DE GRUYTER OPEN
Acta Univ. Sapientiae, Informatica **9,** 1 (2017) 83–96

DOI: 10.1515/ausi-2017-0006

Analysis of Sci-Hub downloads of computer science papers

Darko ANDROČEC
Faculty of Organization and Informatics, University of Zagreb
Pavlinska 2, 42000 Varaždin, Croatia
email: dandrocec@foi.hr

Abstract. The scientific knowledge is disseminated by research papers. Most of the research literature is copyrighted by publishers and available only through paywalls. Recently, some websites offer most of the recent content for free. One of them is the controversial website Sci-Hub that enables access to more than 47 million pirated research papers. In April 2016, Science Magazine published an article on Sci-Hub activity over the period of six months and publicly released the Sci-Hub's server log data. The mentioned paper aggregates the view that relies on all downloads and for all fields of study, but these findings might be hiding interesting patterns within computer science. The mentioned Sci-Hub log data was used in this paper to analyse downloads of computer science papers based on DBLP's list of computer science publications. The top downloads of computer science papers were analysed, together with the geographical location of Sci-Hub users, the most downloaded publishers, types of papers downloaded, and downloads of computer science papers per publication year. The results of this research can be used to improve legal access to the most relevant scientific repositories or journals for the computer science field.

Computing Classification System 1998: K.7.4, K.3.2
Mathematics Subject Classification 2010: 62P30
Key words and phrases: computer science ethics, journal paywalls, pirated papers, scientific publishing

1 Introduction

Access to the research literature is essential to the successful work of researchers and the education of the general public [1]. Scientists and the general public rely on a wide range of channels to access the research literature [2]: printed issues of journals, interlibrary loans, publishers' online platforms such as Elsevier ScienceDirect, preprint repositories such as ArXiv, institutional and authors' web pages. Many now propose free access to all scientific papers to everybody, including the European Union where consensus of all members' ministers of science, innovation, trade and industry was made in May 2016 that all scientific papers founded by the EU should be freely available by 2020 [3]. People failing to retrieve articles through these channels use article requests from authors by e-mail or using social media such as Academia.edu, Mendeley and ResearchGate. Recently, some online repositories offer most of the scientific papers for free. One of them is Sci-Hub. Sci-Hub website currently hosts more than 47 million pirated research papers and has millions of visitors per month. Online piracy represents a copyright infringement whereby copyright material (scientific articles in this case) is reproduced or distributed without appropriate permission [4]. The Sci-Hub founder Alexandra Elbakyan from Kazakhstan claims that the main aim of the Sci-Hub project is to circumvent the copyright restrictions to speed up the development of science, especially in developing countries where researchers do not have institutional access to publishers' paywalls. Of course, publishers have a different opinion, e.g. Elsevier sued Sci-Hub and its founder, and frequently requested its web domains to be put down. Publishers claim that journals have costs, even if they do not pay researchers - authors and reviewers, e.g. editors (and sometimes copy editors, proofreaders, illustrators) are mostly paid professionals, digital publishing is nowadays expensive, and article usage information is lost when using Sci-Hub and/or similar websites [5]. Sci-Hub website is controversial, but many researchers from all parts of the world (even from countries and institutions that have legal access to research papers) are using its services [6]. "Over the 6 months leading up to March, Sci-Hub served up 28 million documents" [6]. The six-month log data (September 2015-February 2016) from Sci-Hub website are now publicly available at [7].

Bohannon [6] presented an aggregate view that relies on all downloads and for all fields of study, but his findings might be hiding interesting patterns within computer science, and this is still gap in the current literature. In this paper, the mentioned data was analysed in more detail and with computer science papers in the focus. The main research questions are: Who are Sci-

Hub users that download computer science research papers and where are they from? What computer science research papers did the mentioned users download? The results of this research can be used to identify the most relevant scientific repositories, journals, and conference proceedings for the computer science field, and which of the sources is more inaccessible to researchers.

