



Universiteit  
Leiden  
The Netherlands

## **How self-archiving influences the citation impact of a paper: A bibliometric analysis of arXiv papers and non-arXiv papers in the field of information and library science**

Wang, Z.; Glänzel, W.; Chen, Y.

### **Citation**

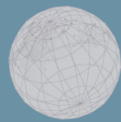
Wang, Z., Glänzel, W., & Chen, Y. (2018). How self-archiving influences the citation impact of a paper: A bibliometric analysis of arXiv papers and non-arXiv papers in the field of information and library science. *Sti 2018 Conference Proceedings*, 323-330. Retrieved from <https://hdl.handle.net/1887/65329>

Version: Not Applicable (or Unknown)

License: [Leiden University Non-exclusive license](#)

Downloaded from: <https://hdl.handle.net/1887/65329>

**Note:** To cite this publication please use the final published version (if applicable).



# STI 2018 Leiden

*23rd International Conference on Science and Technology Indicators  
"Science, Technology and Innovation Indicators in Transition"*

## **STI 2018 Conference Proceedings**

*Proceedings of the 23rd International Conference on Science and Technology Indicators*

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

### **Chair of the Conference**

Paul Wouters

### **Scientific Editors**

Rodrigo Costas  
Thomas Franssen  
Alfredo Yegros-Yegros

### **Layout**

Andrea Reyes Elizondo  
Suze van der Luijt-Jansen

The articles of this collection can be accessed at <https://hdl.handle.net/1887/64521>

ISBN: 978-90-9031204-0

© of the text: the authors

© 2018 Centre for Science and Technology Studies (CWTS), Leiden University, The Netherlands



This ARTICLE is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

## How self-archiving influences the citation impact of a paper: A bibliometric analysis of arXiv papers and non-arXiv papers in the field of information science and library science

Zhiqi Wang<sup>\*</sup>, Wolfgang Glänzel<sup>\*\*</sup>, Yue Chen<sup>\*\*\*</sup>

<sup>\*</sup> [zhiqi\\_wang90@126.com](mailto:zhiqi_wang90@126.com)

WISE Lab, Institute of Science of Science and S&T management, Dalian University of Technology, Dalian 116085 (China)

ECOOM, KU Leuven, Naamsestraat 61, Leuven, 3000 (Belgium)

<sup>\*\*</sup> [Wolfgang.Glanzel@kuleuven.be](mailto:Wolfgang.Glanzel@kuleuven.be)

ECOOM and Dept MSI, KU Leuven, Naamsestraat 61, Leuven, 3000 (Belgium)

Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

<sup>\*\*\*</sup> [chenyuedlut@163.com](mailto:chenyuedlut@163.com)

WISE Lab, Institute of Science of Science and S&T management, Dalian University of Technology, Dalian 116085 (China)

### Introduction

Preprint literature has become an important instrument of scholarly communication. Although several scientific communities, for instance, mathematicians, use traditionally this communication channel for decades, its real emerge is linked to electronic publication and the availability of large repositories. The advantage of preprint literature is obvious: fast dissemination and the opportunity to receive response by the community before the final version of the paper is published. Although the role of preprints has changed during the emergence of preprint archives and self-archiving repositories, their basic function is still the same as has been extended by important new ones, the increased visibility and the possibility of open access and post-prints. Evidence has been given that pre- or post-prints could get more visibility, usage and citations. Harnad & Brody (2004) found that physics articles published in journals also deposited as pre-prints in arXiv generated citations up to 400% higher than papers in the same journals that had not been deposited and similarity results were also found in astrophysics and astronomy (Schwarz et al., 2004; Kurtz et al., 2005). Davis (2011) reported that making articles free access received significantly more downloads and reached a broader audience within the first year in sciences, social sciences and humanities. A recent study found that WoS papers in mathematics that also were arXiv e-prints had the highest impact, but the citation rates gap between them and the articles only published in WoS was decreasing (Larivière et al., 2014). However, systematic study on the effect of self-archiving on impact in Information Science & Library Science (LIS) is still missing. In former research we compared the effect of the pre-prints in LIS and the Robotics filed, and found that OA citation advantage existed in both two fields, but a little more significant in LIS than Robotics, which contained a vast amount of conference proceedings (Chen & Wang et al., 2016). However, the results in that paper need further verification by a larger sample size and a fixed citation window. The main objective of this article is to reveal the citation characteristics of preprint articles, here represented by arXiv papers in LIS. The key research questions are as follows.

