JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 6(4): 06-16, 2018
https://doi.org/10.1633/JISTaP.2018.6.4.1

Citations to arXiv Preprints by Indexed Journals and Their Impact on Research Evaluation

Antonia Ferrer-Sapena*
Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València, Valencia, Spain
E-mail: anfersa@upv.es

Fernanda Peset
Instituto Universitario deMatemática Pura y Aplicada, Universitat Politècnica de València, Valencia, Spain
E-mail: mpesetm@upv.es

Rafael Aleixandre-Benavent
Instituto de Gestión de la Innovación y del Conocimiento-Ingenio (CSIC-Universitat Politècnica de València), UISYS, Universitat de València, Valencia, Spain
E-mail: Rafael.Aleixandre@uv.es

Enrique A. Sánchez-Pérez*
Instituto Universitario deMatemática Pura y Aplicada, Universitat Politècnica de València, Valencia, Spain
E-mail: easancpe@mat.upv.es

ABSTRACT
This article shows an approach to the study of two fundamental aspects of the prepublication of scientific manuscripts in specialized repositories (arXiv). The first refers to the size of the interaction of "standard papers" in journals appearing in the Web of Science (WoS)—now Clarivate Analytics—and "non-standard papers" (manuscripts appearing in arXiv). Specifically, we analyze the citations found in the WoS to articles in arXiv. The second aspect is how publication in arXiv affects the citation count of authors. The question is whether or not prepublishing in arXiv benefits authors from the point of view of increasing their citations, or rather produces a dispersion, which would diminish the relevance of their publications in evaluation processes. Data have been collected from arXiv, the websites of the journals, Google Scholar, and WoS following a specific ad hoc procedure. The number of citations in journal articles published in WoS to preprints in arXiv is not large. We show that citation counts from regular papers and preprints using different sources (arXiv, the journal's website, WoS) give completely different results. This suggests a rather scattered picture of citations that could distort the citation count of a given article against the author's interest. However, the number of WoS references to arXiv preprints is small, minimizing this potential negative effect.

Keywords: preprint, research evaluation, arXiv, impact, bibliometric analysis, citation

Open Access

Accepted date: September 12, 2018
Received date: May 03, 2018

***Corresponding Authors:**
Antonia Ferrer-Sapena
Professor
Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València, Camino de Vera s/n 46022 Valencia, Spain
E-mail: anfersa@upv.es

Enrique A. Sánchez-Pérez
Professor
Instituto Universitario deMatemática Pura y Aplicada, Universitat Politècnica de València, Camino de Vera s/n 46022 Valencia, Spain
E-mail: easancpe@mat.upv.es

1. INTRODUCTION

The prior publication of scientific manuscripts in electronic preprint repositories has proved, since the beginning of this practice, to be a useful way of increasing the visibility and accessibility of research work. Since the last years of the past century a big amount of papers have shown that, as a direct consequence of prepublication, citations of previously posted articles increase. In this paper we are interested in showing a particular aspect of the citation of research documents deposited in arXiv whose consequences would not be so positive. Specifically, we wanted to analyze what is the total number of arXiv documents that get citations in "standard journals"—journals appearing in Web of Science (WoS), now Clarivate Analytics—and what is the citation dynamics of those documents. Since these kinds of citations are, in a way, beyond the reach of typical counting tools, we try to provide quantitative information on how many citations could be missed. Our interest is to measure to what extent these missed citations may harm the interests of authors undergoing a bibliometric evaluation. In this sense, it is well known that some national research evaluation agencies—such as the Polish or the Spanish—use citation counts for research evaluation.

Therefore, we analyzed citations to preprints that appeared in WoS and were previously posted in arXiv. The reason we chose arXiv is that it has become a prototype of a universal e-preprints repository for physics, mathematics, and computer science. Consequently, we restrict our attention to these disciplines. As sources of citations for further analysis, we mainly use WoS and Google Scholar. To complete the chart, other sources have also been used, such as citations provided by the websites of individual journals.

First, let us explain some basic facts about the context of our work and the previous research that has been done on the subject. Since arXiv was an early initiative in the field of e-preprint repositories, some authors have carried out several analyses of the motivation of researchers to use it over the last twenty years. The general opinion is that the main reason for uploading a manuscript to arXiv is the same as the one that caused classical preprint circulation (hard copies). The outline of the main practical motivations of the authors with regard to preprint publication as presented in the paper by Pinfield (2005) should be mentioned here. Although this paper is not recent, an inspection of the authors' reasons for preprint publication in the current literature suggests that they have not changed at all. The first motive explained there for registering a manuscript in arXiv is that this is a way of setting priorities when presenting a new idea or research result. Preprints provide a way to register them without having to wait for standard journal publication. The second objective of e-preprint publication is rapid dissemination. Preprint circulation is clearly faster than formal peer-reviewed publication. The third reason is that the circulation of a preprint is a way of improving the finished article by considering the comments of colleagues for the drafting of the final version. More works about motivation for preprint publication supporting these ideas can be found in Ardichvili, Page, and Wentling (2003), Kim (2011), and Zha, Li, and Yan (2013) (see also the references therein). It should also be mentioned that the new social media and other technological tools of the digital era have changed the role of preprint publication in the scholarly communication process. They have produced a clear diversification of "knowledge objects" (e-preprints, datasets, open access, on-going manuscripts, short letters…). Although there is no discussion here of how this might affect the prepublication of manuscripts, it is clear that the role of traditional documents in the dissemination of science, and therefore of preprints themselves, will change. The reader can find more information in the paper by Haustein (2016) and the references therein. In this regard, other interesting contributions have also been made from sociology. The prepublication of e-prints and, in general, open access initiatives have greatly changed the classic world of scientific publishing in the way that Bohlin (2004) had already noticed: New internet technologies had changed the needs and interest of potential users of scientific publishing, producing a transformation in academic communication. The study of how these changes might also modify the evaluation of the research would also be interesting, and would help to show a general picture of the problem we are facing. However, this issue is outside the scope of this document, where we provide only some bibliometric information and general explanations.

