Pirated Economics

Babutsidze, Zakaria

SKEMA Business School, OFCE Sciences Po

2 June 2016

Pirated Economics

Zakaria Babutsidze

SKEMA Business School &
OFCE Sciences Po

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 to journal publisher revenues. To validate the claim, I examine the data from illegal downloads of economics content from Sci-Hub over five-month period. I conclude that: (a) the extent of piracy in economics is not pervasive; (b) as downloads are coming mostly from under-developed countries; (c) users pirate even the content freely available online. As a result, publishers are not losing much revenues, while the exposure to generated knowledge is being extended.

JEL Code: A1

1. Introduction

The idea of open science has challenged many stake-holders in science and publishing for years. Many have argued that pricing practices by mainstream scientific journal publishers have built walls around the knowledge precluding a large part of researchers and public from accessing public good. Some have even compared this "paywall" to the wall dividing east and west Berlin during the cold war (Oxenham 2016).

This has become particularly problematic when it comes to the knowledge generated by publicly funded research. Some reckon that eliminating scientific journal publishing from the knowledge creation process will save $9.8bln of public money annually (Brembs 2016). Many years of contemplation by public funding bodies have resulted in clear actions in terms of institutionalizing 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).

The main argument made for open access science is the fact that scientific journal publishers turn high profit margins. However, the problem is somewhat more complex and involves understanding the incentives of various stake holders in the knowledge creation process. The discussions around the "new economics of science" advanced in two decades ago demonstrate subtleties of the problem (Partha and David 1994; David 1998).

Notwithstanding, the raise of the "open science" is a fact. This move can be illustrated by three distinct developments. The first is the emergence of open access journals. A good example of this development is *PLoS* suit of journals. In similar vein, many non-open access

journals have also joined the initiative to provide authors with the option to make the published article open access (for a fee).

It is believed that open access to the publication increases the impact of the research. As a result, the number of articles published under open access has skyrocketed over last two decades (Laasko et al. 2011). However, the evidence of greater impact of open access research is not clear-cut. While some researchers find a positive impact of open access on citation count (Antelman 2004, Eysenbach 2006), other researchers find no evidence of open access advantage (Davis et al. 2008, Gaule and Maystre 2011). Open access publications do seem to have a clear-cut advantage in terms of non-academic dissemination, however (Tennant et al. 2016).

The second development along the lines of open science development is the push by journals for openly sharing the data involved in scientific publications. This has become an all-encompassing trend covering journals from open to 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 following in terms of research and innovation. However, significant challenges facing main actors have been identified in this direction too (Perkmann and Schildt 2015, Wainwright at al. 2016).

The third, perhaps the most controversial and radical development has been the development of channels to circumvent the paywalls which usually involve a violation of copyright laws. These range from crowdsourced research sharing (e.g. using a hashtag #icanhazpdf to ask other researchers to download and send an article to which and individual does not have an access) (Caffrey Gardner and Gardner 2016), all the way to the creation of digital piracy engines that provide free access to scientific content illegally.

The most famous of these sort of services is Sci-Hub. Sci-Hub was created in 2011 and by now amounts to tens of thousands of illegal downloads per day. Among the 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.

Up to very recently not much has been known about the size and geographical breakdown of the 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 Sci-Hub service has emerged (Elbakyan and Bohannon 2016).

The analysis of the 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-payed access to scientific content. Based on this finding, the author advances another reason of Sci-Hub popularity – simplicity of use compared to the legal alternatives.

This sheds a new light on the ongoing discussion about the positive and negative impacts of Sci-Hub on science and publishers. To clarify the matter, it is useful make a clear distinction between two types of research. The first part of the scientific research can be commercialized. Most of this research is concentrated in the fields of real sciences. The

second part of the scientific knowledge is not commercializable and represents the bases for the further (public) knowledge generation. This constitutes the most of the research in social sciences.

Therefore, I argue that positive effects of Sci-Hub on research and potential damages inflicted on publishers will strongly depend on whether we are considering research in real or social sciences. Social science has potentially lots to gain from such piracy engines, while publishers in real science journals will have lots to loose.

Bohannon (2016) analysis does not distinguish between real and social sciences. It uses all download requests received by Sci-Hub servers. Given that real science publications are more numerous compared to their social science counterparts (by perhaps as much as an order of a magnitude), these findings risk to be hiding interesting details when it comes to social science.

In this note we examine the Sci-Hub downloads data in order to get a sense of the size of piracy in social sciences on the example of economics. Identifying all social science publications is virtually impossible, while we can approach the problem by concentrating on one sub-field. We choose economics, as it has clear and stable ranking of top scientific journals which allows us to identify the most pirated content and make conclusions about the overall extent of piracy. We also analyze the geographical decomposition of the download requests in order to shed some light on convenience hypothesis in Sci-Hub usage by the economics researchers.

2. Data

We use the data comprising all download requests received by the Sci-Hub servers between October 2015 and February 2016 (Elbakyan and Bohannon 2016). This represents a total of 22 915 621 download requests. The data has been anonymized in order to protect the identity of the user. For this the IP addresses have been aggregated to the nearest city location. Thus the data contains the city and the country from where the download request was received. The data contains the Digital Object Identifier (DOI) of the article requested. There is no other information about the requested article.

Therefore, identifying the articles from economics field represents a challenge. Clearly, all economics articles cannot be identified. Therefore, we proceed as follows. The economics field is dominated by few highly regarded journals. The general consensus is that these top journals aggregate the most robust and cutting-edge research. Therefore, the quality of these articles is the highest in all of the discipline. They also represent general interest journals as opposed to the narrow field-specific journals like the Journal of Economic Growth or the Journal of Labor Economics. Therefore, all else equal, if a researcher wants to download a paper, he/she is more likely to opt for the piece that has been published in the top journal.

