2020

Bibliometric Mapping of Scholar Publishing in Physics: Exploratory Study

Anna Novotna
Institute of Computer Science, Silesian University in Opava, anna.janikova@gmail.com

Jan Novotny
Institute of Physics, Silesian University in Opava, jan.novotny@fpf.slu.cz

Kamil Matula
Institute of Computer Science, Silesian University in Opava, kamil.matula@fpf.slu.cz

Bibliometric Mapping of Scholar Publishing in Physics: Exploratory Study

Anna Novotna[1], Jan Novotny[2], and Kamil Matula[1]

[1]Institute of Computer Science, Faculty of Philosophy and Science in Opava, Silesian University in Opava, Opava – Czech Republic

[2]Institute of Physics and Research Centre for Theoretical Physics and Astrophysics, Silesian University in Opava, Opava – Czech Republic

27th of November 2019

Keywords: academic collaboration, bibliometric mapping, co-authorship, publication team, scholar publishing, exploratory study, data analysis

Disciplines: library and information science, visualization

Abstract: The focus of this study is quite narrow, still we strove to keep our ways of exploring the data extensive to thoroughly examine them. We were not led by specific hypotheses but we tried to find dependencies and interesting points which could be examined further in bigger scope with more data.

We investigated preference of journals based on publications. We also explored publishing groups and their core journals. We also examined quantity of topics in which the scholars publish and their behavior in terms of their uploads of preprints to arXiv.org before the paper gets officially published.

There were some conclusion drawn from these explorations. Based on our data it seems that preference of the journal lies in the scope or preference of the group the author publishes with not in personal preference. Another question we addressed was concerned with how much are scholars involved in more than one topic. It seems that scholars in our dataset concern themselves with 2 to 4 sub-disciplines of physics or disciplines immediately connected to their scope. Out of our next investigation we conclude that physicists are very motivated to keep their record in database arXiv.org up to date. We can assume it is to link their published articles with preprints to acquire more citations. We also tried to estimate the length of publication process in journals in our dataset. We estimate that in our journals (with one exception) the publication time from acquiring review till final publication takes less than 200 days.

Creation of this paper was supported by the Institute of Physics and Research Centre of Theoretical Physics and Astrophysics, at the Silesian University in Opava as well as Development project "Support of International Mobility of Academic Staff and Strengthening of International Ties of FPF 2019".

1 Introduction

There are many ways how scholars communicate, among those publishing articles is the one most common and in many science fields most respected. Achievements among scholars are defined not only by good work and good results[1] it is also a question of good publication policy.[2] In theory publishing in more cited scientific journal should bring more estimation to the author. Still the question where to publish a paper is not easy. Authors have to choose a perfect journal not only based on their bibliometric indices but also based on scope of the journal, their publishing history and so on. Beside other considerations there is also a question of publication in Open Access journal and a need to avoid predatory publications.

We can agree that in reality the question of good publication politics signifies a complex problem. We can also say that it is a question most pressing because scientific grants and projects are usually directly adjoined to publication of scientific results, total count of publications and sometimes even to quality of journals designated via bibliometric indices.[3] It is therefore important to carefully evaluate publication policy of university, faculty, department or even a scientist[4]. To not only keep records of publications but evolve the publication policy to its best state.

The question where or what to publish lies also on a leader of a team of scholars. To help with publication problem, some of the bibliometric scholars from universities, libraries and elsewhere have prepared tools for scientists to find potential coworkers[5, 6], research trends[7], research preferences[8], matching journals[5] or gaps in the topic that could be covered by future research.[1]

There are customs in various academic fields expressing which of the journals, conferences, workshops or scientific venues are more valuable than others. These customs also inflict publishing behavior of scholars.[9] Usual investigation of scholar publishing by bibliometric studies was done based on country[10], age of a scholar[9], academic institution[11] or based on discipline. Studies covering specific publishing behavior was usually reserved to some journal from the field.[12, 13, 14, 15, 16]

In literature also a way of investigating the publishing behavior differs a lot. Some of the scholars address the problem through methods of social sciences[10], some use bibliometric methods[15], some use more matematical approach[17] and some addressed the problem in data exploratory fashion.[8]

Practice which is normally followed in bibliometric studies begins with hypothesis which helps the researcher to narrow down the scope and follow the chosen method. Although we use previous studies conversely we decided to start in a different way. We chose an exploratory approach because we want to vary the methods of investigation. On the other hand we need to narrow down our scope as well. That is the reason we decided to investigate publishing behavior of relatively small group of scholars from one institute. The benefit of this is a close proximity to the group which helps us to add a lot of context to the study. We can also say which of these analysis are beneficial to the group. In this way we can also reveal some domain specific behavior and at the same time reflect on overall implications.

