Pirated Economics

Zakaria Babutsidze

ASECU South-Eastern Europe Journal of Economics 2 (2018) 209-219 SEEJE

PIRATED ECONOMICS

ZAKARIA BABUTSIDZE
SKEMA Business School, Université Côte d'Azur (GREDEG)
and OFCE, Sciences Po Paris

Abstract
I argue that the impact of piracy engines for scholarly content on science depends on the nature of the research. Social sciences are more likely to reap benefits from such engines without inflicting much damage on journal publishers' revenues. To validate the claim, I examine the data from illegal downloads of economics content from Sci-Hub over a five-month period. I conclude that: (a) the extent of piracy in economics is not pervasive; (b) downloads mostly occur in under-developed countries; (c) users pirate even content that is freely available online. As a result, publishers are not losing much revenue, while exposure to generated knowledge is extended.

JEL Classification: A1
Key Words: Economics, Scientific Research, Open-access Publishing, Online Piracy

Acknowledgements: I would like to thank Lionel Nesta and an anonymous referee for comments on the earlier draft, John Bohannon for clarifying a few issues with the original Sci-Hub data and Natalia Timus for encouragement and contributions to a complementary paper.

Corresponding Address: 60 rue Dostoievski, 06560 Sophia Antipolis, France.
E-mail: zakaria.babutsidze@skema.edu

1. Introduction

The idea of open science has challenged many science and publishing stake-holders for years. Many have argued that pricing practices by mainstream scientific journal publishers have built walls around knowledge precluding a large number of researchers and members of the general public from accessing a public good. Some have even compared this 'paywall' to the Berlin Wall, dividing east and west, during the Cold War period (Oxenham 2016).

This becomes particularly problematic in regards to knowledge generated through publicly funded research. Some claim that eliminating scientific journal publication from the knowledge creation process will lead to annual savings of $9.8bln of public money (Brembs 2016). Many years of contemplation by public funding bodies have resulted in clear actions concerning institutionalising open access. Best examples of such cases are the NIH Public Access Policy (National Institutes of Health 2009) and the Guidelines for Open Access to Publications and Data in Horizon 2020 (European Commission 2016).

One of the major arguments made for open access science is the fact that scientific journal publishers have high profit margins. However, the problem is somewhat more complex and involves understanding the incentives of various stakeholders in the knowledge creation process. Discussions around the 'new economics of science' have advanced in the last two decades and demonstrate the subtleties of the problem (Partha and David 1994; David 1998). This stream of literature provides a framework for thinking about complex inter-dependencies between academic science and technological progress, which pass through private and public R&D efforts and the organisation of science.

In any case, the rise of 'open science' is a fact. This move can be illustrated by three distinct developments: the first one is the emergence of open access journals. A good example of this development is *PLoS* suite of journals, the highest impact one among them being PLoS One, which has managed to dramatically increase its attractiveness since its establishment in 2007. The number of articles published by the journal has increased over 20 times in the 10 years of its existence and it has managed to achieve an impact factor of 2.8. In a similar vein, many non-open access journals have also joined the initiative to provide authors with the option of making the published article open access (for a fee). For example, Springer provides this option for most of titles it publishes. There are current collective efforts to further such arrangements between publishers and content consumers (Vogel and Kupferschmidt 2017).

Open access to publication is believed to increase the impact of research. As a result, the number of open access articles published has skyrocketed over the last two decades (Laasko *et al.* 2011). However, evidence supporting the greater impact of open access research is not clear-cut. While some researchers find a positive impact through open access reflected on their citation count (Antelman 2004,

Eysenbach 2006), others find no evidence of open access advantage (Davis *et al.* 2008, Gaule and Maystre 2011). Nevertheless, open access publications do seem to have a clear-cut advantage in terms of non-academic dissemination, (Tennant *et al.* 2016).

