INFSCI 2415

Mid-term project report

Academia Talent Migration in The Recent Decade (2010 – Present)

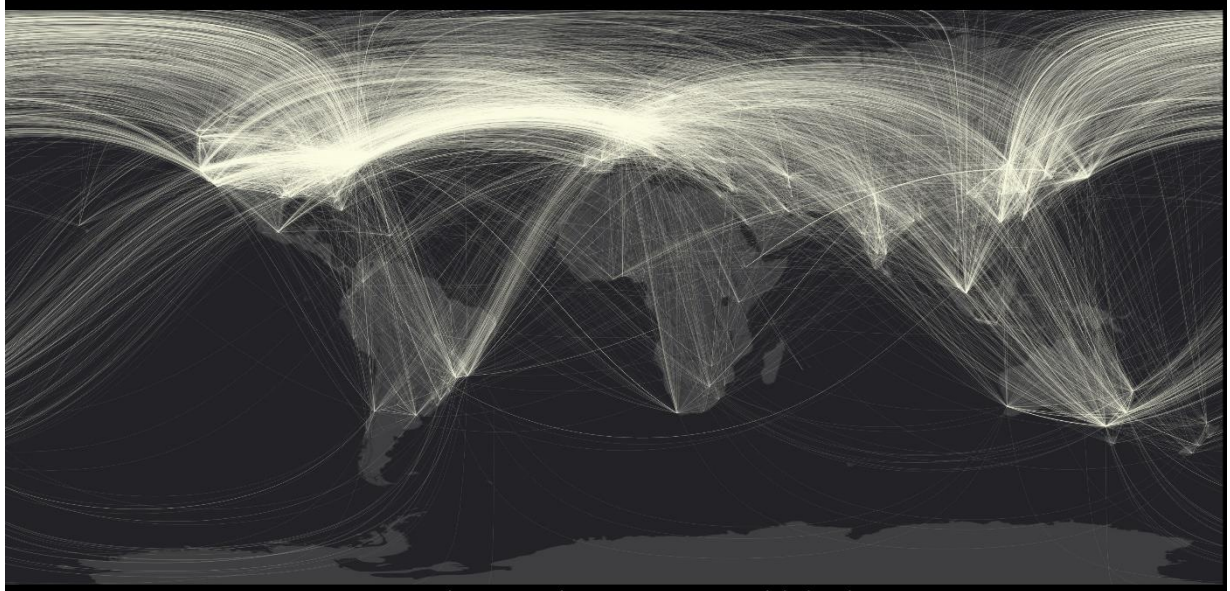
Shangbin Tang

4367695

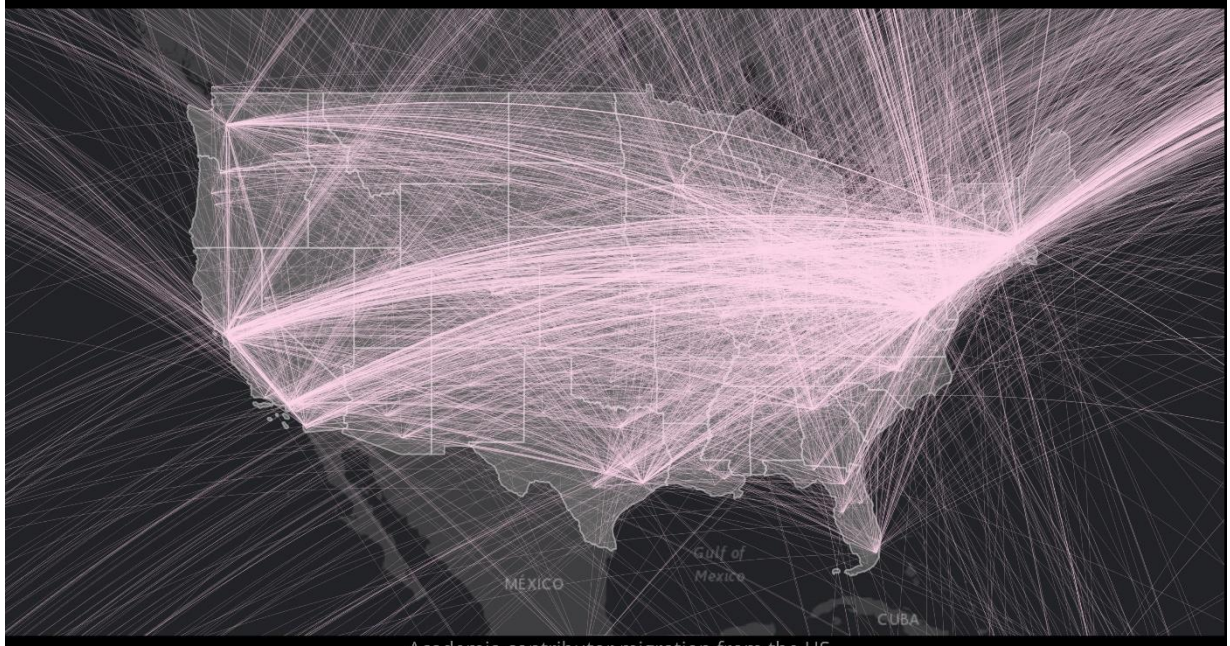
Shangbin.tang@pitt.edu

Academia Talent Migration in the Recent Decade (2010 - Present)

v 0.0.1



Academia contributor migration in a global scale



Academia contributor migration from the US

Data Source: Microsoft Adamemic Graph
Author: Shangbin Tang
shangbin.tang@pitt.edu
shangbin.tang.github.io
Version: 0.0.1

Data and Methods

I downloaded datasets: Papers.txt, PaperAuthorAffliations.txt, and PaperFieldsOfStudy.txt from Microsoft Academic Graph. In the original data Papers.txt and PaperAuthorAffliations.txt, there're a total of 268 million papers and 729 million paper-author- affiliation records. Among the published papers with valid year information, 49% percent are published after the year 2010. Hence, from the perspectives of being meaningful, representative, and practical, papers published after 2010 and their author information were chosen as analysis samples of this project.

After limiting the time range of this project, I used a python dictionary of PaperID published after 2010 to filter the paper-author- affiliation dataset. Then, another dictionary of AuthorID and the publication count of each author were used to filter out the author with more than one publication. They are the potential people who can provide me with "visible" traces – authors can't have multiple affiliation records with only one paper-author- affiliation record. Among the whole dataset, near 28 million authors are relatively "fruitful" – have more than one publication. After that, I built a dictionary with python set as the value type to summary the author and their affiliations. After all these, authors with more than one affiliation and their papers were also picked out using dictionaries.

