Changes in Research Collaborations During the Pandemic

Ziao Wang[1], Kian Ahrabian[2], Casandra Rusti[3], Jay Pujara[4], Kristina Lerman[5]

[1] *wang-za19@mails.tsinghua.edu.cn*
Tsinghua University, Beijing (China)

[2] *ahrabian@isi.edu* [3] *mrusti@isi.edu* [4] *jpujara@isi.edu* [5] *lerman@isi.edu*
USC Information Sciences Institute, Los Angeles, CA (USA)

Abstract
The COVID-19 pandemic impacted people on a global scale, yet its effect on scientists and scientific institutions has yet to be fully understood. We study the effect of the pandemic on scientific research by analyzing newly available bibliographic datasets covering millions of papers and authors. We apply statistical methods to understand how the pandemic disruptions affected the collaborations and productivity of researchers at over 40 top-ranked institutions from around the world. Our analysis shows that although more researchers published during the pandemic, producing more papers, their individual productivity decreased. Further analysis is required to identify factors that promote the resilience of scientific innovation.

Introduction
The COVID-19 pandemic disrupted people's work routines worldwide, including those of scientists (Myers et al., 2020, Morgan et al., 2021, Gao et al., 2021). The pandemic response, however, was highly heterogeneous, with policies varying by country (e.g., China's "zero COVID" policy vs. Sweden's more hands-off approach), US state (e.g., strong vs. weak COVID-19 response), community (e.g., political climate and school closures), and even institution (e.g., remote work policy, vaccine mandates). These diverse responses create conditions that allow us to separate factors affecting the resilience of scientific innovation. As a first step, we study the impact of the pandemic on patterns of scientific collaborations at different institutions. Specifically, we pose the following research questions:

RQ1: How did the pandemic affect the productivity of scientists at different institutions?
RQ2: How did the pandemic affect the patterns of collaboration at different institutions?

To answer these research questions, we use a large-scale bibliographic database called OpenAlex (Priem, Piwowar, & Orr, 2022). OpenAlex contains information about researchers, their affiliations, and the collaborations they belong to, which are indicated by joint authorships of research papers. We study researchers affiliated with 41 top-ranked institutions from around the world. We operationalize measures of interest using their bibliometric information and perform analyses over these variables to estimate the pandemic's impact on collaborations and productivity of researchers from these institutions.

Methodology

Data
Our study relies on the bibliographic dataset OpenAlex, which provides information about entities such as publications, authors, institutions, venues, and the relationships between them. OpenAlex gathers data from various sources, including Microsoft Academic Graph, and is updated every two weeks. We use the dump obtained on *Nov 22nd, 2022*.

Constructing Collaboration Networks

We select 41 high-ranking institutions worldwide to analyze their general performance and internal collaboration networks, which include only authors affiliated with those institutions and collaboration links between them. Prior work has shown that internal collaborations are highly correlated with all collaborations (Burghardt et al., 2021). To reduce noise, we remove all the publications with an OpenAlex designated type with fewer than one million appearances.

We construct and analyze scientific collaborations at these institutions over time. We represent collaborations as an unweighted, undirected network in which nodes are authors and edges represent co-authorship of papers published during a specific aggregation interval (e.g., a year). Although the frequency of co-authorship between two researchers is not directly represented (as edges are unweighted), low-frequency co-authorships are less likely to be captured during a specific time interval. We create a sequence of collaboration networks for an institution over time by considering papers published during different time intervals.

We analyze changes in the following outcomes over time to measure the impacts of the pandemic disruptions on an institution: 1) the number of publications, 2) the number of collaborations, 3) the average team size, and 4) the average clustering coefficient.

Estimating Effect of a Disruption

To estimate the effect of the pandemic disruption, we performed linear regression over a sliding window of 10 years and used the trained model to predict the outcome for the next time point. We measure the deviation, i.e., the difference between the actual and predicted values of the outcome, slide the time window, and repeat the procedure. We use the difference between values for outcomes with a fixed range, like the clustering coefficient. For other outcomes, we calculate the relative deviation, i.e., deviation divided by the predicted value. We calculate an outcome's mean and standard deviation across all selected institutions.