2 Dataset

We are using data of 32 scientists from the field of physics. The number of investigated unique articles is 190. All of the researchers are working for one institute even though they all do not work there for the same time. For the time period of our investigation we chose years from 1991 till 2019. We are using data from Web of Science (WOS) and from arXiv.org, because both of them represent domain specific practice in publishing. For physics only publications listed in Web of Science have any scientific value and at the same time most of the scholars upload their

preprints to arXiv.org which represents a good practice. Those 190 articles do not represent a total number of articles from WOS or arXiv.org, it is a collection of articles found in both databases and linked together. Articles which we were not able to find in both databases or we were not able to link them together were not taken into account.

Selected scholars represent the main group of academics working for the institute. Still not everyone is working for the institute continuously for years. Some scholars are there from the beginning, some started later. Papers published while away were also taken out of selected data. Also all of the data in figures were anonymized to avoid complains.

We downloaded the text data from WOS and in case of arXiv.org we used their API constructed for queries.[19] In both cases data desperately needed cleaning as we discovered that both of the sources are quite noisy. In time we filtered out all the false positive authors and all of the papers we were not concerned with and got the final number that we worked with.

Unfortunately arXiv.org API for queries is not flawlessly constructed and it took us quite a while to match 190 articles in both databases and link them together. In arXiv.org there is no author identification[18] and often any other information beside article name is missing.

For data preparation and data cleaning we used software R[20] and R Studio.[21] Packages helping with data processing were aRxiv[22], visNetwork[23] and ggplot[24].

3 Visualizations

First of all we wanted to find out how much scholars published in the time period. It would also give us a rough idea when scholars started to work for the institute or better said when they have started to work scientifically. That is shown by figure number 1. On horizontal axis is time, on vertical axis is ID of an author. The size of a circle refers to number of publications in period.

In the figure can be seen that out of 32 only 24 scholars were active in publication. Also it shows that author no. 23 is very active in publication, that he was the one of the first people who started to publish for the institute and that he is also the most productive one. This author published an article at least once in two ears and it can be said that frequency of publication speeds up. Similar tendency can be seen in publication frequency of author number 24 even though his record is not so dense. It can be also said that some of the authors are very steady in their publications, some of them are much more variant but beside that nothing too specific can be said about the publication trend.

Now let us investigate where the scholars published, if they have a preferred journal or group of journals. If we take into account number of papers published in a journal and sort them according to the highest value we can acquire in figure number 2. We can see in the figure that preferred journals are Physical Review D, European Physical Journal C, Astronomy & Astrophysics, Classical and Quantum Gravity and so on. Still this ranking can be a bit misleading, because we know that the number of articles written by authors is not the same. This figure shows total number of published articles not scholar's preference. What we will explore now is a personal or professional preference of scientific journal of a scholar.

We can express this preference by visualization in figure number 3. Each of the columns represents an author. To find the most preferred journal by the scholar we need to rescale the data using min-max normalization. In this figure we can see an overall trend. It could be looked at in two ways, trend by journal or trend by scholar. In the first scope the preferred journal for publication in accordance with figure number 2 is still Physical Review D. However on the second place is Astronomy & Astrophysics and there are 3 candidates for the third place which differ based on selection of author.

Another scope we want to investigate are publication teams. We would like to know how

