Examining International Research Collaboration during the COVID-19 Pandemic using arXiv Preprints

Jiangen He[1], Erjia Yan[2], and Chaoqun Ni[3]

[1] *jiangen@utk.edu*
School of Information Sciences, The University of Tennessee, Knoxville (United States)

[2] *erjia.yan@drexel.edu*
College of Computing and Informatics, Drexel University, Philadelphia (United States)

[3] *chaoqun.ni@wisc.edu*
The Information School, University of Wisconsin-Madison, Madison (United States)

Abstract
This research-in-progress paper seeks to understand how the pandemic affected international research collaboration different countries and regions. It collected 333,793 preprints submitted to ArXiv between 2019 and 2020 to compare international research collaboration patterns pre-COVID-19 and COVID-19 eras. The paper finds that international research collaboration has been substantially affected by the pandemic, but the impact is manifested in varied extent over different time periods and in different countries. The project observed 1.55% decrease in international research collaboration in 2020 as compared with 2019. More specifically, there was a significant drop of international research collaboration at the early stage of the pandemic (from January 2020 until May 2020), and a sturdy recovery (to pre-COVID time) after May 2020. The change pattern varies by discipline and country. The results also demonstrate the resilience and adaptiveness of the scientific community in maintaining international research collaboration.

Introduction
International Research Collaboration (IRC) is a powerful driving force for innovation and discovery, as witnessed by the most rapid vaccine development in history: many existing COVID-19 vaccines are the result of IRC. IRC has also been found to increase research productivity (Thelwall & Maflahi, 2020), and impact (Wagner & Jonkers, 2017) for researchers. With the pandemic affecting almost every aspect of society, efforts have been invested in investigating its impact on research productivity (Vincent-Lamarre, Sugimoto, & Larivière, 2020), journal submission patterns (Maghfour, Olson, & Jacob, 2020), and how scientists collaborated on COVID-19 specific research (Fry, Cai, Zhang, & Wagner, 2020; Haghani & Bliemer, 2020). Yet, there's a lack of research examining the impact of the pandemic on IRC patterns. This study aims to fulfill the gap by examining the IRC pattern under the pandemic using submissions to preprint servers.

We collected all preprints submitted to aXriv.org between 2019 and 2020. We used computational methods to extract author affiliations and countries from submitted PDF files. Using 2019 preprint data as the baseline, the study compared the level of IRC during the pandemic at the author-, country-, and discipline-levels to identify the impact of the pandemic on IRC. We aim to answer two research questions:

1. Whether and how IRC changed under the COVID-19 pandemic?
2. Whether and how the change (if any) varies by discipline and country?

Data Collection
Coauthorship is often used to approximate research collaboration. This study considers arXiv.org submissions with more than one author and one country for their affiliations as IRC. Using submissions from preprint servers like arXiv.org is advantageous: it allows us to analyze research output without having to wait for publication delay. Using preprint submissions from arXiv.org, this study leveraged the capabilities of machine learning and cloud computing to

extract metadata and geographic information from PDF files. Figure 1 shows the process of data collection.

1. Download the complete set of processed arXiv PDF files for 2019 and 2020 through arXiv bulk data access that is available from Amazon S3[1].
2. Convert PDF files into XML files in NLM JATS format by using an adapted version of CERMINE, a Java library and a web service for extracting metadata and content from PDF files[2].
3. Extract author affiliation information (including country) from NLM JATS XML files.
4. Map country information into *ISO 3166-1 alpha-3* code (i.e., a set of standard three-letter country codes) using a predefined Python dictionary. If a country code cannot be identified, country, address, and affiliation information is used to request geographic coordinates through Google Geocoding API[3]. Country information is extracted from geographic coordinates.
5. Download arXiv metadata through OAI protocol for metadata harvesting.
6. Extract author and discipline information from arXiv metadata.

