Pre-EPC2022 workshop:

Using bibliometric data for demographic research

Organizers:

Aliakbar Akbaritabar[1]

Xinyi Zhao[1,2]

[1] Max Planck Institute for Demographic Research (MPIDR)

[2] Leverhulme Centre for Demographic Science, Department of Sociology, University of Oxford

Please tweet with hashtag
#BiblioDemography,
a tribute to James W. Vaupel (1945-2022).
Thanks Ilya and Jonas for bringing up Jim's
labeling idea!
and #EPC2022

AGENDA (3:45')

1. Overview of literature on using bibliometric data for demographic research and potential research projects (30 minutes)

2. Panel discussion (90 minutes) with invited speakers

(15 minutes pause/break)

3. Introduction to available data sources, especially publicly available data (15 minutes)

4. Introduction to handling bibliometric data for demographic research: data retrieval, data pre-processing and repurposing based on research questions (including Q&A) (60+15 minutes, includes 10 minutes pause/break)

5. Conclusion and discussion of potential limitations and future directions of using bibliometric data in demographic research (15 minutes)

KEY CONTRIBUTORS TO PROJECTS ON SCHOLARLY MIGRATION AT MPIDR

Emilio Zagheni
(MPIDR)

Aliakbar Akbaritabar
(MPIDR)

Andrea Miranda-Gonzalez
(UC Berkeley)

Samin Aref
(U. of Toronto)

Ebru Sanliturk
(MPIDR)

Maciej J. Dańko
(MPIDR)

Tom Theile
(MPIDR)

Xinyi Zhao
(MPIDR)

Welcome and introduction

Use of bibliometric data for demographic research and study of scholarly migration at the MPIDR:

- Zhao, X., Aref, S., Zagheni, E., & Stecklov, G. (2022). Return migration of German-affiliated researchers: Analyzing departure and return by gender, cohort, and discipline using Scopus bibliometric data 1996-2020. Scientometrics

- Kashyap, R., Rinderknecht, R. G., Akbaritabar, A., Alburez-Gutierrez, D., Gil-Clavel, S., Grow, A., Kim, J., Leasure, D. R., Lohmann, S., Negraia, D. V., Perrotta, D., Rampazzo, F., Tsai, C.-J., Verhagen, M. D., Zagheni, E., & Zhao, X. (2022). *Digital and Computational Demography*. SocArXiv. https://doi.org/10.31235/osf.io/7bvpt

- Zhao, X., Aref, S., Zagheni, E., & Stecklov, G. (2021a). International Migration in Academia and Citation Performance: An Analysis of German-Affiliated Researchers by Gender and Discipline Using Scopus Publications 1996-2020. ArXiv:2104.12380 [Cs]. http://arxiv.org/abs/2104.12380

- Subbotin, A., & Aref, S. (2021). Brain drain and brain gain in Russia: Analyzing international migration of researchers by discipline using Scopus bibliometric data 1996–2020. Scientometrics. https://doi.org/10.1007/s11192-021-04091-x

- Miranda-González, A., Aref, S., Theile, T., & Zagheni, E. (2020). Scholarly migration within Mexico: Analyzing internal migration among researchers using Scopus longitudinal bibliometric data. EPJ Data Science, 9(1), 34. https://doi.org/10.1140/epjds/s13688-020-00252-9

- Aref, S., Zagheni, E., & West, J. (2019). The Demography of the Peripatetic Researcher: Evidence on Highly Mobile Scholars from the Web of Science. In I. Weber, K. M. Darwish, C. Wagner, E. Zagheni, L. Nelson, S. Aref, & F. Flöck (Eds.), Social Informatics (pp. 50–65). Springer International Publishing. https://doi.org/10.1007/978-3-030-34971-4_4

- Alburez-Gutierrez, D., Zagheni, E., Aref, S., Gil-Clavel, S., Grow, A., & Negraia, D. V. (2019). Demography in the digital era: New data sources for population research.

- Abel, G. J., Muttarak, R., Bordone, V., & Zagheni, E. (2019). Bowling Together: Scientific Collaboration Networks of Demographers at European Population Conferences. European Journal of Population, 35(3), 543–562. https://doi.org/10.1007/s10680-018-9493-1

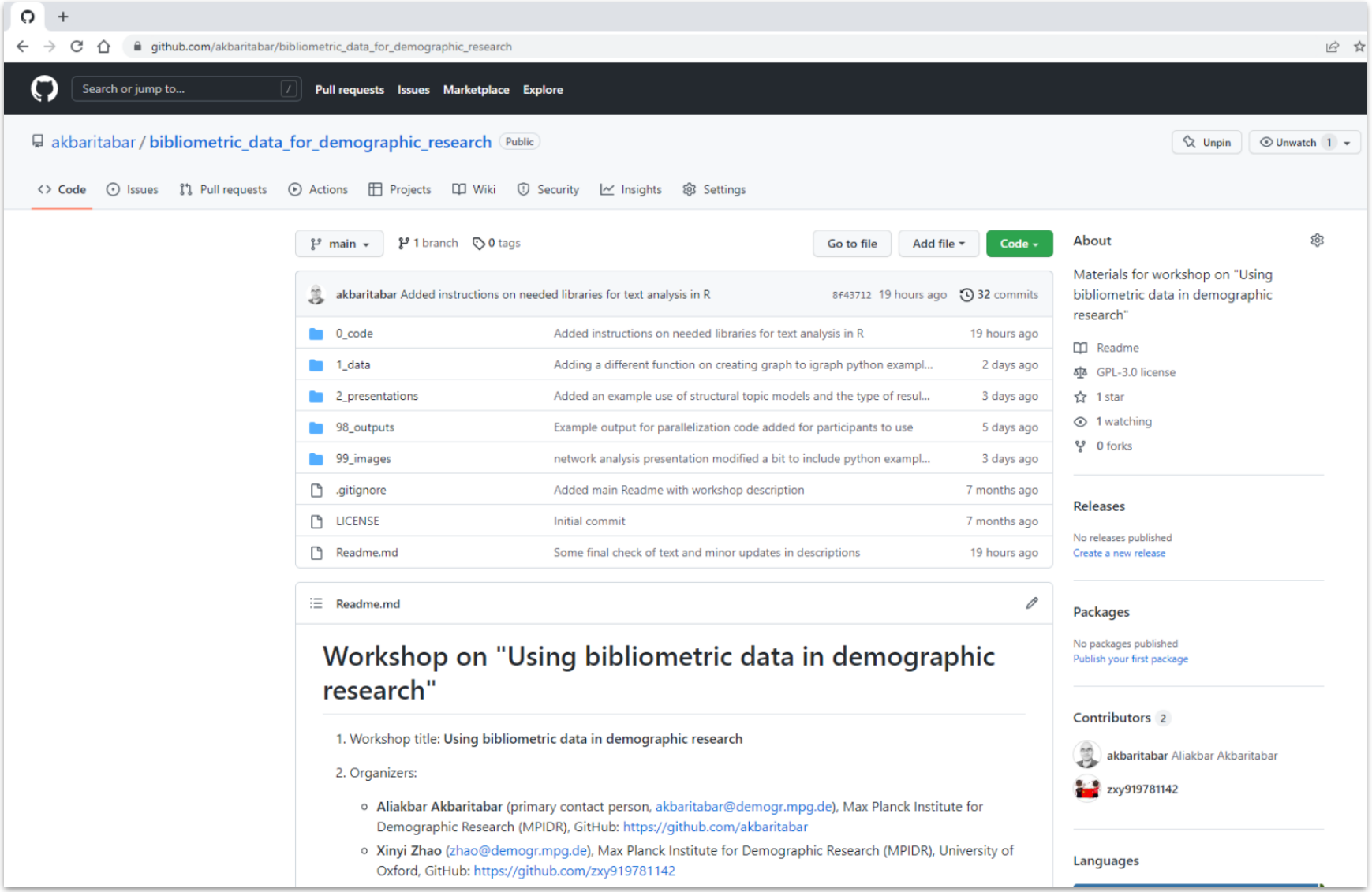
Materials publicly available

Related links:

- **IUSSP**: https://iussp.org/en/using-bibliometric-data-demographic-research

- **EPC2022**: https://eaps.nl/page/side-meetings-and-pre-conference-events

- Interview and references on using bibliometric data for demographic research: https://www.demogr.mpg.de/en/news_events_6123/news_press_releases_4630/news/how_to_use_bibliometric_data_for_demographic_research_10784



Workshop materials: https://github.com/akbaritabar/bibliometric_data_for_demographic_research

Target audience:

Postgraduate students, early career researchers, those interested in high-skilled and scholarly mobility/migration and those interested in science of science, bibliometrics, scientometrics and informetrics.

Aims:

The aim of this workshop is to enable attendees to:

- Familiarize themselves with the availability of bibliometric data as a type of digital data and its potentials for demographic research
- Familiarize themselves with the most commonly used and available bibliometric databases
- Learn about pre-processing steps and handling bibliometric data, including:
  - Author/organisation name disambiguation
  - Gender identification
  - Scholarly mobility and movement identification
- Learn about practical applications of bibliometric data in demographic research, including:
  - Use of some relevant scientometric indicators
  - Calculation and use of demographic indicators *(e.g., Crude Migration Intensity, Migration Effectiveness Index and the Aggregate Net Migration Rate)*
  - Network analysis of scientific collaborations
  - Text analysis of bibliometric information

1. Overview of literature on using bibliometric data for demographic research and potential research projects (30 minutes)

What is bibliometric data?

Bibliometric data: metadata on scientific documents and on citation links between these documents.

- **Authorship profile**
- **Affiliation addresses**
- **Publication type (article, review article, proceeding paper)**
- **Subject classification (often based on journal assignment)**
- **Descriptors of publications (keywords/subject headings)**
- **Title, abstract and keywords (for text mining)**
- **Complete reference list (for citation analysis)**
- **Acknowledgement (for funding information, sub-authorship)**
- **Open Access information**

What is bibliometric data?