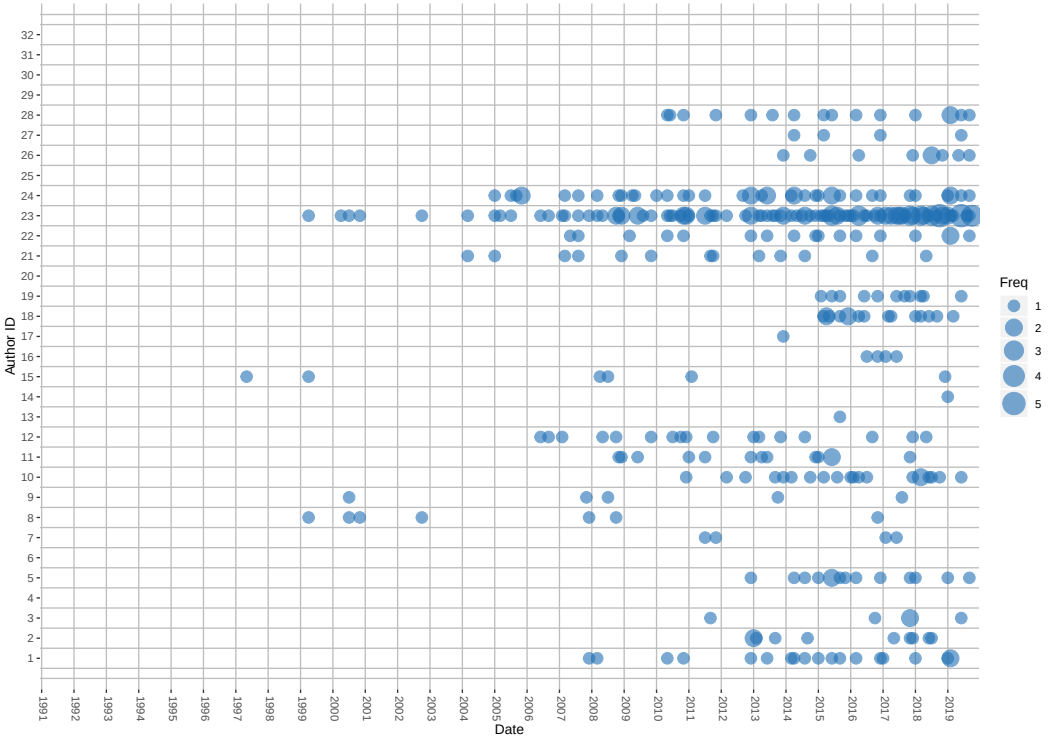

Figure 1: Number of publications by authors in years.