This paper proceeds as follows. First, in Section 2, the related work is listed. The next section presents the used research methodology. The results are shown in Section 4. The conclusions are provided in the final section.

2 Related work

Bohannon [6] presented the statistics of the Sci-Hub usage based on extensive server log data [7] supplied by the Sci-Hub creator, Alexandra Elbakyan. To protect the privacy of Sci-Hub users, their geographical locations were aggregated to the nearest town using Google Maps, so IP addresses are not contained in the publicly released data set [7]. Bohannon's first conclusion is that Sci-Hub users are not limited to the developing world, e.g. a quarter of the Sci-Hub requests came from OECD members that should have the best journal access.

Parkhill [8] loaded the top 100 downloads from the Sci-Hub data set [7] into tool PlumX. Most of the downloaded papers are from 2015, so Sci-Hub users were downloading the most recently published papers. Physical sciences and engineering, together with life sciences, garnered most of the downloads. Babutsidze [9] examined the data from illegal downloads of economic content from Sci-Hub, based on the log data from [7]. He concentrated on downloads of the top five economics journals: American Economic Review, Quarterly Journal of Economics, Journal of Political Economy, Econometrica and Review of Economic Studies. Babutsidze [9] concluded that there is a very small number of downloads of economics papers from Sci-Hub, most of the downloads are from under-developed countries, and open access economics papers were downloaded from Sci-Hub because of convenience.

Cabanac [1] reveals that 36% of all digital object identifier (DOI) articles are available for free at Library Genesis platform (LibGen). For the three major publishers (Elsevier, Springer and Wiley) the percentage is even higher (68%). As of January 2014, LibGen hosted and distributed 25 million digital documents, mainly for educational purposes. Users crowdsource articles to LibGen directly or indirectly through services such as Reddit Scholar and Sci-Hub. Cabanac [1] also claims that people crowdsource articles through various

other channels such as Reddit, Sci-Hub, #icanhazpdf hashtag on Twitter, personal websites and text-sharing platforms.

Swab and Romme [10] analysed the use of #icanhazpdf hashtag as a means of obtaining health science literature. They have used RowFeeder software to monitor #icanhazpdf requests between 1 February and 30 April 2015, and they concluded that there were 302 requests for health sciences literature in this period. This number accounts for a relatively small proportion of paper sharing compared with other online platforms.

Gardner and Gardner [11] surveyed users of Twitter, Reddit Scholar and Facebook about crowdsourced research, their demographic information, frequency of use and motivations. They concluded that primary platforms used to organize crowdsourcing of research articles are Twitter, Facebook and Reddit, and the primary websites to host the content were AvaxHome, LibGen and Sci-Hub. Elsevier, Springer and Wiley account for 83% of all LibGen's content. The majority of respondents claim to obtain articles for utilitarian reasons, and crowdsourcing is their preferred alternative to using interlibrary loans.

Timus and Bautsidze [4] examined the Sci-Hub downloads data to uncover patterns in piracy in the European Studies research. They relied on the information provided by the University Association for Contemporary European Studies (UACES) that provides the list of European Studies journals. For their analysis, they have chosen the journals with ISI impact factor greater than 1. Their analysis reveals that the readers are mostly interested in subjects reflecting the current European challenges such as populism, extremism and the economic crisis.

3 Research methodology

The publicly available Sci-Hub's server log data available at [7] was downloaded and imported into MySQL database system. The mentioned data set was already anonymised (there are no IP addresses, IP addresses were aggregated to the nearest city location). For the efficiency reasons, the data was put into six tables (scihub_data1 - scihub_data6), one table for each monthly server log data. The data consists of date and time, digital object identifier (DOI), country, town and geographical position. DOIs are not tagged by subject or keyword so it is not possible to return a set of DOIs for the computer science field. Therefore, identifying the articles from computer science field represents a challenge. For this reason, all the DOIs from DBLP were extracted in XML

format on May 23, 2016, and imported into another MySQL table. DBLP service provides open bibliographic information on major computer science journals and proceedings. "The DBLP Computer Science Bibliography of the University of Trier has grown from a very specialized small collection of bibliographic information to a major part of the infrastructure used by thousands of computer scientists" [12]. As of May 2016, DBLP indexes over 3.3 million publications, published by more than 1.7 million authors. To this end, DBLP indexes about 32,000 journal volumes, more than 31,000 conference or workshop proceedings and more than 23,000 monographs.