Publication - Article

Bowling Together: Scientific Collaboration Networks of Demographers at European Population Conferences

European Journal of Population, 35(3), 543-562 - June 2018
https://doi.org/10.1007/s10680-018-9493-1 ↗

Authors

Guy J. Abel - Shanghai University; Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW, WU), International Institute for Applied Systems Analysis, Schlossplatz 1, 2361, Laxenburg, Austria

Raya Muttarak - Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW, WU), International Institute for Applied Systems Analysis, Schlossplatz 1, 2361, Laxenburg, Austria; University of East Anglia

Corresponding Author

Valeria Bordone - Ludwig-Maximilians-Universität München

Emilio Zagheni - University of Washington; Max Planck Institute for Demographic Research

1 less

Abstract

Studies of collaborative networks of demographers are relatively scarce. Similar studies in other social sciences provide insight into scholarly trends of both successful scientists. Exploiting a unique database of metadata for papers presented at six European Population Conferences, this report explores factors e demographers. We find that (1) collaboration among demographers has increased over the past 10 years, however, among co-authored papers, collaboratio unchanged over the period, (2) papers based on core demographic subfields such as fertility, mortality, migration and data and methods are more likely to in teams that are all female are less likely to co-author with colleagues in different institutions. Potential explanations for these results are discussed alongsid

More

Document history

2019-07 Published print
2018-06-18 Published online

Research Categories

Fields of Research

16 Studies in Human Society
1603 Demography

Data gathering techniques in studying scholars as a population

- **Surveys** of scholars (Cañibano et al., 2020; Franzoni et al., 2012a, 2012b, 2014, 2015; Gibson & McKenzie, 2014; Guthrie et al., 2017; Jöns, 2011; Nascia et al., 2021; Netz & Jaksztat, 2017; Scellato et al., 2015, 2017; Wilson & Gaston, 1971)
- **Interviews** (Cole & Zuckerman, 1987; Jonkers, 2011; Schaer, 2021; Schaer et al., 2017, 2020; Shen et al., 2017)
- **Administrative** and **census** data (Fenton et al., 2000; ippedico, 2021; Shauman & Xie, 1996)
- **Online sources** (e.g., LinkedIn data or website of universities) (Park et al., 2019; Yuret, 2017).
- Collection of **CVs** and **ORCID** (Aman, 2018b; Bohannon, 2017; Cañibano et al., 2008; Lawson & Shibayama, 2015; Li & Tang, 2019; Mejias & Eyre, 2019; Yan et al., 2020; Z. Zhao et al., 2020)

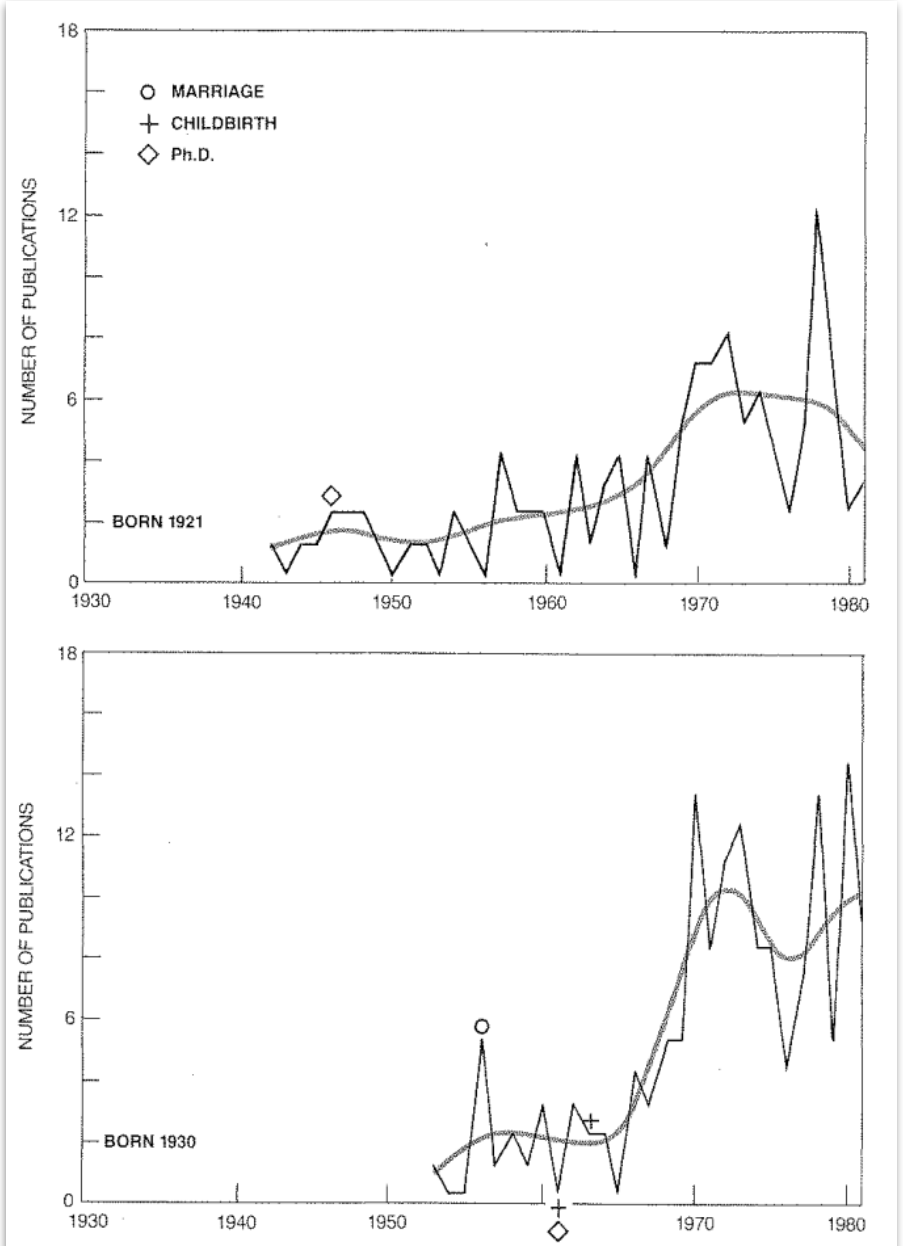
Two groups of literature in scientific mobility (scholarly migration)

- **Bibliometric** data (Alburez-Gutierrez et al., 2019, Kashyap et al. 2022, Aman, 2018a; Laudel, 2003; Moed et al., 2013; Moed & Halevi, 2014)
- Divided based on the **study focus** to two groups.
  - Focuses on the **geographic scale of academic mobility**.
    - Internal or international migration to/from a country (see the case of Russia (Subbotin & Aref, 2021), Mexico (Miranda-González et al., 2020), and Germany (X. Zhao et al., 2021))
    - Others have focused on the global mobility of scholars (Chinchilla-Rodríguez, Miao, et al., 2018; Czaika & Orazbayev, 2018; Robinson-Garcia et al., 2019).
  - Focuses on **(dis)advantages of mobility**, or its contributions to the field, knowledge transfer, institutional, national or global productivity and innovation.
  - Some research has focused on the performance and impact of mobile scientists or the so called "**mover's advantage**" (Aman, 2020; Bernstein et al., 2018; Franzoni et al., 2014; Halevi et al., 2016; J. A. Jacobs & Mizrachi, 2020). **Downsides** of scholarly mobility and costs that academics bear by leaving an academic context for another (Ackers & Gill, 2005; Schaer et al., 2017, 2020) or the potentials for (in)stability of scientific collaborations or difficulties of finding a job during or after mobility (Baruffaldi & Landoni, 2012; Z. Zhao et al., 2020).
  - **Policy changes** and how they can inspire (ippedico, 2021) or inhibit mobility of the general population and more specifically academics (Chinchilla-Rodríguez, Bu, et al., 2018; Sugimoto et al., 2017).

A very innovative interview set-up

- **Using bibliometric data alongside demographic life events**

Cole, J. R., & Zuckerman, H. (1987). Marriage, Motherhood and Research Performance in Science. Scientific American, 256(2), 119–125. JSTOR. https://doi.org/10.1038/scientificamerican0287-119



LOWER RATES OF PUBLICATION in the early part of a career are characteristic of both married men and single women. The publication profile of a distinguished woman biologist (*top*) who never married shows the same pattern of oscillations and an overall increase as the graphs of women who married and had children. The same pattern can be seen in the profile of an eminent male chemist (*bottom*). He published at a much slower pace when his children were young, although his domestic responsibilities were minimal.

Some of the main topics of research

- **Research assessment**
  - Researcher
  - Institution
  - National science system
- **Scientific migration**
  - National
  - International
- **Collaboration networks**
  - Researcher
  - Institution
  - National science system

Research assessment

Springer Link

Open Access | Published: 06 May 2010

Characterizing a scientific elite: the social characteristics of the most highly cited scientists in environmental science and ecology

John N. Parker, Christopher Lortie & Stefano Allesina

Scientometrics **85**, 129–143 (2010) | Cite this article

3744 Accesses | **47** Citations | **22** Altmetric | Metrics

Abstract

In science, a relatively small pool of researchers garners a disproportionally large number of citations. Still, very little is known about the social characteristics of highly cited scientists. This is unfortunate as these researchers wield a disproportional impact on their fields, and the study of highly cited scientists can enhance our understanding of the conditions which foster highly cited work, the systematic social inequalities which exist in science, and scientific careers more generally. This study provides information on this understudied subject by examining the social characteristics and opinions of the 0.1% most cited environmental scientists and ecologists. Overall, the social characteristics of these researchers tend to reflect broader patterns of inequality in the global scientific community. However, while the social characteristics of these researchers mirror those of other scientific elites in important ways, they differ in others, revealing findings which are both novel and surprising, perhaps

ELSEVIER

Journal of Informetrics

Volume 12, Issue 4, November 2018, Pages 1031-1041

Do females create higher impact research? Scopus citations and Mendeley readers for articles from five countries

Mike Thelwall

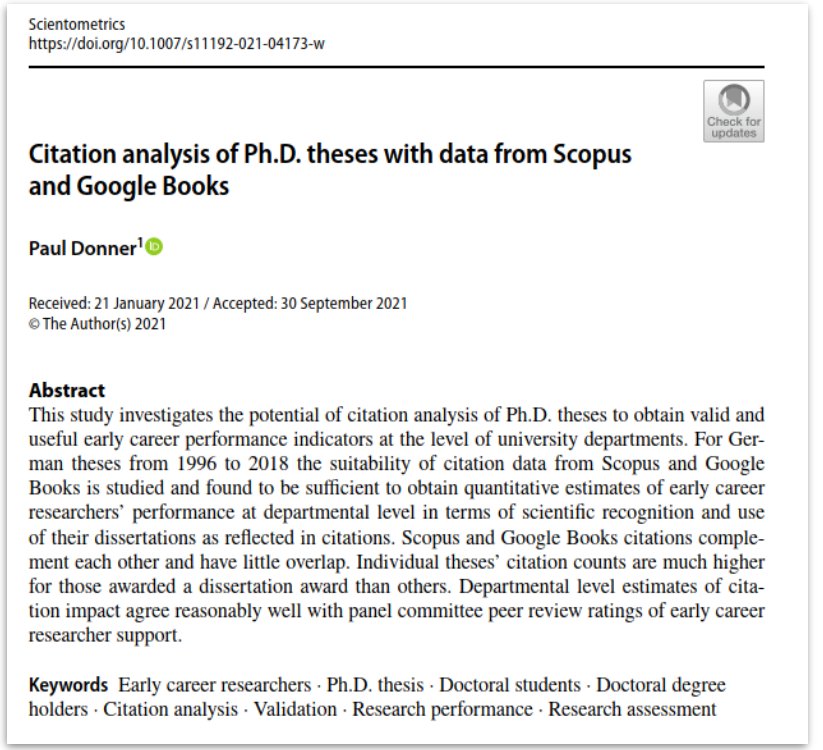
Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.joi.2018.08.005 Get rights and content

Highlights

- Female-authored research is less cited in Turkey and India.
- Female-authored research is marginally more cited in Spain, the UK, and the USA.