The second development along the lines of open science development is the trend of journals pushing for openly sharing data contained in scientific publications. This has become an all-encompassing phenomenon, covering journals from both open and closed access sides of the spectrum, as well as universities and other public and private institutions. Similar to open access publishing, open access data is thought to facilitate the advancement of science by promoting further research and innovation (Nature 2015, Silva 2014). However, significant challenges faced by main actors have been identified in this direction, too (Perkmann and Schildt 2015, Wainwright *et al.* 2016). The main challenge here is to overcome the disincentive of private institutions to share their scientific output, because of their unwillingness to share their proprietary data.

The third and perhaps most controversial and radical development has been the development of channels to circumvent paywalls, which usually entails a violation of copyright laws. These range from crowd-sourced research sharing (e.g. using a hashtag #icanhazpdf to ask other researchers to download and send an article to which not all individuals have access) (Caffrey Gardner and Gardner 2016) all the way to creating digital piracy engines that provide free access to scientific content illegally. Publishers have pushed back hard on such developments (Singh Chawla 2017).

The most famous of this sort of services is Sci-Hub. Sci-Hub was created in 2011 and now notches tens of thousands illegal downloads a day. Among researchers, the service is seen as a portal giving a chance to scholars from poorer countries to access cutting-edge research in all fields of study (Greshake 2016).

Up until very recently not much has been known about the size and geographical breakdown of Sci-Hub operations. Thus, the poor-country enabler status of Sci-Hub could not have been verified. However, recently the data on five months of downloads from the Sci-Hub service have emerged (Elbakyan and Bohannon 2016). These data show that Sci-Hub contains 68.9% of all published scholarly articles (Himmelstein *et al.* 2017).[1]

1. According to the interactive browser available at https://greenelab.github.io/scihub/, the coverage of the top five economics journals (which are at the focal point of this article), comes to around 73.1%. Four journals (*American Economic Review, Econometrica, Quarterly Journal of Economics* and *Review of Economic Studies*) have coverage rates above 97%. The coverage of the *Journal of Political Economy* is estimated at 36.4%, which is relatively low. However, due to the complications presented by the DOI assignment policy of the publisher, which is discussed later in the paper, this coverage value might be severely under-estimated.

The analysis of raw server data allows Bohannon (2016) to conclude that the service is used not only by researchers in less-developed countries, but also in the developed world, where researchers usually have institutionally-paid access to scientific content. Based on this finding, the author advances another reason for Sci-Hub popularity – simplicity of use when compared to legal alternatives.

This sheds new light on the ongoing discussion about the positive and negative impacts of Sci-Hub on science and publishers' revenue. To clarify the matter, it is useful to make a clear distinction between two types of research. The first part of the scientific research can be commercialised. These are studies that report scientific advances which companies can use to generate revenue streams. As a result, owning (and enforcing) copyright for these studies and charging high fees for accessing the content is justified. Most of this research focuses on natural sciences. The second part of scientific knowledge is not for commercial purposes and becomes the basis for further (public) knowledge generation.[2] These are findings which do not have immediate revenue-generating applications. Such research only creates footing for further advances, yielding higher future research output and possible monetising opportunities. Most research in social sciences belongs to the latter category.

Therefore, I argue that the positive effects of Sci-Hub on research and potential damage inflicted on publishers will strongly depend on whether the research in question concerns natural or social sciences. Social science has a lot to potentially gain from such piracy engines, while publishers in natural science journals have a lot to lose.

Bohannon's (2016) analysis makes no distinction between natural and social sciences. He uses all download requests received by Sci-Hub servers. Given that natural science publications are more numerous, when compared to their social science counterparts (by, perhaps, as much as one order of a magnitude), these findings may be hiding interesting details, when it comes to social science. The analysis of Sci-Hub data by Gershake (2017) further reveals that there is no single social science journal that appears among the top 20 most pirated journals.

Here, I examine Sci-Hub download data in order to get a sense of the scale of piracy in social sciences as an example of economics. Identifying all social science publications is virtually impossible, but problem can be approached by concentrating on one sub-field. I chose economics, due to the clear and long-standing ranking of relevant top scientific journals, which allows us to identify the most pirated content and draw conclusions about the overall extent of piracy. I also analyse the geographical decomposition of download requests in order to shed some light on the convenience hypothesis concerning Sci-Hub usage by economics researchers.