The DBLP data was downloaded in the XML format. The Java classes were developed to parse the DBLP's XML file and extract the DOIs data into CSV file that was used to import the DBLP's DOIs data into MySQL database. The list of all DOIs from DBLP computer science bibliography was saved into a new table (dblp_dois), and it is assumed that all the main computer science works are included. Of course, some works may not be in this catalogue, which is the limitation of this study, but the situation in other science fields is even worse, e.g. comprehensive meta-indices are missing for most areas of science [12]. The DOIs fields in both tables were defined as database indexes, to enable better query performance. Next, the SQL queries were used to find an intersection of two data sets (Sci-Hub log data and DBLP's DOIs).

First, the views containing the intersection of each of the six scihub_data tables with dblp_dois were created.These tables were very big, and it was impractical (time-consuming) to analyse the data, because SQL queries on the views in some cases run for several hours. Table 1 shows the number of rows in each of the six tables containing Sci-Hub data, and the number of rows of views showing the intersection with DBLP data. Only 5.95% of the data downloaded from Sci-Hub were computer science papers. For this reason, separate tables for each of the six views were created.

The creation of each table took several hours, but after that the complete intersection data was in tables, indexes were put on relevant fields (DOI and country), and SQL queries needed to analyse the results took reasonable execution time (a few minutes). This data (Sci-Hub downloads of computer science papers) is publicly available at https://github.com/dandrocec/in the SciHub-ComputerScience repository in the form of SQL scripts.

Table	Total	in DBLP	%
scihub_data1	3 759 219	217 689	5.79
scihub_data2	6,017,112	407,387	6.77
scihub_data3	4,774,085	279,952	5.86
scihub_data4	1,837,701	97,042	5.28
scihub_data5	5,876,395	325,795	5.54
scihub_data6	4,752,852	280,770	5.91
Sum	27,017,364	1,608,635	5.95

Table 1: Intersection of Sci-Hub and DBLP data

4 Results

The analysis was done on the six tables (data1-data6) that represent the intersection of Sci-Hub log data and DBLP DOIs data, i.e. the computer science papers downloaded at Sci-Hub web page in a six-month time frame. One of the research questions of this paper is: What computer science research papers did the mentioned users download? First, the ranking of the most downloaded computer science papers from Sci-Hub had to be obtained. For the purpose of performing queries on all data, one view with all the data was created Then, the most downloaded computer science papers were obtained by using the SQL query.

In total, 607,023 computer science papers were downloaded from Sci-Hub website in a six-month period. Table 2 shows twenty most downloaded computer science papers from Sci-Hub. Fourteen of the mentioned papers are journal papers, three are conference papers, two are chapters in scientific books, and one is a technical report. Eleven papers are published in IEEE's publications, three in Springer's publications, two in Elsevier's, and one paper in MIT's, Wiley's, SIAM's and ArXiv's publications, respectively. Most of the downloaded papers are new, the only exception being "How to Construct Pseudorandom Permutations from Pseudorandom Functions" from 1988. The most downloaded papers are aligned with currently popular themes in computer science, e.g. titles of five papers contain the phrase "Internet of things". Some of the papers are from open-access journals, but readers still decide to use Sci-Hub instead of journals' official web pages, so paid access is not the only problem. The reason might be that Sci-Hub users do not bother with which of the articles is open-access, they just use Sci-Hub website to retrieve all the necessary articles from one place.