However, there are other reasons for using the "electronic version" of this classic practice, which is represented by repositories such as arXiv. In some cases, and also depending on the scientific field, articles are deposited in arXiv in the author's version after their acceptance and even after their publication in a standard journal. In fact, this practice is proposed and accepted by major publishers such as Springer. In the "Self-archiving Policy" section of the website, the following sentence appears in the Copyright Transfer Statement: "Authors may also deposit this version (the author's version) of the article in any repository, provided it is only made publicly available 12 months after official publication or later." This should be understood in

the context of the open access movement, to facilitate the dissemination of research results outside the business of scientific publishers (Klein Broadwell, Farb, & Grappone, 2016). Although the reasons why authors use the arXiv repository in this way is also an interesting topic of study, we will not analyze it in this article, since a priori it does not seem to interfere too much with article citations.

From the point of view of bibliometric parameters, the advantages of prepublication have been explained in terms of the following facts, which are widely accepted. The reader can find the following classification in the paper by Kurtz et al. (2005): Open Access Postulate: Free access papers can be read more easily, and so get cited more frequently. Early Access Postulate: posted preprints are available sooner and thus gain primacy, increasing citations. Self-selection Bias Postulate: The authors select their most important (and so more citable) papers to post them. This explanation serves to justify the empirical fact that preprint publication indeed increases the total number of citations. Some early studies have already noted this (see for example Fig. 4 in Henneken et al. [2006] and the references therein); however, other works warn that this is not always the case (Kurtz & Henneken, 2007). A 2010 report lists 27 studies in which this positive conclusion is found, compared to four studies in which the conclusion is the opposite (Swan, 2010).

Considering all these issues, we study how the publication of a manuscript in arXiv has effects in terms of benefits for the authors, from the point of view of increasing the number of citations. Of course, prepublication ensures a better opportunity in the diffusion of the work, but it is not easy to know to what extent this practice can actually improve some of the bibliometric parameters of authors, such as the number of citations of their papers or the publication of their articles in journals with higher citation rates. It is already well understood that prepublication affects the citation dynamics of a given paper, and should be taken into account in any comprehensive citation analysis (Neuhaus & Daniel, 2008). In particular, some specific statistical studies have been carried out on arXiv. The main current references are the exhaustive papers by Larivière et al. (2014) and Li, Thelwall, and Kousha (2015), but also the earlier works by Kurtz et al. (2005), Henneken et al. (2006), and Kurts et al. (2007). The statistical studies presented there—mainly the first one by Larivière—give a clear idea of the relationship between arXiv and the main databases of scientific articles. This type of analysis is not reproduced here: Our aim is to provide more specific information on the aspects of this relationship explained above and to discuss them together with some empirical opinions often expressed by researchers in the fields of physics and mathematics.

Let us finish this section by explaining the main conclusions presented in the existing literature on the subject. Some studies confirm that documents deposited in arXiv receive more citations and are cited before (see p. 2053 in the paper by Moed, 2007). According to this reference, the main advantage of using arXiv from the point of view of bibliometric parameters is that citations occur earlier. The author explains that, although the number of citations does not seem to increase due to the use of arXiv, the scientific community begins to process the information earlier, so the citations appear earlier. This obviously means an improvement in the promotion of the document. However, there are other studies on particular contexts in which this effect is not detected (Davis & Fromerth, 2007), although they are in the minority. There is also evidence that the quality of papers previously published in arXiv is generally above average (Moed, 2007; Davis & Fromerth, 2007); measuring quality is always delicate, so these results must be considered in the appropriate context. More studies on arXiv and the dissemination of the manuscripts deposited in it can be found in the papers by Haque and Ginsparg (2009, 2010), Manuel (2001), and Youngen (1998). In general, it must be said that all of them demonstrate some aspects of the advantages of prepublication that we have explained above: impact, parallel form of distribution, independence from the delays produced by the standard publication process, etc. A different methodology has been used in the present document. We have considered only the total set of preprints that appear in arXiv and that have been cited at least once in a regular WoS journal.

