IMPRS-PHDS 2022 course (IDEM187) on Topics in Digital and Computational Demography – Day 4 (November 14th 2022)

BiblioDemography:

Using large-scale bibliometric data for demographic research; Advantages and pitfalls of using Scopus data to trace internal and international scholarly migration worldwide

Aliakbar Akbaritabar[1]

[1] Max Planck Institute for Demographic Research (MPIDR)
Akbaritabar@demogr.mpg.de

Please tweet with hashtag
#BiblioDemography,
a tribute to James W. Vaupel (1945-2022).
Thanks Ilya and Jonas for bringing up Jim's
labeling idea!

AGENDA

1. Introduction (15 minutes, [video 1])
   - Welcome and introduction
   - What is bibliometric data?
   - What type of questions can be studied using bibliometric data?
   - What type of demographic questions can be studied using bibliometric data?
   - Limitations and pitfalls of using bibliometric data.

2. Data Science skills to use bibliometric data (45 minutes, [videos: 2_1, 2_2])

- [video 2_1] Introduction to parallelised analysis of large-scale bibliometric, text and network data (with Dask in Python, DuckDB and DBeaver in SQL)

- [video 2_2] Hands-on example of parallelised analysis of bibliometric data

- [video 2_2] Hands-on example on use of text and network analysis

3. Example empirical study using bibliometric data (30 minutes, [video 3])
   - Internal and international migration of scholars worldwide: Trends, patterns, and inter-relationships

Section on more advanced data processing using parallelization

Introductory course to R or Python?!

Please check this repository by Vincent Traag and others, for an introductory course and code to familiarize yourself with Python:

https://github.com/vtraag/intro-python

Or this one by Data Carpentry:

https://datacarpentry.org/python-ecology-lesson/

This course by Data Carpentry provides basics in R:

https://datacarpentry.org/R-genomics/index.html

Parallelised analysis of large-scale bibliometric data (with Dask in Python, DuckDB and DBeaver in SQL); Using example of ORCID 2019 XML files

Materials under code, data and output folders of repository, Python users see:
https://github.com/akbaritabar/BiblioDemography_IMPRS_PHDS_2022_IDEM187/blob/main/0_code/03_parallelization_with_dask_duckdb_dbeaver.md

1. Why do we need to learn about parallelization and out of memory computation?
2. (if you are convinced), Required installation and set-up
3. Data preparation using Dask in Python
4. Further processing and analyzing data with SQL, using DuckDB and DBeaver

Section on text analysis

Code:

https://github.com/akbaritabar/BiblioDemography_IMPRS_PHDS_2022_IDEM187/blob/main/0_code/05_text_analysis_exact_and_anchor_words.py

https://github.com/akbaritabar/BiblioDemography_IMPRS_PHDS_2022_IDEM187/blob/main/0_code/07_text_analysis_noun_phrase_clauses.py

Data and methods (1/2)

Clarivate's Web of Science[1] (WOS), 1990-(end of) 2018, "article" and "review" publications

Corpus construction: search the lowercased "**plasticity**" in title, abstract and keywords

To further limit the usages of this concept to the areas closer (but not limited to) neurosciences, we follow **three** strategies

1) **Subject categories** by WOS (excluding less relevant fields, but they are covered in more precise strategies)

2) Specific **keyword combinations** (word pairs):

- **'adult neurogenesis', 'synaptic plasticity', 'cortical plasticity', 'cortical map plasticity', 'receptive field plasticity', 'heterosynaptic plasticity', 'hebbian plasticity', 'structural plasticity', 'neuronal responses', 'plasticity of neuronal responses', 'bidirectional plasticity', 'functional plasticity', 'developmental plasticity', 'critical period plasticity', 'adult brain plasticity', 'occular dominance plasticity', 'neuronanatomical plasticity', 'cross modal plasticity'**

[1]Provided by the German competence center for bibliometrics Data is obtained from Kompetenzzentrum Bibliometrie, which is funded by the Federal Ministry for Education and Research (BMBF), Germany with grant number 16WIK2101A. http://bibliometrie.info/

Data and methods (2/2)

Two previous approaches could be driven by

- a) the way WOS defines the subject categories (*journal based*)
- b) the keyword combinations highlighted previously in the literature could be **driven by the area of focus (niche naming/labeling of similar things, i.e., academic dialects)**

3) **A parallel attempt**, we try to apply a more objective criteria using text and content analysis (found words).

- **nltk** library in Python (Bird, Loper and Klein, 2009), first tokenize the words used in the title and abstract of publications. We decided to tokenize the text as our first step since we need the punctuation and sentence structure in determining the words. We then remove punctuations from these tokenized text, and lower case all the text elements. We then use plasticity as an “anchor word” and see if it is used once or more in the title and abstract. We identify which are the words pairing with it as being used before it.
- There were specific cases where the word used before plasticity was not a proper adjective or a scientific term, we excluded those from the following visualizations.