Scientometrics
https://doi.org/10.1007/s11192-021-04173-w

Check for updates

Citation analysis of Ph.D. theses with data from Scopus and Google Books

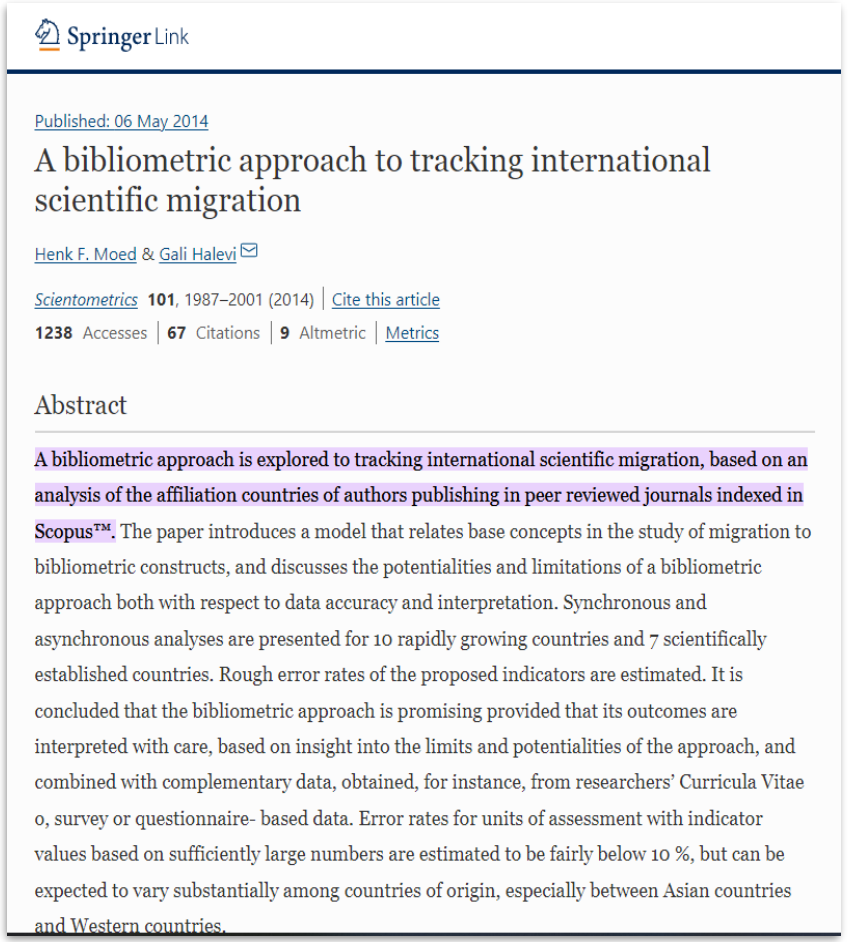
Paul Donner[1]

Received: 21 January 2021 / Accepted: 30 September 2021

Abstract
This study investigates the potential of citation analysis of Ph.D. theses to obtain valid and useful early career performance indicators at the level of university departments. For German theses from 1996 to 2018 the suitability of citation data from Scopus and Google Books is studied and found to be sufficient to obtain quantitative estimates of early career researchers' performance at departmental level in terms of scientific recognition and use of their dissertations as reflected in citations. Scopus and Google Books citations complement each other and have little overlap. Individual theses' citation counts are much higher for those awarded a dissertation award than others. Departmental level estimates of citation impact agree reasonably well with panel committee peer review ratings of early career researcher support.

Keywords Early career researchers · Ph.D. thesis · Doctoral students · Doctoral degree holders · Citation analysis · Validation · Research performance · Research assessment

Scientific migration



Springer Link

Published: 06 May 2014

A bibliometric approach to tracking international scientific migration

Henk F. Moed & Gali Halevi

Scientometrics **101**, 1987–2001 (2014) | Cite this article

Abstract

A bibliometric approach is explored to tracking international scientific migration, based on an analysis of the affiliation countries of authors publishing in peer reviewed journals indexed in Scopus™. The paper introduces a model that relates base concepts in the study of migration to bibliometric constructs, and discusses the potentialities and limitations of a bibliometric approach both with respect to data accuracy and interpretation. Synchronous and asynchronous analyses are presented for 10 rapidly growing countries and 7 scientifically established countries. Rough error rates of the proposed indicators are estimated. It is concluded that the bibliometric approach is promising provided that its outcomes are interpreted with care, based on insight into the limits and potentialities of the approach, and combined with complementary data, obtained, for instance, from researchers' Curricula Vitae o, survey or questionnaire- based data. Error rates for units of assessment with indicator values based on sufficiently large numbers are estimated to be fairly below 10 %, but can be expected to vary substantially among countries of origin, especially between Asian countries and Western countries.

ELSEVIER

Journal of Informetrics
Volume 13, Issue 1, February 2019, Pages 50-63

The many faces of mobility: Using bibliometric data to measure the movement of scientists ☆

Nicolás Robinson-Garcia [a, b], Cassidy R. Sugimoto [c, d], Dakota Murray [c], Alfredo Yegros-Yegros [d], Vincent Larivière [d, e, f], Rodrigo Costas [d, g]

Show more

+ Add to Mendeley | Share | Cite

https://doi.org/10.1016/j.joi.2018.11.002

Highlights

- We present a novel taxonomy of mobility types of scientists based on bibliometric data.
- Four mobility classes are presented: not mobile researchers, directional travelers, non-directional travelers and migrants.
- Travelers are defined as those who have affiliation ties with more than one country, always maintaining a link with their country of origin.
- We present a brief analysis worldwide to illustrate the potential of these mobility classes to inform science policy.

EPJ Data Science

About | **Articles** | Submission Guidelines

Regular article | Open Access | Published: 05 November 2020

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

Andrea Miranda-González, Samin Aref, Tom Theile & Emilio Zagheni

EPJ Data Science **9**, Article number: 34 (2020) | Cite this 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

Collaboration networks

Scientometrics (2020) 124:2361–2382
https://doi.org/10.1007/s11192-020-03555-w

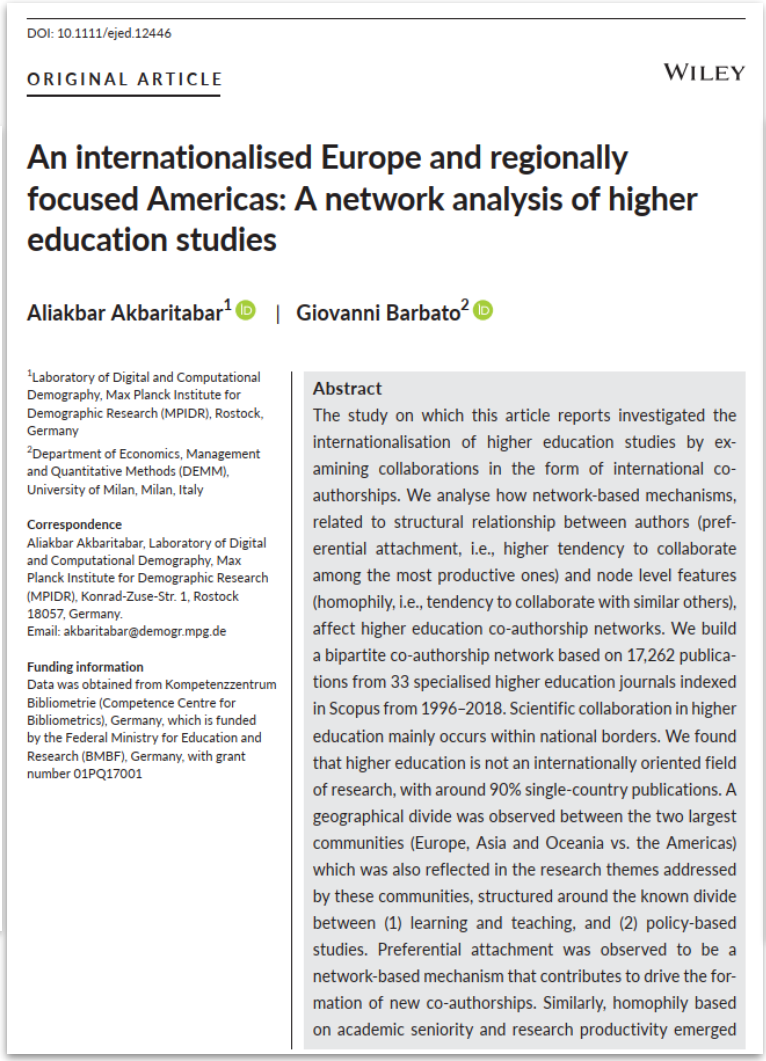
Check for updates

Italian sociologists: a community of disconnected groups

Aliakbar Akbaritabar[1] · Vincent Antonio Traag[2] · Alberto Caimo[3] · Flaminio Squazzoni[4]

Received: 14 February 2020 / Published online: 13 June 2020

Abstract
Examining coauthorship networks is key to study scientific collaboration patterns and structural characteristics of scientific communities. Here, we studied coauthorship networks of sociologists in Italy, using temporal and multi-level quantitative analysis. By looking at publications indexed in Scopus, we detected research communities among Italian sociologists. We found that Italian sociologists are fractured in many disconnected groups. The giant connected component could be split into five main groups with a mix of three main disciplinary topics: sociology of culture and communication (present in two groups), economic sociology (present in three groups) and general sociology (present in three groups). By applying an exponential random graph model, we found that collaboration ties are mainly driven by the *research interests* of these groups. Other factors, such as *preferential attachment*, *gender* and *affiliation homophily* are also important, but the effect of gender fades away once other factors are controlled for. Our research shows the advantages of multi-level and temporal network analysis in revealing the complexity of scientific collaboration patterns.



DOI: 10.1111/ejed.12446

ORIGINAL ARTICLE

WILEY

An internationalised Europe and regionally focused Americas: A network analysis of higher education studies

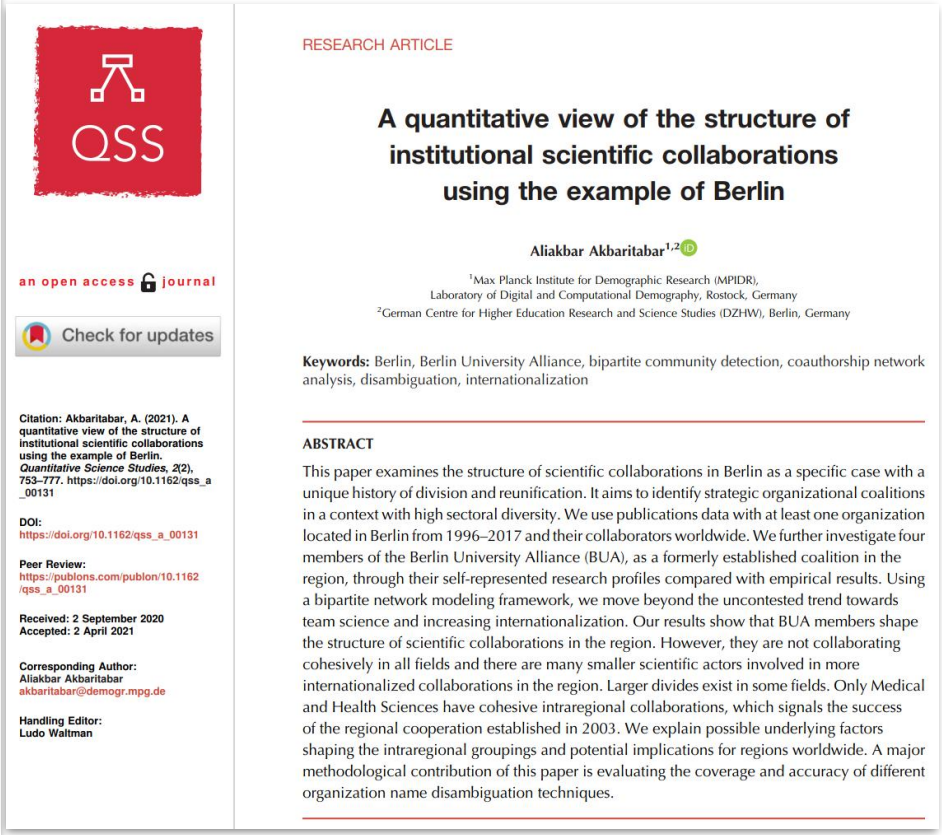
Aliakbar Akbaritabar[1] | Giovanni Barbato[2]