1. What is the share of papers published in the core LIS journals that are in arXiv and what time lags occur in the publication process?
2. Do aging characteristics and citation impact of journal articles deposited in arXiv compared to those not deposited differ?

In order to answer these questions, all citations received by papers indexed in Clarivate Analytics Web of Science Core Collection (WoS) are collected.

### Data sources and methods

Two data sources are used, the arXiv e-print service and the WoS database. In a first step, all citable documents (article, letter or review) assigned to the Subject Category *Information Science & Library Science* indexed in the WoS volumes 2008–2017 have been retrieved. The number of documents amounts to 39,173 papers. In a second step, only papers published in the 25 out of the 50 most relevant journals (accounting for nearly 80% of the all documents) could be found in arXiv by matching journal information. Finally, the top three journals with largest amount of arXiv papers, *Scientometrics* (SCIM), *Journal of the Association for Information Science and Technology* (JASIST) and *Journal of Informetrics* (JOI) (with 129, 121 and 106 arXiv papers respectively), are selected for a comparative analysis between arXiv papers and non-arXiv papers. The representation of the other 22 journals in arXiv does not allow any statistically reliable analysis.

Using this dataset, links are made between the citable documents published in the three journals published during 1991 and 2017 to the arXiv database metadata. The matching procedure used two sets of links, one was the direct correspondence between the arXiv titles and WoS titles, the other one was a fuzzy matching between the titles and first author and between titles and the abstracts, respectively. Finally, 570 papers published in the three journals are found being deposited in arXiv database. For the identification of citations to the arXiv-version of these published papers, the arXiv references formats were used (which changed on 1 April 2007). These identifiers make it possible to use the cited reference search in WoS.

### Results

#### *Share of WoS papers in arXiv and the distribution of time lag*

In the period 2005–2017, 8.7% papers published in the three journals are also deposited in arXiv (called *arXiv papers* in the following), which is more than the average share 3.6% of WoS papers in arXiv in 2010–2011, but much lower than in mathematics (21%), physics (20%) and earth and space science (12%) (see Larivière et al., 2014). Among the three journals, JOI has the highest share (20.52%) papers but the least absolute number of papers.

Table 1. The proportion of WoS papers on arXiv.

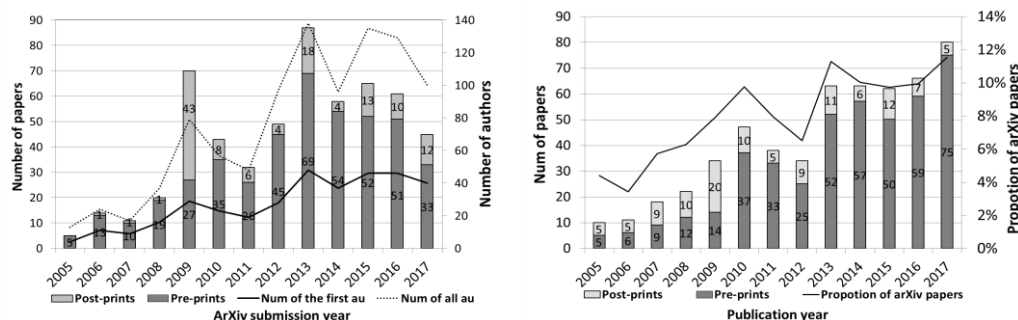
| Journal (JIF 2016)          | ArXiv papers<br>(1997,2002-2018) <sup>3</sup> | ArXiv papers<br>(2005-2017) | All WoS papers<br>(2005-2017) | Percentage |
|-----------------------------|-----------------------------------------------|-----------------------------|-------------------------------|------------|
| SCIM (2.147)                | 203                                           | 194                         | 3181                          | 6.10%      |
| JASIST <sup>1</sup> (2.322) | 202                                           | 198                         | 2457                          | 8.06%      |
| JOI <sup>2</sup> (2.920)    | 165                                           | 158                         | 770                           | 20.52%     |
| Total                       | 570                                           | 550                         | 6332                          | 8.69%      |

<sup>1</sup> JASIST: “*Journal of The American Society for Information Science and Technology*” and “*Journal of The Association for Information Science and Technology*”

<sup>2</sup> JOI: *Journal of Informetrics*, indexed in WoS since the year 2007.