Example 1 of text analysis using exact word pairs

PK_ITEMS	SOURCETITLE	UT_EID	PUBYEAR	DOCTYPE	SC_DESCRIPTION	OECD_DESCRIPTION
35583791345	PLOS COMPUTATIONAL BIOLOGY	000463877900036	2019	Article	BIOCHEMICAL RESEARCH METHODS	NATURAL SCIENCES

Exact word pairs

Title

Synaptic plasticity onto inhibitory neurons as a mechanism for **ocular dominance plasticity**[1]

Abstract

Ocular dominance plasticity is a well-documented phenomenon allowing us to study properties of cortical maturation. Understanding this maturation might be an important step towards unravelling how cortical circuits function. However, it is still not fully understood which mechanisms are responsible for the opening and closing of the critical period for ocular dominance and how changes in cortical responsiveness arise after visual deprivation. In this article, we present a theory of ocular dominance plasticity. Following recent experimental work, we propose a framework where a reduction in inhibition is necessary for ocular dominance plasticity in both juvenile and adult animals. In this framework, two ingredients are crucial to observe ocular dominance shifts: a sufficient level of inhibition as well as excitatory-to-inhibitory synaptic plasticity. In our model, the former is responsible for the opening of the critical period, while the latter limits the plasticity in adult animals. Finally, we also provide a possible explanation for the variability in ocular dominance shifts observed in individual neurons and for the counter-intuitive shifts towards the closed eye. Author summary During the development of the brain, visual cortex has a period of increased plasticity. Closing one eye for multiple days during this period can have a profound and life-long impact on neuronal responses. A well-established hypothesis is that the absolute level of inhibition regulates this period. In light of recent experimental results, we suggest an alternative theory. We propose that, in addition to the level of inhibition, synaptic plasticity onto inhibitory neurons is just as crucial. We propose a model which explains many observed phenomena into one single framework. Unlike theories considering only the level of inhibition, we can account for both the onset as well as the closure of this period. Furthermore, we also provide an explanation for the small fraction of neurons that show counter-intuitive behaviour and provide some testable predictions.

[1] I had a typo in ocular in my exact words definition, writing it as “occular”, and this very accident shows how this method is prone to accidental exclusion (later my coauthor shared that “occular” is correct!)

Example 1 of text analysis using anchors

PK_ITEMS	SOURCETITLE	UT_EID	PUBYEAR	DOCTYPE	SC_DESCRIPTION	OECD_DESCRIPTION
35583791345	PLOS COMPUTATIONAL BIOLOGY	000463877900036	2019	Article	BIOCHEMICAL RESEARCH METHODS	NATURAL SCIENCES

Word used before — Anchor — Word used after

Title

Synaptic plasticity onto inhibitory neurons as a mechanism for ocular dominance plasticity

Abstract

Ocular dominance plasticity is a well-documented phenomenon allowing us to study properties of cortical maturation. Understanding this maturation might be an important step towards unravelling how cortical circuits function. However, it is still not fully understood which mechanisms are responsible for the opening and closing of the critical period for ocular dominance and how changes in cortical responsiveness arise after visual deprivation. In this article, we present a theory of ocular dominance plasticity. Following recent experimental work, we propose a framework where a reduction in inhibition is necessary for ocular dominance plasticity in both juvenile and adult animals. In this framework, two ingredients are crucial to observe ocular dominance shifts: a sufficient level of inhibition as well as excitatory-to-inhibitory synaptic plasticity. In our model, the former is responsible for the opening of the critical period, while the latter limits ~~the~~ plasticity in adult animals. Finally, we also provide a possible explanation for the variability in ocular dominance shifts observed in individual neurons and for the counter-intuitive shifts towards the closed eye. Author summary During the development of the brain, visual cortex has a period of ~~increased~~ plasticity. Closing one eye for multiple days during this period can have a profound and life-long impact on neuronal responses. A well-established hypothesis is that the absolute level of inhibition regulates this period. In light of recent experimental results, we suggest an alternative theory. We propose that, in addition to the level of inhibition, synaptic plasticity onto inhibitory neurons is just as crucial. We propose a model which explains many observed phenomena into one single framework. Unlike theories considering only the level of inhibition, we can account for both the onset as well as the closure of this period. Furthermore, we also provide an explanation for the small fraction of neurons that show counter-intuitive behaviour and provide some testable predictions.

Example 1 of text analysis using exact Noun-Phrase-Clauses

PK_ITEMS	SOURCETITLE	UT_EID	PUBYEAR	DOCTYPE	SC_DESCRIPTION	OECD_DESCRIPTION
35583791345	PLOS COMPUTATIONAL BIOLOGY	000463877900036	2019	Article	BIOCHEMICAL RESEARCH METHODS	NATURAL SCIENCES

Relevant noun-phrase clauses found

Title

Synaptic plasticity onto inhibitory neurons as a mechanism for ocular dominance plasticity[1]

Abstract

Ocular dominance plasticity is a well-documented phenomenon allowing us to study properties of cortical maturation. Understanding this maturation might be an important step towards unravelling how cortical circuits function. However, it is still not fully understood which mechanisms are responsible for the opening and closing of the critical period for ocular dominance and how changes in cortical responsiveness arise after visual deprivation. In this article, we present a theory of ocular dominance plasticity. Following recent experimental work, we propose a framework where a reduction in inhibition is necessary for ocular dominance plasticity in both juvenile and adult animals. In this framework, two ingredients are crucial to observe ocular dominance shifts: a sufficient level of inhibition as well as excitatory-to-inhibitory synaptic plasticity. In our model, the former is responsible for the opening of the critical period, while the latter limits the plasticity in adult animals. Finally, we also provide a possible explanation for the variability in ocular dominance shifts observed in individual neurons and for the counter-intuitive shifts towards the closed eye. Author summary During the development of the brain, visual cortex has a period of increased plasticity. Closing one eye for multiple days during this period can have a profound and life-long impact on neuronal responses. A well-established hypothesis is that the absolute level of inhibition regulates this period. In light of recent experimental results, we suggest an alternative theory. We propose that, in addition to the level of inhibition, synaptic plasticity onto inhibitory neurons is just as crucial. We propose a model which explains many observed phenomena into one single framework. Unlike theories considering only the level of inhibition, we can account for both the onset as well as the closure of this period. Furthermore, we also provide an explanation for the small fraction of neurons that show counter-intuitive behaviour and provide some testable predictions.