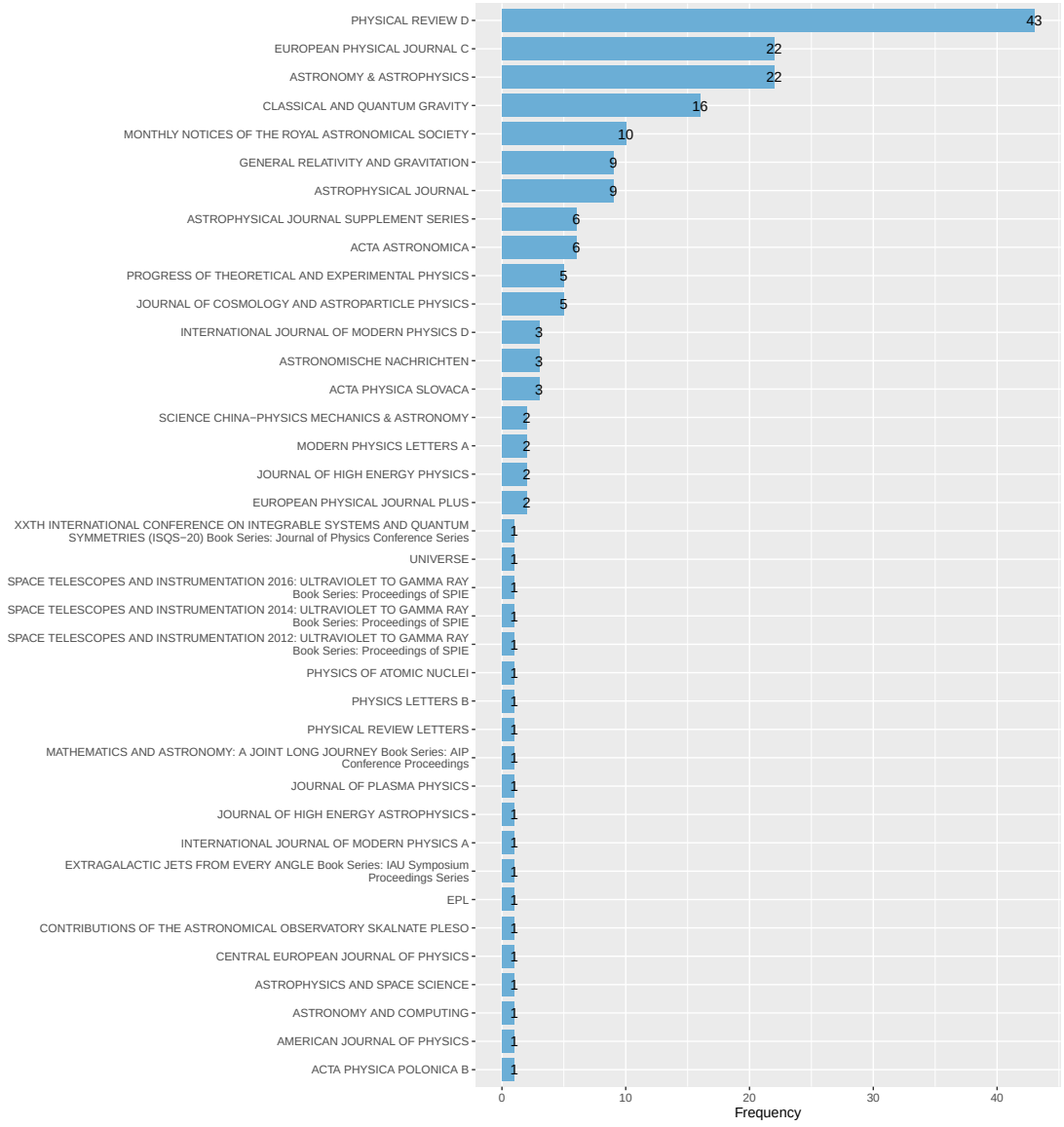

Figure 2: Number of papers published in a particular journal.

Figure 3: Journals preferred by authors for publication.

intertwined the publication net is and how far teams are one from the other. We created a network shown by figure number 4 which shows a net of co-authorship. Blue dots represent stable workers at the institute, yellow dots represent co-authors from outside of the institute. The figure shows that there are some scientists at the institute who publish mainly with people outside of the institute but the majority of the people at the institute publish with other members. The big yellow flower-like structures, which have only one connection to the people from the institute are co-authors from big publishing groups who publish either some observational data or data from particle accelerators.

In the figure can be seen several people residing in the middle of publishing groups. These are working as publication hubs, they connect publishing teams by their well established collaborations. Proximity of the blue dots either from each other or from other groups represent the amount of publications that the authors wrote together. We can also see some detached teams, or some very close to one another. We can find small publication teams as well as large ones. On the edge of the figure can be seen 8 blue dots which represent authors without publications in our scope.

The similar way how to demonstrate the ties of authors to journals is shown by figure number 5. We left out all of the outside co-authors and tied together journals with scholars who published in them together. It illustrates strong ties of the author number 23 to some scientists and weaker to others. Author number 23 represents founder of the institute and therefore person who works for the institute for the longest time. The graph shows strength of the relationships between authors as well as strength of relationship author–journal. The figure also and suggest division of publication groups.

Some of the journals are more strongly incorporated into the net of authors like Physical Review D and Astronomy and Astrophysics, some are more on the edge of interest like Physics of Atomic Nuclei. Still sometimes these journals represent core of interest of certain publication groups. For example journal Progress of Theoretical and Experimental Physics or Journal of Cosmology and Astroparticle Physics.

Figure 4: Publication teams shown in a net of publications.

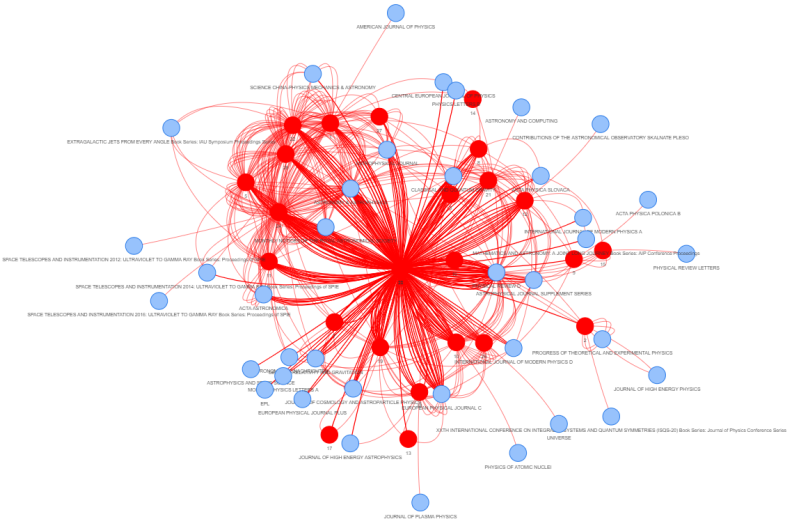

Figure 5: A net of authors who publish together with journals they are using.

If we put together figure number 3 and figure number 5 from the scholars point of view preferred journal reflects the scope of scholar's expertise or area of expertise of the co-author group. Personal preference of the journal is a bit aside. There are some exceptions of course. First of them is author number 1, who's publication activity is dispersed between several journals. It could be caused by co-author's preferences or changing area of interest. For example author number 5 is also a bit dispersed between the journals but it is caused by frequent co-authorship with number 1. The scholar which is most dispersed between the journals is author number 23. The reason is he works with many co-author groups and his scientific record is also the longest. Out of 190 papers included, he was listed as a co-author in 130. Reason for disperse of publication activity in bigger number of journals of author number 10 and 28 is membership in two co-author groups which both prefer different journals. Another dispersed author is author number 18 who is a member of co-author group outside of the institute and came to the main team later. That means he is also influenced by both author groups, the one he was member before he came and the one from the institute. The last author who is dispersed between the journals is author number 12. He is a member of 3 main publication groups, out of these groups two are outside of the institute and all of the groups have their own preferred journals.