<sup>3</sup> In the three journals, the earliest publication time of papers stored in arXiv is 1997, which was submitted to arXiv on 25/01/2010, about 23 years later than its publication date, however, from 1997 to 2001, no published papers are found in arXiv and the latest publication time of the arXiv paper is 05/03/2018 (the retrieval time of this research is 06/03/2018).

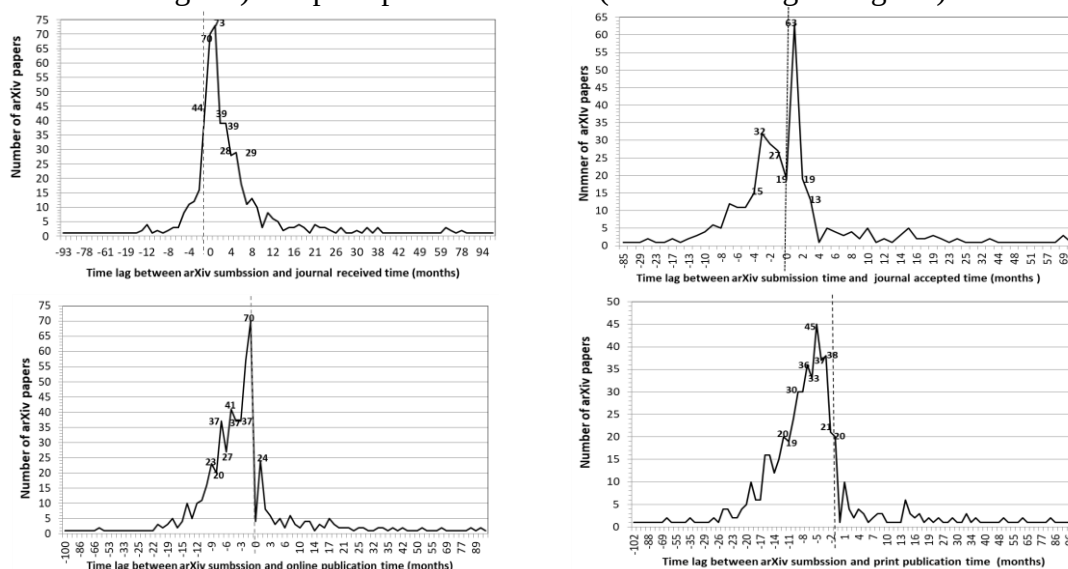
Figure 1. The distribution of submission time (A – left diagram) and publication time (B – right diagram) of arXiv papers during the period of 2005–2017



The number of first authors increased from 4 in 2005 to 38 in 2017 (cf. Figure 1), and the number of all authors from 13 to 100, indicating the reception is increasing in the LIS community. Note that the slight discrepancy between these two growth rate may reflect minor structural changes in the composition of co-author lists. However the increasing trend of arXiv papers over the submission time is more fluctuant than over the publication time. There is a significant overall growth and the two peaks in Figure 1 (A) have different explanation. The peak in 2009 is caused by a large submission from one individual author, while in 2013 it is caused by more authors submitting their papers to arXiv. Figure 1 (B) suggests an overall increasing trend, and for the latest five years, the share of journal papers deposited in arXiv even reaches or exceeds 10.5% on average. The high share of pre-prents more than 70% since 2010 means that readers can have an earlier view of the manuscripts of the journal papers from arXiv.