[1]Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research (MPIDR), Rostock, Germany
[2]Department of Economics, Management and Quantitative Methods (DEMM), University of Milan, Milan, Italy

Correspondence
Aliakbar Akbaritabar, Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research (MPIDR), Konrad-Zuse-Str. 1, Rostock 18057, Germany.
Email: akbaritabar@demogr.mpg.de

Funding information
Data was obtained from Kompetenzzentrum Bibliometrie (Competence Centre for Bibliometrics), Germany, which is funded by the Federal Ministry for Education and Research (BMBF), Germany, with grant number 01PQ17001

Abstract
The study on which this article reports investigated the internationalisation of higher education studies by examining collaborations in the form of international co-authorships. We analyse how network-based mechanisms, related to structural relationship between authors (preferential attachment, i.e., higher tendency to collaborate among the most productive ones) and node level features (homophily, i.e., tendency to collaborate with similar others), affect higher education co-authorship networks. We build a bipartite co-authorship network based on 17,262 publications from 33 specialised higher education journals indexed in Scopus from 1996–2018. Scientific collaboration in higher education mainly occurs within national borders. We found that higher education is not an internationally oriented field of research, with around 90% single-country publications. A geographical divide was observed between the two largest communities (Europe, Asia and Oceania vs. the Americas) which was also reflected in the research themes addressed by these communities, structured around the known divide between (1) learning and teaching, and (2) policy-based studies. Preferential attachment was observed to be a network-based mechanism that contributes to drive the formation of new co-authorships. Similarly, homophily based on academic seniority and research productivity emerged



QSS

an open access journal

Check for updates

Citation: Akbaritabar, A. (2021). A quantitative view of the structure of institutional scientific collaborations using the example of Berlin. *Quantitative Science Studies*, *2*(2), 753–777. https://doi.org/10.1162/qss_a_00131

DOI:
https://doi.org/10.1162/qss_a_00131

Peer Review:
https://publons.com/publon/10.1162/qss_a_00131

Received: 2 September 2020
Accepted: 2 April 2021

Corresponding Author:
Aliakbar Akbaritabar
akbaritabar@demogr.mpg.de

Handling Editor:
Ludo Waltman

RESEARCH ARTICLE

A quantitative view of the structure of institutional scientific collaborations using the example of Berlin

Aliakbar Akbaritabar[1,2]

[1]Max Planck Institute for Demographic Research (MPIDR), Laboratory of Digital and Computational Demography, Rostock, Germany
[2]German Centre for Higher Education Research and Science Studies (DZHW), Berlin, Germany

Keywords: Berlin, Berlin University Alliance, bipartite community detection, coauthorship network analysis, disambiguation, internationalization

ABSTRACT
This paper examines the structure of scientific collaborations in Berlin as a specific case with a unique history of division and reunification. It aims to identify strategic organizational coalitions in a context with high sectoral diversity. We use publications data with at least one organization located in Berlin from 1996–2017 and their collaborators worldwide. We further investigate four members of the Berlin University Alliance (BUA), as a formerly established coalition in the region, through their self-represented research profiles compared with empirical results. Using a bipartite network modeling framework, we move beyond the uncontested trend towards team science and increasing internationalization. Our results show that BUA members shape the structure of scientific collaborations in the region. However, they are not collaborating cohesively in all fields and there are many smaller scientific actors involved in more internationalized collaborations in the region. Larger divides exist in some fields. Only Medical and Health Sciences have cohesive intraregional collaborations, which signals the success of the regional cooperation established in 2003. We explain possible underlying factors shaping the intraregional groupings and potential implications for regions worldwide. A major methodological contribution of this paper is evaluating the coverage and accuracy of different organization name disambiguation techniques.

Q&A

Please raise any comments, [clarification or else] questions, or points on the slides and content that were presented!

2. Panel discussion (90 minutes) with invited speakers

- **Welcome and introduction**
- **Emilio Zagheni, MPIDR**: Bibliometric data for migration research
- **Andrea Miranda Gonzalez, UC-Berkeley**: Internal migration in Mexico
- **Asli Ebru Sanlitürk, MPIDR**: Brexit's effect on scholarly migration to and from the UK
- **Maciej J. Dańko, MPIDR**: Development and international scholarly migration
- **Tom Theile, MPIDR**: Trends of return international migration worldwide
- **Xinyi Zhao, MPIDR**: Gender perspective in international scholarly migration
- **Aliakbar Akbaritabar, MPIDR**: Integrating internal and international scholarly migration worldwide
- Open discussion between speakers and Q&A

(Switch to the other slides)

PAUSE - BREAK (15 MINUTES)

We will reconvene in the same room and online link!

3. Introduction to available data sources, especially publicly available data (30 minutes)

Aims:

- Familiarize themselves with the availability of bibliometric data as a type of digital data and its potentials for demographic research
- Familiarize themselves with the most commonly used and available bibliometric databases

Historical Review

Less standardised bibliographic data: individual CVs, publication lists, institutional reports …

- statistical study on scientific bibliographies 1901-1913 (*Hulme,1923*)
- frequency distribution of scientific productivity, index of Chemical Abstracts 1907-1916 (*Lotka, 1926*)
- citation-based study on 3,633 citations from the 1926 volume of the Journal of the American Chemical Society (*Gross & Gross, 1927*)

‘Foundation of modern research evaluation techniques ‘Little Science - Big Science’ (Price, 1963)

- using the Science Citation Index (***SCI***) database of the Institute for Scientific Information (***ISI***) as a tool in quantitative analysis of science
- ***SCI***: first large multidisciplinary bibliographic database that allowed advanced bibliometric studies

Standardised databases.

- ***Bibliometrics***, as the application of mathematical and statistical methods to books and other media of communication (*Pritchard, 1969*)
- ***Scientometrics*****,** all quantitative aspects of science and scientific research” (*Sengupta, 1992*)

Clarivate's Web of Science (WoS)

ISI (including SCI) , now part of Clarivate Analytics

Main products of WoS core Collection:

- Science Citation Index Expanded (SCIE)
- Social Sciences Citation Index (SSCI)
- Arts & Humanities Citation Index (A&HCI)
- Conference Proceedings Citation Index- Science (CPCI-S)
- Conference Proceedings Citation Index- Social Science & Humanities (CPCI-SSH)
- Book Citation Index (BKCI)
- Emerging Sources Citation Index (ESCI)
- Journal Citation Reports (JCR) – now part of Clarivate Analytics InCites

Elsevier’s Scopus

- **Largest abstract and citations database of peer-reviewed research literature., provided by Elsevier since 2004**

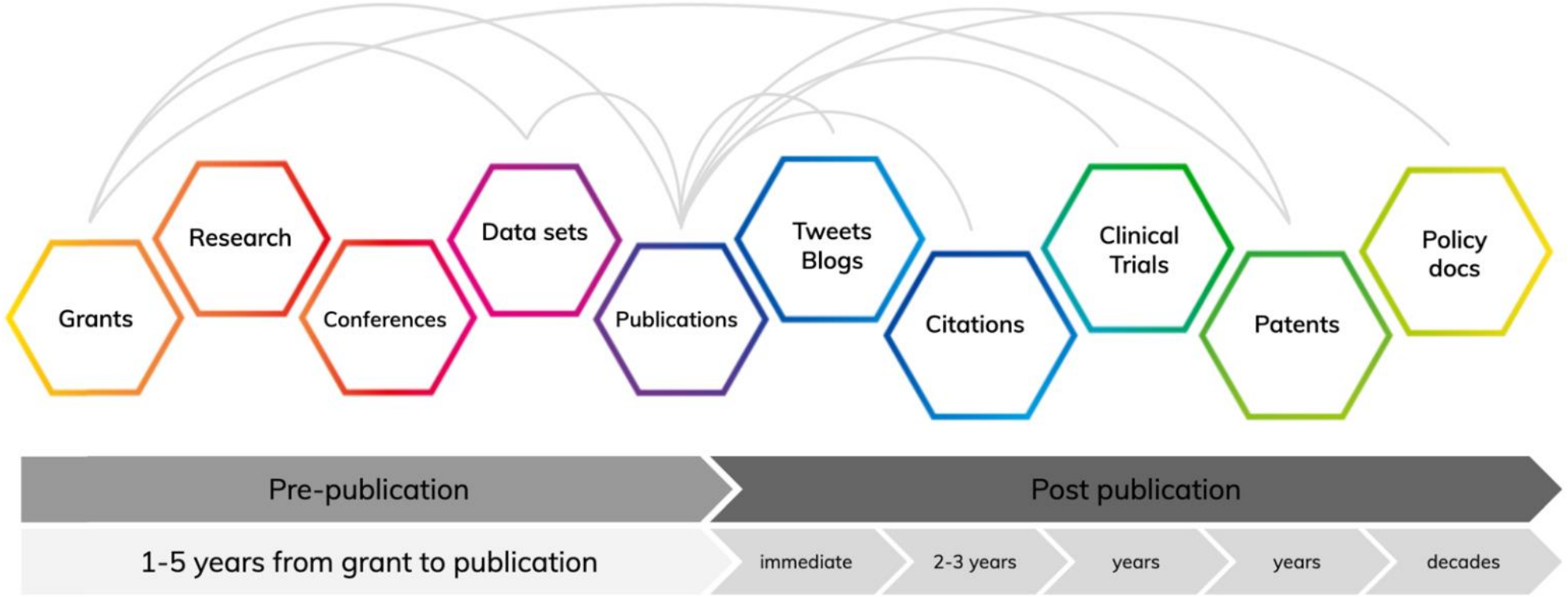
- **With over 22,000 titles from more than 5,000 international publishers: Reviews, Articles, short surveys, books, chapters, Letters and notes, conference proceedings, etc.**

- **CiteScore** (average number of citations received from a publication in a certain year with respect to the average of the cites of the previous three years)
- **SJR** (Scimago Journal and Country Rank)
- **SNIP** (Source Normalized Impact per Paper)
- **Index h** (average between the number of citations and the number of publications of an author)