In the next part of our paper we will investigate author's involvement in various topics. Data about journals and their topics were taken from WOS. WOS offers similar graphs of scope dispersion of authors between journals and topics but as we said earlier the data are quite noisy. There is no possibility to filter out only authors from one institute and authors not matched only by name. Nevertheless data used to construct figure number 6 are clean. The graph shows how much is each author concerned with multiplicity of topics. The whole circle represents all of the publication activity of that author in the dataset. Each topic is represented by different color. Each circle represents ratio of multiple topics in each author's publications. The figure shows that with some exceptions the main topic in which everyone publishes is *Astronomy and astrophysics*. Another area in which everyone publishes at least to some degree is *Physics, particles and fields*. 10 of 24 authors published also in *Quantum science and technology*, and 17 out of 24 publishes in field marked as *Physics, multidisciplinary*. There were also some exceptional publications from fields of *Education, scientific disciplines*, *Nuclear physics*, *Physics, fluids and plasma*, *Mathematical physics* and *Computer science*.

The results based on figure number 6 are that all of the scholars are involved at least in 2 topics (sub-disciplines in most cases) and maximally with 4. Out of these topics 2 are non physical which is *Computer science* and *Education, scientific disciplines*.

Because we know that the main scope of our scholars is physics the results do not surprise us but we must be wary of one flaw in results. We took into our dataset only publications noted in WOS and arXiv.org both. During working with the dataset we found that one of the scientists was publishing two papers in co-authorship with paramedics in medical science. These papers had no chance of appearing in arXiv.org as arXiv.org registers only physics, mathematics, computer science and to some degree quantitative biology, statistics, economics and electrical engineering and systems science.

The reason why we dwell on connection of the data from both of the data sources is this. We wanted to investigate submit-publish behavior of the authors. The custom in the field of physics is to post an article on arXiv.org at an initial stage, before it gets reviewed. After corrections from review there is a habit to update the file in arXiv.org before final corrections after which the paper gets published. We matched the dates of submission and print in both databases and found kind of behavior shown by figure number 7.

For a histogram in the figure number 7 to show the data right we needed to do some data cleaning. ArXiv.org was founded in 1991[25] but it was not used so extensively right away. Therefore some of the papers were submitted quite some time later. Also some authors were not

Figure 6: Authors and their involvement in topics.