The distributions of different time lags are showed in Figure 2 to explore the authors' behaviour in using arXiv service in the LIS field. The values of X-axis represent the time lag between papers' arXiv submission and the different times in the journal publication process, including received date, accepted date, online publication date and print publication date. The time lag is interpolated and converted to months. Numbers are negative or positive according as paper was submitted to arXiv earlier or later than its received/accepted/online/print date for the journal publication.

Figure 2. Distribution of the time lag between arXiv submission and journal received (A – top left diagram), journal accepted (B – top right diagram), online publication (C – bottom left diagram) and print publication time (D – bottom right diagram)



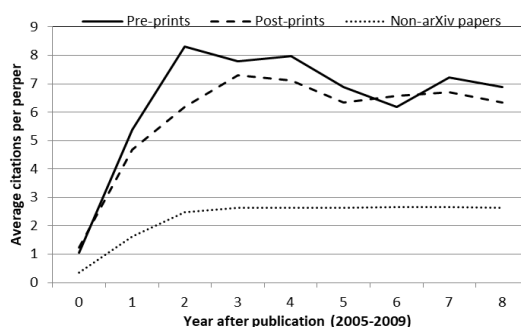
Note: Accepted time is not available from the website of SCIM, so Figure 2 (B) only contains the statistics of the time lag of all of the 346 papers published in JOI and JASIST from 2005–2017; 15 papers published in SCIM and 9 papers published in JASIST with no received time are excluded from Figure 2 (A).

Figure 2 (A) shows that most papers are posted to arXiv on the same day or within one month later than their journal received time, and 18 papers just one day later than its received time. About 50% of the authors submit their papers to arXiv before being accepted by the journal (see Figure 2 (B)), while most authors submit it very close to be the date accepted but no more than 2 months before or after. Figure 2 (C) and Figure 2 (D) shows that 78% of all the arXiv papers' submission date is earlier than journal online publication and 84% are earlier than print publication date, indicating that the initial arXiv papers deposited in arXiv are always the author's version of the article before peer-review taking place. Furthermore, what we should also notice is that about 39% of all arXiv papers are updated after the first submission, among which, 117 out of the 140 arXiv papers published in JOI and JASIST updated their arXiv papers with new versions after journal acceptance, and about 23.5% of all arXiv papers in the three journals updated the first submission version after publish online on the publishers' website.

### *The aging characteristics and the citation impact of arXiv papers*

In this part, we analyse the distribution and aging characteristics of citations in several ways. To do so, the dataset has been subdivided into three different document types, pre-prints (papers submitted to arXiv before published online), post-prints (papers submitted to arXiv after published online) and non-arXiv papers. First the overall citation processes are studied, then, in a second step, looking from a micro perspective, the citation advantage of arXiv papers at the individual level, on the basis of authors is analysed. Finally, the citations to the arXiv-version of the journal articles are further explored.

Figure 3. Age distribution of average citations to pre-prints, post-prints and non-arXiv papers published during 2007–2009



For the first measurable level, we select the papers published from 2005 to 2009 (Fig.3) so that we could have a 9-year citation window, sufficiently long to adequately capture of the aging characteristics of citations to the three distinct types of papers, pre-prints (the number of them is 51), post-prints (the number of them is 46), non-arXiv papers (the number of them is 1579). It shows that the three different data sets follow distinct citation patterns. During the first three years after publication (including the publication year), the average citations to pre-prints are higher and increase more rapidly than post-prints, peaking at about the third year after publication, but then, decay faster than post-prints. However, there is no significant different in the long-term citation patterns of pre- or post-prints, especially from the seventh year after publication. While citations to non-arXiv papers have a slower growth and decay, peaking at around the sixth year after publication. It is suggested that the boost in the early citations to pre-prints may due to the “early view” effect, that is, the pre-prints are posted to

arXiv before publication online in a journal, and thus have a longer effective citation life-time. To test the hypothesis, we further quantify the citation advantage of different types of arXiv papers during different citation time window.