Thus, the sample of papers is not the same as the one that has been analyzed in other works. The results of the use, citations, and journal publication of the articles in the selection will be explained and some conclusions will be presented. Mainly, the dynamics of the citations will be explained, considering preprints as if they were regular papers in standard journals, as well as the statistical data on the areas to which these papers belong. A recent paper that studies the dynamics of publication/citation in arXiv in comparison to other sources and that is related to our methodology in a sense is the one found in the paper by Bar-Ilan (2014). This work is dedicated to the area of astrophysics. It analyzed the work of one hundred European astrophysicists indexed by Scopus, including the number of manuscripts deposited in arXiv and the number or brands of Mendeley readers. Although arXiv is widely used in astrophysics, it shows that more documents appear in Scopus than in arXiv; it also shows that the number of

marks in Mendeley is significantly lower than the number of citations in Scopus. In this case, the comparison between the data sources was made based on the names of the authors and the titles of the publications, thus being more related to our methodology.

In order to facilitate easy understanding of the arguments in this paper, we recall that the term "standard publication" of an arXiv manuscript will be used when it is published in a journal appearing in WoS. The term "standard citation" from an arXiv document will also be used if the journal in which the citation appears belongs to the WoS Core Collection. In general, the word "standard"—or "regular"—will be used for citations, journals, and articles that are measured and covered by journals in the WoS Core Collection. We have adapted the terminology found in the paper by Kling, Spector, and McKim (2002).

Specifically, our bibliometric analysis is guided by the following general questions.

Q1. "arXiv to standard" publication dynamics: How many documents in arXiv are cited in WoS? Which are the scientific fields in which research preprints posted in arXiv—with at least one citation in WoS—are most cited in standard journals? What is the proportion of papers that meet this requirement and are finally published in standard journals? What about the delay in publication?

Q2. "non-standard citation" of arXiv manuscripts as non-standard documents: How can citation of documents in arXiv with at least one citation in WoS be measured outside the WoS context?

2. MATERIAL AND METHODS

2.1. Data Collection

Our study followed the steps explained below. The data collection procedure started by setting the end date: December 2015. We have collected all article citations in arXiv that have appeared in WoS up to this date. Using the option Cited Reference Search in WoS, a search was made of the word "arXiv" in the field Cited Work. This provided the total amount of papers that, coming from the repository—and therefore accessible by the scientific community without peer review—enter the world of standard publications by appearing in a list of references of a published paper. It must be said that we searched these references one-by-one, attending to the specific properties of each of them in order to decide whether or not they were acceptable for the sample. The reason is that the way researchers cite preprints in arXiv is not homogeneous, and there are no fixed rules for doing this. This implies that the process of identification of a paper is in general difficult, if not impossible. This is the case if the final published version of the article does not have the same title, in which case it is difficult to realize that this article actually coincides with an earlier preprint. Although arXiv allows you to upload updated versions, this is not always done.

For instance, references with the following structure "MAYOR M, 2008, ARXIV," or "BEIRAO, ARXIV ASTROPHYSICS" were difficult to find. To detect the first one in arXiv, the name "MAYOR" has been introduced in the field "Authors," limiting also the date of storage. The result obtained in which "Mayor" appeared as the first author—also with the initial of the name "M"—was considered as the document referred to. If it appeared as the author, and there are no more preprints, it was also considered as such. In the event that there were two or more preprints with these characteristics, the paper was classified as "untraceable." The easiest references to find were those that appeared as follows: "Compere G, 2007, ARXIV07083153HEPTH."

After setting the correct reference, Google Scholar was used to determine whether the article was already published in a regular journal. To check this, the DOI number was used if it was in arXiv; otherwise, the title was used for this purpose. This gave us a set of 561 preprints as a working sample. Some of them were later withdrawn for other reasons—for example, some were classified as biology papers—and so the final sample was set at 554 manuscripts. We will present only the most relevant data to support our arguments.

2.2. Citation Analysis

Once the total set of relevant preprints was identified, several analyses were conducted.

a. The first was to calculate the proportion of manuscripts deposited in arXiv that appeared in references of articles published in journals that are listed in WoS. This analysis was carried out after grouping some of the different scientific fields determined by arXiv, in order to have a relevant number of papers in each group. The proportion of articles cited in this way that were eventually published in WoS journals was also calculated.
b. The difference between the year of standard publication and the year in which the preprint was deposited in arXiv was also calculated.

c. Citations of articles published in a standard journal were also counted: number of citations recorded in arXiv, number of citations recorded on the website of the regular journal that published the manuscript, and the difference between these amounts. In case the journal did not provide the number of citations for the articles, Google Scholar was used.
d. Finally, we also counted the number of citations of articles that did not appear in any regular journal: number of citations registered in arXiv of the preprints, number of citations registered in Google Scholar, and the difference between these amounts.