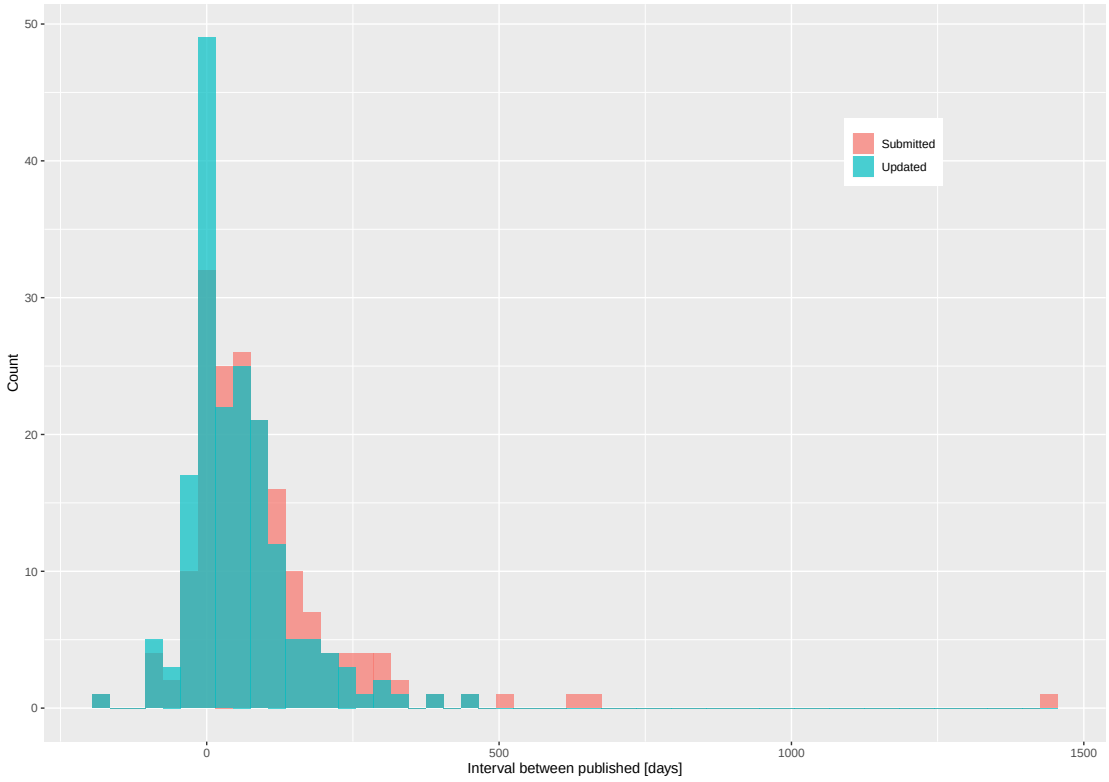

Figure 7: Histogram showing difference in time of publishing in a journal and preprint update on arXiv.org.

too rigorous in uploading the paper to the database. These were exceptions we needed to leave out.

In the figure number 7 x-axis represents interval between date the paper was firstly uploaded to arXiv.org and the date being published according to WOS. One bin represents 30 days. Y-axis represents occurrences of the bin. Red color stands for first submission to arXiv.org, the aquamarine color stands for last update of the paper in arXiv.org before publication.

Bins from zero to the left signify that the paper was uploaded to arXiv.org after being published. This behavior could be caused by journal policy or author decision. The highest peak corresponds with bin at zero. It means that in the interval between 0 to 30 days from official publication there was the highest occurrence in both updates and publishing. More interesting are data after first bin. There the red color represents longer waiting time for a response from the journal. The elevation of the aquamarine color above red in the bin zero signifies that after the author got the information from the journal about his paper being published, he updated that paper in arXiv.org.

The elevation of the aquamarine color in the bin zero and in one bin left from zero also signifies that most of the authors updated their papers on arXiv.org up to 2 months from publication. We can say that the overall tendency of physicists is to keep their papers on arXiv.org updated. It could mean either to keep the paper itself up to date or keep the metadata about the paper up to date for possible citation.

The difference between date updated and date published can lead us to the length of review period. We expect that update in arXiv.org was done due to review results that the author received. Hence we took the dates of update and official publication and matched them together in figure number 8. It is a representation of time between update of an article in arXiv.org and its official publication in a journal.

Whiskers in boxplot in figure number 8 show maximal and minimal amount of time between update in arXiv.org and publication date. The edges of a purple bar represent one 25th and 75th percentile of the distribution which means first and third quartile. Black line inside the bar denote median. Black dots in the graph mark outliers. If there is a black line without a bar it means there is too little data. It could signify only one or two papers published there. Before we plotted the graph we took out papers in proceedings because the publication process in this case can work in a different way.

Out of the figure number 8 could be assumed that the biggest variability in update–publish dates were recorded with journal General Relativity and Gravitation. On the other side of the spectrum smallest variability in time could be seen with journal Astronomishe Nachrichten. If we skip the records with too little data it would be Classical and Quantum Gravity. Still the overall trend (except for journal General Relativity and Gravitation) is that since update in arXiv.org the article is published in less than 200 days.

4 Discussion

What the dataset primarily shows is that the team of the authors who scientifically publish at the institute is growing. Also overall frequency of submitted articles is growing steadily. We have also found out how much popular scholars from the institute find each journal covering the field of physics. The main scope of the institute is published in Physical Review D both by the number of papers published in there or by the professional preference. Another differ by an author or better said by the group of authors. We have also found that the co-author group is the main deciding factor in terms of journal preference.

We also tried to investigate what is the scope of the researchers by sorting the topics they

