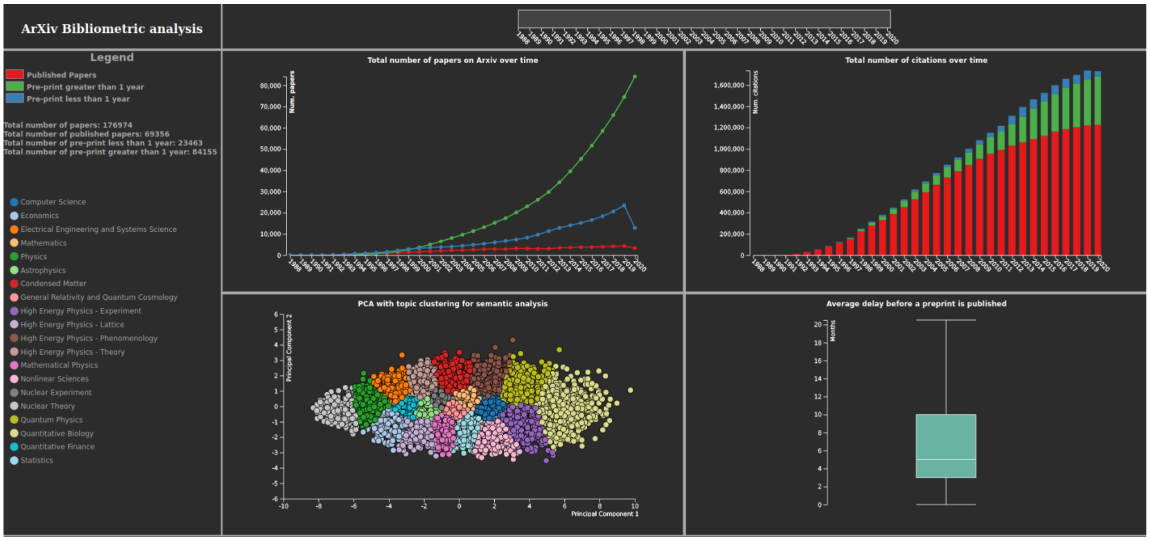
Arxiv Bibliometric Analysis

Giulia Muscarà, Claudia Medaglia, Francesco Molfese

Sapienza University of Rome



ArXiv Bibliometric analysis

Legend

🟥 Published Papers
🟩 Pre-print greater than 1 year
🟦 Pre-print less than 1 year

Total number of papers: 176974
Total number of published papers: 69356
Total number of pre-print less than 1 year: 23463
Total number of pre-print greater than 1 year: 84155

Figure 1: *Arxiv Bibliometric Analysis Visualization*

Abstract

Arxiv is a pre-print server born in 1989 as a High-energy Physics repository, allowing researchers to self-distribute scientific paper drafts. It was expanded to comprise astronomy, computer science, mathematics, quantitative biology, statistics, and quantitative finance. While some of these papers are never published in official journals, many are accepted for publication at academic venues after a double-blind peer-review process. Authors of these papers are faced with the decision to distribute their papers on Arxiv before or after acceptance at their target publication venues.

We will exclusively refer to papers posted on Arxiv, distinguishing pre-prints, those papers that have not been accepted yet, and published articles. The goal is to analyze the impact of this phenomenon onto the scientific community since its creation to these days, identifying the most affected areas.

The following questions will be addressed: what is the total number of published papers and pre-prints over the years? How many citations will the pre-print articles that are not likely to be accepted in the future get? Do pre-prints deposited in Arxiv for many years get more citations than articles that were already peer-reviewed? Does the delay between publication on Arxiv and official acceptance of a paper vary from area to area?

1 Introduction

It is certain that the Internet speeded up the publication process, which is likely to continue to quicken in the future. Arxiv, just as other pre-print servers, has the main goal of bringing the communication of research results forward to readers. For authors, this need arises from the usually long publication delays. Overall, there are both benefits and drawbacks of publishing a paper on Arxiv. On one hand, it allows to rapidly spread results across the scientific community, accelerating communication and the exchange of feedbacks, useful to improve the quality of the work. Then, it can help to avoid having to wait for the submission's response in order to share research results or a new idea, preventing the submission of double-works.