Dimensions

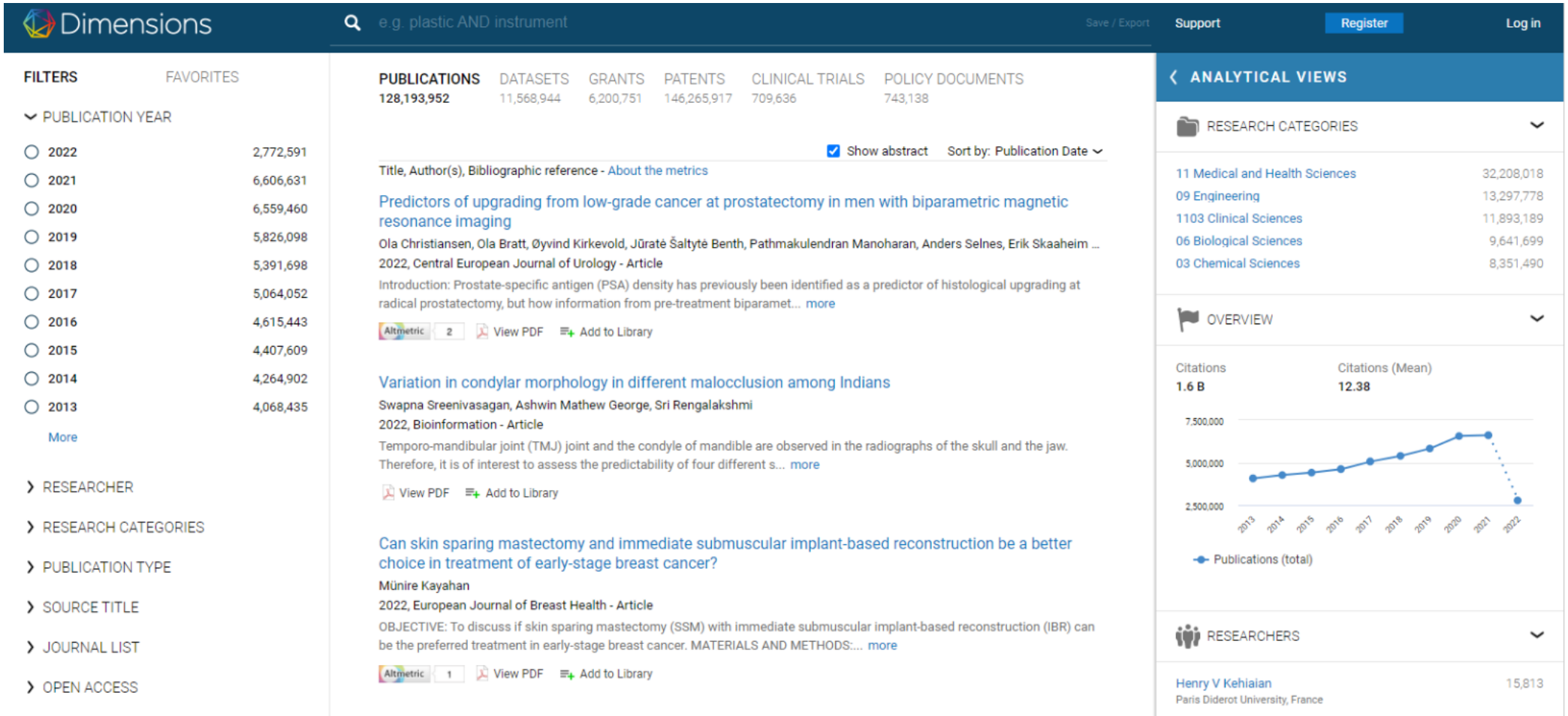
- **A new multidisciplinary data platform for scholarly journals, patents, grants, clinical trials and policy documents, provided by Digital Science since 2018.**

- **Be open to encompassing all scholarly content that is available for inclusion**



Dimensions

- **Dimensions publications core search freely available to everyone**



Google Scholar

- **Utilizing the massive document index from a web search engine to achieve comprehensive coverage of contemporary scholarly materials, many of which are not published and distributed through traditional channels and not assigned DOIs**

- **Strictly speaking not a bibliographic database**
- **No direct subject classification (implemented via author profiles)**
- **Proper documentation and crucial information about coverage is missing**
- **No acceptable disambiguation**
- **About half of Google Scholar unique citations are not from journals**

Other data sources and software packages

- **Crossref**
- **Unpaywall**
- **Semantic scholar**
- **Microsoft Academic→ OpenAlex**
- **JSTOR Data for Research**
- **PubMed**
- **Grobid**

API, software and packages:

- Scopus (rscopus package r),
- WOS(wosR package in R),
- OpenAlex (based on MAG),
- ORCID API,
- Dimensions (API),
- Google Scholar (publish perish, scholar package R),
- Bibliometrix (package in R),
- VOSviewer,
- CitenetExplorer
- CiteSpace

[1]Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, *2*(1), 20–41. https://doi.org/10.1162/qss_a_00112

[2]https://datalab.ucdavis.edu/bibliometrics/#:~:text=As%20data%2C%20bibliometrics%20typically%20measure,over%20a%20certain%20period%20of

Q&A

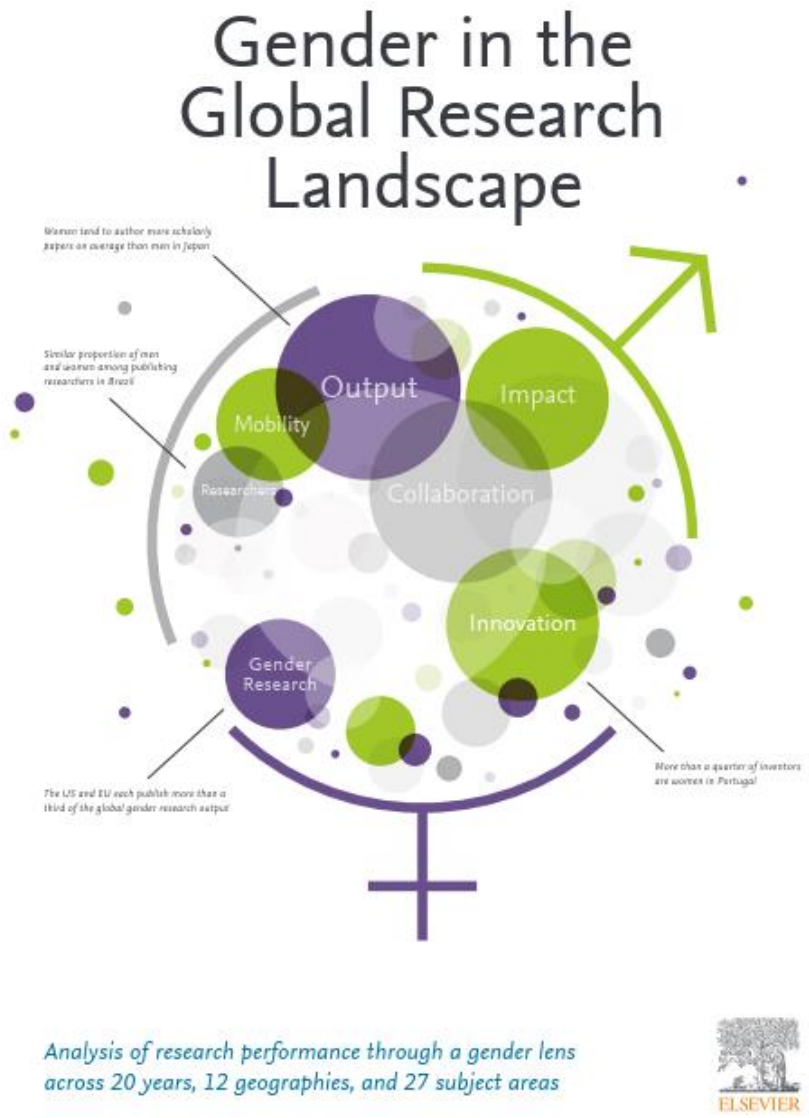
Please raise any comments, [clarification or else] questions, or points on the slides and content that were presented!

4. Introduction to handling bibliometric data for demographic research: data retrieval, data pre-processing and repurposing based on research questions (60 minutes)

Aims:

- Learn about pre-processing steps and handling bibliometric data, including:
  - Author/organisation name disambiguation
  - Gender identification
  - Scholarly mobility and movement identification
- Learn about practical applications of bibliometric data in demographic research, including:
  - Use of some relevant scientometric indicators
  - Calculation and use of demographic indicators *(e.g., Crude Migration Intensity, Migration Effectiveness Index and the Aggregate Net Migration Rate)*
  - Network analysis of scientific collaborations
  - Text analysis of bibliometric information

Gender detection



Scientometrics (2016) 106:143–162
DOI 10.1007/s11192-015-1775-3

Gender differences in research performance and its impact on careers: a longitudinal case study

Peter van den Besselaar[1] · Ulf Sandström[2,3]

Gender differences in research areas, methods and topics: Can people and thing orientations explain the results?[1]

Mike Thelwall, Carol Bailey, Catherine Tobin, [School of Mathematics and Computer Science, International Academy, School of Sciences], University of Wolverhampton;
Noel-Ann Bradshaw, Department of Mathematical Sciences, University of Greenwich.

Historical comparison of gender inequality in scientific careers across countries and disciplines

Junming Huang[a,b,c,1], Alexander J. Gates[a,1], Roberta Sinatra[d,e], and Albert-László Barabási[a,f,g,h,2]

[a]Network Science Institute and Department of Physics, Northeastern University, Boston, MA 02115; [b]CompleX Lab, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; [c]Paul and Marcia Wythes Center on Contemporary China, Princeton University, Princeton, NJ 08540; [d]Department of Computer Science, IT University of Copenhagen, 2300 Copenhagen, Denmark; [e]ISI Foundation, 10126 Turin, Italy; [f]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115; [g]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115; and [h]Department of Network and Data Science, Central European University, 1051 Budapest, Hungary

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved January 22, 2020 (received for review August 15, 2019)

Gender detection

Scientometrics (2016) 106:143–162
DOI 10.1007/s11192-015-1775-3

CrossMark

Gender differences in research performance and its impact on careers: a longitudinal case study

Peter van den Besselaar[1] · Ulf Sandström[2,3]

1. **Manual detection:**
   Internet search including searching the personal homepage, curricula vitae, online profiles, biographies in publications, and other online sources

   limitation: Difficult to scale up or apply for large datasets.

   The largest dataset which was manually coded for gender used 1, 059, 939 articles (*A. Paul-Hus et al., 2014*)

G Halevi, Bibliometric studies on gender disparities in science in Handbook of Science and Technology Indicators, eds. W Glänzel, H Moed, U Schmoch, M Thelwall. (Springer International Publishing), pp. 563–580 (2019)

Gender detection

2. **Questionnaires:**
   Used to verify authors' gender, seniority, and more personal information, when the study spanned several universities across various location.

   limitation: the sample size is limited considering the response rate; the gender difference in response.

   Sidhu et al. (*2009*) investigates gender differences in first and senior authorship in six peer-reviewed British journals and factors that are associated with publication rates, identifying the gender of authors through a web-based self-selected on-line questionnaire .

Gender detection

3. National/ Institutional database:

Examples are The Spanish Council for Scientific Research, The American Doctoral Dissertations Database, The Indian Directory of CSIR Scientists, and so forth.

limitation: more suitable for studies focusing on institutional or national levels

Abramo and D'Angelo (*2015*) measured the differences of the performance and the rank of all Italian research institutions by considering the gender information from the Italian Ministry of Education, Universities and Research (MIUR) .