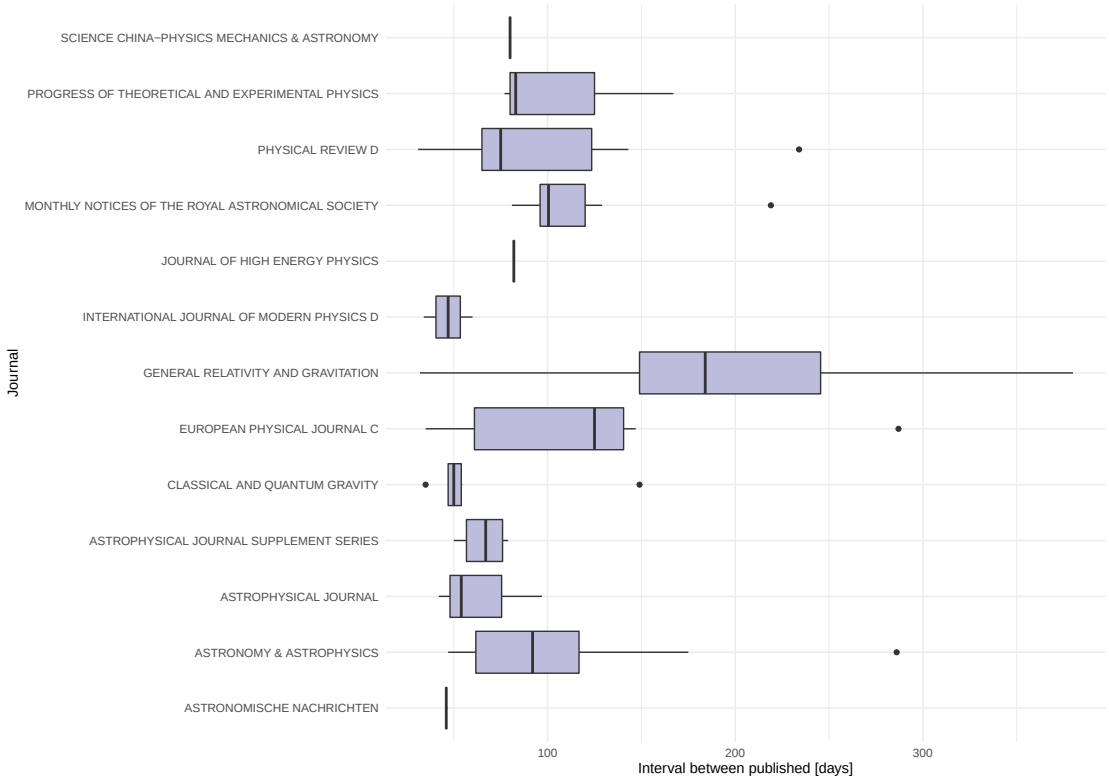

Figure 8: Graph showing difference in time between update of a paper in arXiv.org and actual publishing date according to Web of Science.

are covering based on journals they published in. The scope of all authors lies in physics but interestingly in all of the profiles appeared beside *Astronomy and astrophysics* also topic *Physics, particles and fields* which is probably caused by publications in Physical Review D. Because the journal has these categories assigned by WOS and all of the authors have at least one publication in there.

Two of the scholars have published works with topics slightly multidisciplinary with scope to *education, scientific disciplines* and *computer science*. More specific division in topics would have to be done by frequency analysis of the papers. Further analysis could be used to recommend topics and groups of authors which have similar scope.

We have found out that physicists consider important to update their information in preprint database arXive.org probably to increase possibility of citation. Some of the scholars updated their papers in arXive.org after the paper was published. It would be interesting to find out if that behavior is dependent on journal politics or scholars decision.

We have also estimated the length of publication process based on time of update o the paper in arXive.org. In this estimation we assume that all of the papers (with one exception) are published in less than 200 days from update in arXive.org which could mean receiving the review. It would be very interesting to investigate this point further.

5 Conclusion

With the methods combined we can say that we caught a glimpse of some overall publication behavior in the field of physics. To say something more general about publishing behavior in the whole field of physics we would need to gather more data of more institutes involved in this area.

It would be also beneficial to investigate this behavior together with bibliometric indices so the scholars and their team leaders can manage their ratings better and based on their results acquire more finances for their research.

The benefit of this study lies in possible conclusion for management of the institute and for leaders of scientific groups. It would be also very useful to extend the study in the areas that the scholars themselves would like to know.

References

[1] Fergnani, A. (2019). Mapping futures studies scholarship from 1968 to present: A bibliometric review of thematic clusters, research trends, and research gaps [Online]. Futures, 105(1), 104-123. https://doi.org/doi.org/10.1016/j.futures.2018.09.007

[2] Chávez-garcía, M. (2017). Strategies for Publishing in the Humanities: A SENIOR PROFESSOR ADVISES JUNIOR SCHOLARS [Online]. Journal Of Scholarly Publishing, 48(4), 199-220. https://doi.org/10.3138/jsp.48.4.199

[3] Gomez-jauregui, V., Gomez-jauregui, C., Manchado, C., & Otero, C. (2014). Information management and improvement of citation indices [Online]. International Journal Of Information Management, 34(2), 257-271. https://doi.org/10.1016/j.ijinfomgt.2014.01.002

[4] Lindahl, J. (2018). Predicting research excellence at the individual level: The importance of publication rate, top journal publications, and top 10% publications in the case of early career mathematicians [Online]. Journal Of Informetrics, 12(2), 1327-1329. https://doi.org/doi.org/10.1016/j.futures.2018.09.007