On the other hand, the peer-review process plays a fundamental role in checking the validity of a reasearch and cannot be put aside. Especially, it should remain a standard in some fields, where wrong papers could lead to serious consequences. Anyway, since the communication times are accelerating, this phenomenon is growing over time and apparently the benefits of publishing a pre-print paper on these platforms outweight the drawbacks.

A pre-print paper is a paper submitted for journal publication but not accepted yet, or uploaded for peer consideration but not submitted to any scientific journal yet. The present work is based on the distinction of three kinds of papers uploaded on Arxiv: published papers, those accepted after a peer-review process, papers in pre-print for less than one year and papers in pre-print for more than one year. The above division stems from the analysis of the average publication delay, from which it was possible to infer that papers in pre-print from more than one year are not likely to be accepted.

Moreover, only papers that were first published on Arxiv as pre-prints and then subjected to a peer-review process are taken into account in the present analysis. In particular, this assumption is relevant in the boxplot graph.

2 Dataset and pre-processing

The dataset used for the analysis was retrieved from a free, open pipeline on Kaggle to the machine-readable Arxiv dataset, a repository of 1.7 million articles. The dataset provides a metadata file in json format, with an entry for each paper, containing the following fields:

- ID: ArXiv ID (can be used to access the paper)
- SUBMITTER: Who submitted the paper
- AUTHORS: Authors of the paper
- TITLE: Title of the paper
- JOURNAL-REF: Information about the journal the paper was published in
- COMMENTS: Additional info, such as number of pages and figures
- DOI: Digital Object Identifier
- ABSTRACT: The abstract of the paper
- CATEGORIES: List of related topic tags in the Arxiv system in order of relevance
- VERSIONS: A versions history
- REPORT-NO: Report Number
- LICENSE: Copyright license
- UPDATE-DATE: Date of the last update
- AUTHORS_PARSED: Parsed names of authors

As it was previously stated, Arxiv comprehends papers related to scientific areas and in particular, this dataset refers to the following 20 main topics: *Computer Science, Economics, Electrical Engineering and System Science, Mathematics, Physics, Astrophysics, Condensed Matter, General Relativity and Quantum Cosmology, High Energy Physics - experiment, High Energy Physics - lattice, High Energy Physics - phenomenology, High Energy Physics - theory, Mathematical Physics, Nonlinear Science, llear Experiment, Nuclear Theory, Quantum Physics, Quantitative Biology, Quantitative Finance, Statistics.*

Given the big dimension of the dataset and the fact that it contained superfluous information that was not relevant four our work, it was subjected to a phase of pre-processing, where the fields ID, SUBMITTER, COMMENTS, TEXT, ABSTRACT, LICENSE, AUTHORS, REPORT-NO and UPDATE-DATE were removed. However, the field DOI was kept in order to be able to retrieve the number of citations of a certain paper.

Even after removing these fields for each entry of the dataset, its dimension was still not feasible to be processed for visualization. This is why we decided to sample a representative subset of the dataset such that the portion of papers in a given area and year was the 10% of the original.

Each paper was associated to the area represented by the first topic in its categories list, precisely because it is the most relevant. As far as the publication date on Arxiv is concerned, it was extracted from the creation date of the first version in VERSIONS.

3 Visualizations and interactions

The visualization fits in a single page and comprehends four graphs, showing data in a coordinated and coherent way. Indeed, a filter on the Arxiv's topics allows the user to select and deselect a subset of them from the list on the left. After the selection, every visualization will display information about the papers related to the selected topics.

The legend on the left represents the three different kinds of papers: published papers (in red), papers in pre-print from less than a year ago with respect to the considered year (in blue) and papers in pre-print since more than one year (in green). It is referred only to the line chart and to the stacked bar chart, as they are the only graphs to make this distinction among papers. Analogously, a zooming brush in the top area of the page allows to choose a smaller range of years and to visualize the restricted results in both graphs.

On the left four counters are also indicating the total number of papers visualized at any moment and the total number of papers of each category.