Next, the authors of the top 20 most downloaded papers were extracted to analyse how many total downloaded articles these authors have. Data contains

DOI	Title	Downloads
10.1109/ISPASS .2015.7095803	Nyami: a synthesizable GPU architectural model for general-purpose and graphics-specific workloads	1,118
10.1007/s11948-014-9521-4	Penetrating the Omerta of Predatory Publishing: The Romanian Connection	746
10.1109/TKDE .2013.109	Data mining with big data	725
10.1007/978-3-319-28658-7_55	Detection of Copy-Move Forgery in Images Using Segmentation and SURF	656
10.1162/jocn_a _00880	The Role of Dopamine in Temporal Uncertainty	457
10.1109/ACCESS. 2014.2362522	Information Security in Big Data: Privacy and Data Mining	451
10.1016/j.comnet .2010.05.010	The Internet of Things: A survey	343
10.1109/CTS .2014.6867550	Defining architecture components of the Big Data Ecosystem	307
10.1109/TMI .2013.2265603	Deformable medical image registration: a survey	307
10.1016/j.neunet .2014.09.003	Deep Learning in Neural Networks: An Overview	305
10.1137/0217022	How to Construct Pseudorandom Permutations from Pseudorandom Functions	301
10.1109/JIOT .2014.2306328	Internet of Things for Smart Cities	291
10.1007/978-3-642-55032-4_6	Do Personality Traits Work as Moderator on the Intention to Purchase Mobile Applications Work? - A Pilot Study	288
10.1109/COMST .2015.2444095	Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications	281
10.1109/MobileCloud .2015.40	Cloud Computing for Emerging Mobile Cloud Apps	281
10.1002/asi.23445	Bibliogifts in LibGen? A study of a text-sharing platform driven by biblioleaks and crowdsourcing	271
10.1016/j.future .2013.01.010	Internet of Things (IoT): A vision, architectural elements, and future directions	259
10.1109/TIE .2006.878356	Power-Electronic Systems for the Grid Integration of Renewable Energy Sources: A Survey	227
10.1109/JPROC .2014.2371999	Software-Defined Networking: A Comprehensive Survey	224
10.1109/JIOT .2014.2312291	Research Directions for the Internet of Things	214

Table 2: Twenty most downloaded computer science papers

only DOIs, so the Java program was created to parse the DBLP text file to extract all DOIs of the mentioned authors and to create SQL query to check how many times these articles where downloaded from SciHub during time period of publicly available SciHub's log data. The analysis have shown that

other articles of the same authors were not downloaded in great numbers. So, the SciHub users mostly search for a particular article, not all papers of the particular author. The most downloaded are articles from the following authors: Xindong Wu (1280 downloads), Timothy N. Miller (1124), Jeff Bush (1118), Philip Dexter (1118), Aaron Carpenter (1118), Xingquan Zhu (944), Wei Ding (865), Rajkumar Buyya (795), Gong-Qing Wu (776), Dragan Djuric (760), V. T. Manu (672), B. M. Mehtre (656), Lei Xu (519), Jian Wang (513), etc. Most of the authors have small number of papers, and one paper (the one from the list of the 20 most downloaded articles) has more the 90% of authors' total SciHub's downloads. If we look at Google Scholar's citations (on 14th February 2017) of the mentioned authors there is also no correlation: some authors are often cited (e.g. Rajkumar Buyya - 56076 citations and Xindong Wu - 15611 citations), and some have low number of citations (e.g. V. T. Manu - 11 citations). Of course, some authors are junior researchers with small number of recent publications, and the high number of citations of their works cannot be expected yet.

Next analysis, the list of twenty top countries per Sci-Hub downloads of computer science papers is presented in Table 3. The top five countries with the most downloads were India, Iran, China, the United States and Indonesia. Most of the countries on this list are developing countries, but the OECD countries are also there.