2. Of course, commercially useful knowledge also constitutes such a basis.

2. Data

I use the data comprising all download requests received by Sci-Hub servers between October 2015 and February 2016 (Elbakyan and Bohannon 2016). This entails a total of 22,915,621 download requests. Data have been anonymised in order to protect users' identity. To this end, IP addresses have been aggregated to the nearest city location. Thus, data contain the city and the country from which the download request was received. The data contains the Digital Object Identifier (DOI) of the article requested. There is no other information about the article requested.

Therefore, identifying articles from the economics field is a challenge. Clearly, not all economic articles can be identified. Therefore, I proceeded as follows. The field of economics is dominated by few highly regarded journals. The general consensus is that these top journals contain the most robust and cutting-edge research. Therefore, the quality of these articles is the highest in the entire relevant field. They also represent general interest journals, as opposed to narrow field-specific journals, such as the *Journal of Economic Growth* or the *Journal of Labor Economics*. Therefore, all else being equal, if a researcher wants to download a paper, they are more likely to opt for the piece that has been published in a top journal.

Therefore, I argue that content downloads from top economic articles will fairly approximate downloads received by the field of economics. This is definitely so for top economic content downloads, i.e. top journals pirated, which is very likely emerging from analysing the origin of the download. As a consequence, I concentrated on downloads from the top five economics journals. These journals are *American Economic Review* (AER), *Quarterly Journal of Economics* (QJE), *Journal of Political Economy* (JPE), *Econometrica* (ECTA) and *Review of Economic Studies* (REStud). The publishers of four of these five journals use a journal-specific DOI assignment procedure, which allows us to identify articles belonging to these journals fairly easily. One publisher, The Chicago University Press, which publishes JPE, assigns DOI across all of its journals, in what seems to be a random manner. This complicates the identification of JPE articles. To overcome this, I generated citation reports for all JPE articles available on ISI Web of Science, which collects all articles starting from 1956. These reports include the DOI for each article, which allows us to identify JPE articles in the data.[3]

This clearly reduces the working dataset drastically to 2,147 observations and represents only less than 0.01% of the entire dataset.

3. I am still missing JPE articles prior to 1956. However, our analysis shows that researchers are overwhelmingly interested in recent articles in Economics. This confirms Greshake's observation (2016) about the level of all scientific fields; he finds that Sci-Hub searches are dominated by recent content. Therefore, missing articles published over 60 years ago are not likely to generate a significant number of illegal downloads.

Before carrying out the analysis I removed duplicate downloads from the raw data, something that had not been done by Bohannon (2016), as confirmed by the author in a private e-mail. It should be noted that these are raw server log file data. They contain all page load requests received by Sci-Hub servers. Because Sci-Hub operation directly depends on the operation of the Internet, which is known to be problematic in many under-developed countries, duplicate downloads are likely. When the user refreshes the browser that is still in the process of loading the article, the server registers an additional download request. If I had the original IP data, these kinds of downloads could have been completely screened out. However, given the data anonymisation, I had to work with the download time-download location pair of variables. In order to screen out multiple records for one actual download, I identified groups of downloads for the same paper that occurred from the same city within five minutes from one another. When the most downloaded economics article has only been downloaded 18 times during the five-month period, with three downloads from a small town in Iran within a few seconds from each other, it is clearly suspicious. For each of these identified groups I retained only one download in our final dataset. This eliminated 64 observations and left us with the final dataset of 2,083 downloads for 1,096 distinct papers.

3. Analysis

A number of 2,083 downloads over the span of five months implies an average of about 417 downloads per month for all the content generated by the five economics journals in our sample. This means that economics piracy numbers are not all that impressive. This can be explained by the fact that researchers in economics do not need to pirate (much). A large portion of published economic content is available in pre-print versions on SSRN or exists in the public domain in various working paper formats aggregated by RePEc. However, it might also be that Sci-Hub is not that widespread in the field under investigation.