[1] I had a typo in ocular in my exact words definition, writing it as “occular”, and this very accident shows how this method is prone to accidental exclusion (later my coauthor shared that “occular” is correct!)

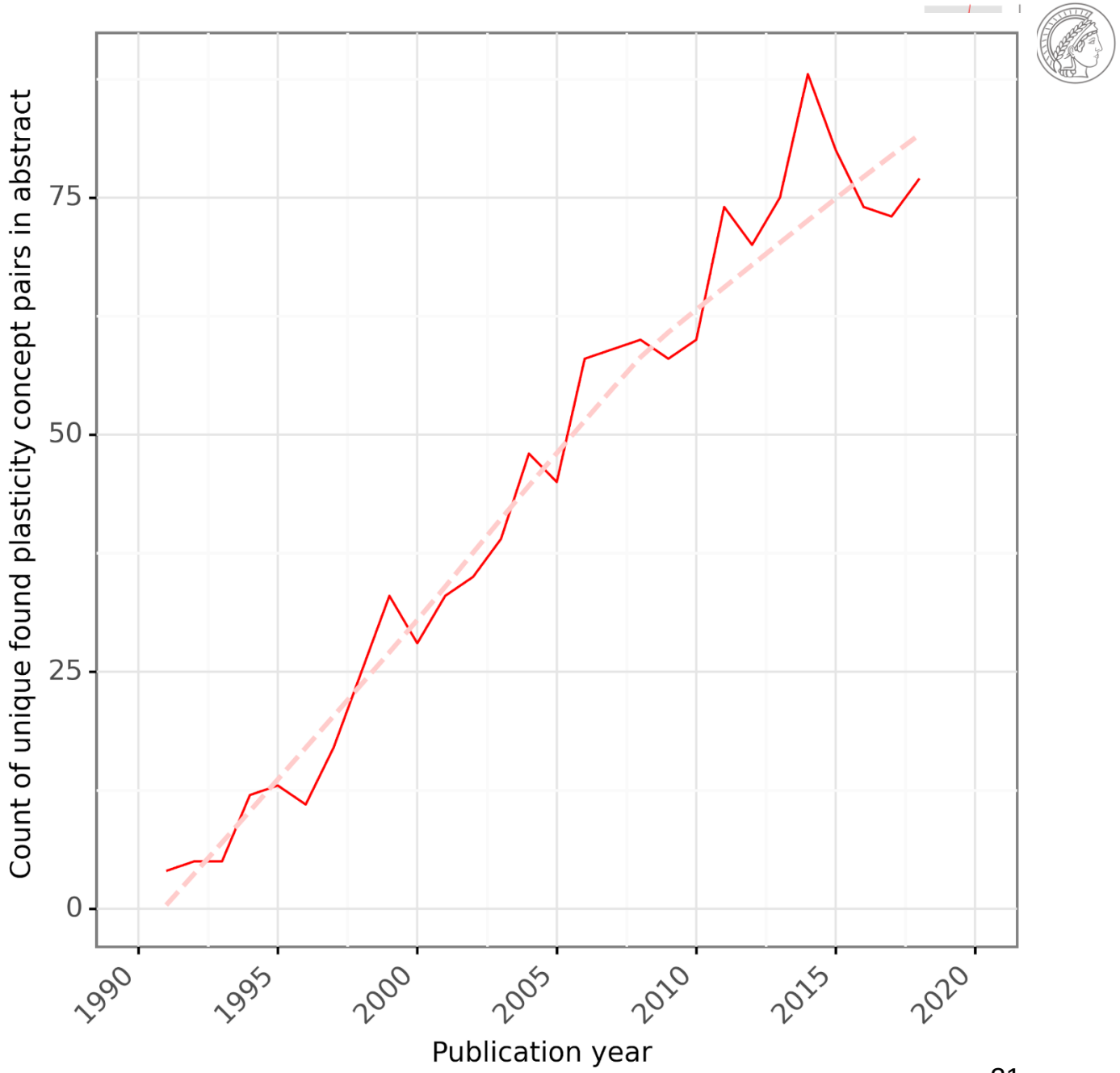
Plasticity as a concept in the neurosciences, a word pair analysis

- Plasticity is an anchor concept

- Trend of unique word (or words in case multiple words are used in the same abstract) used before plasticity

- Red line shows the count of unique words per year and smooth trend based on mean is shown on light red dashed line.

- The count of words pairing with plasticity has been constantly on the rise since recent years. Year 2014 has the highest count of unique words (88) paired with plasticity.



