fast and furious (at publishers): The motivations behind crowdsourced research sharing.** *Coll Res Libr.* 2017; **78**(2): 131–149.
Publisher Full Text
22. Ottaviani J: **Correction: The Post-Embargo Open Access Citation Advantage: It Exists (Probably), It's Modest (Usually), and the Rich Get Richer (of Course).** *PLoS One.* 2016; **11**(10): e0165166.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Mounce R: **Comparing oup to other publishers.** (Accessed 4th of april, 2017), 2017.
Reference Source
24. Mounce R: **Hybrid open access is unreliable.** (Accessed 4th of april, 2017), 2017.
Reference Source
25. National Center for Science and Engineering Statistics: **2013 survey of doctorate recipients.** (Accessed 4th of april, 2017), 2014.
Reference Source
26. Greshake B: **Data and Scripts for Looking into Pandora's Box: The Content of Sci-Hub and its Usage.** *Zenodo.* 2017.
Data Source

Open Peer Review

Current Referee Status: ✓ ? ✓ ?

Version 1

Referee Report 12 May 2017

doi:10.5256/f1000research.12270.r22120

?

Balázs Bodó

Institute for Information Law, University of Amsterdam, Amsterdam, Netherlands

The analysis of shadow libraries usage data is not a trivial matter, and requires some caution, especially when someone tries to understand the processes that produce these usage numbers. The article is very modest in its aims, and hopes to present only a very basic analysis of the Sci-Hub usage, but I believe more could have been done in terms of the analysis, and more caution should have been used in when offering explanations.

During the analysis, I think the logic of Sci-Hub allows us to distinguish between two processes: the one that *produces* the collection, and one that *consumes* the collection. Articles get into the Sci-Hub collection when someone bumps into a paywall, and turns to Sci-Hub to circumvent it. This means that the corpus of Sci-Hub is indicative of works that have limited accessibility. When analyzing the corpus, the distribution of publishers, and topics, one should look at it from this perspective, and check, for example the open access policies of the most highly represented publishers, or journals, and analyse the results not just within the sci-hub universe, but against the whole population of articles/journals/publishers/topics, including those with widespread open access policies.

The download numbers, on the other hand, represent the demand for an article. I would argue that articles with only 1 download only inform about the accessibility (someone met a paywall, and downloaded the article from sci-hub), while articles with more than 1 downloads actually suggest some things about the demand (how many individuals were interested in that article/discipline).

On that note I missed the geographic analysis, especially as some data on the location of the download was also available in the original dataset.

Regarding the interpretation of the data. I think the analysis in the Who’s reading? section is not substantiated by the data in any manner. On the contrary, while the data covers all downloads, across all the globe, the interpretation relies on a US census. I don't think that is appropriate. Local usage is structured and explained by local characteristics of higher education, research, and economy. One should not generalize a US explanation to the whole dataset.

The analysis in the *Non solus* section is also misleading. It makes claims about the academic publishing space in general, while the sci-hub data is biased, as it only contains articles with accessibility problems. Articles, journals and publishers with no accessibility problems are probably missing from, or are heavily underrepresented in the dataset, thus one cannot come to any conclusion on the state of academic publishing. Take the case of PLOSone as an example, on why the current analysis is flawed.

As a result, the validity of the overall conclusions is limited.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 05 May 2017

doi:10.5256/f1000research.12270.r22123

✓

Jill Emery
Portland State University, Portland, ON, USA

Bastian Greshake has done a good job in presenting his argument and providing supporting documentation. He may want to consider: Mark Ware's 2015 STM Report noted below[1] in regards to the research behaviour & motivation, as there may be information in this report that help further augment why SciHub is used & who is "reading". Greshake's graphs readily illustrate the points he is making regarding regarding the represented journals & publishers. His use of the publicly available data and noting both where the data is located and scripts used in order to perform his study lend to the transparency of his study. Lastly, these findings are of use and interest to librarians and information scientists as well as to product and resource developers looking to develop mechanisms to counter the "SciHub phenomena."

References

1. Ware M: The STM Report An overview of scientific and scholarly journal publishing. *STM*. 2015.
Reference Source

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: scholarly communication and scholarly publishing

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 02 May 2017

doi:10.5256/f1000research.12270.r22122

?

Gabriel Gardner[1] , **Stephen McLaughlin**[2]
[1] California State University, Long Beach, CA, USA
[2] School of Information, University of Texas at Austin, Austin, TX, USA

In general, this is a clearly written and argued paper on a developing topic affecting the scholarly communications ecosystem. The author has engaged with much of the recent literature on the topic of which we are aware. The underlying data is freely available and thus possible to replicate. The quantitative analysis proceeds logically and is easy to understand. There are a few areas where we would like to see discussion expanded (noted below) though overall this paper is a very valuable contribution to the literature on this topic.