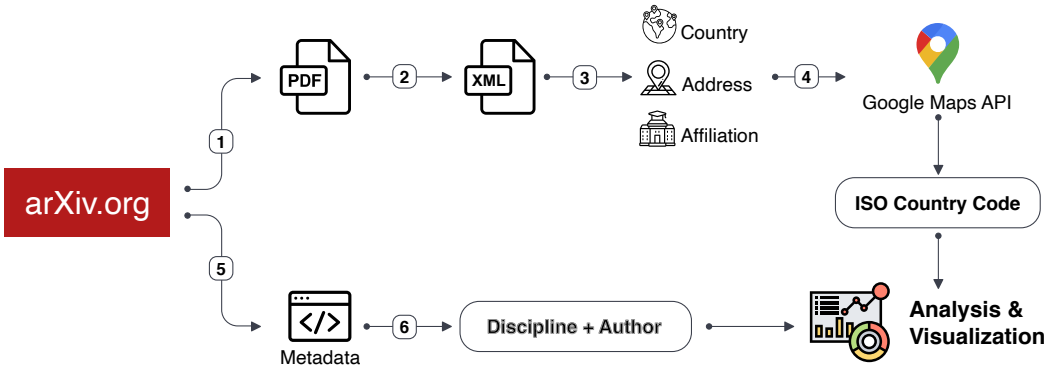

Figure 1. The process of data collection

We collected 155,464 and 178,329 preprints submitted to arXiv in 2019 and 2020, respectively. Country information for author affiliations was not available for some submissions due to missing, typo, etc. Figure 2 shows the number of collected preprints and the percentage of preprints with and without detected country information for each month. Country information was not detected for approximately 20% of harvested preprints. The percentage of preprints missing country information is roughly the same across months. Since most of our analysis is conducted by months, we assume limited impact from the missing on our conclusion. Ultimately, the final analytical sample include 240,648 preprints, 111,257 from 2019 and 128,391 from 2020.

The productivity patterns were also similar between the two years, but different in several months. June was the most productive month in 2020, but October was the one in 2019. The burst might be because scientific activities resumed and rebounded after the first wave of the pandemic in Spring 2020.

[1] https://arxiv.org/help/bulk_data_s3
[2] https://github.com/CeON/CERMINE
[3] https://developers.google.com/maps/documentation/geocoding/start

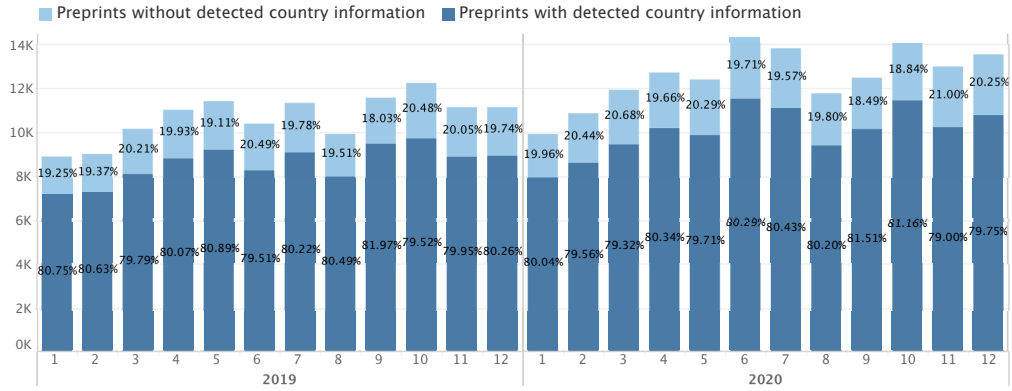

Figure 2. Collected data

Result

Collaboration at the author level

We first analyzed the collaboration at the author level. Figure 3 shows the percentage of papers with multiple authors. 90.34% (99,512) and 89.44% (115,990) of preprints had coauthors in 2019 and 2020, respectively. Compared with coauthorship patterns in 2019, more preprints in 2020 were authored by more than two, three, or four authors, and the pattern was consistent across 12 months. For example, 69.72% of preprints (89,513) were authored by at least three authors in 2020, but 67.95% of preprints (75,598) were authored by at least authors in 2019. We didn't see any evidence showing that multiple-author collaboration was affected by the pandemic.