[5] Lewis, D. m., & Alpi, K. m. (2017). Bibliometric Network Analysis and Visualization for Serials Librarians: An Introduction to Sci2 [Online]. Serials Review, 43(3/4), 239-245. https://doi.org/10.1080/00987913.2017.1368057

[6] Wrigley, J., Carden, V., & Von isenburg, M. (2019). Bibliometric mapping for current and potential collaboration detection [Online]. Journal Of The Medical Library Association, 107(4), 597-600. https://doi.org/10.5195/jmla.2019.764

[7] Börner, K., Klavans, R., Patek, M., Zoss, A. m., Biberstine, J. r., Light, R. p., et al. (2012). Design and Update of a Classification System: The UCSD Map of Science [Online]. Plos One, 7(7), 1-10. https://doi.org/10.1371/journal.pone.0039464

[8] Lin, G., Hu, Z., & Hou, H. (2018). Research preferences of the G20 countries: a bibliometrics and visualization analysis [Online]. Current Science (00113891), 115(8), 1477-1485. https://doi.org/10.18520/cs/v115/i8/1477-1485

[9] Laudel, G., & Bielick, J. (2019). How do field-specific research practices affect mobility decisions of early career researchers? [Online]. Research Policy, 48(9). https://doi.org/doi.org/10.1016/j.respol.2019.05.009

[10] Suchá, L. Z., & Steinerová, J. (2015). Journal publishing models in the Czech Republic. Learned Publishing, 28(4), 239-250. https://doi.org/10.1087/20150403

[11] Wu, D., Li, J., Lu, X., & Li, J. (2018). Journal editorship index for assessing the scholarly impact of academic institutions: An empirical analysis in the field of economics [Online]. Journal Of Informetrics, 12(2), 448-460. https://doi.org/doi.org/10.1016/j.joi.2018.03.008

[12] Zou, X., & Vu, H. l. (2019). Mapping the knowledge domain of road safety studies: A scientometric analysis [Online]. Accident Analysis, 132, N.PAG. https://doi.org/10.1016/j.aap.2019.07.019

[13] Popescu, G., Istudor, N., & Zaharia, A. (2019). Sustainable Food Research Trends in EU During 2009 and 2018: Bibliometric Analysis and Abstract Mapping [Online]. Quality – Access To Success, 20, 511-516. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=e5h&an=135437261&scope=site

[14] Piotrowski, C. (2019). Contemporary Research Emphasis in Personality Assessment: A Bibliometric Analysis Mapping Investigatory Domain (2009-2018) [Online]. Sis Journal Of Projective Psychology, 26(2), 97-103. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&an=137787590&scope=site

[15] Wen, B., Horlings, E., Van der zouwen, M., & Van den besselaar, P. (2017). Mapping science through bibliometric triangulation: An experimental approach applied to water research [Online]. Journal Of The Association For Information Science, 68(3), 724-738. https://doi.org/10.1002/asi.23696

[16] Albuquerque, P. c., De paula fonseca e fonseca, B., Girard-dias, W., Zicker, F., De souza, W., & Miranda, K. (2019). Mapping the Brazilian microscopy landscape: A bibliometric and network analysis [Online]. Micron, 116, 84-92. https://doi.org/10.1016/j.micron.2018.10.005

[17] Van eck, N. jan, Waltman, L., Dekker, R., & Van den berg, J. (2010). A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS [Online]. Journal Of The American Society For Information Science, 61(12), 2405-2416. https://doi.org/10.1002/asi.21421

[18] Rieger, O. Y. (2019). 2019 arXiv Roadmap [Online]. Retrieved November 30, 2019, from https://confluence.cornell.edu/display/arxivpub/2019+arXiv+Roadmap

[19] ArXiv's Feedback on the Guidance on the Implementation of Plan S [Online]. (c2019). Retrieved November 1, 2019, from https://blogs.cornell.edu/arxiv/2019/02/04/arxivs-feedback-on-the-guidance-on-the-implementation-of-plan-s/

[20] The R Project for Statistical Computing [Online]. Retrieved November 3, 2019, from http://www.r-project.org/

[21] R Studio [Online]. (c2019). Retrieved November 3, 2019, from https://rstudio.com/

[22] Karthik Ram and Karl Broman (2019). aRxiv: Interface to the arXiv API. R package version 0.5.19. https://CRAN.R-project.org/package=aRxiv

[23] Almende B.V., Benoit Thieurmel and Titouan Robert (2019). visNetwork: Network Visualization using 'vis.js' Library. R package version 2.0.8. https://CRAN.R-project.org/package=visNetwork

[24] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016

[25] About arXiv [Online]. (2019). Retrieved November 1, 2019, from https://arxiv.org/about