Table 1 presents the ranking of the most downloaded papers. The most pirated economics article (Helpman *et al.* 2010) has only been downloaded 18 times over a five-month period. It is also noticeable that people pirate recent articles. Four out of nine papers on the list are from 2015 and the oldest paper is from 2004. *Quarterly Journal of Economics* accounts for four papers on the list, *Journal of Political Economy* accounts for three.

Table 2 presents the analysis at the journal level. In order to compare journals properly, I have to acknowledge that journals have generated different sizes of article stock. Obviously, more articles imply more potential downloads. In order to take this into account, I gathered data from ISI Web of Science (WoS) about the total number of articles published by each journal to date. Even though the WoS coverage is not complete, it is rather extensive for all five journals. I used the number of articles on

WoS platform to estimate the total output of each of the journals, assuming that journal output has stayed constant over time. As JSTOR completely covers all five journals and the moving wall is rather short in all cases, I can be certain that one has access to all publications from these five journals on Sci-Hub. The last two columns normalise downloaded data by using information on the journals' total output.

It is apparent from Table 2 that users are not interested in the great majority of articles published by the top five economics journals. This is not surprising, as most scientific articles (even in top journals) do not receive any citations. Even though *American Economic Review*'s piracy numbers are the highest in absolute terms (365 articles downloaded at least once during the period between October 2015 and February 2016), the *Journal of Political Economy* seems to be the most attractive outlet for Sci-Hub users (over 0.4% of the journal's output has been downloaded at least once during the five-month period).

Numbers show that JPE tops the rankings in both relevant measures, namely, the number of downloads per published article and the pirated articles as a share of the journal's total output.

Table 1. Top downloaded economics articles

Authors	Year	Title	Journal	# of downloads
E. Helpman, O. Itskhoki & S. Redding	2010	Inequality and Unemployment in a Global Economy	ECTA	18
M. Gentzkow & J. Shapiro	2011	Ideological Segregation Online and Offline	QJE	17
D. Acemoglu, G. Egorov & K. Sonin	2015	Political Economy in a Changing World	JPE	15
I. Welch	2004	Capital Structure and Stock Returns	JPE	15
K. Manova	2012	Credit Constraints, Heterogeneous Firms, and International Trade	REStud	13
N. Voigtlander & H.-J. Voth	2012	Persecution Perpetuated: The Medieval Origins of Anti-Semitic Violence in Nazi Germany	QJE	12
H. Cronqvist & S. Siegel	2015	The Origins of Savings Behavior	JPE	12
M. Aguiar, M. Amador, E. Farhi & G. Gopinath	2015	Coordination and Crisis in Monetary Unions	QJE	11
A. Akerman, I. Gaarder & M. Mogstad	2015	The Skill Complementarity of Broadband Internet	QJE	11

Table 2. Top downloaded economics journals

Journal	# of downloads	# of articles downloaded	# of downloads / journal's total output (%)	# of downloaded articles / journal's total output (%)
American Economic Review	527	365	0,018	0,012
Journal of Political Economy	463	226	0,838	0,409
Econometrica	450	227	0,770	0,389
Quarterly Journal of Economics	415	154	0,815	0,302
Review of Economic Studies	228	124	0,448	0,244

Table 3 presents the countries where the content has most frequently been downloaded from. As one can see, similar to the aggregate analysis by Bohannon (2016), developed countries like the US, Germany and France, make it into the top 10 countries pirating economic content. Gershake (2016) also reports the positive correlation between a country's GDP level and piracy activity on Sci-Hub.

Tabel 3. Top downloading countries

Country	# of downloads	# of yearly downloads / 1mln inhabitants	# of yearly downloads / # of registered economics institutions
China	266	0,470	2,014
Indonesia	264	2,535	5,510
United States	160	1,204	0,122
Iran	140	4,338	5,695
Russia	131	2,191	0,847
Brazil	83	0,994	0,862
Pakistan	83	1,094	2,075
Malaysia	65	5,249	2,137
France	64	2,326	0,354
Germany	60	1,786	0,201

Therefore, the analysis based on absolute numbers points to the same direction as that indicated by Bohannon (2016) – everyone is downloading pirated papers. However, a more accurate picture has to take into account the size of the research bodies in each of the countries. The best measure for this would be the number of economics researchers in each country. However, such data is not available. We can follow Gershake (2016) and use country population to proxy such a measure. The yearly downloads normalised by the population are presented in Table 3.