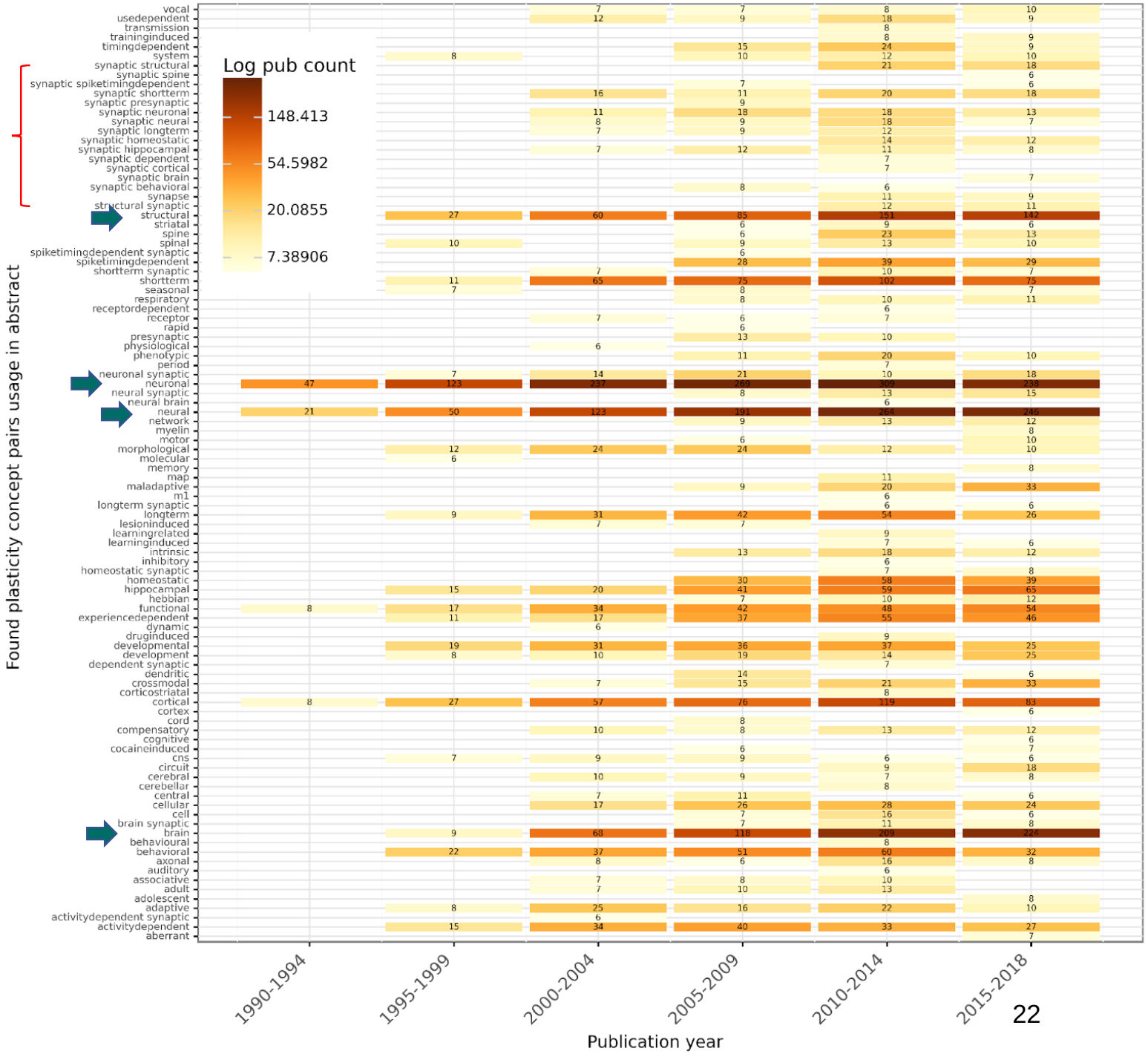
Plasticity as a concept in the neurosciences, a word pair analysis