3.1 Line Chart

The first element of the visualization is a line chart representing the total number of papers on Arxiv in time. The graph shows quantitative - quantitative information: on the x axis, the range of years going from 1988 to 2020 is represented, while the y axis keeps track of the total number of papers related to the selected topics.

To plot the red line, only papers that were accepted were considered. For every year, the total number of papers published in that year was plotted.

To plot the blue line, for each year only papers that were published on Arxiv less than a year before the considered one were plotted.

Finally, for the green line only papers that had been in pre-print on Arxiv for more than one year were plotted in a given year. Obviously, the total number of papers in pre-print for more than one year will grow progressively in time.

On this chart, four interactions can be triggered by the user:

- For each year, when passing the mouse over one point of a line, a tooltip shows the number of papers of that kind for that year;
- As stated above, it is possible to plot only information about the papers belonging to the selected subset of topics;
- As stated above, it is possible to zoom over a sub range of years;
- By clicking on an entry in the legend, the corresponding series in the line chart will appear or disappear.

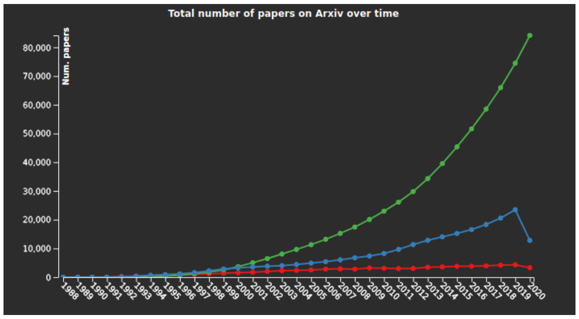


Figure 2: *Line chart*

3.2 Stacked Bar Chart

The stacked bar chart displays quantitative - quantitative information about the the total

number of citations over the years, for each of the three types of papers specified in the legend. On the x axis, the range of years going from 1988 to 2020 is represented, while the y axis keeps track of the number of total citations of papers related to the selected topics.
Given that the original dataset did not contain any information about citations, the DOI was used to retrieve the number of citations of a certain paper from an external website. This number was summed to the total number of citations of papers in the same category.
Also on this chart, four interactions can be triggered by the user:

- For each year, when passing the mouse over one section of a stacked bar, a tooltip shows the number of citations of papers in that category;
- As stated above, it is possible to plot only information about the papers belonging to the selected subset of topics;
- As stated above, it is possible to zoom over a sub range of years;
- By clicking on an entry in the legend, the stacked bar chart will be recomputed considering only the categories of papers chosen in the legend.

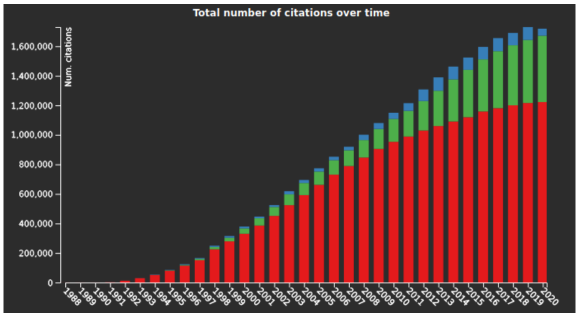


Figure 3: *Stacked bar chart*

3.3 Boxplot

In the boxplot, only published papers (those with a non-empty journal-ref field) were taken into account, such that they were published only after being posted on Arxiv as preprints. The graph shows the average delay in time between their creation on Arxiv and their acceptance after a peer-review process.
On the axis, time is measured in months. In this case, the following interactions can be triggered by the user:

- When passing the mouse over the box plot, a tooltip shows the median value of the time, the first quartile, the third quartile, the minimum value of the distribution and its maximum value;
- As stated above, it is possible to plot only information about papers belonging to the selected subset of topics;

