ISSI2021 Tutorial on Studying migration and mobility among scholars using bibliometric data with practical session in Jupyter Python

Tutorial organizer: Samin Aref

Laboratory of Digital and Computational Demography,
Max Planck Institute for Demographic Research, Rostock 18057, Germany
https://saref.github.io/
aref@demogr.mpg.de

Bibliometric data give us unprecedented opportunities for understanding patterns of mobility among published researchers by analyzing affiliation addresses and other meta-data. In a series of studies [1,2,3], we leverage large-scale bibliometric data to measure and model the migration of scholars in different contexts and answer fundamental questions in the intersection of high-skilled migration and scientometrics.

The **first part** of this tutorial explores three recent case studies within the scope of migration in academia and analysis of individual-level bibliometric data. These three case studies have different substantive themes and are in different context and geographies: *return migration* among highly mobile researchers [1], *internal migration* of researchers in Mexico [2], and *international migration* of researchers in Russia [3]. These three studies showcase the insight that we can obtain from bibliometric data and different approaches on modeling scholarly migration for a range of research questions which are extendable to other geographical contexts. A brief summary of each case study is provided in the appendix.

The **second part** of this tutorial is a hands-on lab session on pre-processing large-scale individual-level bibliometric data for studying migration and mobility in academia. We use *Python* as the programming language and *Jupyter* (from Anaconda[1]) as the environment for processing the data.

No background knowledge in Python programming or migration studies is expected. Instead, an interactive tutorial file is sent to the participants at least two weeks in advance which allows them to set up the requires software, open Jupyter on their computers, explore basic operations[2] in Python, and get familiar with the syntax for programming in Python. Completing the interactive tutorial takes around 1 hours (and probably a bit more for participants without data or computing background). Completing the interactive tutorial before the day of the tutorial **is required and expected** for all participants who would

[1] Freely accessible from https://www.anaconda.com/products/individual

[2] • Modules • Variables and types • Operators and comparisons • Compound types: Strings, List, Tuples and Dictionaries • Control Flow (conditional statements) • Loops • Functions

like to attend the practical session. All general programming and setup questions can be sent to the tutorial instructor by email[3] which will be answered before the day of the tutorial.

The lab session is based on practicing synchronous tasks on cleaning and pre-processing a dataset of 1.7 million author-publication linkages (authorship records) from Springer Nature SciGraph[4]. Step by step, the instructor and the participants, apply a framework of data pre-processing on the raw bibliometric data to prepare a dataset of internationally mobile researchers and extract key variables from the data. The pre-processing framework includes the following technical tasks:
- data wrangling and data cleaning,
- handling missing values,
- identifying outliers,
- inferring gender from first names,
- extracting countries of academic origin and destination,
- extracting academic age,
- inferring modal countries of affiliation, and
- categorizing mobile and non-mobile researchers.

Upon completion of the tutorial, participants will have achieved the following learning outcomes:
- several analysis themes on studying migration of researchers,
- current methodological approaches for modeling scholarly migration,
- current methodological approaches for measuring scholarly migration,
- practical skills for pre-processing individual-level bibliometric data,
- practical skills for re-purposing affiliation addresses for studying migration, and
- general skills in Python programming and big data analytics.

References

1. Aref, S., Zagheni, E., West, J.: The demography of the peripatetic researcher: Evidence on highly mobile scholars from the Web of Science. In: International Conference on Social Informatics. pp. 50–65. Springer (2019), doi: 10.1007/978-3-030-34971-4-4
2. Miranda-González, A., Aref, S., Theile, T., Zagheni, E.: Scholarly migration within Mexico: Analyzing internal migration among researchers using Scopus longitudinal bibliometric data. EPJ Data Science **9**, 1–26 (2020). https://doi.org/10.1140/epjds/s13688-020-00252-9
3. Subbotin, A., Aref, S.: Brain drain and brain gain in Russia: Analyzing international mobility of researchers by discipline using Scopus bibliometric data 1996-2020. ArXiv (2020), https://arxiv.org/pdf/2008.03129

[3] aref@demogr.mpg.de
[4] `https://scigraph.springernature.com/`

Appendix

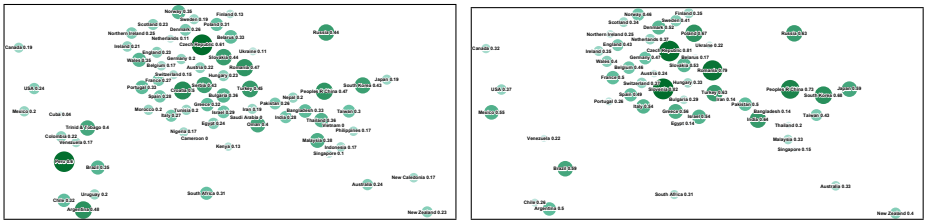
Case study 1: Highly mobile researchers around the world

In order to understand the scholarly migration, determining the extent to which researchers have worked in more than two countries is essential. We focus on the subgroup of highly mobile researchers whom we refer to as "super-movers." More specifically, we track the international movements of researchers who have published in more than two countries through changes in the main affiliation addresses of over 62 million publications indexed in the Web of Science database over the 1956-2016 period.