_id	ANNO	COD_AT...	NOME_ATENEO	Reg_ATENEO	AREA_GEO	GENERE	COD_Q...	DESC_...	N_PERS
1	2015	101	Torino - Università degl...	Piemonte	NORD-OVEST	F	1PO e 2PA	Professo...	419
2	2015	101	Torino - Università degl...	Piemonte	NORD-OVEST	M	1PO e 2PA	Professo...	748
3	2015	101	Torino - Università degl...	Piemonte	NORD-OVEST	F	3RU e 3...	Ricercat...	395
4	2015	101	Torino - Università degl...	Piemonte	NORD-OVEST	M	3RU e 3...	Ricercat...	384
5	2015	101	Torino - Università degl...	Piemonte	NORD-OVEST	F	4AR	Titolare ...	250
6	2015	101	Torino - Università degl...	Piemonte	NORD-OVEST	M	4AR	Titolare ...	146
7	2015	101	Torino - Università degl...	Piemonte	NORD-OVEST	F	5PC	Personal...	368
8	2015	101	Torino - Università degl...	Piemonte	NORD-OVEST	M	5PC	Personal...	377
9	2015	101	Torino - Università degl...	Piemonte	NORD-OVEST	F	6CR	Collabor...	984
10	2015	101	Torino - Università degl...	Piemonte	NORD-OVEST	M	6CR	Collabor...	610
11	2015	101	Torino - Università degl...	Piemonte	NORD-OVEST	F	7CL	Collabor...	31

Gender detection

4. Large-scale name-gender database:

- Popular given names being M/F based on US names from the US Social Security Administration (SAA): 18,953 female names, 13,959 male names

- Given names being M/F based on inventors of GB patent applications (1978-2015) from the Intellectual Property Office of UK government :
  63,121 female names,39,656 male names

- Worldwide gender-name dictionary (WGND) including 6.2 million names for 182 different countries to disambiguate the gender (*Lax Martínez et al., 2016*).

- WGND 2.0 including more than 26 million records linking given names and 195 different countries and territories (*Raffo, Julio, 2021*)

Gender detection

5. Softwares:

Gender detection tool	Classified as female physicians n (%)	Classified as male physicians n (%)	Nonclassified physicians n (%)
Gender API			
Female physicians	3006 (97.4)	67 (2.2)	12 (0.4)
Male physicians	23 (0.8)	3014 (98.9)	9 (0.3)
NamSor			
Female physicians	3031 (98.2)	54 (1.8)	0 (0.0)
Male physicians	70 (2.3)	2976 (97.7)	0 (0.0)
Wiki-Gendersort			
Female physicians	2832 (91.8)	85 (2.8)	168 (5.4)
Male physicians	43 (1.4)	2895 (95.0)	108 (3.6)
genderize.io			
Female physicians	2519 (81.7)	59 (1.9)	507 (16.4)
Male physicians	17 (0.6)	2529 (83.0)	500 (16.4)

Sebo P. (2021). Performance of gender detection tools: a comparative study of name-to-gender inference services. Journal of the Medical Library Association : JMLA, 109(3), 414–421. https://doi.org/10.5195/jmla.2021.1185

Gender detection

Challenges:

1. Gender neutral names (such as Alex, Taylor, Austin…)
2. Gender determination depending on countries (Andrea which is a typically a female name in north America and Eastern Europe but a male name in Italy)
3. Names with regional traits or in other languages (Asian, Middle Eastern, and African names)
4. Gender self-identification
5. Problematic names in bibliometric data (such as only initials in authorship, same author ID with different names)

Q&A

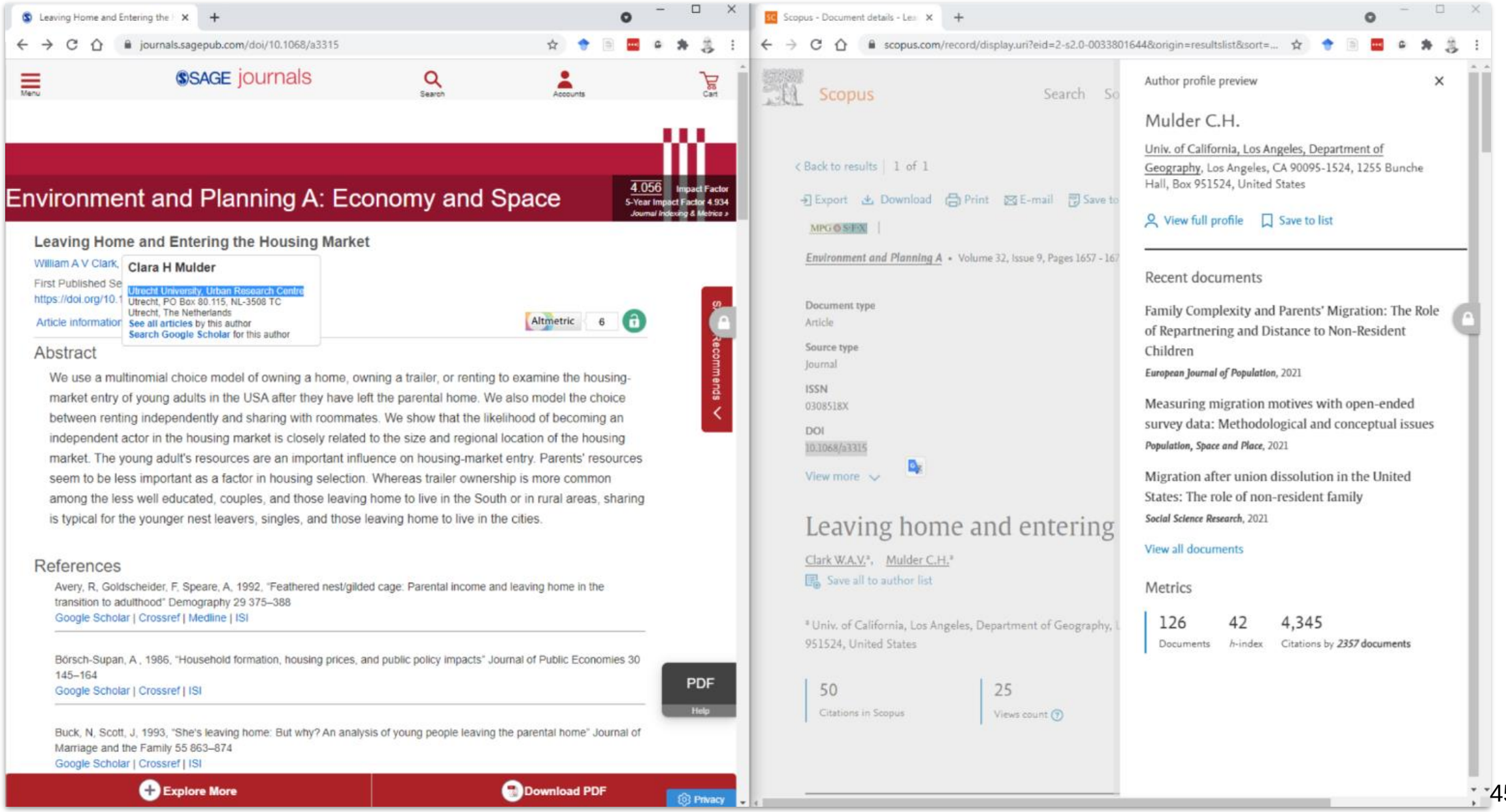
Please raise any comments, [clarification or else] questions, or points on the slides and content that were presented!

Data quality and need for organisation name disambiguation

authorsid	authorswithaffiliations	year	country
23494432900	Zagheni, E., Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research, Rostock, Germany	2020	DEU
23494432900	Zagheni, E., Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research, Rostock, Germany	2020	DEU
23494432900	Zagheni, E., Max Planck Institute for Demographic Research Rostock, Germany	2020	DEU
23494432900	Zagheni, E., Max Planck Institute for Demographic Research, Germany	2020	DEU
23494432900	Zagheni, E., Max Planck Institute for Demographic Research, Germany	2020	DEU
23494432900	Zagheni, E., Max Planck Institute for Demographic Research, Rostock, Germany	2020	DEU
23494432900	Zagheni, E., Max Planck Institute for Demographic Research, Rostock, Germany	2019	DEU
23494432900	Zagheni, E., Max Planck Institute for Demographic Research, Konrad-Zuse-Straße 1, Rostock, 18057, Germany	2019	DEU
23494432900	Zagheni, E.	2019	
23494432900	Zagheni, E., Department of Sociology, University of Washington, Seattle, United States, Max Planck Institute for Demographic Research, Rostock, Germany	2019	DEU
23494432900	Zagheni, E., Max Planck Institute for Demographic Research, Germany	2019	DEU
23494432900	Zagheni, E., Max Planck Institute for Demographic Research, Rostock, Germany	2019	DEU
23494432900	Zagheni, E., Max Planck Institute for Demographic Research, Germany	2019	DEU
23494432900	Zagheni, E., Max Planck Institute for Demographic Research, Konrad-Zuse-Str. 1, Rostock, 18057, Germany	2019	DEU
23494432900	Zagheni, E., Qatar Computing Research Institute, Doha, Qatar	2019	QAT
23494432900	Zagheni, E., Max Planck Institute for Demographic Research	2019	
23494432900	Zagheni, E., Max Planck Institute for Demographic Research, Rostock, Germany	2018	DEU
23494432900	Zagheni, E., Max Planck Institute for Demographic Research, Rostock, Germany	2018	DEU
23494432900	Zagheni, E., University of Washington, Max Planck Institute for Demographic Research, United States	2018	USA
23494432900	Zagheni, E., University of Washington and Max Planck Institute for Demographic Research, United States	2018	USA
23494432900	Zagheni, E.	2017	
23494432900	Zagheni, E., Department of Sociology, University of Washington, Seattle, 211 Savery Hall, Box 353340, Seattle, WA 98195-3340, United States	2017	USA
23494432900	Zagheni, E., University of Washington, Seattle, United States	2017	USA
23494432900	Zagheni, E., University of Washington, United States	2017	USA
23494432900	Zagheni, E., University of Washington, Seattle, United States	2016	USA
23494432900	Zagheni, E., Department of Sociology, University of Washington at Seattle, 211 Savery Hall Box 353340, Seattle, WA 98195-3340, United States	2016	USA
23494432900	Zagheni, E., University of Washington, Seattle, WA, United States	2015	USA
23494432900	Zagheni, E., Department of Sociology, University of Washington, United States	2015	USA
23494432900	Zagheni, E., University of Washington, Seattle, United States	2015	USA
23494432900	Zagheni, E., Department of Sociology, University of Washington, Seattle, United States	2015	USA
23494432900	Zagheni, E., Department of Sociology, University of Washington, Seattle, WA, United States	2015	USA

Check https://doi.org/10.1162/qss_a_00131 for a description and https://github.com/ror-community/ror-api

A publication on publisher's website (left) and Scopus (right)



Need for author name disambiguation

Out[43]:

	AUTHOR_ID	PK_AUTHORS	LAST_NAME	FIRST_NAME	INDEXED_NAME
342531	6503908779	5159491198	Jan	Mendling	Jan M.
1230731	6503908779	12326655929	Jan	Mendling	Jan M.
8960149	6503908779	15956230109	Jan	Mendling	Jan M.
12752651	6503908779	20665874236	Jan	Mendling	Jan M.
13345007	6503908779	5817946947	Jan	Mendling	Jan M.

In [44]:
```
papers_new[papers_new["AUTHOR_ID"]==6503908779][["AUTHOR_ID","PK_AUTHOR
```

Out[44]:

	AUTHOR_ID	PK_AUTHORS	LAST_NAME	FIRST_NAME	INDEXED_NAME
62161	6503908779	15391353604	Mendling	NaN	Mendling J.
82568	6503908779	20744642022	Mendling	Jan	Mendling J.
122612	6503908779	2725380135	Mendling	Jan	Mendling J.
128880	6503908779	5107144514	Mendling	Jan	Mendling J.
159362	6503908779	11448344549	Mendling	Jan	Mendling J.