Specific Criticisms:
The abstract brings up the question of who uses Sci-Hub and why. However, there is relatively little discussion of this in the paper. By our reading of the literature, the question has not been rigorously addressed to date. But some have taken steps toward an answer. Specifically Travis, (2016) is a data point worth discussing < http://www.sciencemag.org/news/2016/05/survey-most-give-thumbs-pirated-papers>. (The survey had a large response rate but should be viewed with the skepticism that would normally apply to any “open” internet survey.)

The Supplementary Figures are worth incorporating into the text. S2, in particular, is an informative chart. It should be improved by matching the colors for each publisher in the legend. That is, “other” should appear as the darkest blue in both bars, rather than being assigned different shades of blue as it is presently. That will allow readers to observe the important differences easily.

Your methods section should include some additional discussion of what you mean by “expected number of downloads for each publisher.” You are using “expected” in a mathematical sense that diverges from the word’s everyday meaning, so you should spell this out for the reader.

We find the use of the term “Black Open Access” in the discussion section puzzling. “Guerilla open access” is more widely used, as suggested by Google Trends < https://trends.google.com/trends/explore?q=%22black%20open%20access%22,%22guerilla%20open%2
>. Additionally, there are important issues of “respectability politics” to consider here; there are vocal OA advocates and practitioners who condemn Sci-Hub and do not want the OA movement to be associated with it or with copyright violation. Using the word “black” may be interpreted as implying that Sci-Hub is compatible with so-called green and gold OA publishing. Librarians in particular are loath to associate Sci-Hub with the OA movement, due to professional norms that often include upholding intellectual property restrictions on ethical grounds (e.g., <http://crln.acrl.org/content/78/2/86.full>, <https://thewinnower.com/papers/3489-signal-not-solution-notes-on-why-sci-hub-will-not-open-access>. On the other end of the spectrum, Sci-Hub’s supporters and sympathizers may object to negative connotations conjured by the term “black.” None of the above comments are meant to imply that your usage of “Black Open Access” is wrong. However, if you are going to use the less familiar term, you should explain why and note that this is a contested issue.

In the Introduction section, your remarks on Sci-Hub’s legal status are well made, but another aspect of this is the fact that credential sharing is explicitly prohibited by many publishers (and some libraries) in their terms of use. This is worth mentioning. Elsevier’s and Wiley’s Terms are clear on this issue. < https://www.elsevier.com/legal/elsevier-website-terms-and-conditions> < http://onlinelibrary.wiley.com/termsAndConditions>
Due to the ambiguous legality of copying factual and educational works under various copyright regimes, we prefer the terms “potentially illegal” or “likely illegal” when describing Sci-Hub’s activities. A recent ruling in India, for instance, suggests that Sci-Hub may not violate the law in that country.
<
https://hughstephensblog.net/2016/09/27/the-indian-high-court-decision-on-delhi-universitys-copy-shop-a
>

Also in the Introduction, the citation for the sentence discussing #icanhazpdf refers to Cabanac, 2015. However, #icanhazpdf is mentioned in that article only in passing. A more thorough analysis can be found in Gardner & Gardner, 2015. <http://eprints.rclis.org/24847/>

Bodó deserves to be cited, but there are better sources on long-term changes in the academic publishing industry. Thompson (2005) is an especially good candidate. And Royster’s slides on the history of the OA movement [3] strikes us as insufficiently authoritative. Willinsky (2006) and/or Suber (2012) are potential alternatives.

Under “Data Sources,” you should credit Elbakyan (not Hahnel) with releasing the list of DOIs in Sci-Hub.
<https://sci-hub.cc/downloads/doi.7z>
<https://twitter.com/Sci_Hub/status/843546352219017218>

Suber, Peter. 2012. Open Access. MIT Press Essential Knowledge Series. Cambridge, Mass: MIT Press.
Thompson, John B. 2005. Books in the Digital Age: The Transformation of Academic and Higher Education Publishing in Britain and the United States. Cambridge, UK ; Malden, MA: Polity Press.
Willinsky, John. 2006. The Access Principle: The Case for Open Access to Research and Scholarship. Cambridge, MA: MIT Press.