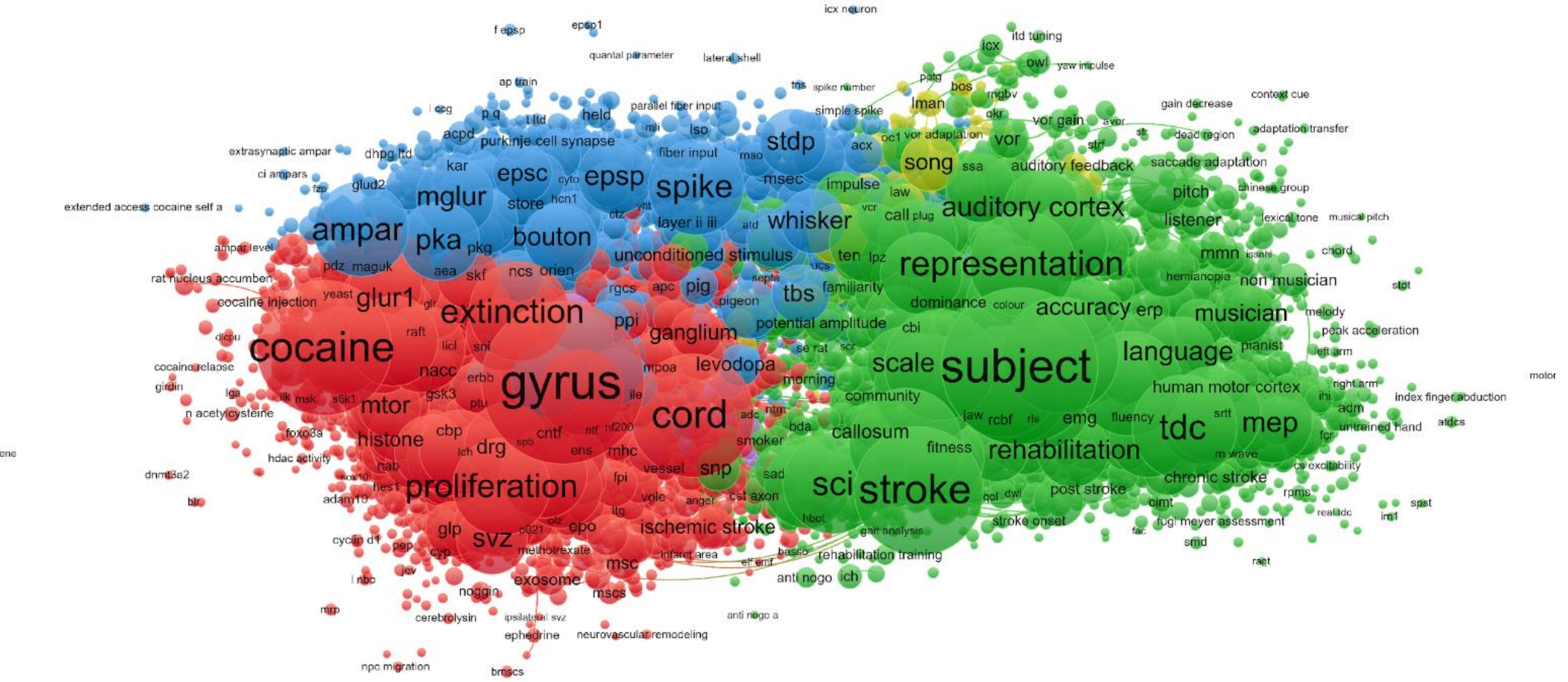
- Synaptic plasticity is the central anchor concept

- Found words used at least 5 times before plasticity as an anchor word and temporal use of them in the abstract of scientific publications in the neurosciences (based on WOS subject categories)

- If two words are used before plasticity in the same abstract, they are presented together on the Figure. We exclude “synaptic” as it dominates the visualization (see red curly bracket for some) and it is covered before

- *Neural*, *neuronal*, *structural*, and *brain* are the most frequently used words before plasticity (green arrows)





Data oriented analysis without domain-knowledge gives less relevant results

VOSviewer

Text analysis in R?!

If you are an R user, consider Monica Alexander's writeup and RMD file below that takes demography papers and does cool text analysis on them.

Link to Monica's slides/files:

https://github.com/MJAlexander/demopop-workshop/blob/main/rmd/2_articles.Rmd

In addition, you might want to check out tm (text mining) and stm (structural topic models) in R.

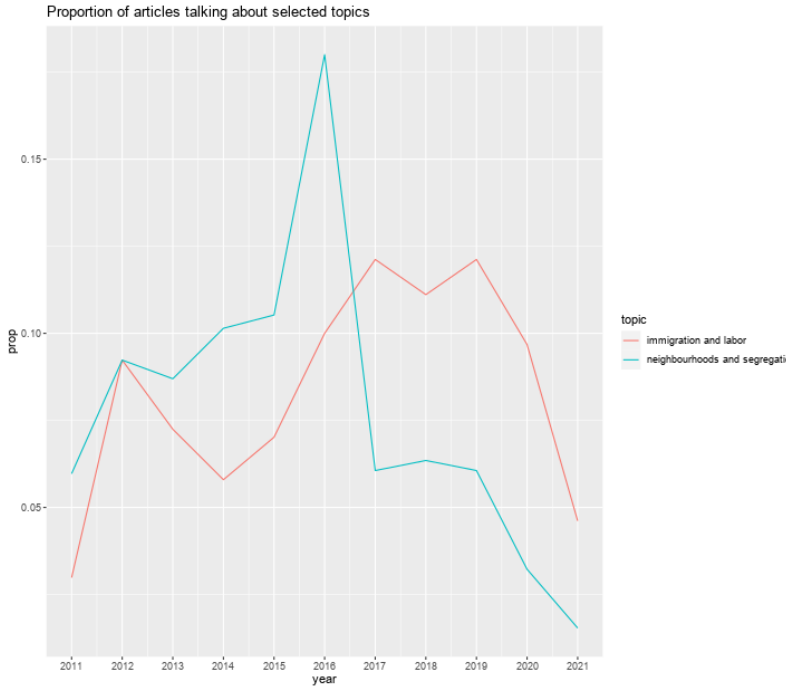
```
r$> bigrams_separated <- bigrams %>%
      separate(bigram, c("word1", "word2"), sep = " ")

    bigrams_filtered <- bigrams_separated %>%
      filter(!word1 %in% stop_words$word) %>%
      filter(!word2 %in% stop_words$word)

    bigrams_united <- bigrams_filtered %>%
      unite(bigram, word1, word2, sep = " ") %>%
      filter(bigram!="NA NA")

    bigrams_united %>%
      group_by(bigram) %>%
      tally() %>%
      arrange(-n) %>%
      filter(!str_detect(bigram, "table\ "), bigram!= "online appendix",
             !str_detect(bigram, "al "),
             !str_detect(bigram, "resource")) %>%
      top_n(20)
Selecting by n
# A tibble: 20 x 2
   bigram                      n
   <chr>                   <int>
 1 statistically significant  1925
 2 fixed effects              1866
 3 life expectancy            1734
 4 labor market               1643
 5 standard errors            1551
 6 birth weight               1263
 7 labor force                1193
 8 family structure           1126
 9 sex couples                1085
10 family size                1018
11 data set                    954
12 infant mortality            928
13 race ethnicity              923
14 mortality rates             871
15 dummy variables             856
16 age specific                848
17 cross sectional             834
18 foreign born                822
19 sex ratio                   800
20 low income                  739

r$>
```