To get more insight, the countries' download data was normalized by downloads per 100,000 population. The population was obtained from Wikipedia data (last estimated value) on 9 September 2016. The results are shown in Table 4. In this case, the top five countries include Tunisia, Iran, Greece, Morocco and Jordan. The only European country in top five is Greece, a country that has been in serious financial crisis for many years now. The reason why Sci-Hub is so popular in Greece could be that the access to scientific database is worse than it was before the country started to experience serious financial crisis. Also, some central European countries such as Serbia, Croatia and Slovenia are high on the list. Mapping IP addresses to real-world locations can paint a false picture if people hide behind web proxies or anonymous routing services. But according to Elbakyan, fewer than 3% of Sci-Hub users are using those [6].

Next, the detailed information of computer science papers was retrieved by their DOIs. For this purpose, a Java class was developed which invokes CrossRef DOI REST application programming interface. The CrossRef REST call returns JSON information about a specific resource (paper identified by its DOI), e.g. for the identified most downloaded paper (Nyami: a synthesizable

Country	Downloads
India	424,652
Iran	231,691
China	94,422
The United States	67,336
Indonesia	56,751
Egypt	52,264
Morocco	51,326
Russia	46,659
Pakistan	45,859
Germany	43,332
Tunisia	33,917
Vietnam	26,913
Brazil	26,255
France	25,640
Malaysia	21,596
Greece	19,783
Algeria	17,405
Canada	13,677
Jordan	11,553
The Netherlands	10,513

Table 3: Top downloads per country

GPU architectural model for general-purpose and graphics-specific workloads). In total, 607,023 computer science papers were downloaded from Sci-Hub website in a six-month period, so the invocation of the CrossRef REST web service took some time. The list of DOIs was split into six partitions, and the Java class was executed during several nights to retrieve and parse all the data. The JSON response was parsed with the aim to receive publication date, publisher and paper type information. The results were stored in CSV files and imported into single Microsoft Excel file for further analysis. First, the analysis by publishers of computer science papers downloaded at Sci-Hub site was performed. The results are shown in Figure 1. The most downloaded publishers were Institute of Electrical and Electronics Engineers (IEEE) with 169,313 papers, Elsevier BV with 132,098, and Springer Nature with 109,586 papers. These results can be compared with Bohannon's general (all scientific fields) analysis [6] where he concluded that three most downloaded publishers were Elsevier, Springer and IEEE. The fact that IEEE is the most downloaded publisher in computer science field is expected, because the computer science and electrical engineering is its main focus.

The next analysis involves downloaded papers by years. The results in Figure

Country	Downloads	Population	Normalized
Tunisia	33,917	10,982,754	308.8205381
Iran	231,691	79,200,000	292.5391414
Greece	19,783	10,955,000	180.5842081
Morocco	51,326	33,848,242	151.6356448
Jordan	11,553	9,710,752	118.9712187
Switzerland	7,968	8,341,000	95.52811413
Portugal	9,136	10,341,330	88.34453595
Serbia	4,889	7,041,599	69.43025299
Malaysia	21,596	31,192,000	69.23570146
Croatia	2,672	4,190,700	63.76023099
Singapore	3,513	5,535,000	63.46883469
The Netherlands	10,513	17,000,059	61.84096185
Lithuania	1,765	2,866,935	61.56400476
Latvia	1,177	1,973,700	59.63418959
Ireland	3,676	6,378,000	57.63562245
Egypt	52,264	91,688,000	57.00200681
Hong Kong	3,907	7,234,800	54.00287499
Germany	43,332	82,175,700	52.7309168
Slovenia	1,022	2,063,077	49.53765662
Algeria	17,405	40,400,000	43.08168317

Table 4: Countries' downloads normalized per population

2 show that the most recently published papers were the most popular to download from Sci-Hub site. Most papers were downloaded in 2015, which is not surprising since the available Sci-Hub dataset spans downloads from September 2015 till February 2016.

Finally, the types of downloaded papers were analysed (Figure 3). Journal papers were dominant (58% of computer science papers downloaded from Sci-Hub website were journal papers). The reason for this can be seen in limited accessibility of researchers to chosen journals and perceived quality of papers published in established journals. The next types of papers are conference proceeding articles (31%) and book chapters (10%).