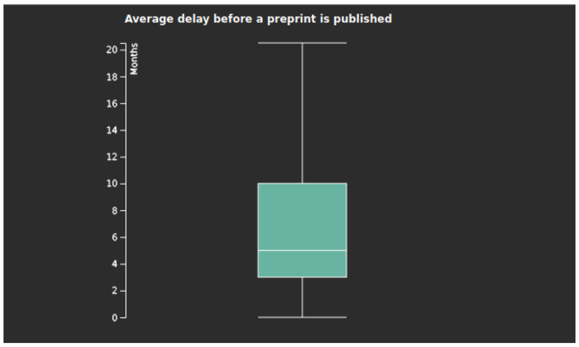


Figure 4: *Boxplot*

3.4 PCA Scatterplot

The scatterplot shows a set of clusters, each associated to one of the selected topics, computed through K-means after applying PCA. PCA was executed over a subset of the attributes: the title of the paper, the names of the authors, the first topic of the categories' list. Each paper, after PCA, will be represented by a two-dimensional point, so as to plot it in a bidimensional space.
K-means algorithm was then applied on the projected points. The algorithm will produce a number of clusters equal to the number of selected topics.
For the scatterplot, the following interactions can be triggered by the user:

- As stated above, it is possible to plot only clusters containing the papers belonging to the selected subset of topics;

- When passing the mouse over a cluster, a tooltip shows the associated topic and the total number of points in that cluster;
- When clicking on a point of a given cluster, all the other visualizations will be recomputed so as to represent only the papers related to the topic of the selected cluster.

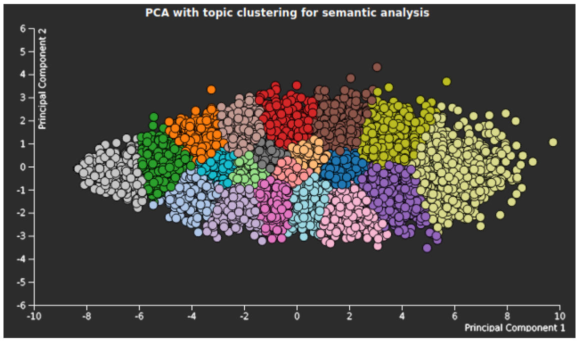


Figure 5: *Scatterplot*

4 Related works

Since pre-print servers like Arxiv became popular, many researchers focused their works on the impact of this new way of sharing knowledge, that influences both the world of research and the academical world. Pre-prints enable not only an unlimited and free access to relevant information both for students and researchers, but they also allow the convenient dissemination of results. Students often turn primarily to Arxiv to find the most up-to-date papers, before considering published journals. Nonetheless, the final publication in a journal is still inevitable and common practice for most researchers.

Given that the first graph's analysis is based on measuring the evolution of Arxiv, it was compared to related papers that concerned this aspect. In particular, accordingly to the studies of Larivière et al. [1] and Aman V. [2], the website is becoming an increasingly popular venue for authors to share their work, which is indeed reflected by our trends. An analogous result was proved for other similar pre-print repositories, such as Biorxiv, as reported by Abdill R. et al. [3]. The matter of what is the area of gretest interest in terms of number of uploads was questioned as well. For instance, the latter paper shows that the topic neuroscience collection has had more submissions than any bioRxiv category in every month since September 2016. Indeed, for Inglis J. and Sever R. [7], evolutionary biology, bioinformatics, and genomics were attracting the largest proportions of the total until 2016.

As far as Arxiv is concerned, according to Aman V. [2], the highest share is in High Energy Physics field, where more than 90% of journal articles can be found as pre-print versions. Astrophysics and mathematics follow with a share of 50% to 60%, whereas researchers publishing in Quantitative Biology do not yet make excessive use of Arxiv. Also, mathematicians are making a transition to publishing pre-prints because of the longevity of mathematical journal articles and the long publication delays.

A significant factor in the evolution of pre-print servers was the Covid-19 crisis. During the first months of 2020 medical pre-print servers saw an increasing number of submissions due to the need of sharing data on the pandemic. As far as Arxiv is concerned, "the total number of publications in the q-bio field had almost doubled from January 2020" according to Aviv-Reuven A. and Rosenfeld A. [9]. Indeed, the quantitative biology topic is the only one in Arxiv related to medicine.