In this research, the arXiv Citation Impact Differential (CID), an optimized calculated function put forward by Moed (2007) on the basis of the OA versus non-OA Impact Ratio (IR) put forward by Harnad & Brody in 2004 is calculated, defined as:

$$CID = 100 * (CPP_a - CPP_{na}) / ((CPP_a + CPP_{na}) / 2)$$

$CPP_a$ : the number of received citations per paper in arXiv

$CPP_{na}$ : the number of received citations per paper not in arXiv

Figure 4. ArXiv Citation Impact Differential (CID) using variable citation windows (2005–2017)

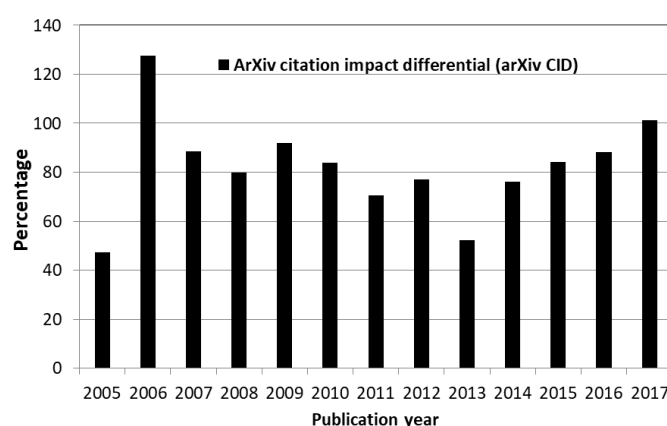


Figure 4 gives the Citation Impact Differential of the ArXiv documents on the basis of variable citation windows, in which citations are cumulated from the publication year till 2017. Since the CID is a relative indicator, the length of the actual citation windows do not cause any bias and CID values can be compared over time. The observed fluctuation during the period 2005–2012 may be caused by the small number of pre-prints, which makes the indicator sensitive to extreme citation rates. The stable upward trend of arXiv CID in the last five years is based on larger sets and supports the early view effect hypotheses, because the time period over which the CID is calculated shortens as the publication year becomes more recent.

In addition, we calculate the arXiv CID in two 3-year citation windows as well: (a) the first three years beginning with the publication year, and (b) the 3-year window subsequent to the previous one. The last row in Table 2 gives the CID values calculated for the complete collection of the three journals and shows large differences among the three journals. JASIST has the highest CID values, while JOI has the lowest. The change of CID values during the two citation windows for SCIM and JOI are negligible, but not so for JASIST, where there is a roughly 20%  $((108.36 - 87.3) / 108.36)$  decrease. Overall, papers in the three journals posted to arXiv are cited more than 2.5 times than those are not.

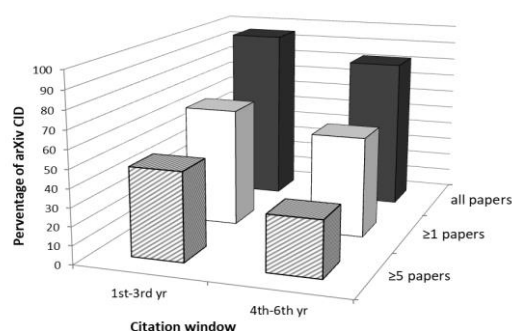
Table 2. ArXiv Citation Impact Differentials (CID) at the journal level

| JR     | Amount of arXiv papers(Percentage)<br>(2005-2012) | Average citations                                                   |                                                                     | ArXiv CID                                      |                                                |
|--------|---------------------------------------------------|---------------------------------------------------------------------|---------------------------------------------------------------------|------------------------------------------------|------------------------------------------------|
|        |                                                   | 1 <sup>st</sup> -3 <sup>rd</sup> yrs after pub<br>(arXiv/non-arXiv) | 4 <sup>th</sup> -6 <sup>th</sup> yrs after pub<br>(arXiv/non-arXiv) | 1 <sup>st</sup> -3 <sup>rd</sup> yrs after pub | 4 <sup>th</sup> -6 <sup>th</sup> yrs after pub |
| SCIM   | 65(5.62%)                                         | 13.48/4.64                                                          | 18.54 /8.02                                                         | 77.90                                          | 79.24                                          |
| JASIST | 86(7.39%)                                         | 14.92/4.43                                                          | 21.21 /8.31                                                         | 108.36                                         | 87.30                                          |
| JOI    | 55(18.73%)                                        | 14.43 /7.21                                                         | 19.95 /10.12                                                        | 66.76                                          | 65.38                                          |
| ALL    | 216(7.88%)                                        | 13.49 /4.77                                                         | 20.08 /8.33                                                         | 95.47                                          | 82.65                                          |