Same author's first and last names are swapped in Scopus while author identification number is similar!

On Scopus' author disambiguation, see: https://doi.org/10.1162/qss_a_00019

ORCID
Connecting research and researchers

https://orcid.org/
0000-0002-7260-524X

Is this you? Sign in to start editing

Name
Jan Mendling

Biography

This author profile is generated by Scopus Learn more

Mendling, Jan

Mendling, Jan ; Jan, Mendling ; Mendling, J.
Wirtschaftsuniversität Wien, Vienna, Austria

Affiliation history

2004 - 2022 Wirtschaftsuniversität Wien, Vienna, Austria
2008 - 2022 Humboldt-Universität zu Berlin, Berlin, Germany
2013 - 2022 Univerza v Ljubljani, Ljubljana, Slovenia
2021 KU Leuven, 3000 Leuven, Belgium
2008 - 2020 Universität Wien, Vienna, Austria
2016 - 2017 Institute for Information Business, Vienna, Austria
2016 Vrije Universiteit Amsterdam, Amsterdam, Netherlands
2007 - 2009 Queensland University of Technology, Brisbane, Australia

Subject Areas

Computer Science • Business, Management and Accounting • Mathematics • Decision Sciences • Engineering • Arts and Humanities • Biochemistry, Genetics and Molecular Biology • Psychology • Social Sciences • Economics, Econometrics and Finance • Medicine • Chemical Engineering • Nursing

Show less author info

6503908779 https://orcid.org/0000-0002-7260-524X Is this you? Connect to Mendeley account

DATA QUALITY

- **Author disambiguation in Scopus:**

  98.1% of author profiles do not include publications written by someone else**;

  94.4% of author profiles include all publications written by the author.

- **Organization disambiguation:**

  Research Organization Registry (ROR) API***

- **Subset of data with the highest quality:**

  Period:1996-2020; Type of publications: Articles and Reviews;

  → 36+ Million publications for 16+ Million unique authors

**Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, *1*(1), 377–386. https://doi.org/10.1162/qss_a_00019

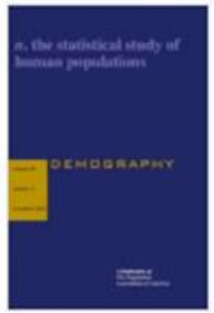
*** See for comparison and description of ROR: Akbaritabar, A. (2021). A quantitative view of the structure of institutional scientific collaborations using the example of Berlin. Quantitative Science Studies, 2(2), 753–777. https://doi.org/10.1162/qss_a_00131

Q&A

Please raise any comments, [clarification or else] questions, or points on the slides and content that were presented!

PAUSE - BREAK (10 MINUTES)

We will reconvene in the same room and online link!



Demography

November 2012, Volume 49, Issue 4, pp 1307–1333 | Cite as

Migration Systems in Europe: Evidence From Harmonized Flow Data

Authors | Authors and affiliations

Jack DeWaard 1
Email author

Keuntae Kim 2

James Raymer 3

1. Department of Sociology, Center for Demography and Ecology, University of Wisconsin–Madison, Madison, USA
2. Department of Sociology, Center for Demography and Ecology, University of Wisconsin–Madison, Madison, USA
3. Division of Social Sciences, University of Southampton, Southampton, England

Demography

November 2012, Volume 49, Issue 4, pp 1307–1333 | Cite as

Migration Systems in Europe: Evidence From Harmonized Flow Data

Authors | Authors and affiliations

Jack DeWaard 1
Email author

Keuntae Kim 2

James Raymer 3

1. Department of Sociology, Center for Demography and Ecology, University of Wisconsin–Madison, Madison, USA
2. Department of Sociology, Center for Demography and Ecology, University of Wisconsin–Madison, Madison, USA
3. Division of Social Sciences, University of Southampton, Southampton, England

Demography
November 2012, Volume 49, Issue 4, pp 1307–1333 | Cite as

Migration Systems in Europe: Evidence From Harmonized Flow Data

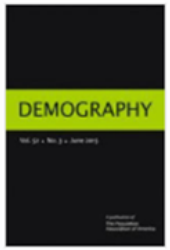
Authors | Authors and affiliations

Jack DeWaard 1
Email author

Keuntae Kim 2

James Raymer 3

1. Department of Sociology, Center for Demography and Ecology, University of Wisconsin–Madison, Madison, USA
2. Department of Sociology, Center for Demography and Ecology, University of Wisconsin–Madison, Madison, USA
3. Division of Social Sciences, University of Southampton, Southampton, England



Demography
June 2015, Volume 52, Issue 3, pp 1035–1059 | Cite as

Bayesian Population Forecasting: Extending the Lee-Carter Method

Authors | Authors and affiliations

Arkadiusz Wiśniowski 1
Email author

Peter W. F. Smith 1

Jakub Bijak 1

James Raymer 2

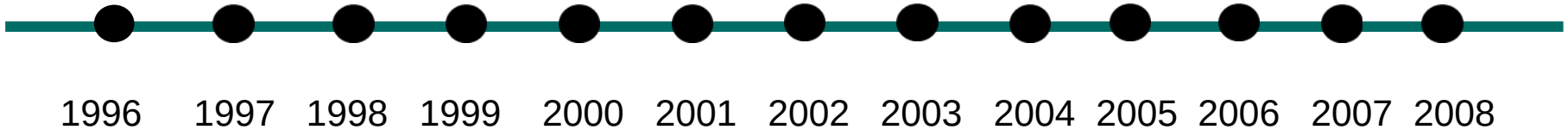
Jonathan J. Forster 1

1. Economic and Social Research Council Centre for Population Change, University of Southampton, Southampton, UK
2. Australian Demographic & Social Research Institute, The Australian National University, Acton, Australia

A SIMPLE IDEA

Changes in institutional affiliations can be used to infer changes in residence over time for individual scholars and for populations

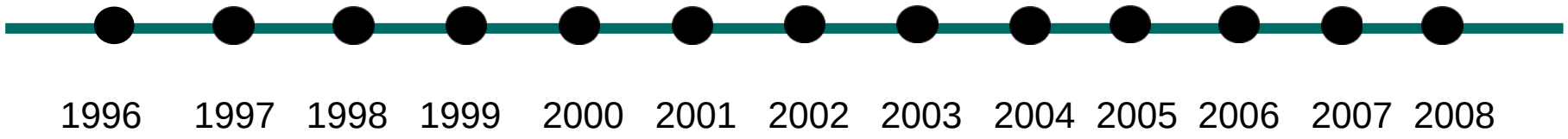
IDENTIFYING MIGRATION EVENTS: ILLUSTRATIVE EXAMPLES



IDENTIFYING MIGRATION EVENTS: ILLUSTRATIVE EXAMPLES

Maria's first publications are with a Canadian affiliation in 1998

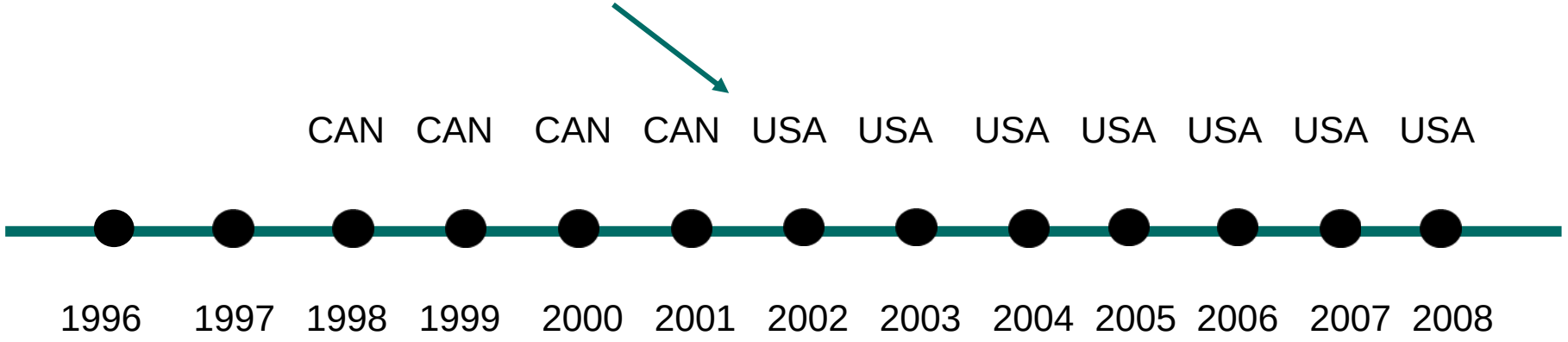
Maria's country of "academic" birth is Canada, regardless of her citizenship or country of actual birth



The country of residence is inferred as the modal country of publications in a given year

IDENTIFYING MIGRATION EVENTS: ILLUSTRATIVE EXAMPLES

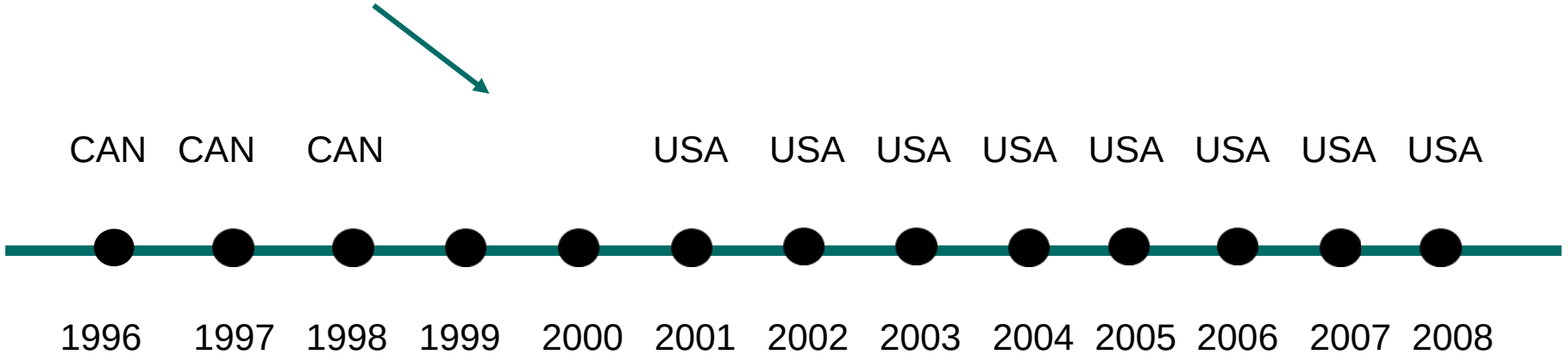
Maria changes residence from Canada to the US between 2001 and 2002



Maria's modal countries of publications over time

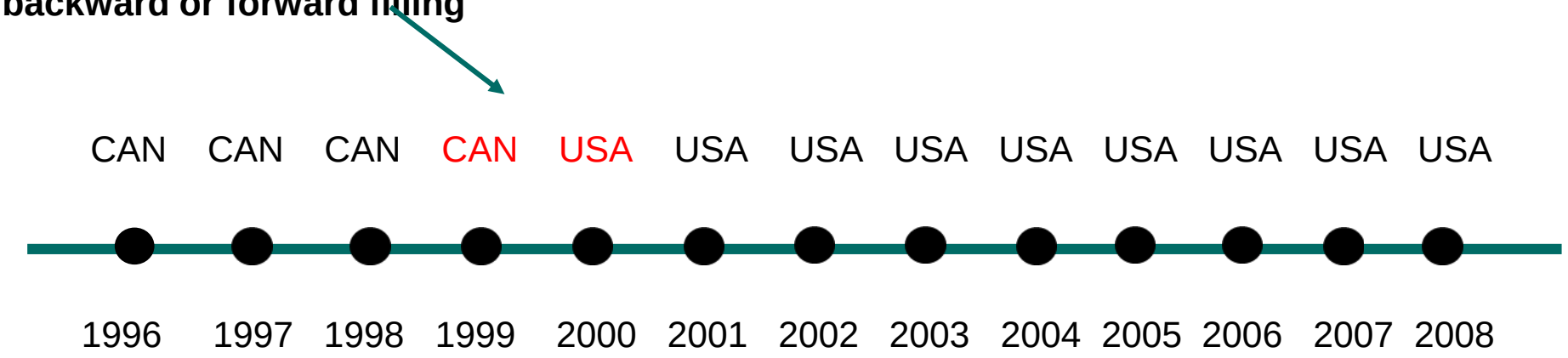
IDENTIFYING MIGRATION EVENTS: ILLUSTRATIVE EXAMPLES

Joe has no publications in our database for 1999 and 2000



IDENTIFYING MIGRATION EVENTS: ILLUSTRATIVE EXAMPLES

We impute the country values for up to 2 years from the closest observation with backward or forward filling



Q&A

Please raise any comments, [clarification or else] questions, or points on the slides and content that were presented!