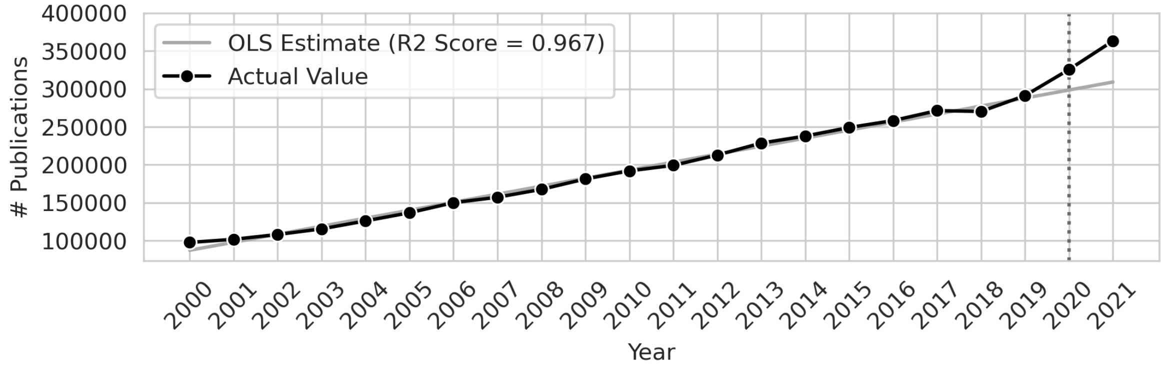
Analyses

RQ1: How did pandemic affect research productivity of institutions?

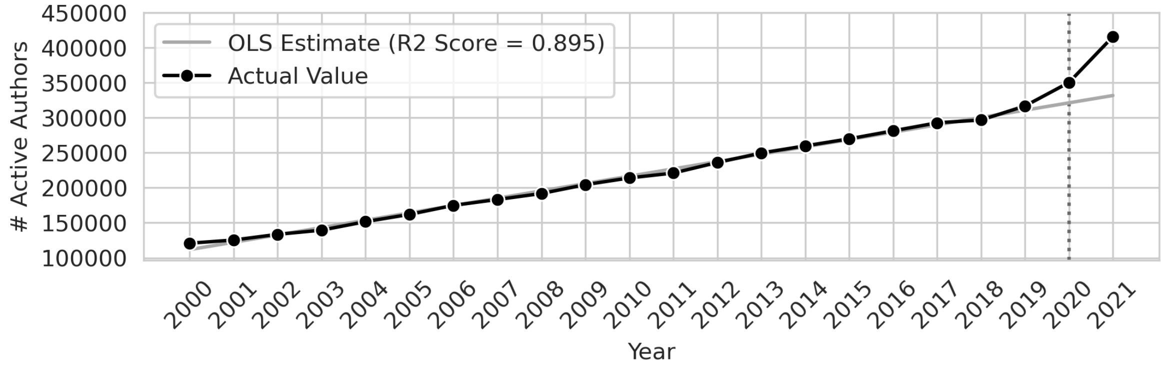
At first glance, we hypothesize that the pandemic reduced research productivity due to aftereffects such as extended lockdowns and work-from-home transition that disrupted the normal flow of everyday work. To test our hypothesis, we define three metrics of research productivity and examine their behavior at 41 high-ranking institutions.

The most straightforward metric of research productivity is the *number of publications*. As a second proxy, we define the *number of active authors*. An author is designated active if they publish at least one paper during the aggregation time interval. Since both metrics correlate with the institution size, we use the *average number of papers per active author* as the normalized research productivity metric. Using these three metrics, we compare each institution with itself and other institutions in the past and the present.

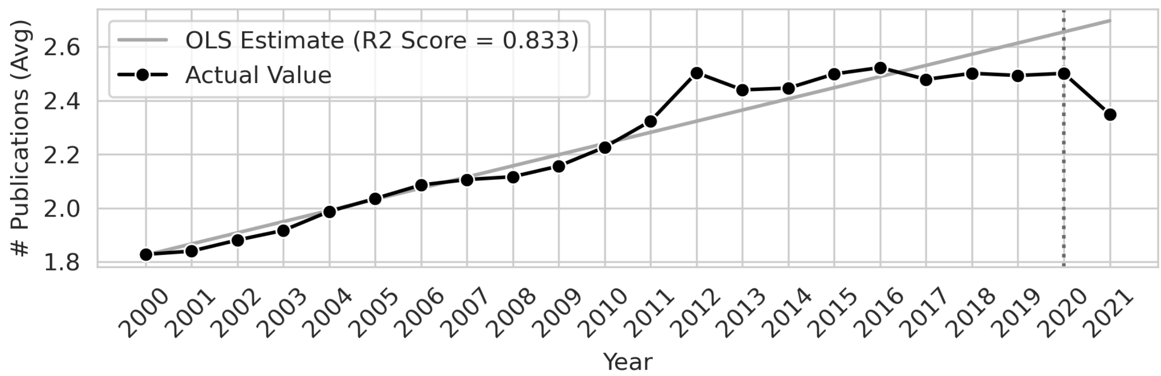
We also calculate the ordinary least squares (OLS) estimates with respect to time on data from 2000 to 2019 and project all three metrics for 2020 and 2021. Figures 1a, 1b, and 1c illustrate the results of our analysis.