A combined data table was built using the filtered paper-author-affiliation dataset and the paper data, sorted with AuthorID and the publish time information. The "switches" between adjacent affiliations were recognized as trips, and the affiliation dataset provided the location of starting point and destination of each trip.

Finally, I used ArcGIS Pro to map out the trips as a flow map.

Current Results and Discussion

There're over 3 million trips in the current result dataset, which is too dense to show on a map. Consequently, I randomly sample 1% of the data to present on the map as plot shown on the previous page.

On the aspect of a global scale, the two major of talent/contributor migration routes, no specific in and out direction here, are Europe – the US, and Eastern Asia – the US. Most of the beginnings and destinations lie on the coastal lines of continents, which coincide with the development view of geography. This phenomenon is also indicated in the US: coastal cities/states tend to have adequate academic talent resources.

The Next Steps

Although the main results are presented on the map, there's still more work to do. One major task is to filter out those "fake" trips. Some scholars have multiple affiliations, and along time they have affiliation lists like ["A", " A", " A", "B", "A", "A", "C", "D"], where the trip between "A" and "B" might not be a true migration, they might temporally switch affiliation or use multiple affiliations on same publication; and "A" to "C" to "D" would be a true trace because it's one-way. Also, I'm trying to make sub-plots of migrations of different fields, which might be another big challenge waiting ahead for me.