Among other findings, our results point to the emergence of a global system that includes the USA and China as two large hubs, and England and Germany as two smaller hubs for highly mobile researchers [1].

The bibliometric data on super-movers allow us to investigate return migration and compare it across countries. Our results in Fig. 1 show return migration by country (fraction of super-movers who return to their country of academic origin) for the early-career and the mid-career super-movers. This analysis points to low return migration in Iran, Singapore, Ukraine, and Venezuela. In contrast, Czech Republic, China, Romania, Russia, South Korea, and Turkey seem to have high return migration in the sub-population of highly mobile researcher.

This research contributes to the literature by offering a snapshot of the key features of highly mobile researchers, including their patterns of migration and return migration by academic age, the relative frequency of their disciplines, and the relative frequency of their countries of origin and destination [1].



(a) Early-career super-movers (b) Mid-career super-movers

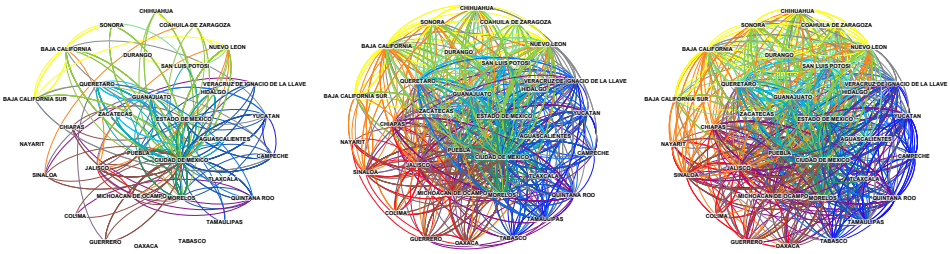
Fig. 1. Return migration among super-movers by country: (a) early-career and (b) mid-career (see the figure on screen for high resolution).

Case study 2: Internal migration of researchers in Mexico

Our understanding of internal migration among researchers is quite limited partly due to lack of data aggregated at a suitable sub-national level. We repurpose bibliometric data using a neural network which provides a sub-national

level for aggregating affiliation data and tracking changes of affiliations. The neural network takes an affiliation address as input and predicts with a high accuracy the state in Mexico associated with that affiliation address.

We analyze internal migration based on over 1.1 million authorship records from the Scopus database to trace the movements of over 250,000 scholars in Mexico and provide measures of internal migration such as net migration rates for all states over the period 1996-2019 [2].



(a) 1997-1998 (b) 2007-2008 (c) 2017-2018

Fig. 2. Networks of internal migration among researchers in Mexico based on selected one-year periods. Directions of edges are clock-wise and their colors are the mix of respective origins and destinations (see the figure on screen for high resolution).

Migration patterns between states in Mexico appear to be heterogeneous in size and direction across regions. However, while many scholars remain in their regions, there seems to be a preference for Mexico City and the surrounding states as a destination. Over the past two decades, we observed a general decreasing trend in the crude migration intensity. The origins and destinations of internal migrants have become more diverse over time, including greater exchange between states along the Gulf of Mexico and the Pacific Coast [2].

Case study 3: Brain drain and brain gain in Russia

Debates on international migration in academia often consider scientists and the scientific community as one unit which could be either a net loss (brain drain) or a net gain (brain gain) for a given country disregarding the fact that the impact of migration on a national science system could vary for different fields of scholarship.

We use data from 2 million publications in Scopus to analyze international migration of researchers in Russia as a commonly debated subject of brain drain. We analyze origins and destinations of migrant researchers with respect to their fields and performance and compute net migration rates.

Our analysis shows that Russia has overall suffered a net loss in human capital due to a lack of balance between incoming and outgoing flows of researchers. Also, we observe that the total citations of researchers immigrating to Russia is

substantially lower than that of researchers emigrating from Russia except for the case of social sciences in which the difference is negligible. Our results on net migration rates indicate that while Russia has been a donor country in the late 1990s and early 2000s, in more recent years Russia has experienced more balanced flows of incoming and outgoing researchers [3].

We also develop a new methodology to quantify the impact of migration on each field of scholarship and implement it to the case of Russia. Using the subjects associated with the publication venues where migrant researchers have published, we quantify brain drain in Russia for different fields of scholarship according to the All Science Journal Classification (ASJC).

Our results in Fig. 3 suggest that Russia has suffered a loss in almost all disciplines and more notably in neuroscience, decision sciences, mathematics, and biochemistry genetics and molecular biology. Our substantive results reveal new aspects of international mobility in academia and its impact on a national science system which speak directly to policy development [3].

Fig. 3. Field-based net brain drain for all disciplines of scholarship quantified based on movements of different migrant researchers in Russia and their publications