a) Number of publications



b) Number of active authors



c) Average number of publications per active author

Figure 1. Productivity metrics for 2000 – 2021 period

Comparing the projected metrics to the actual values in 2020 and 2021, we find that the growth of the selected high-ranking institutions systematically deviates from expected values on the first two metrics (Figure 1a, 1b), showing a somewhat unintuitive productivity acceleration. However, the normalized metric (Figure 1c) shows a steady trend in 2020 and a relatively significant decline in 2021. Results show that since the pandemic began, more authors have published their research, which has accelerated the increase in the sheer number of publications; however, there was a decline in the average productivity of researchers. From these observations, institutions have compensated for decreased individual productivity by having more researchers actively participate.

Due to seasonal fluctuations in the number of publications over the course of a year, we study the effect of the pandemic by splitting yearly data into early-year and late-year intervals, and tracking the expected number of publications over time as a function of active authors. Specifically, for a given time point between 2010 and 2021, we extract these values for the previous ten years, train an OLS model and project the value for the next time point. We apply this procedure to each institution and average the relative deviations across institutions.

Figure 2 shows the relative deviation of the expected number of publications with respect to the number of active authors, separately for the early-year and late-year aggregation periods. The results for 2020 show a slight increase in individual productivity, whereas, for early 2021, there is a slight increase, and for late 2021, a severe decline. Both results are consistent with Figure 1c; however, it is difficult to conclude with only one data point. Adding more data for 2022 will help us better understand the underlying effects. We leave this question to future work as the current data is incomplete for 2022.

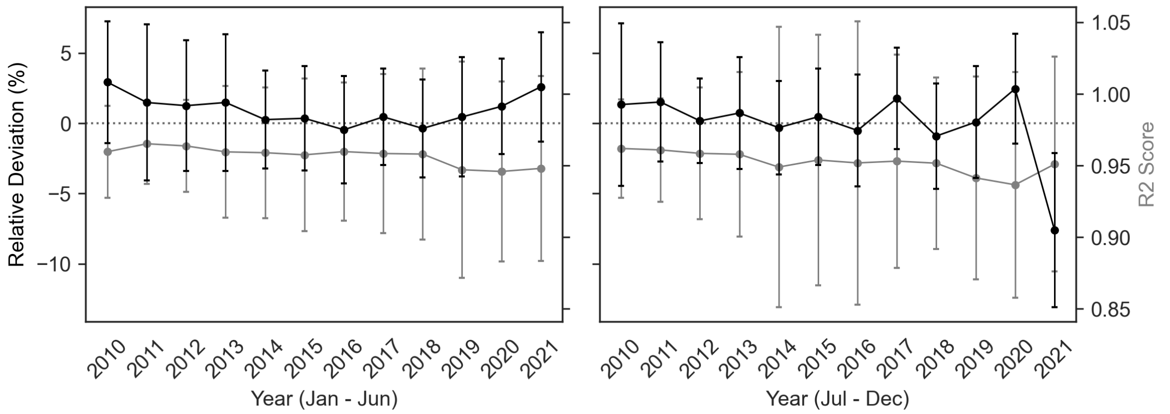


Figure 2. Deviation from the predicted number of publications over time

RQ2: How did the pandemic affect the patterns of collaboration at different institutions?

To analyze how the pandemic impacted patterns of research collaborations, we looked at the deviations in the *number of collaborations*, *average team size*, and the c*lustering coefficients* (cliquishness) across institutions.