The arXiv CID calculated above is on the basis of journal level, however, as mentioned above, the impact of arXiv papers from individual authors' perspective may add to the picture. Therefore we calculate CID values separately for authors, who have published at least one paper each deposited in arXiv and not deposited in arXiv (the number is 287), and for authors who have published at least five papers deposited in arXiv and one or more papers not deposited in arXiv (the number is 39), so authors with no arXiv papers are excluded from the two sets. The results are shown in Figure 5. Here "all papers" denotes the arXiv CID for the complete document set of the three journals, " $\geq 1$  papers" (" $\geq 5$  papers") denotes the authors publishing at least one (five) paper deposited in arXiv.

As has already been observed in Figure 3 and Table 2, the CID calculated for the fourth to sixth year after publication tends to be lower than that calculated for the first three years after publication, and for authors in " $\geq 1$  papers" (" $\geq 5$  papers"), the relatively decrease is about 15% (37%), bigger than the 13% decrease for the complete document set, indicating that the early view effect becomes stronger when CID calculated at the author level. Another obviously visible pattern is that the CID calculated for authors is lower than for the journals as a whole, and decreases with increasing author productive.

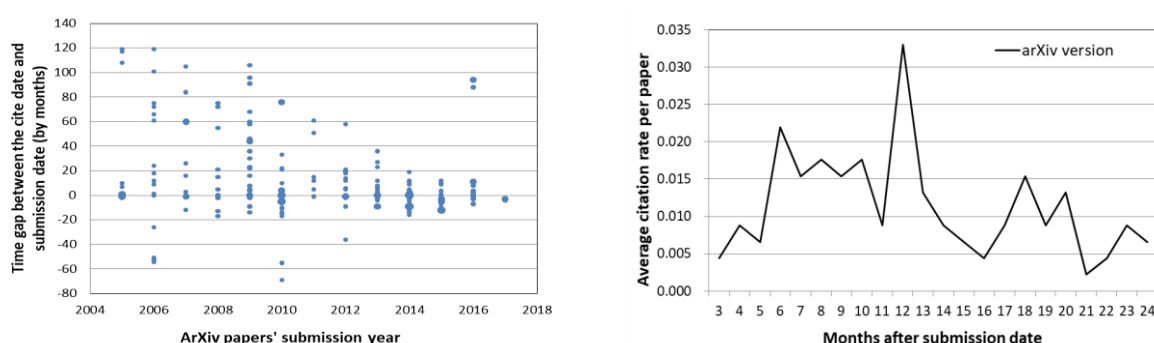
Figure 5. ArXiv Citation Impact Differential (CID) over authors publishing



Similarly to journal articles with a unique DOI provided by the publisher, papers deposited in arXiv identifier, providing a complete and unique citation for arXiv paper. As mentioned above, the identifier can be used to collect citations to the arXiv version of a published paper. This could help understand the position of e-prints in scholar communication in LIS. The content of these arXiv versions are essentially the same as that of the published journal articles (Schwarz & Kennicutt, 2004; Moed, 2007; Larivière et al., 2014). The distributions and aging characteristics of citations in the WoS database to the arXiv version of a paper are shown in Figure 6. The X-axis in the chart of Figure 6 (A) presents the arXiv submission year of a paper, and the Y-axis represents the time lag in months between the publication date of the citing paper and the cited arXiv-version paper. 112 arXiv-version papers were found