Among the others, these papers address whether it is beneficial or not for authors to share their articles before they have been accepted for publication. Generally, they all agree that pre-print repositories let users increase their visibility by opening their results to the community, so that they can be accessed, used, and cited. Authors can bypass the longsome publication process and get their results published as early as possible. However, there is the risk of spreading incorrect data that could be referenced in other works. Larivière et al. [1], observing that about 50% of Arxiv submissions were also findable in WoS, raised the problem that people may not be willing to pay for journals with such great amounts of free material online. Anyway, "publishers should not fear Arxiv as long as they are aware of the value-adding role of peer review" [2], in which case the benefits of posting even incomplete work outweights the downsides.

Just like the previous analysis shows, the average delay of the publication of a pre-print paper was also investigated and it was the focus of many others as well. An interval similar to our results was found from Larivière et al. [1] showing pre-prints on Arxiv were most frequently published within a year of being posted there. Specifically, in this paper the data computed with the whole of the Arxiv database show that the works related to mathematics, statistics and nonlinear sciences have a publication delay of more than 1.4 years. On the other hand, the value estimated by Ferrer-Sapena A. et al. [6], is higher (2.5 years). For Aman V. [2], whereas in High Energy Physics and Astrophysics it takes two to six months for a paper to get published, one to three years can elapse in mathematics.

Moreover, the comparable study of Serghiou and Ioannidis [4] examining bioRxiv pre-prints found that “the probability of publication in the peer-reviewed literature was 48% within 12 months” and according to Abdill R. et al. [3], among pre-prints that are eventually published, 75% have appeared in a journal after approximately 8 months on bioRxiv, with a median delay was of 166 days. The contrast with the median delay time found by Fraser N. and Momeni F. [5] of 154 days, can likely be explained by the different points of publication used. In general, it reveals that pre-prints effectively shorten the time to public dissemination of an article by 5-6 months compared to the traditional journal publication route.

Finally, in Inglis J. and Sever R. [7] work, the interval between a manuscript’s appearance on bioRxiv and its publication in a journal may exceed 400 days, with a median interval is 134 days.

Aman V. [2] demonstrated furthermore that the publication delay between the submission in Arxiv and the publication in a Scopus-indexed peer-reviewed journal has not significantly decreased over the years. In fact, given that over the years authors started to post on Arxiv more frequently, requests for acceptance increased along with acceptance times. However, this may depend also on the subject field, the journal publisher, the peer-review process, the quality of the submitted paper, and the need for fast publication. In the end, if the journal publication delay were not as long, Arxiv would not have been so powerful.

Another relevant estimate was that of the number of citations of both pre-prints and published papers.

Being easily accessible, the former get cited more frequently: they are available sooner in time and thus are likely to receive a higher amount of citations in the end.

Basing on the findings of Aman V. [2], 69% to 84% of pre-prints in High-Energy Physics receive their first citation prior to publication, while in astrophysics only a quarter of all pre-prints do. Pre-prints in both topics can accumulate many citations before they are published as a journal article. It was also shown that papers published in astrophysical journals with a citation window of two years received on average three more citations by the end of two years, if they had a previous pre-print in Arxiv.

Finally, Feldman S. et al. [8], observed that papers submitted to Arxiv before acceptance have, on average, 65% more citations in the following year compared to papers submitted after.

As a matter of fact, Arxiv contributes to gaining more citations since papers are available earlier (Ferrer-Sapena A. et al. [6], Fraser N. and Momeni F. [5], Serghiou and Ioannidis [4]). In this way, readers can start citing pre-prints from the beginning. This has not always a positive effect, as incorrect assumptions could be considered valid even if they are not proved to be true yet.

In general, all the cited works shared goals with our study and, even if they could differ in the way the objectives were pursued, it was still insightful to compare results. The main difference is that none of them proposed interactive visualizations, but only static graphs or tables. Despite the common subject of study, this work proposed an innovative and intuitive approach to visualize and understand data more easily.

5 Conclusion

Conclusions of our analysis agree with the results presented by considered related works.