The *number of collaborations* is measured by the number of co-author pairs that have published together in each aggregation interval (six-month periods between 2000 and 2021). Figure 3 shows the relative deviation of the number of collaborations over time. We see a systematic positive deviation in the *number of collaborations* during the second half of 2020 and the first half of 2021, representing a larger-than-expected number of collaborations. Large error bars during this one year suggest that there was much more variation at the individual institution level within this period. In the second part of 2021, we see the relative deviation return to neutral (i.e., the actual number of collaborations returned to projected values), and the variation across institutions decreased, suggesting the pandemic disrupted this collaboration metric for about one year before returning to its prior trend.

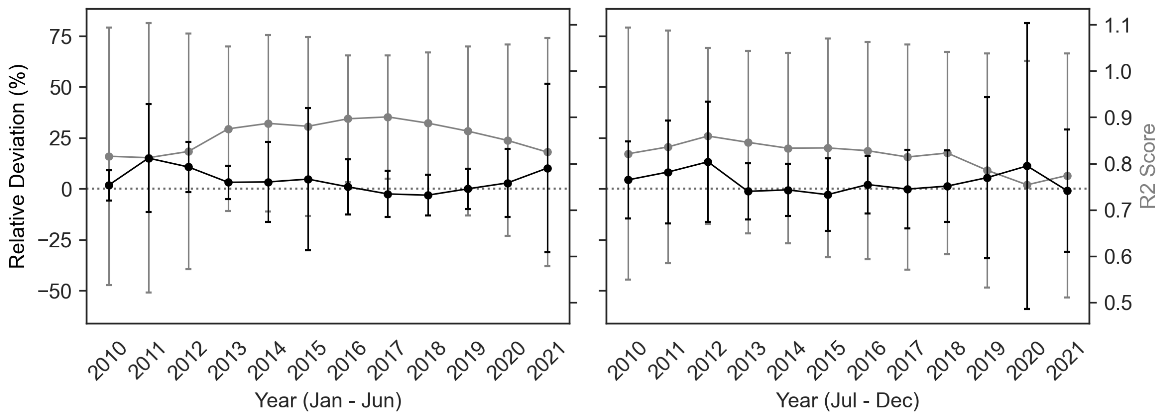


Figure 3. Deviation in the number of collaborations at an institution

The *average team size* is the average number of authors (at an institution) per paper. Here we see that the pandemic did not noticeably affect this collaboration metric. Generally, the trends remain in line with the pre-pandemic period. We also do not see a spike in standard deviation as we did in the number of collaborations, indicating slight variation in team size trends.

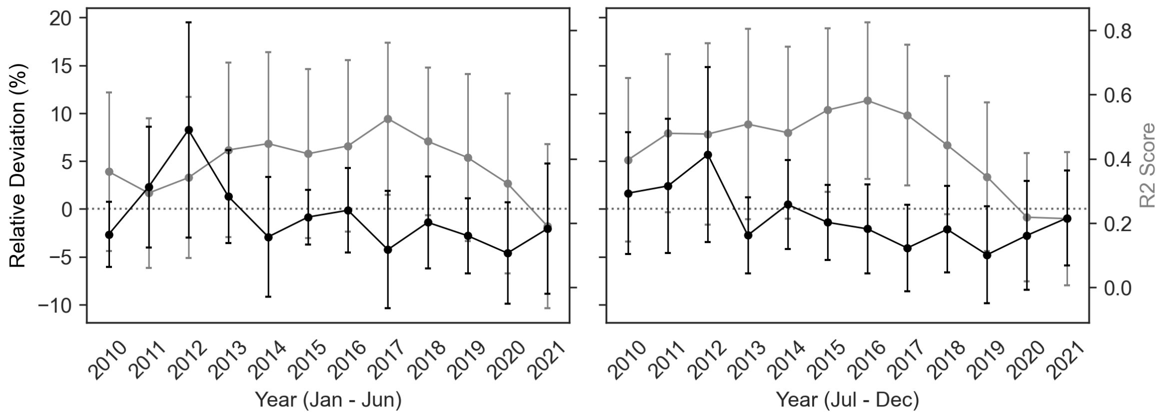


Figure 4. Average team size

The *clustering coefficient* of an author represents the fraction of that author's collaborators who collaborate with each other. A clustering coefficient close to one indicates a cliquish network, i.e., a tight community. We plot the time series of the difference between the actual clustering coefficient and the one predicted by linear regression on the institution's historical data. Figure 5 shows a systematically negative deviation in the average clustering coefficient in late 2020 and early 2021, with recovery by late 2021, although the change may not be statistically significant.