Structural topic models (one avenue of further modeling text data)

- **Was an introduction to “text as data”, using examples from scientific publications’ abstracts from our previous research. Examples were in nltk, python base libraries and a bit on noun-phrase-clauses identification and extraction using Apache’s OpenNLP and in python with spacy pre-trained models.**
- **Now, brief on modelling possibilities (e.g., structural topic models, stm package in R)**
  - See example slides here: https://github.com/akbaritabar/BiblioDemography_IMPRS_PHDS_2022_IDEM187/blob/main/2_presentations/10_Structural_topic_models_an_example.pdf
- **Other topic modeling possibilities (e.g., LDA and SBM-like methods**)**

**** Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018). A network approach to topic models. Science Advances, 4(7), eaaq1360. https://doi.org/10.1126/sciadv.aaq1360**

[Brief, self-study] Introduction to Social Network Analysis (SNA)

[Brief] Section on Network analysis of scientific collaborations

[self-study] Introduction to Social Network Analysis (SNA)

Materials under code, data and output folders of workshop repository, R users see:
https://github.com/akbaritabar/BiblioDemography_IMPRS_PHDS_2022_IDEM187/blob/main/0_code/08_HU_seminar_network_analysis.Rmd

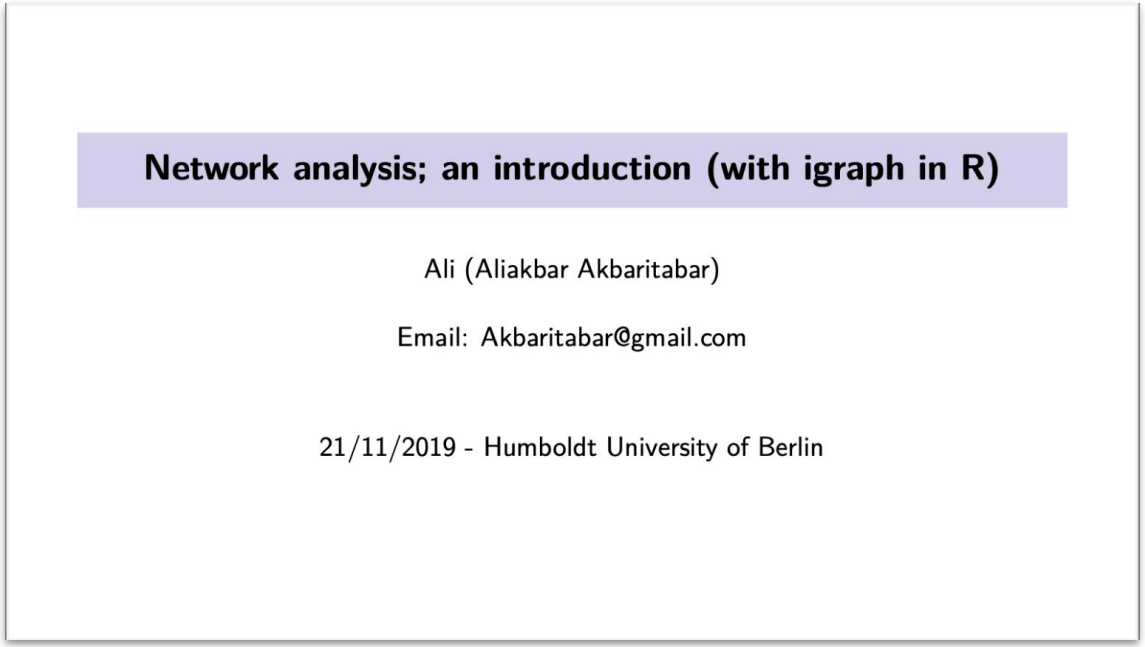
(Python users, can see an example in:
https://github.com/akbaritabar/BiblioDemography_IMPRS_PHDS_2022_IDEM187/blob/main/0_code/09_example_network_igraph_python.py **)**

Presentation file:

https://github.com/akbaritabar/BiblioDemography_IMPRS_PHDS_2022_IDEM187/blob/main/2_presentations/08_HU_seminar_network_analysis.pdf

Outline of brief introduction to Network Analysis:

- **What is relational view and network analysis?**
- **Ethnography of network ties! Context of interactions**
- **How to gather and use network data?**
- **Possible questions to ask!**
- **A real life example from science studies!**
- **Where to next?!**