Therefore, I argue that the downloads of the content from the top economics articles will fairly approximate the downloads received by the economics field. Definitely so for the top

economics downloads or top journals pirated. Very likely so when it comes from the analysis of the origin of the download. As a consequence, we concentrate on the downloads of 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). Publishers of four of these five journals use a journal-specific DOI assignment procedure, that allows us to identify the articles belonging to these journals fairly easily. One publisher, The Chicago University Press, that publishes JPE assigns DOI across all of its journals seemingly randomly. This complicates the identification of JPE articles. To overcome this, we generate citation reports to all JPE articles available on ISI Web of Science. This collects all articles starting from 1956. These reports include the DOI for each article which allows us to identify JPE articles in the data.[1]

This, clearly reduces the working dataset drastically to 2147 observations. This represents only less than 0.01% of the whole dataset.

Before carrying out the analysis we remove duplicate downloads from the raw data that has not been done by Bohannon (2016), as confirmed by the author in a private e-mail. Notice that this is a raw server log file data. It contains all page load requests received by Sci-Hub servers. Because Sci-Hub's functioning depends directly on the functioning of the Internet, which is known to be problematic in many under-developed countries, there is a potential for duplicate downloads. When the user refreshes the browser that is in the process of loading the article, the server registers an additional download request. If we had the original IP data, these kinds of downloads could have been perfectly screened out. However, given the anonymized data we have to work with download time - download location pair of variables. In order to screen out multiple records for one actual download, we identify groups of downloads for the same paper that occur 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-months period, receiving three downloads from a small town in Iran within few seconds from one another is clearly suspicious. For each of these identified groups we retain only one download in our final dataset. This eliminates 64 observations and leaves us with the final dataset of 2083 downloads for 1096 distinct papers.

3. Analysis

2083 downloads over the span of five months implies about 417 downloads on average per month for all the content generated by the five economics journals in our sample. This means that economics piracy numbers are not that impressive. This can be explained by the fact that researchers in economics do not need to pirate (much). Large portion of published economics content is available in pre-print versions on SSRN or exists in public domain in

[1] We are still missing the JPE articles prior to 1956. However, our analysis shows that researchers are overwhelmingly interested in recent articles. Therefore, missing articles published over 60 years ago are not likely to generate a significant number of illegal downloads.

various working paper formats that get aggregated by RePEc. However, it might also be that Sci-Hub is not that widespread in the discipline.

Table 1 presents the ranking of the most downloaded papers. The most pirated economics article (Helpman et al. 2010) has collected only 18 downloads over five-months 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 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 presents the analysis on the journal level. In order to compare journals properly we have to acknowledge that journals have generated different size of article stock. Obviously, more articles imply more potential downloads. In order to take this into account we gather the data from ISI Web of Science (WoS) about the total number of articles published by each journal as of today. Even though the WoS coverage is not complete, it is rather extensive for all five journals. We use the number of articles on WoS platform to estimate the total output of each of the journals, by assuming that journal output has stayed constant over time. As JSTOR covers completely all five of the journals and the moving wall is rather short in all cases, we can be sure that one has access to all publications from these five journals on Sci-Hub. The last two columns normalize download data by using the information about journals' total output.

It is apparent from table two that users are not interested in great majority of the articles published by top five economics journals. This is not surprising as most of scientific articles (even if 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 journals output has been downloaded at least once during the five-month period).

The numbers show that JPE tops the rankings in both relative measures: the number of downloads per published article and the pirated articles as the share of the journal's total output.

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 been most frequently downloaded. As one can see, similar to the aggregate analysis by Bohannon (2016), the developed countries like the US, Germany and France make into top 10 countries pirating economics content.

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 Bohannon (2016) – everyone is downloading the 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 use country population to proxy the measure. The yearly downloads normalized by the population are presented in table.

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

4. Discussion

All in all, even if there are few downloads coming from virtually every country in the world, we see that Sci-Hub does benefit mostly developing countries when it comes to economics. This is in some contrast by the general findings reported by Bohannon (2016). Downloads coming from developing countries are arguably for the reason that Sci-Hub is very easy to use compared to the usual university subscriptions. In order to examine the validity of this claim I have also looked at the downloads 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 have requested its content 177 times, which is comparable to the similar statistic from the top five economics journals from table 2. This looks to confirm the hypothesis of the convenience usage.

In fact, a quick Google search for nine most pirated economics articles from table 1 also points to convenience as being 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, overall impact of Sci-Hub on economics (including publishing) can be evaluated as being positive. Researchers in under-developed parts of the world are getting 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 downloads is very low, perhaps due to a large number of popular working paper distribution services. The economics is not the only sub-discipline where advantages of Sci-Hub hugely exceed its costs. Similar findings 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 across social and real sciences and weighting of its costs and benefits should take into account the real sciences. In this respect important to be precise about what sort of service Sci-Hub provides to its users. It allows to view and download the article, but the right for any legal use of the content remain with the publisher

(Priego 2016). Therefore, Sci-Hub cannot inflict any losses on publishers other than un-sold journal subscriptions. As a result, one can argue that Sci-Hub scientific journal publishers (not only authors) by popularizing their content and generating an additional channel for dissemination (Priego 2016), much like Google's book previews or journal's 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

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

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

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. *F1000Research* 5:632. http://dx.doi.org/10.12688/f1000research.8460.1

Timus N., Babutsidze. Z. (2016) Pirating European Studies. *Journal of Contemporary European Research*. Forthcoming.

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.