accounting for 20% of all 570 arXiv papers and having received 206 citations in total. The number of citing papers is 186. Among these papers, 105 are pre-prints, indicating that the preprints started the citation circle earlier than post-prints or non-arXiv papers. 57 pre-prints are cited before its publication time, and a large amount of citations generated within two years after the arXiv submission time. Although the citations received by the arXiv version tends to decay fast after it is published in a journal, the total amount of this part of citations accounts for more than 50% of the total citations, high enough to be not ignored. The longest citation lag is 119 months, indicating that the arXiv-version papers could still be cited after having been published five years later. The citation trend during the first two years after submission of the arXiv-version papers is presented Figure 6 (B). The majority of citations to arXiv-version papers are received during the six months to one year after being posted to arXiv, then followed by an overall decline. The decline may be caused by the publication of the paper. The published version of a paper is more likely to be cited than its arXiv version when people have access to both versions. Thus for papers published in a journal and deposited in arXiv as well, either pre-prints or post-prints, could enjoy broader readership, higher visibility, and thus more potential citations.

Figure 6. Distribution (A – left diagram) and aging characteristics (B – right diagram) of citations to the document's arXiv-version



## Conclusions and discussion

Despite the relatively low share of papers published in the three journals with largest amount of arXiv papers in LIS, both the absolute number of arXiv papers and their authors increased, especially during the most recent years. Nearly 80% of all arXiv papers are preprints, and authors are more likely to post the pre-version of a paper to arXiv when they submit it to a journal, or when the paper is accepted by a journal, and about 40% of the authors update it with a new version during the peer-review process of a journal.

The arXiv e-prints archive have to some extent changed the way of scholarly communication in LIS. They seem to boost the citation rates of the posted papers to about 2.5 times of those of unposted ones. Papers published in journals that have a pre-print version in arXiv, could enjoy higher initial citation rates, e.g., in the first three years after publication, which could be explained by the “early view” effect. On the other hand, a decline in the citation advantage of arXiv papers could be observed at the individual author level, when compared with at the complete document set, and arXiv CID values also decrease with increasing author productivity.

Though many authors prefer to cite the journal version of the paper, if accessible, the arXiv version of a published still can be cited independently whether the journal version appeared before or after self-archiving. However, the citation rate decreases faster after it is published in a journal. The slight increase in the first two year after the submission date and the citations

to post-prints can be explained by the OA effect of arXiv papers, which apparently yields the advantage of a high visibility.

The analysis of other relevant aspects of visibility and impact, such as the effect of publication delay, the publishers' embargo period for self-archiving or the deposition in institutional and different open repositories, remain tasks of future research.

### Acknowledgement

Zhiqi Wang acknowledges the support from China Scholarship Council (CSC).

### References

- Chen, Y., Wang, Z., Tan, J., Liu, Z. (2017). The position of preprint in scholarly communication: A bibliometric and empirical study of arXiv. *Proceedings of ISSI 2017 Wuhan: 16th International Society of Scientometrics and Informetrics Conference*, 799–809. Accessible at: <http://www.issi-society.org/publications/issi-conference-proceedings/proceedings-of-issi-2017/>
- Davis, P. M. (2011). Open access, readership, citations: a randomized controlled trial of scientific journal publishing. *The FASEB Journal*, 25(7), 2129–2134.
- Harnad, S., & Brody, T. (2004). Comparing the impact of Open Access vs. non-OA articles in the same journals. *D-Lib Magazine*, Volume 10(6), 1–6.
- Kurtz, M., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., & Henneken, M. (2005). The effect of use and access on citation. *Information Processing and Management*, 41(6), 1395–1402.
- Larivière, V., Sugimoto, C. R., Macaluso, B., Milojevic, S., Cronin, B., & Thelwall, M. (2014). ArXiv e-prints and the journal of record: An analysis of roles and relationships. *JASIST*, 65(6), 1157–1169.
- Moed, H. F. (2007). The effect of “open access” on citation impact: An analysis of ArXiv's condensed matter section. *JASIST*, 58(13), 2047–2054.
- Schwarz, G. J., & Kenicutt, R. C. (2004). Demographic and citation trends in astrophysical journal papers and preprints. *Bulletin of the American Astronomical Society*, 36(5), 14. Accessible at: <http://arxiv.org/abs/astro-ph/0411275>