To start with, the average publication delay of pre-prints among all years was computed. From the results, it was possible to infer that after one year, it is not likely that a pre-print will ever get published. So, in the first two graphs, we decided to compare analytics distinguishing between papers in pre-print for less and more than one year. Anyway, after trying to plot one boxplot per year, we noticed that the median value for the delay did not vary significantly from one to another, as it was found by related works as well. This is why our visualization shows a unique boxplot.

Analyzing results of the boxplot, papers related to mathematics are those affected by a higher publication delay, while High Energy Physics and Astrophysics ones are those with the lowest delay, coherently with the above-mentioned related works.

Another point of agreement with other researchers was that one of the reasons why the delay did not decrease in time may be due to the growing number of submissions on Arxiv. Indeed, in our first visualization it can be deduced how the number of both papers in preprint for more than one year and papers in preprint for less than one year posted on Arxiv is growing every year. In 2020 the number of both pre-print and published papers posted on Arxiv has decreased. This may be due to the impossibility to pursue research projects at the usual pace caused by the pandemic. However, the quantitative biology field was the only one to witness an upward trend, being the most affected area in the last year.

Also, it was interesting to discover when submissions in an area started to catch on with respect to other topics. For instance, articles related to Nuclear Experiment and Theory started to be posted more recently than those in the High Energy Physics' field. Other areas like Computer Science are characterized by a strong growth in recent years, despite related papers started being posted on Arxiv even before than that.

However, the number of published papers has not a steady pace and does not grow as fast as the number of submissions. In the end, from the analysis carried out on both visualizations follows that posting the pre-print of a paper on Arxiv does not speed up the peer review process.

As far as citations are concerned, it was interesting to compare the number of citations of published papers to that of pre-prints. The former is neatly higher than the latter, just as expected. However, the number of citations of papers that are not likely to be accepted anymore should be taken into account to notice that there may be too many works founded on papers being either non-validated yet or rejected after submission, whose thesis may be wrong with high probability. Consequently, basing a study on these unreliable assumptions may cause problems in especially critical fields.

Finally, the field with the highest impact on Arxiv turned out to be Mathematics, followed by Computer Science and High Energy Physics.

References

[1] Larivière, V., et al., *arXiv e-prints and the journal of record: An analysis of roles and relationships*, School of Technology, University of Wolverhampton, 2014.

[2] Aman, V., *The potential of pre-prints to accelerate scholarly communication*, Humboldt University of Berlin, https://arxiv.org/ftp/arxiv/papers/1306/1306.4856.pdf, 2013.

[3] Abdill, J., et al., *Tracking the popularity and outcomes of all bioRxiv pre-prints*, University of Minnesota, https://www.biorxiv.org/content/10.1101/515643v2.full.pdf, 2019.

[4] Serghiou, S., Ioannidis, J. P. A., *Altmetric scores, citations, and publication of studies posted as pre-prints*, Journal of the American Medical Association, 319(4), 402, https://jamanetwork.com/journals/jama/fullarticle/2670247, 2018.

[5] Fraser, N., Momeni, F. *The effect of bioRxiv pre-prints on citations and altmetrics*, Leibniz Information Centre for Economics, Kiel, Germany, 2019.

[6] Ferrer-Sapena, A., *Citations to arXiv Pre-prints by Indexed Journals and Their Impact on Research Evaluation*, Journal of Information Science Theory and Practice, http://koreascience.or.kr/article/JAKO201810760747187.page 2018.

[7] Inglis J., Sever R., *bioRxiv: a progress report*, Cold Spring Harbor Laboratory, https://asapbio.org/biorxiv 2016.

[8] Feldman S.,et al., *Citation Count Analysis for Papers with Pre-prints*, Allen Institute of Artificial Intelligence, Seattle, WA, https://arxiv.org/pdf/1805.05238.pdf 2018.

[9] Aviv-Reuven, A., Rosenfeld, A., *Publication Patterns' Changes due to the COVID-19 Pandemic: a longitudinal and short-term scientometric analysis*, Department of Information Sciences, Bar-Ilan University, Israel, https://arxiv.org/pdf/2010.02594.pdf 2021.