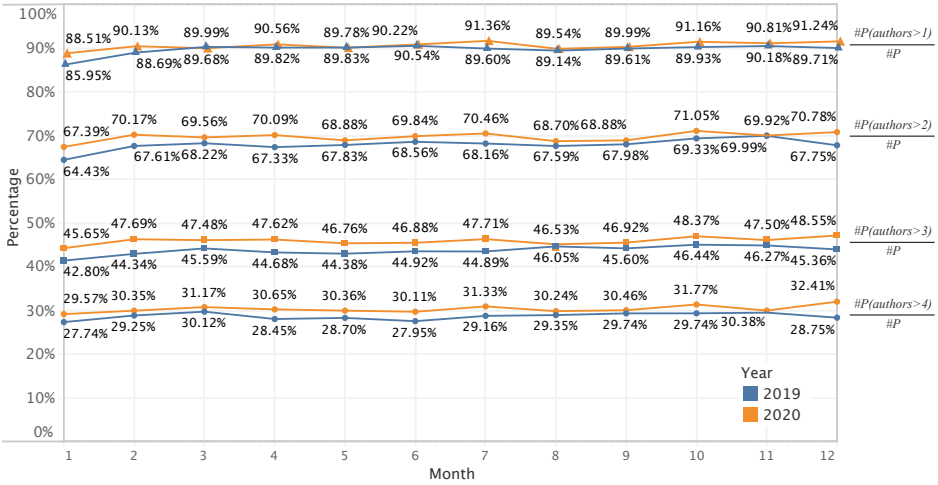

Figure 3. Collaboration at author level. *#P* is the number of papers, and *#P(author>n)* is the number of papers with more than *n* authors.

Collaboration at the country level

The IRC patterns were different between 2019 and 2020. In 2019, 38.81% of collaborative (more than one author) preprints (37,960) was contributed by authors from more than one country. This number declined to 37.26% (43,219) in 2020. Figure 4 shows the percentages of internationally collaborated preprints involving more than one, two, and three countries. Compared with 2019, we observed decreases in internationally collaborated preprints during the first four months of 2020. Since then, the percentage of IRC in 2020 was similar yet slightly lower to that of 2019. One possible explanation for this is that scientists may have adapted to new models of collaboration under the pandemic after the chaos and fluster in the first few

months. Universities have been reported to invest more in online meeting tools and other capacities to ensure better communication and collaboration for both students and researchers. By comparing the results presented in Figure 3 and Figure 4, we found changes of IRC intensity in 2019 and 2020 (especially in the first few months of 2020, see Figure 4), though the changes were not noticeable at the author level when both domestic and IRCs are combined (Figure 3). The results suggest that when international collaborations were affected by the pandemic, scientists may have participated more in domestic collaborations to compensate for the impact of the pandemic on their research agenda.

The lower three panels in Figure 4 show the change of IRC with two, three, and more than three countries involved in 2019 and 2020. We observe that the most significant decline happened in the IRC involving more than three countries, especially for the first three months of 2020. This was likely due to the different levels of coordination needed based on the number of countries involved, and the different developing patterns of COVID-19 in different countries during that time.