We also have to acknowledge that developed countries spend more on education and, therefore, are likely to have more scientists per inhabitant. Therefore, I created another proxy, which is the number of economic institutions registered with the RePEc service. These measures clearly show that downloads from the US, Germany and France are a tiny fraction of their scientific operations. However, downloads from Iran and Indonesia, as well as those from Malaysia, Pakistan and China are one order of magnitude higher.

4. Discussion

All in all, even if there are a few downloads in virtually every country in the world, I see that Sci-Hub is beneficial to, mostly, developing countries, when it comes to economics. This is in some contrast to the overall findings reported by Bohannon

(2016). Downloads in developed countries arguably occur because Sci-Hub is very easy to use, when compared to usual university subscriptions. In order to examine the validity of this claim, I also looked into the download activity generated by the content of the *Journal of Economic Perspectives* (JEP). JEP is an open access journal and, therefore, requires no piracy. Yet, over the five-month period, Sci-Hub users requested its content 177 times, which is comparable to similar statistics from the top five economics journals shown in Table 2. This seems to confirm the hypothesis of convenience usage.

In fact, a quick Google search for the nine most pirated economics articles from Table 1 also points to convenience as the main motivator behind Sci-Hub usage. Google search results, presented in Table 4, reveal that either journal typeset articles or working paper versions are freely available online for all top pirated economics articles.

Table 4. Online accessibility of most pirated economics articles

Article	Availability online
Helpman et al. (2010)	pdf freely available on Stephen Redding's webpage
Gentzkow and Shapiro (2011)	pdf of a version freely available as an NBER working paper
Acemoglu et al. (2015)	pdf freely available on MIT economics department webpage
Welch (2004)	pdf freely available on Ivo Welch's webpage
Manova (2012)	pdf freely available on Kalina Manova's webpage
Voigtlander and Voth (2012)	pdf freely available on Nico Voigtlander's webpage
Cronqvist and Siegel (2015)	pdf of a working paper version freely available on SSRN
Aguiar et al. (2015)	pdf of a working paper version freely available on Minneapolis FED website
Akerman et al. (2015)	pdf of a working paper version freely available on IZA website

Ultimately, the overall impact of Sci-Hub on economics can be evaluated as positive. Researchers in under-developed parts of the world get access to important content. At the same time, there is no indication that publishers are not losing (much) revenues. Firstly, elimination of Sci-Hub would hardly result in any subscriptions from underdeveloped country university libraries. Secondly, the extent of downloading is very low, perhaps due to a large number of popular working paper distribution services. Economics is not the only sub-discipline in which advantages of Sci-Hub hugely exceed its costs. Similar findings were reported by Timus and Babutsidze (2016) with respect to European Studies. One could argue that this is a general pattern for social sciences.

Yet, Sci-Hub does not discriminate between social and natural sciences and weighing its costs and benefits should take into account natural sciences. In this respect, it is important to be precise about what sort of service Sci-Hub provides to its users. It allows them to view and download the article, but the right for any legal use of the content remains with the publisher (Priego 2016). Therefore, Sci-Hub cannot inflict any losses on publishers other than un-sold journal subscriptions. As a result, one may argue that Sci-Hub is beneficial to scientific journal publishers (not only authors) by popularising their content and creating an additional dissemination channel (Priego 2016), much like Google's book previews or journals' free access issues.

References

Antelman K. (2004) Do open-access articles have a greater research impact? *College and Research Libraries* 65(5): 372-382.