3. RESULTS

3.1. Global Impact of Standard Citations to arXiv Documents

A total of 554 documents were considered from our search, after clearing references to documents that were impossible to fix due to deficiencies in the citation. The set is small, compared to the total amount of documents that can be found in arXiv. Taking into account that the number of documents in arXiv in the date of completion of the research was about 1,150,000 (see https://arxiv.org/stats/monthly_submissions), the overall impact of the citations that we are studying is not relevant. However, the number of documents is large enough to analyze some of the properties of these citations.

3.2. Publication Ratio and Publication Delay

As we have explained, the subject classification provided by arXiv was followed, unifying some fields by subject proximity if necessary for getting statistically meaningful results. The way the areas are grouped is the following. Areas with a big number of preprints are considered separately (astrophysics, computer science, and condensed matter).

Table 1. Deposited manuscripts and finally published papers

Scientific areas	Deposited	Published	Ratio (%)
Astrophysics	208	154	74
Computer science	50	36	72
Condensed matter	101	76	75.2
Mathematics, statistics, nonlinear sciences	56	36	64.3
Physics	139	72	51.8
Total	554	374	67.5

The rest of the areas were grouped in a standard way under the names "mathematics, statistics, nonlinear sciences." and "physics." The amount of deposited papers depends strongly on the area, and also the publication rates. Table 1 shows the total number of deposited and published papers, respectively, for some scientific fields that are particularly relevant for our study. The complete tables with all the disciplines can be seen in the attached datasets.

It can be seen that the result depends greatly on the subjects. However, our result coincides broadly with the ones obtained in Larivière et al. (2014). It is shown that about 64% of all arXiv preprints are published in a WoS-indexed journal. In our case the rate is 67.2%. There is a small deviation, probably due to the bias produced by our selection criteria. Indeed, since the set of manuscripts which have citations from standard journals has been chosen, this already means that they are in a sense more relevant than non-cited ones. The results reinforce the idea suggested by the value of the total rate computed in Larivière et al. (2014). It could be interpreted in terms of the coherence of the authors' publication policies: The more citations in standard journals, the greater the likelihood that the paper will be published in a standard way. This could mean that the manuscript is considered a standard scientific document both by the authors and by the rest of the researchers of the scientific field. Publication in arXiv would be just a first step in the standard publication process, not an alternative form of dissemination of information. The value of the ratio itself suggests this conclusion: At least two of each three papers published in arXiv—that is, most authors—understand arXiv as the first step in the publication process, and not as a final publication medium. However, this subtracts some of the potential standard citations, contrary to the interests of authors who need to pass an evaluation process. In return, a rapid and early dissemination of the work would help the authors to gain prestige in the field. Each researcher must find the right balance between these two factors.

The results for specific arXiv specialties follow a similar rule, and are compatible with those published earlier; see Fig. 1 in Larivière et al. (2014). Again, our results show a higher standard publication rate, due to the relevance argument explained above with respect to the results presented in Table 1. Note also that the results are given for the grouped specialties, which does not allow for a direct comparison with previously published material. However, there are some interesting differences in two opposite directions that should be noted. Although the following arguments cannot be considered conclusive, we believe they may provide some ideas for interpretation.

a. In the grouped area "mathematics, statistics and nonlinear sciences" of our study, the ratio obtained is 64.3%, while for the total amount of papers considered in Larivière et al. (2014) it is less than 50%. As was explained before, this could suggest that mathematicians agree in publishing in standard journals independently of prepublication in arXiv, but mainly of those papers that are considered to be relevant enough to be cited. Alternatively, both facts can be considered as independent, and then this deviation would mean that authors that previously publish in arXiv are more actively involved in diffusion of their work, being also the ones with a bigger rate of standard publication. This publication habit may be specific for some scientific fields, but it seems to be the most general behavior.
b. The opposite trend can be observed with regard to the proportion of publication in our grouped area "physics," which in our case is significantly lower than in the general study of Larivière et al. (2014). This would mean that, to some extent, some authors feel that uploading a manuscript to arXiv is good enough to ensure the visibility of their work, and then prepublication and standard publication are two different tools for diffusion. This would be coherent with the hypothesis that documents in arXiv and papers in standard journals are in fact different enough to make it difficult to link the preprint and the final publication.

Other interesting bibliometric information that can be obtained refers to the average time that is needed for publishing a paper after it is posted in arXiv. As an initial approach, authors are supposed to upload their paper to arXiv when they complete their research work, so that from this point onwards the delay can be interpreted directly as exclusively due to the publication process. The results are shown in Table 2 for the grouped specialties; again the reader can find the complete information in the attached datasets. It should be remarked that the dispersion of the result is very high (high variance).