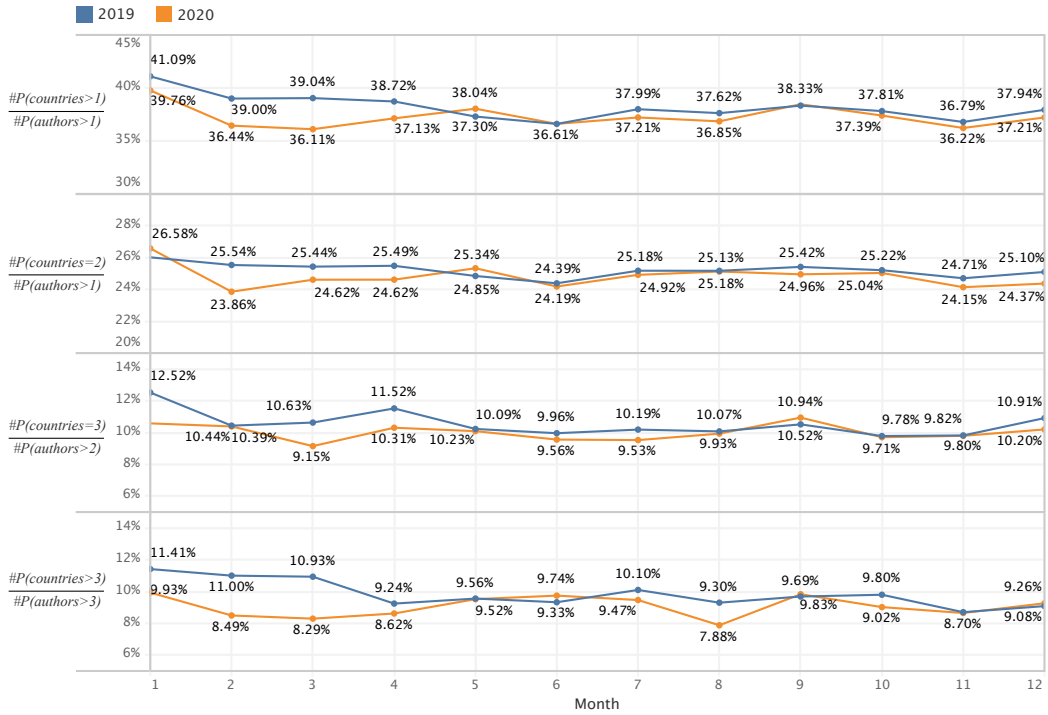

Figure 4. Collaboration at the country level. *#P(countries>n)* is the number of papers with authors from more than *n* countries.

Disciplines

We also compared the changes of IRC under the pandemic across disciplines. arXiv includes preprints from eight disciplines, with the majority of preprints (91.1%) are in physics (33.2%), computer science (29.3%) mathematics (20.6%%), and statistics (8.0%).

Figure 5 shows the percentage of IRC (preprints with authors from at least two countries) over all collaborative preprints in these four disciplines. The patterns in computer science, math, and physics were similar: IRC decreased for the first few months in 2020 and recovered to a similar level in 2019 afterward. The pattern in statistics was slightly different: after the decrease at the beginning of 2020, the IRC rebounded to a higher level than that in 2019.

The time and the extent of being influenced by the pandemic was different across disciplines due to the different nature of the disciplines. In comparison with other disciplines, IRC in computer science was influenced earlier. It takes varying time for scientists in a different discipline to adapt to the pandemic to recover IRC to the level before pandemic. It is also worth noting that some disciplines tended to have even higher level of IRC in the second half of 2020.

For example, scientists in statistics had much more IRC after July 2020 than in 2019. Another observation was that all disciplines except for physics had a dramatic increase in IRC in October 2020, but we need to further investigate this concerted increase.