5 Conclusion

Many scientists, especially from developing countries, cannot access the research papers they need. Researchers who fail to retrieve articles through publisher's paywalls, use other, possibly illegal methods to get the needed materials. One of them is Sci-Hub website, hosting more than 47 million pirated research papers. In this work, publicly available data of a six-month

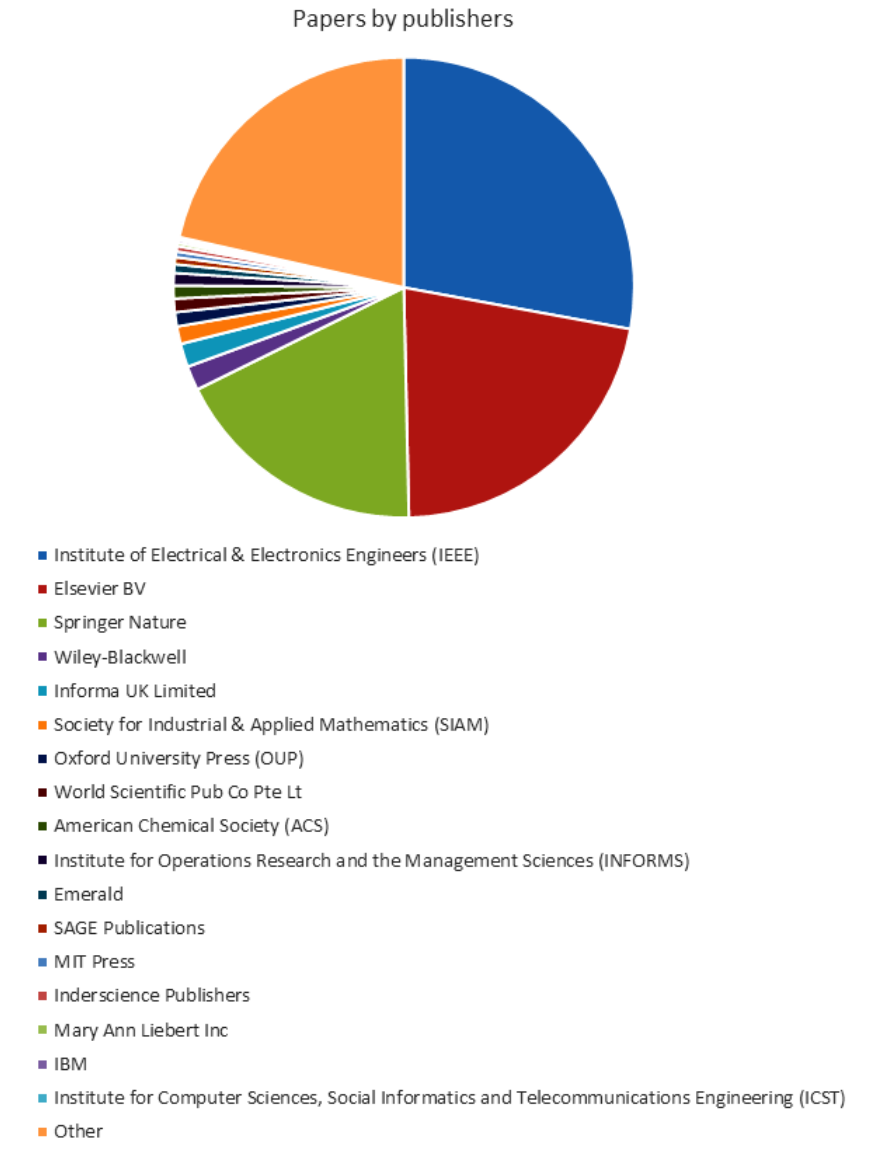

Figure 1: Downloaded papers by publishers

server log data was used to analyse the most downloaded computer science papers in this period with the aim to identify the most important repositories and journals for computer science community. To achieve this task, the