Table 2. Publication delay for grouped specialties

	Publication delay (year)	Publication ratio (%)
Astrophysics	0.8	74
Computer science	1.1	72
Condensed matter	0.8	75.2
Mathematics, statistics, nonlinear sciences	2.5	64.3
Physics	1.5	51.8

A comparison with other publication delay data that can be found in the literature makes sense. In Larivière et al. (2014), the data computed with the whole of the arXiv database show that the specialties grouped in our case with the label "mathematics, statistics and nonlinear sciences" have a publication delay of more than 1.4 years (see Fig. 5 in the referred paper). The value estimated in the present investigation is however higher (2.5 years). Also the time elapsed for publication in areas of the grouped variable "physics" is shown shorter in the analysis in Larivière et al. (2014) than in ours (1.5), which almost doubles the expected value (0.8). Therefore, a fairly large difference has been found with the previous analysis. A suitable explanation of the reason for the higher delay could be the bias produced by our selection method of arXiv manuscripts. Citations to the arXiv version of a manuscript that will be published later may mean that there is a delay on publication. Otherwise, it seems natural to cite the standard published version if possible, or both versions; it is known that this facilitates citation counts by the WoS—what benefits the authors—and also ensures to the potential reader of the citing paper that its reference have been peer reviewed.

Another aspect that should be mentioned is the relationship observed between the delay in publication and the publication rate in each grouped specialty. Delay in publication increases when the proportion of publications decreases, as can be seen in Table 2, although only a weak correlation can be observed. The topic "quantitative biology," which is found in the original sample, has been eliminated due to the small sample size. Each discipline seems to have its own delay/ratio characteristics.

There is an inverse relationship between the number of articles deposited in arXiv and the length of the publication process: the greater the publication ratio, the shorter the delay in publication. However, this relationship is weak. For example, the area "mathematics, statistics and nonlinear sciences" shows its own particular values. It seems to be again a consequence of the authors' publication policy together with the characteristics of the journals that publish in different scientific fields. At one extreme we find "condensed matter" and "astrophysics," with low delay and high ratio, while at the other extreme we find "mathematics, statistics and nonlinear sciences" and "physics," with different proportions between these terms. There are long delays in publication along with relatively small publication rates, which is consistent with what appear to be different philosophical views on the role of arXiv. For example, it could mean that for mathematicians, the final publication is made even if the results are made available to

the scientific community some years earlier. On the other hand, the standard peer reviewed publication would also be important for astrophysicists, but also the rapid presentation of the results facilitated by arXiv.

3.3. Measuring Impact of the Papers with Non-Standard Tools

The second part of our study is dedicated to analysing how to measure the impact of documents previously deposited in arXiv. Due to the nature of the documents—which are not considered as "citable objects" by WoS for the calculation of their impact—an alternative way of measuring the influence of the article other than the number of citations from WoS has been developed. The number of citations of all the documents has been considered in two different sources, which in a sense are complementary. For all manuscripts, citations were calculated in arXiv (provided by High Energy Physics information system). Then, two different procedures were applied, depending on whether the work was finally published or not.

1. If the preprint was finally published, the number of citations was found in the journal where it was published, or in Google Scholar instead in case the journal did not

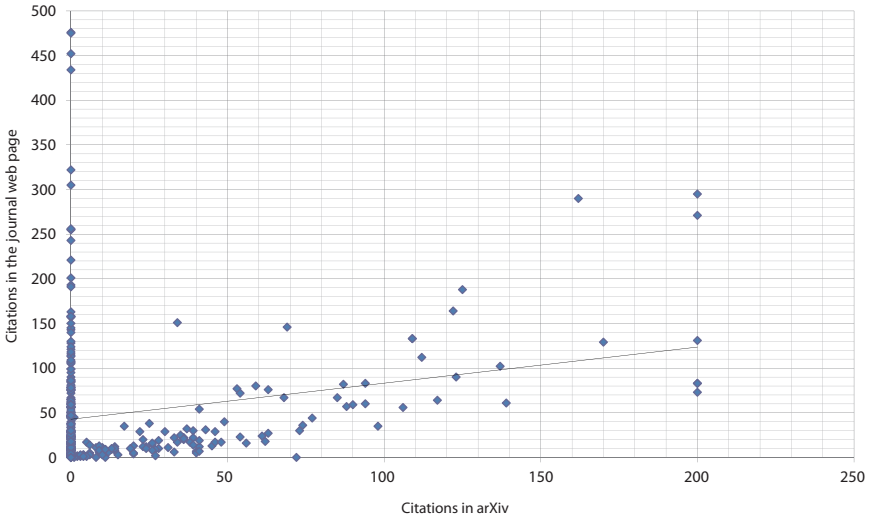

Fig. 1. Citations in arXiv (by documents in arXiv) versus citations registered in the website of the journal where the paper was finally published (Five points out of scale were removed for the representation).

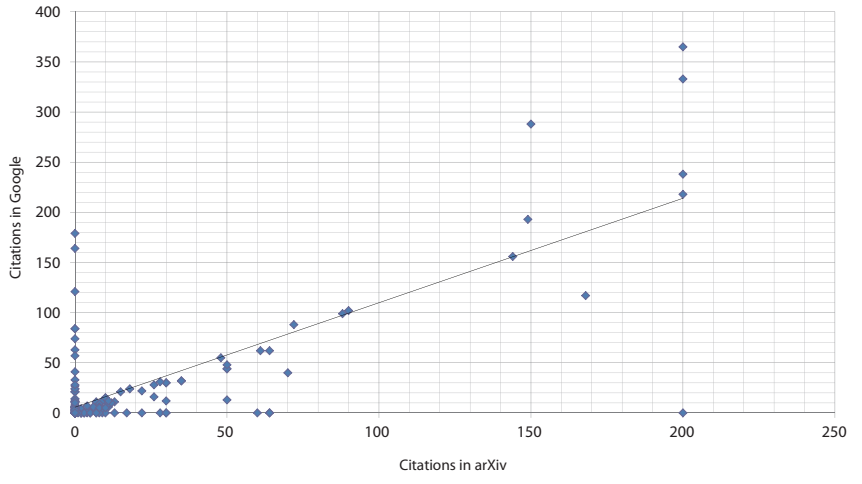

Fig. 2. Citations in arXiv (by documents in arXiv) versus citations in Google Scholar (Two points were removed for the representation).