While the *number of collaborations* (Figure 3) is derived from OLS projections with the goodness-of-fit (R2) values between 0.75 to 0.9, the R2 values for *average team size* (Figure 4) and *average clustering coefficient* (Figure 5) vary between 0.2-0.6, suggesting more complex models may be necessary in future work to better fit the data.

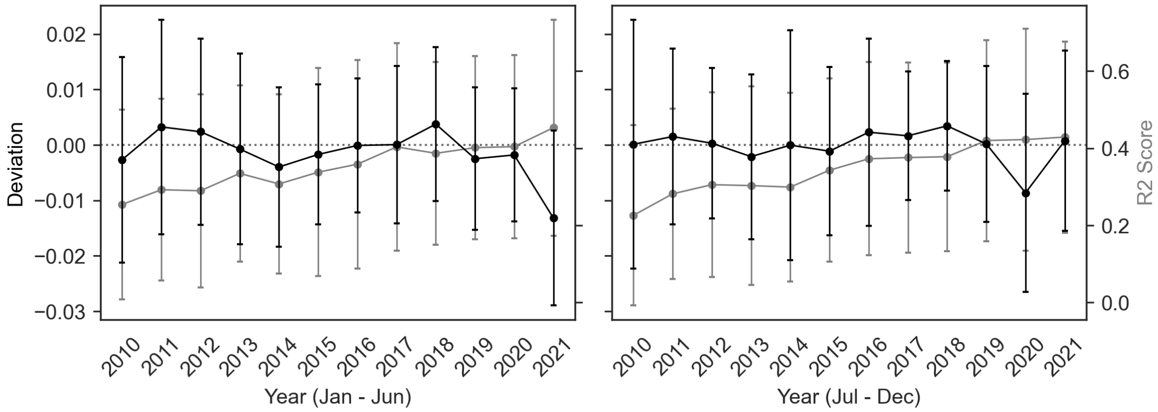


Figure 5. Deviation of average clustering coefficient of collaboration networks over time.

Summary

We explored the effect of the global pandemic on high-ranking institutions, focusing on productivity and collaboration patterns. We operationalized metrics to track these quantities over time and then presented analyses based on these metrics to study the phenomenon.

Our results indicate that high-ranking institutions have shown resilience throughout the pandemic, even showing accelerated growth on some measures. Our most significant observations are 1) an increase in the number of active authors and publications, contrary to intuition; 2) slightly fewer than expected collaborations; 3) a smaller than expected average clustering coefficient of the collaboration network, with both metrics returning to pre-pandemic trends by the second half of 2021. We await 2022 data to verify the trends.

Limitations and Future Directions

Our work focused only on high-ranking institutions to study pandemic disruptions; however, the effects of the pandemic could be further investigated on a much more inclusive range of institutions. Moreover, we mainly analyzed our metrics from a holistic standpoint. Future works could examine the effects more granularly by adding perpendicular variables such as fields of study, gender, etc. Finally, our presented analyses rely heavily on the OpenAlex data quality, and there are known data quality issues in the November 2022 data dump used, specifically related to linking works to authors. These issues should be taken into consideration when interpreting our results. We are currently investigating these issues and working towards improving the accuracy of our data and results. As a result, any issues in the data-gathering process are propagated into our analyses. Although we have spot-checked various samples with other sources, such as DBLP, this should be further examined systematically.

References

Gao, J., Yin, Y., Myers, K. R., Lakhani, K. R., & Wang, D. (2021). Potentially long-lasting effects of the pandemic on scientists. Nature communications, 12(1), 1-6.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:2205.01833.

Burghardt, Z. He, A. G. Percus, and K. Lerman. The emergence of heterogeneous scaling in research institutions. Communications Physics, 4(1):1–6, 2021.

Morgan, A. C., Way, S. F., Hoefer, M. J., Larremore, D. B., Galesic, M., & Clauset, A. (2021). The unequal impact of parenthood in academia. *Science Advances*, 7(9), eabd1996.

Myers, K. R., Tham, W. Y., Yin, Y., Cohodes, N., Thursby, J. G., Thursby, M. C., ... & Wang, D. (2020). Unequal effects of the COVID-19 pandemic on scientists. *Nature Human Behaviour*, *4*(9), 880-883.