Constructing co-authorship edge-list (& network) from authorship list in publications

Scientometrics
https://doi.org/10.1007/s11192-022-04351-4

Return migration of German-affiliated researchers: analyzing departure and return by gender, cohort, and discipline using Scopus bibliometric data 1996–2020

Xinyi Zhao[1,2] · Samin Aref[1,3] · Emilio Zagheni[1] · Guy Stecklov[4]

Received: 15 October 2021 / Accepted: 10 March 2022

Abstract
The international migration of researchers is an important dimension of scientific mobility, and has been the subject of considerable policy debate. However, tracking the migration life courses of researchers is challenging due to data limitations. In this study, we use Scopus bibliometric data on eight million publications from 1.1 million researchers who have published at least once with an affiliation address from Germany in 1996–2020. We construct the partial life histories of published researchers in this period and explore both their out-migration and the subsequent return of a subset of this group: the returnees. Our analyses shed light on the career stages and gender disparities between researchers who remain in Germany, those who emigrate, and those who eventually return. We find that the return migration streams are even more gender imbalanced, which points to the need for additional efforts to encourage female researchers to come back to Germany. We document a slightly declining trend in return migration among more recent cohorts of researchers who left Germany, which, for most disciplines, was associated with a decrease in the German collaborative ties of these researchers. Moreover, we find that the gender disparities for the most gender imbalanced disciplines are unlikely to be mitigated by return migration

Undirected Edge-lists

Zhao -- Aref	Miranda -- Aref
Zhao -- Zagheni	Miranda -- Theile
Zhao -- Stecklov	Miranda -- Zagheni
Aref -- Zagheni	Aref -- Zagheni
Aref -- Stecklov	Aref -- Theile
Zagheni -- Stecklov	Theile -- Zagheni

Miranda-González et al. *EPJ Data Science* (2020) 9:34
https://doi.org/10.1140/epjds/s13688-020-00252-9
EPJ Data Science
a SpringerOpen Journal
EPJ.org
REGULAR ARTICLE Open Access

Scholarly migration within Mexico: analyzing internal migration among researchers using Scopus longitudinal bibliometric data

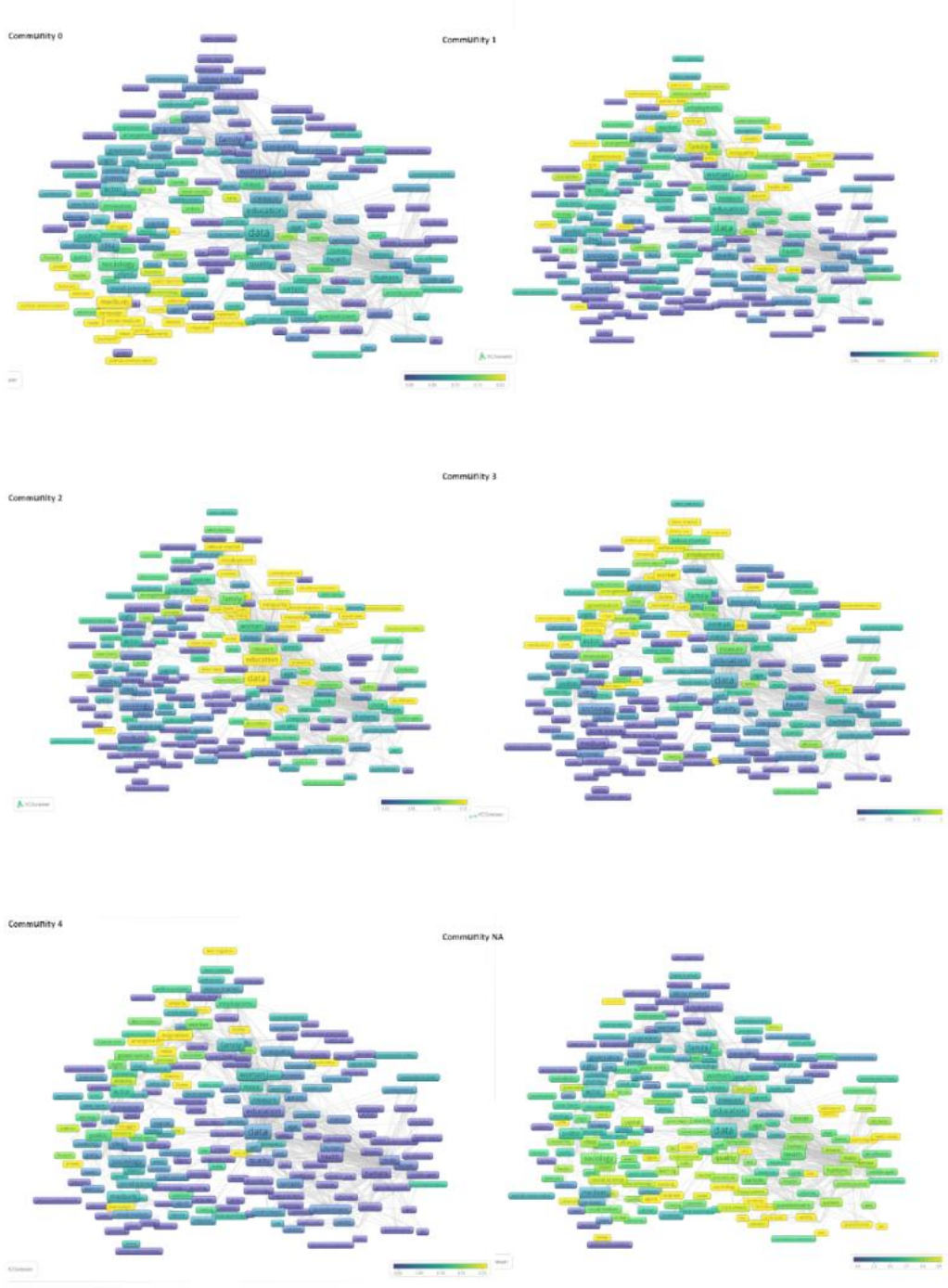
Andrea Miranda-González[1,2], Samin Aref[1], Tom Theile[2] and Emilio Zagheni[2]