provide this information. Fig. 1 shows the result.

2. If the article was never published, an external citation measure was used: the number given by Google Scholar. As explained in the methodology section, the reason for choosing this option is to provide an adequate measure of the citations that do not come from WoS, where only the standard published version is considered as a scientific document. The result can be seen in Fig. 2. This procedure differs from previous intensive studies on the subject that we have used as main references for our analysis.

Although a weak correlation may be observed in both cases, the large amount of papers on the axis suggests that the three citation computations are completely different. For example, there are many papers without citations inside arXiv that have a big citation rate when they are published in a standard journal, or in Google Scholar (see the axis OY in both Figs. 1 and 2). This reinforces our hypothesis about citation behaviors by the three methods considered—arXiv, journals' websites, and Google Scholar—and somewhat independent, and the reasons why authors use arXiv differ in each case.

4. DISCUSSION: HOW AND WHY RESEARCHERS USE arXiv

The standard interpretation of why researchers use arXiv is that it follows the essential aspects of a traditional form of research dissemination and scientific information exchange: e-preprints are the current version of the classic manuscript that was shared with colleagues to communicate research results (see for example the paper by Confrey, 1996; see also the references in the paper by Larivière et al., 2014). Material of this type (electronic preprints) should be understood as the same type, and contained in the broad category sometimes referred to as research manuscripts.

arXiv is mainly devoted to physics, mathematics, and computer science. In recent work (Larivière et al., 2014) it is shown that 64% of all arXiv papers are finally published in WoS. Previous studies have shown that the ratio of deposited manuscripts in arXiv with respect to total published papers in mathematics was 81% in 2010 (Fowler, 2011). Also 75% of the papers on physics of condensed matter are deposited in arXiv (Moed, 2007). A complete review of the existing literature on the high ratio of deposition in arXiv by physicists and mathematicians can be found in Li et al. (2015). Summing up, it can be said that arXiv provides a standard tool for prepublication and post-publication in those fields in which rapid communication is at least as important as the fact that the publication is peer reviewed. It should also be noted that in mathematics the backlog associated with the process of editing a paper is generally quite large, so the academic community is committed to disseminating its results in this way, although peer review is also seen as fundamental.

However, it seems that scientists in these areas know that citing an arXiv manuscript is not the best way to refer to a published paper, as it is supposed to have some sort of temporary value only until the last peer reviewed version is published. This would be supported mainly by the small rate of references to arXiv preprints that we have found in WoS. Also, some editors request that these references be updated in the latest versions of the paper if possible, and some even reject the original submission of papers containing such citations. From the authors' point of view, there is a conflict of interest regarding the balance between the rapid dissemination of a manuscript and the "quality" of citations to this manuscript.

Let us explain a suitable scheme of authors' motivations in a more detailed way, when the problem of the possible dispersion of cites is taken into account. It is based mainly on the analysis of the studies quoted above and our bibliometric data.

a. In the first scenario, the manuscript is supposed to provide a rapid communication of research results and no further publication is expected, or in case publication is done it is of secondary importance. Then, a big rate of citations in arXiv was expected if the citing papers follow the same rule, that is, if the scientific group interested in the topic considers arXiv as a primary and reputable source of relevant information. As already announced in Haustein (2016), a parallel system of scholarly communication is supposed to exist in this case, based on arXiv type documents. Since we have shown that the total amount of documents of this kind referred to in WoS is small, the system would work independently of the standard publication. The extreme cases of this expected behavior are the papers appearing in the OX axis of Figs. 1 and 2. From our personal experience as researchers, it must be said that some publishers refuse—explicitly or implicitly—to publish articles that refer to unreviewed documents.
b. In a second scenario, the paper is deposited in arXiv to ensure authorship of the research or rapid presentation of results, but this version is not assumed to be the final support for the investigation presented in it. Again

the extreme case would give no citations in arXiv and many citations in other sources—a journal website or Google Scholar. These are the manuscripts that appear in the OY axis of Figs. 1 and 2.