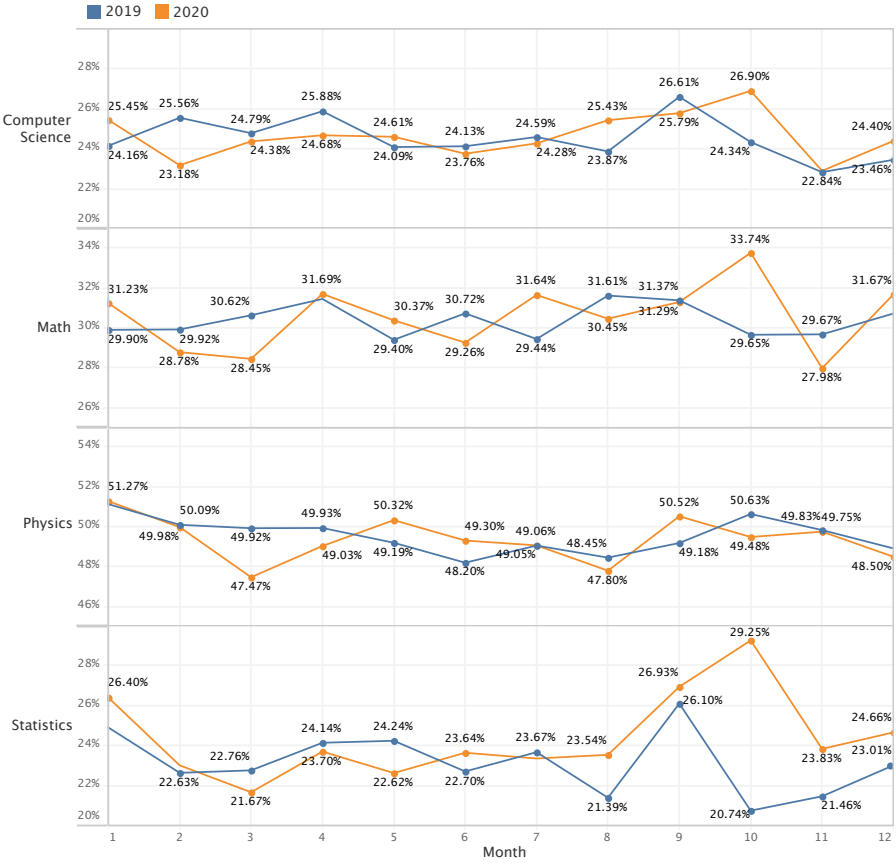

Figure 5. IRC in different disciplines.

Countries

Based on our analysis in previous sections, IRC was mainly hampered by the pandemic in the first four months of 2020. Thus, we examined the percentage change of IRC ($\#P(countries > 1)/\#P(authors > 1)$) in the first four months between 2019 and 2020 (see Figure 6). The IRC of most countries decreased, though with varied country-level differences. The IRC increased in some countries in Eastern Europe, Latin America, the Middle East, and Southeast Asia.

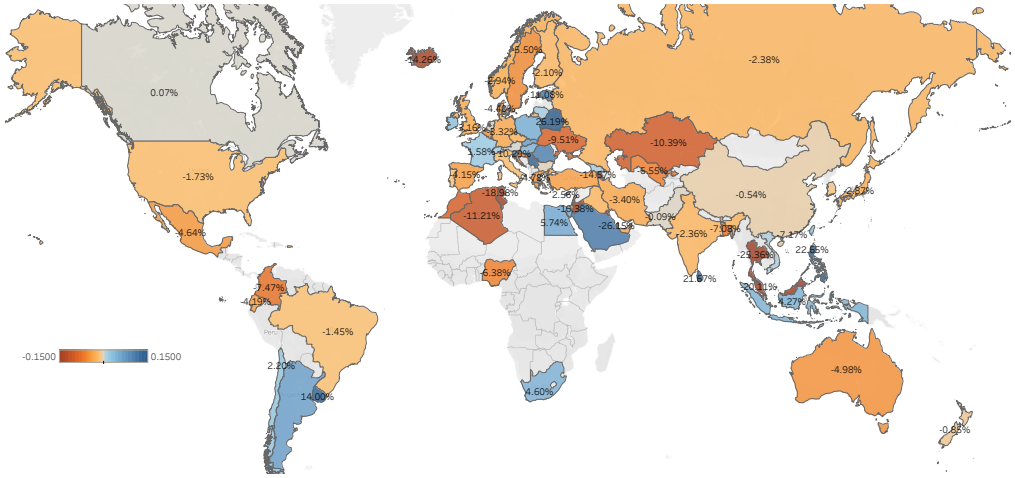

Figure 6. The change of percentage of IRC in the first four months between 2019 and 2020 across countries and regions submitted more than 100 arXiv preprints.

We also analyzed six countries with the most arXiv preprints (see Figure 7). The IRC decreased in the first few months in 2020 and recovered later. Some countries had a higher level of IRC in the second half of 2020 than 2019, such as the United Kingdom and Italy. The exception is France, where IRC was affected by the pandemic throughout the entire year in 2020.