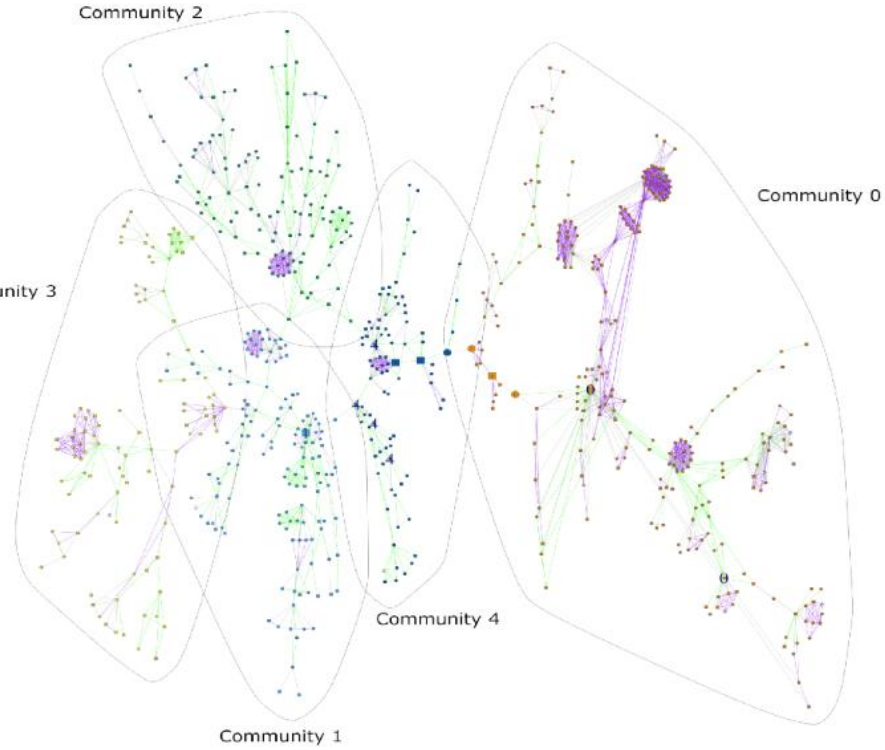
*Correspondence:
aref@demogr.mpg.de
[2]Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research, Rostock, Germany
Full list of author information is available at the end of the article

Abstract
The migration of scholars is a major driver of innovation and of diffusion of knowledge. Although large-scale bibliometric data have been used to measure international migration of scholars, our understanding of internal migration among researchers is very limited. This is partly due to a lack of data aggregated at a suitable sub-national level. In this study, we analyze internal migration in Mexico based on over 1.1 million authorship records from the Scopus database. We trace the movements of scholars between Mexican states, and provide key demographic measures of internal migration for the 1996–2018 period. From a methodological perspective, we develop a new framework for enhancing data quality, inferring states from affiliations, and detecting moves from modal states for the purposes of studying internal migration among researchers. Substantively, we combine demographic and network science techniques to improve our understanding of internal migration patterns within country boundaries. The 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 migration destinations. We observed that over the past two decades, there has been a general decreasing trend in the crude migration intensity. However, the migration network has become more dense and more diverse, and has included greater exchanges between states along the Gulf and the Pacific Coast. Our analysis, which is mostly empirical in nature, lays the foundations for testing and developing theories that can rely on the analytical framework developed by migration scholars, and the richness of appropriately processed bibliometric data.

Keywords: High-skilled migration; Internal migration; Computational demography; Science of science; Network science; Brain circulation

Contrasting co-authorship network's structure (and communities) with substantive content of publications in focus of each community





Read here: https://doi.org/10.1007/s11192-020-03555-w

More networks?

Other type of networks that are frequently built/analyzed: citation networks, bibliographic coupling, etc.

See some examples here: Gao, D., Akbaritabar, A. Using agent-based modeling in routine dynamics research: a quantitative and content analysis of literature. *Rev Manag Sci* 16, 521–550 (2022). https://doi.org/10.1007/s11846-021-00446-z

Hands-on session with examples

Please tweet with hashtag
#BiblioDemography,
a tribute to James W. Vaupel (1945-2022).
Thanks Ilya and Jonas for bringing up Jim's
labeling idea!

Aliakbar Akbaritabar

Research Scientist

Akbaritabar@demogr.mpg.de

https://akbaritabar.github.io/

https://twitter.com/Akbaritabar

THANK YOU!