The final picture would be given by all the intermediate cases between the two extreme situations mentioned above. When an author considers depositing a paper in arXiv, the arguments supporting the decision may be, in a sense, a mixture of those explained. On the one hand, he or she wants to offer the result of his scientific work to the community as soon as possible, while ensuring his authorship. This would provide a long-term benefit—prestige, but could be dangerous in terms of the citation count. It has been shown how this prepublication would affect this count by producing "poor quality" citations to the arXiv document—from standard bibliometric measuring tools. On the other hand, the author may consider having arXiv as a permanent support for his results. Depending on the scientific area and the uses in each research community, each of these arguments becomes the main reason. But in many selection/evaluation processes, a paper in arXiv is not a paper—even if it has a hundred citations—and so prepublication could damage the professional career of the researcher since he cannot put it in his list of publications.

5. THE CURRENT SITUATION: AN INCREASE IN THE NUMBER OF NEW REPOSITORIES AND PREPRINT UPLOADS

We have focused our attention on arXiv because of its recognized position in the world of scientific publishing. However, a long list of new platforms for preprints has appeared, which have been consolidated in recent years. The reader can find in the 'researchpreprints' platform a list of repositories, in which it is easy to see that there are many new records (e.g., AgriXiv, ChemArxiv, ChinaXiv, LIA Scholarship Archive, and OSF preprints). Preprints publishing has also grown very rapidly in the last two years, making it increasingly convenient to analyze the role of scientific manuscript prepublication (see Lin, 2018). It seems clear that this practice benefits the authors in terms of dissemination of their work, but as we have observed in the present study, it also produces some dispersion of citations, against the interest of authors to the extent that this may harm their careers as researchers.

In any case, in view of the growing tendency to deposit preprints in repositories, it seems that authors are increasingly concerned about this alternative form of distribution of their research results. For this practice to be consolidated and useful, applying the main conclusions of our analysis seems urgent: A standard citation method and regulated bibliometric rules must be imposed for the evaluation of the research. Along with the crisis of the peer review system, the recognition of the value of all kinds of scientific material (including data, preprints, projects, etc.) seems to be the main current problem of the global scientific information system. We cannot expect these changes to be promoted by large publishing companies, as preprint publication could affect their business. Bearing in mind that the same companies that own the publishing houses are sometimes also owners of the bibliometric platforms, it does not seem that the changes will come from this part. This will probably be done by national research evaluation agencies or international bodies.

However, it seems that the consequences of prepublication for authors depend to a large extent on the field of research, and researchers generally know very well how this may affect their scientific activity. Therefore, the regulation of prepublication seems to be an issue that will depend on each particular field, although some standard rules should be imposed.

6. CONCLUSIONS

We have added some bibliometric explanation regarding the citation-based interaction between the standard publication world and preprint publication to the existing ones. Our aim was to understand the behavior of authors with regard to the prepublication of their scientific results. It seems that there are no universal trends that can explain this behavior, which seems to depend on each scientific specialty. This may be a consequence of: 1) the existence of prepublication rules implicitly accepted by all researchers only in local communities associated with specific scientific areas, as well as 2) the result of the lack of reliable specific tools for measuring preprint citations. The second aspect critically affects the evaluation of authors and conflicts with the benefits of preprint publication. We must say that these benefits are solid and have been proved in various works, and probably counteracts the loss-citation-problem analyzed in this paper. The small amount of references to arXiv papers in WoS that we have found supports the idea that the damage caused to authors by citation to arXiv preprints is, in any case, small.

Our work also confirms—although with information

collected from a small sample—the previous analysis proving the benefit of preprint publication. The direct relationship between the upload of manuscripts and rapid communication—which is not covered by the standard publication—would fit in with the results that we have found.

The existing literature on the topic shows that many researchers consider arXiv to be an autonomous network for scientific dissemination in some disciplines. We therefore believe that a rapid development of recognized tools to measure citations in the world of prepublishing is necessary to facilitate impact assessment. Although the evaluation of research is done using bibliometric indicators, measuring the impact of preprints seems to be the only way to support and reinforce the prepublication of manuscripts, especially if this is going to be the final form of publication of a relevant part of these articles. This issue is not only about the dissemination of science, but also about open science initiatives, as preprint publication is often free of charge. A new paradigm including preprints and other "non-standard" sources of information as valuable scientific documents for research evaluation is needed. This is already done by some national agencies for research assessment, but in other cases (such as Spain or Poland, for instance) standard bibliometric tools still play a fundamental role. A researcher's career is developed through a sequence of evaluation processes, at all levels. Some new tools—such as downloads of associated electronic files and other altmetrics—are beginning to be considered in these evaluations, but even so, citation count still plays a relevant role. Therefore, it seems natural to think that in the near future the consideration of non-peer-reviewed articles will also enter evaluation systems. In fact, the need for these new rules is evident, as preprint publication seems to be increasing exponentially, as we mentioned in the previous section. The impact of preprints can be measured in terms of, for example, citations or downloads, but our analysis suggests that some of the existing tools are not adequate yet. For example, both Google Scholar and arXiv have been shown to provide a citation counter of all documents appearing in any standard search, but the results obtained are somewhat random (Figs. 1 and 2). No uniformity is observed when different scientific areas or even individual works are considered.