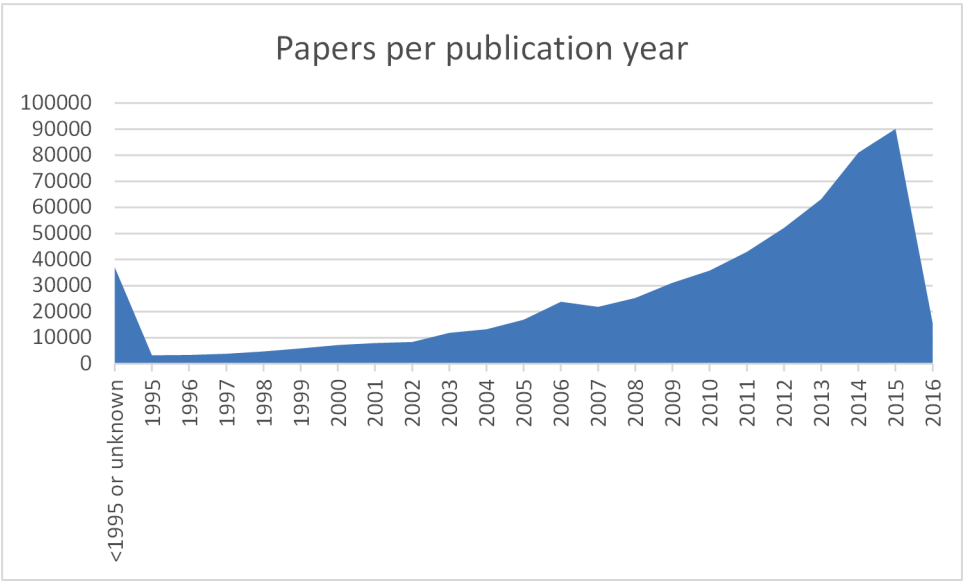

Figure 2: Downloaded computer science papers per publication year

DOIs from the available Sci-Hub log server data were intersected with the DOIs from DBLP (computer science bibliography of the University of Trier), to analyse only log entries of computer science papers. It was assumed that all main computer science works are included in DBLP. The limitation of this study is that some works may not be in DBLP catalogue, but the situation is worse in other science fields because there are no meta-indices for most areas of science. In total, 607,023 computer science papers were downloaded from Sci-Hub in the six-month period.

Table 2 of this work shows the twenty most downloaded computer science papers from Sci-Hub. Next, the top five countries with the most downloads were India, Iran, China, the United States and Indonesia. If the data is normalized by downloads per population, then the ranking of the countries is different and top five countries include Tunisia, Iran, Greece, Morocco and Jordan. Next, the detailed information of computer science papers was retrieved by their DOIs and CrossRef DOI API. The most downloaded publishers were Institute of Electrical and Electronics Engineers (IEEE) with 169,313 papers, Elsevier BV with 132,098 and Springer Nature with 109,586 papers. If the countries, institutions or libraries tried to improve computer scientists' work conditions, they should consider enabling access to repositories or journals of the men-