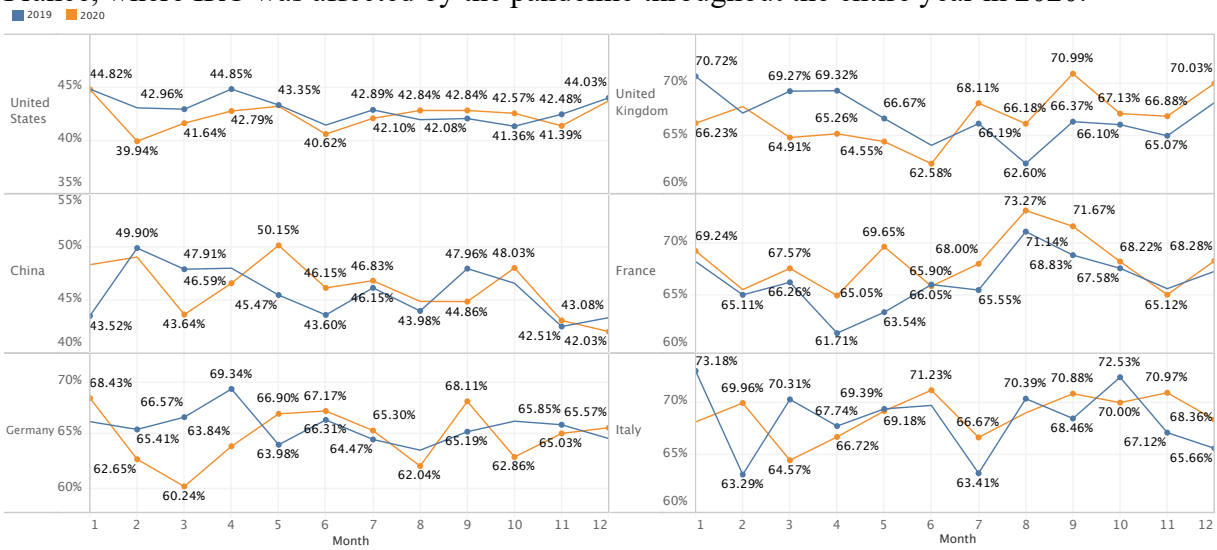

Figure 7. IRC from countries that submitted most preprints to arXiv.

Conclusion and Future Work

In this research-in-progress work, we found changes to IRCs in 2020, which is likely due to the COVID-19 pandemic. IRC rate declined during the first few months of 2020 and bounced back to a level similar to 2019 after May 2020. Yet, the change varied by discipline and country. Quick adoptions to research collaboration norms under the pandemic may have helped IRC resume during COVID-19. Different developing patterns as well as measures and efforts curtailing with COVID-19 by country is likely related to the varying changing patterns of IRC for countries.

Due to the varying impact of the pandemic across time, disciplines, and countries, we will examine factors affecting the impact of the pandemic in future research. For example, we will investigate the factors that alleviate or intensify the impact of the pandemic at the country level and examine the role of geographical and economic proximity in IRC during a public health crisis. This study, along with the future work, will provide a comprehensive and timely description of IRC patterns during the pandemic and provide an in-depth understanding of factors affecting IRC.

References

Fry, C. V., Cai, X., Zhang, Y., & Wagner, C. S. (2020). Consolidation in a crisis: Patterns of international collaboration in early COVID-19 research. *PLoS ONE*, *15*(7 July), 1–15.

Haghani, M., & Bliemer, M. C. J. (2020). Covid-19 Pandemic and the unprecedented mobilisation of scholarly efforts prompted by a health crisis: Scientometric comparisons across SARS, MERS and 2019-nCoV literature. In *Scientometrics* (Vol. 125).

Maghfour, J., Olson, J., & Jacob, S. E. (2020). COVID-19 impacts medical journal submissions. *International Journal of Women's Dermatology*, *6*(4), 255–256.

Thelwall, M., & Maflahi, N. (2020). Academic collaboration rates and citation associations vary substantially between countries and fields. *Journal of the Association for Information Science and Technology*, *71*(8), 968–978.

Vincent-Lamarre, P., Sugimoto, C. R., & Larivière, V. (2020). The decline of women's research production during the coronavirus pandemic. *Nature Index*, *19*.

Wagner, C. S., & Jonkers, K. (2017). Open countries have strong science. *Nature*, *550*(7674), 33.