Minor corrections:

- Page 2, first sentence of 2nd paragraph: Change “was gotten from” to “was obtained from.”
- Page 2, last sentence of 2nd paragraph (and throughout): “peer-to-peer” is preferable to “peer2peer.”
- Page 2, first sentence of last paragraph: Change “publications is actually” to “publications are actually.”
- Page 2, last sentence of 3rd paragraph: Change “lead” to “led” (past tense).
- Page 2, first sentence of Data Sources (and throughout): Change “DOI” to “DOIs” for plural use.
- Page 2, second sentence of Data Sources: Change “downloads” to “download requests”
- Page 2, second sentence of Resolving DOIs: Change “meta data” to “metadata.”
- Page 2, second sentence of Results (and throughout): Insert comma after “i.e.”
- Page 3, first sentence of “Which Journals are Being Read?” and first sentence of “Are Publishers Created Equal?”: Change “at least a single paper downloaded” to clarify that you’re referring to the 6 months included in the log dataset.
- Page 3, first paragraph of the Discussion section: Change “large” to “largely.”
- Page 3, first paragraph of the Discussion section: “the whole corpus of Sci-Hub” implies you used the articles themselves. Change to “metadata for the whole corpus” or something similar.
- Page 3, second paragraph of the Discussion section: Change “more and more surpassed” to “more and more by.”
- Page 3, last paragraph: errant comma after ‘the long tail of’.
- Page 6, “Competing interests”: Change “SciHub” to “Sci-Hub.”
- Reference [1] should read “Balázs Bodó” instead of “Bodó Balázs.” “Bodó” is both his legal surname and his familiar name, so he occasionally flips the order.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Referee Report 25 April 2017

doi:10.5256/f1000research.12270.r22119

✓ **April Hathcock**
Specialized Research Services, New York University , New York, NY, USA

This is a clear and well-researched paper on a very timely topic for science communication. I have just a few issues with some of the conclusions reached and with some of the literature represented in the review.

So far it is not well studied how and why its users are accessing publications through Sci-Hub.
This isn't necessarily true. The last year has seen a lot of articles pop up in the science communication and library literature about SciHub and the whys and hows of its use, including last year's widely shared Science article by John Bohannon, which you briefly mention. This statement should be a bit tempered.

Speaking of the whys of Sci-Hub, you discuss the founder's description of how it is done but did not include any discussion from her about why she chose to develop the database. Her main occupation is as a scientist and she chose to develop SciHub because of being unable to access the literature in her field. I think that story is a compelling backdrop to your own research here.

Again, Bohannon's Science article from April 2016 "Who's downloading pirated papers? EVERYONE," gets very little mention in your paper. In any case, it certainly warrants a bit more discussion in your work. What did Bohannon do right in his analysis? Wrong? How does your work build on or diverge from his findings? In addition to Bohannon's work, there have been a number of scholarly communication experts who have explored and written about they hows and whys of Sci-Hub usage, particularly in the library and information science field. I think a review of some of that literature would really help to ground your work.

The analyses presented here suggest that Sci-Hub is used for a variety of reasons, by different populations. You argue that your study shows that users use Sci-Hub for a "variety of reasons" but I don't know that your research really supports that. Certainly you've shown what is being accessed and revealed interesting findings in terms of disciplinary, publisher, and publication date distribution, but your results can hardly be said to reveal the underlying motivations of users accessing materials from Sci-Hub. You posit some interesting theories that could explain the numbers you found (lack of access because of lack of well-funded institutional affiliation, etc.), but they are just that: theories. I'd be a bit more cautious in the conclusions you draw from your data, as interesting as they may be.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Referee Expertise: Scholarly Communication

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Discuss this Article

Version 1

Author Response 19 May 2017

Bastian Greshake,

Hey Ernesto,
thanks for the interest in the paper!

I just tried to download the data from Zenodo using the link you gave in your comment and it worked on my end without any issues (with Chrome, using the University's internet connection in my office). So either it was a temporary issue with Zenodo or the issue must be somehow with your connection.

I vaguely remember that someone had issues with Zenodo and their connection as well at some point. Could you try another connection for the download? Otherwise I'd be happy to find another way to get the data to you, i.e. if it helps I can deposit the data somewhere else for comparison.

Cheers,
Bastian

Competing Interests: No competing interests were disclosed.

Reader Comment 19 May 2017

Ernesto Priego, City University London, UK

It is commendable the author has published the article here, and that the code and raw data has been made available open access as well on both Github and Zenodo. Very good practice. Refreshing and inspiring!

I tried to download the data from Zenodo but I get the following message on Firefox:

"The site at https://zenodo.org/record/472493/files/data.tar.gz has experienced a network protocol violation that cannot be repaired.

The page you are trying to view cannot be shown because an error in the data transmission was detected.

Please contact the website owners to inform them of this problem."

Is this me or my system or is there an issue with Zenodo or the upload?

Thought I'd ask here...

Cheers.

Competing Interests: None