Bohannon J. (2016) Who's downloading pirated papers? Everyone. *Science* 352(6285): 508-512. http://dx.doi.org/10.1126/science.352.6285.508

Brembs B. (2016) Sci-Hub as necessary, effective civil disobedience. Bjorn Brembs' Blog. Accessed on 18 July 2016. Available at: http://bjoern.brembs.net/2016/02/sci-hub-as-necessary-effective-civil-disobedience/

Caffrey Gardner C., Gardner G. (2016) Fast and Furious (at Publishers): The Motivations behind Crowdsourced Research Sharing. *College and Research Libraries*. Forthcoming.

David P. (1998) Common agency contracting and the emergence of 'open science' institutions. *American Economic Review* 88(2): 15-21

Davis P., Lewenstein B., Simon D., Booth J., Connolly M. (2008) Open access publishing, article downloads, and citations: randomised controlled trial. *British Medical Journal*. http://dx.doi.org/10.1136/bmj.a568

Elbakyan A. and Bohannon J. (2016) Data from: Who's downloading pirated papers? Everyone. *Dryad Digital Repository*. http://dx.doi.org/10.5061/dryad.q447c

European Commission (2016) Guidelines on open access to scientific publications and research data in Horizon 2020. Accessed 18 July 2016. Available on: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Eysenbach G (2006) The Open Access Advantage. *Journal of Medical Internet Research*. http://dx.doi.org/10.2196/jmir.8.2.e8

Gaule P., Maystre N. (2011) Getting cited: Does open access help? *Research Policy* 40(10):1332–1338

Greshake B. (2016) Correlating the Sci-Hub data with World Bank Indicators and Identifying Academic Use. *The Winnower* 4:e146485.57797

Greshake B. (2017) Looking into Pandora's Box: The Content of Sci-Hub and its Usage. *F1000 Research* 6:541.

Himmelstein DS, Romero AR, McLaughlin SR, Greshake Tzovaras B, Greene CS. (2017) Sci-Hub provides access to nearly all scholarly literature. *PeerJ Preprints* 5:e3100v2 https://doi.org/10.7287/peerj.preprints.3100v2

National Institutes of Health (2009) NIH public access policy. Accessed 18 July 2016. Available on: https://publicaccess.nih.gov/policy.htm

Laakso M., Welling P., Bukvova H., Nyman L., Björk B/-C. and Hedlund T. (2011) The Development of Open Access Journal Publishing from 1993 to 2009. *PLoS ONE* 6(6): e20961. doi:10.1371/journal.pone.0020961

Nature (2015) Open data and reproducible research. Availably on https://www.nature.com/open-research/about-open-access/open-data/

Oxenham S. (2016) Meet the Robin Hood of Science. *Big Think*. Accessed on July 18. Available at: http://bigthink.com/neurobonkers/a-pirate-bay-for-science

Partha D. and David P. (1994) Toward a new economics of science. *Research Policy* 23(5): 487-521.

Perkmann M., Schildt H. (2015) Open data partnerships between firms and universities: The role of boundary organizations. *Research Policy* 44(5):1133-1143.

Priego E. (2016) Signal, Not Solution: Notes on Why Sci-Hub Is Not Opening Access. *The Winnower*. http://dx.doi.org/10.15200/winn.145624.49417

Silva, L. (2014) PLOS' New Data Policy: Public Access to Data. PLOS ONE community blog. Available on http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/

Singh Chawla D. (2017) Publishers take ResearchGate to court, alleging massive copyright infringement. *Science*. http://dx.doi.org/10.1126/science.aaq1560

Tennant J., Waldner F., Jacques D., Masuzzo P., Collister L., Hartgerink C. (2016) The academic, economic and societal impacts of Open Access: an evidence-based review. *F1000 Research* 5:632. http://dx.doi.org/10.12688/f1000research.8460.1

Timus N., Babutsidze. Z. (2016) Pirating European Studies. *Journal of Contemporary European Research* 12(3): 783-791.

Vogel G., Kupferschmidt K. (2017) A bold open-access push in Germany could change the future of academic publishing. *Science*. http://dx.doi.org/10.1126/science.aap7562

Wainwright T., Huber F., Rentocchini F. (2016) Open Innovation: revealing and engagement in Open Data organisations. Presented at the Governance of Complex World conference, Valencia, Spain.