ACKNOWLEDGMENTS

The work of the first, second, and third author was supported by Ministerio de Economía, Industria y Competitividad, Spain, under Research Grant CSO2015-65594-C2-1R Y 2R (MINECO/FEDER, UE). The work of the fourth author was supported by Ministerio de Economía, Industria y Competitividad, Spain, and FEDER, under Research Grant MTM2016-77054-C2-1-P. The authors would also like to thank the referees for their useful comments and references, which helped them to improve the work, especially in Section 5.

REFERENCES

Ardichvili, A., Page, V., & Wentling, T. (2003). Motivation and barriers to participation in virtual knowledge-sharing communities of practice. *Journal of Knowledge Management*, 7(1), 64-77.

Bar-Ilan, J. (2014). Astrophysics publications on arXiv, Scopus and Mendeley: A case study. *Scientometrics*, 100(1), 217-225.

Bohlin, I. (2004). Communication regimes in competition: The current transition in scholarly communication seen through the lens of the sociology of technology. *Social Studies of Science*, 34(3), 365-391.

Confrey, E. A. (1996). The information exchange groups experiment. *Publishing Research Quarterly*, 12(3), 37-39.

Davis, P. M., & Fromerth, M. J. (2007). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics*, 71(2), 203-215.

Fowler, K. K. (2011). Mathematicians' views on current publishing issues: A survey of researchers. *Issues in Science and Technology Librarianship*, 67. Retrieved November 2, 2018 from https://doi.org/10.5062/F4QN64NM.

Haque, A. U., & Ginsparg, P. (2009). Positional effects on citation and readership in arXiv. *Journal of the American Society for Information Science and Technology*, 60(11), 2203-2218.

Haque, A. U., & Ginsparg, P. (2010). Last but not least: Additional positional effects on citation and readership in arXiv. *Journal of the American Society for Information Science and Technology*, 61(12), 2381-2388.

Haustein, S. (2016). Grand challenges in altmetrics: Heterogeneity, data quality and dependencies. *Sciento metrics*, 108(1), 413-423.

Henneken, E. A., Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Thompson, D., & Murray, S. S. (2006). Effect of e-printing on citation rates in astronomy and

physics. *Journal of Electronic Publishing*, 9(2). Retrieved November 2, 2018 from https://quod.lib.umich.edu/j/jep/3336451.0009.202/--effect-of-e-printing-on-citation-rates-in-astronomy?rgn=main;view=fulltext.

Kim, J. (2011). Motivations of faculty self-archiving in institutional repositories. *Journal of Academic Librarianship*, 37(3), 246-254.

Klein, M., Broadwell, P., Farb, S. E., & Grappone, T. (2016). Comparing published scientific journal articles to their pre-print versions. In N. R. Adam, B. Cassel, & Y. Yesha (Eds.). *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries* (pp. 153-162). New York: ACM.

Kling, R., Spector, L., & McKim, G. (2002). Locally controlled scholarly publishing via the Internet: The Guild Model. *Proceedings of the American Society for Information Science and Technology*, 39(1), 228-238.

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., & Murray, S. S. (2005). The effect of use and access on citations. *Information Processing & Management*, 41(6), 1395-1402.

Kurtz, M. J., & Henneken, E. A. (2007). Open Access does not increase citations for research articles from The Astrophysical Journal. Retrieved November 2, 2018 from https://arxiv.org/ftp/arxiv/papers/0709/0709.0896.pdf.

Larivière, V., Sugimoto, C. R., Macaluso, B., Milojević, S., Cronin, B., & Thelwall, M. (2014). arXiv E-prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*, 65(6), 1157-1169.

Li, X., Thelwall, M., & Kousha, K. (2015). The role of arXiv, RePEc, SSRN and PMC in formal scholarly communication. *Aslib Journal of Information Management*, 67(6), 614-635.

Lin, J. (2018, May 31). Preprints growth rate ten times higher than journal articles. Retrieved November 2, 2018 from https://www.crossref.org/blog/preprints-growth-rate-ten-times-higher-than-journal-articles/.

Manuel, K. (2001). The place of e-prints in the publication patterns of physical scientists. *Science & Technology Libraries*, 20(1), 59-85.

Moed, H. F. (2007). The effect of "open access" on citation impact: An analysis of ArXiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, 58(13), 2047-2054.

Neuhaus, C., & Daniel, H. D. (2008). Data sources for performing citation analysis: An overview. *Journal of Documentation*, 64(2), 193-210.

Pinfield, S. (2005). Self-archiving publications. In G. E. Gorman, & F. Rowland (Eds.). *International yearbook of library and information management 2004/2005: Scholarly publishing in an electronic era* (pp. 118-145). London: Facet Publishing.

Swan, A. (2010). *The Open Access citation advantage: Studies and results to date*. Southampton: University of Southampton Institutional Repository.

Youngen, G. K. (1998). Citation patterns to traditional and electronic preprints in the published literature. *College & Research Libraries*, 59(5), 448-456.

Zha, X., Li, J., & Yan, Y. (2013). Understanding preprint sharing on Sciencepaper Online from the perspectives of motivation and trust. *Information Development*, 29(1), 81-95.