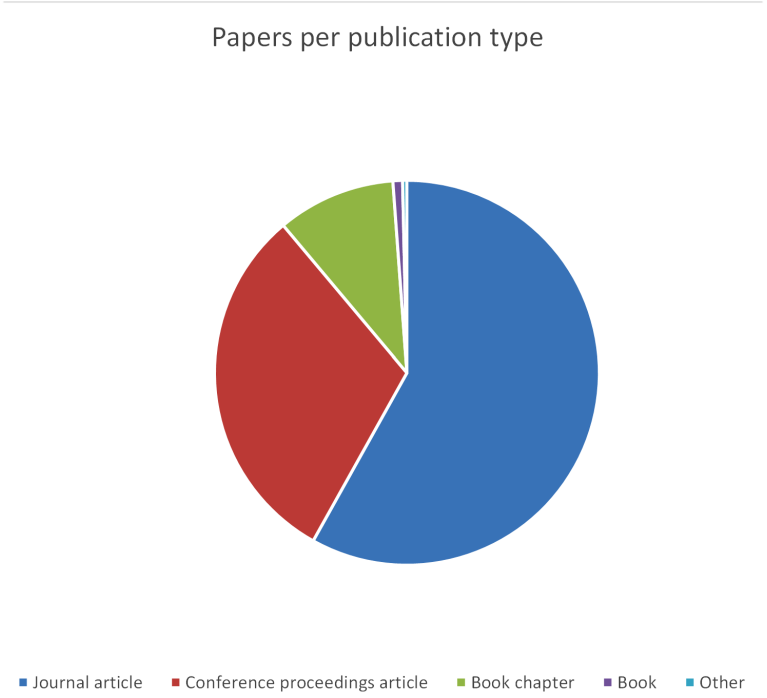

Figure 3: Types of downloaded papers

tioned publishers first. Furthermore, the most recently published computer science papers were the most popular to be downloaded from Sci-Hub site. If the types of computer science papers downloaded from Sci-Hub were observed, the journal papers were dominant (58%), followed by conference proceeding articles (31%) and book chapters (10%). The conclusion can be made that access to quality journal papers is the most important for computer scientists. The findings of this paper show who Sci-Hub users that download computer science research papers are and where they are from; and what computer science research papers the mentioned users downloaded. This information is useful for libraries and policies that make decisions on the most important sources (online repositories and journals) for computer science community. It is also important for publishers who can consider new methods how to make access to the most wanted content more accessible and cheaper for their end-users.

References

[1] G. Cabanac, Bibliogifts in LibGen? Study of a text-sharing platform driven by biblioleaks and crowdsourcing, *Journal of the Association for Information Science and Technology* **67,** 4 (2016) 874–884. ⇒84, 85

[2] P. M. Davis, W. H. Walters, The impact of free access to the scientific literature: a review of recent research, *Journal of the Medical Library Association* 99, 3(2011) 208–217. ⇒84

[3] M. Enserink, In dramatic statement, European leaders call for 'immediate' open access to all scientific papers by 2020, *Science,* May. 27, 2016. ⇒84

[4] N. Timus, Z. Babutsidze, Pirating European Studies, *Journal of Contemporary European Research* **12,** 3 (2016) 783–791. ⇒84, 86

[5] M. McNutt, My love-hate of Sci-Hub, *Science,* **352,** 6285 (2016) 497–497. ⇒84

[6] J. Bohannon, Who's downloading pirated papers? Everyone, *Science.* April 28, 2016 . ⇒84, 85, 90, 91

[7] A. Elbakyan, J. Bohannon, Data from: Who uses Sci-Hub? Everyone, *Dryad Digital Repository.* April 28, 2016. ⇒84, 85, 86

[8] M. Parkhill, Sci-Hub: The academic cat is out of the bag, *Plam Analytics*, May 16, 2016 ⇒85

[9] Z. Babutsidze, Pirated economics, *MPRA paper* 71703. 2016. ⇒85

[10] M. Swab, K. Romme, Scholarly sharing via Twitter: #icanhazpdf requests for health sciences Literature, *Journal of the Canadian Health Libraries Association/Journal de l'Association des bibliotheques de la sante du Canada* **37,** 1 (2016) 6–11. ⇒86

[11] C. C. Gardner, G. J. Gardner, Fast and furious (at publishers): the motivations behing crowdsourced research sharing, *College & Research Library* January 2017, pp. 1–24. ⇒86

[12] M. Ley, The DBLP computer science bibliography: evolution, research issues, perspectives, in *String Processing and Information Retrieval*, LNCS 2476 (2002) 1–10 (eds. A. H. F. Laender, A. L. Oliveira), Springer Berlin Heidelberg. ⇒87

Received: January 14, 2017 • Revised: February 14, 2017