[Brief, self-study]

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

[self-study] 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 workshop repository, Python users see:
https://github.com/akbaritabar/bibliometric_data_for_demographic_research/blob/main/0_code/03_parallelization_with_dask_duckdb_dbeaver.md

Recorded Video of advanced tutorial:

https://youtu.be/pYDVrBcIuYI

- **00:00 - 02:45; Introduction**
- **2:45 - 11:04; Requirements and installation (skip this if you have already installed everything based on instructions in the ReadMe on repository)**
- **11:05 - 38:00; Steps in using Dask/Python and DuckDB/DBeaver and results**

Calculation and use of demographic indicators using bibliometric data

- **Measures**: Net Migration Rate, Crude Migration Intensity, Migration Effectiveness Index, Aggregate Net Migration Rate[5,10]:

$$NMR_{i,t} = 1000 \times \frac{I_{i,t} - E_{i,t}}{N_{i,t}}$$

$$CMI_{i,t} = 100 \times \frac{M_{i,t}}{\sum_i N_{i,t}}$$

$$MEI_{i,t} = 100 \times \frac{\sum_i |I_{i,t} - E_{i,t}|}{\sum_i (I_{i,t} + E_{i,t})}$$

$$ANMR_{i,t} = 100 \times \frac{0.5 \sum_i |I_{i,t} - E_{i,t}|}{\sum_i N_{i,t}}$$

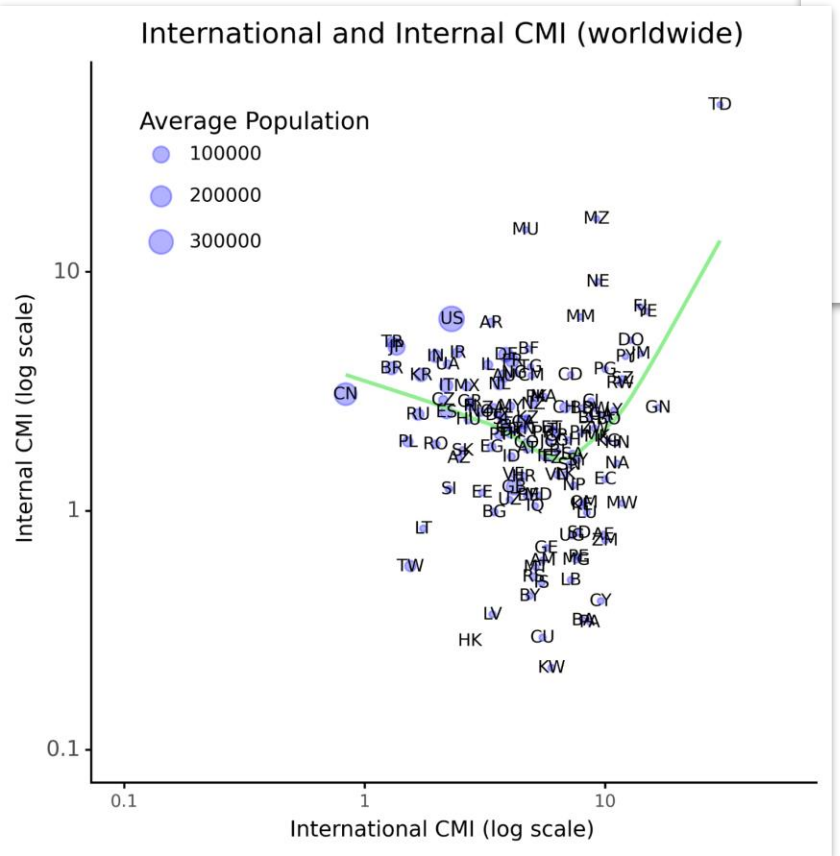
$$CMI_{i,t} = 100 \times \frac{ANMR_{i,t}}{MEI_{i,t}}$$

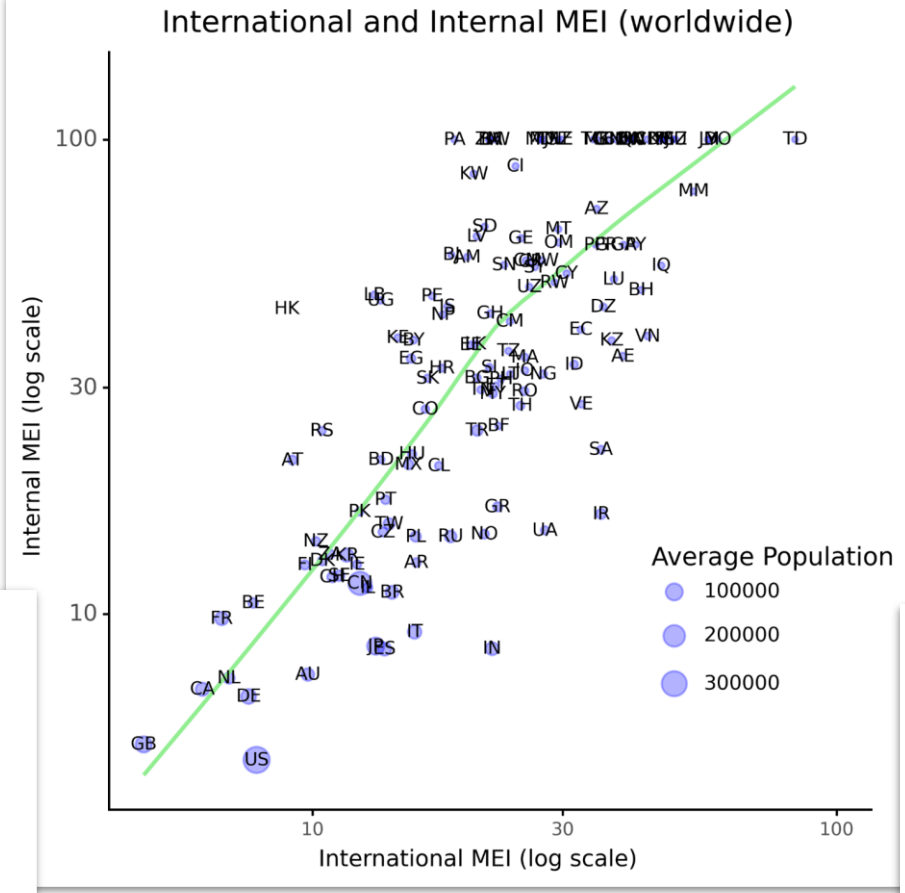
$I_{i,t}$ = Incoming scholars

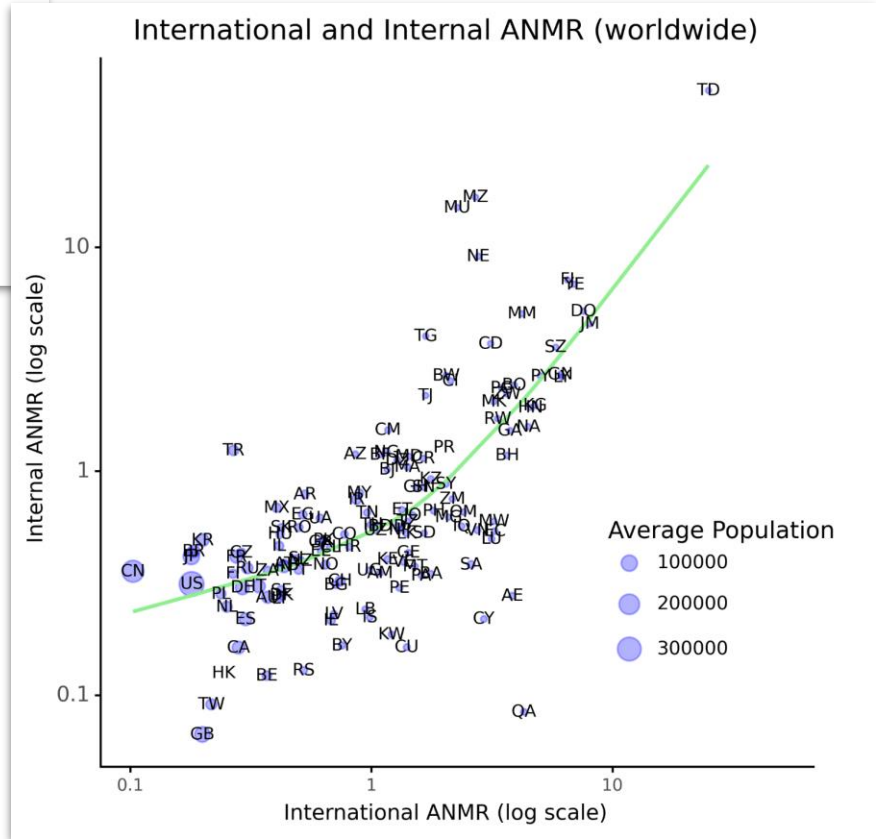
(to a given region/country in a given year)

$E_{i,t}$ = Outgoing scholars (same as $M_{i,t}$)

$N_{i,t}$ = Population of scholars


International and Internal CMI (worldwide)
Average Population
100000
200000
300000
Internal CMI (log scale)
International CMI (log scale)
0.1
1
10


International and Internal MEI (worldwide)
Average Population
100000
200000
300000
Internal MEI (log scale)
International MEI (log scale)
10
30
100


International and Internal ANMR (worldwide)
Average Population
100000
200000
300000
Internal ANMR (log scale)
International ANMR (log scale)
0.1
1
10

Other relevant demographic and [only a few] scientometrics measures

- N and population of active scholars
- N of institutions (how large is the science system of a country)
- Stock and flows of scholars
- Academic birth
- Exiting academia (e.g., mortality and prospects of survival)
- Academic kinship (mentorship ties and supervision, collaboration)

- N of publications,
- N of co-authors,
- Citations,
- 3-year citations,
- FSS (fractional scientific strength)

Q&A

Please raise any comments, [clarification or else] questions, or points on the slides and content that were presented!

[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)

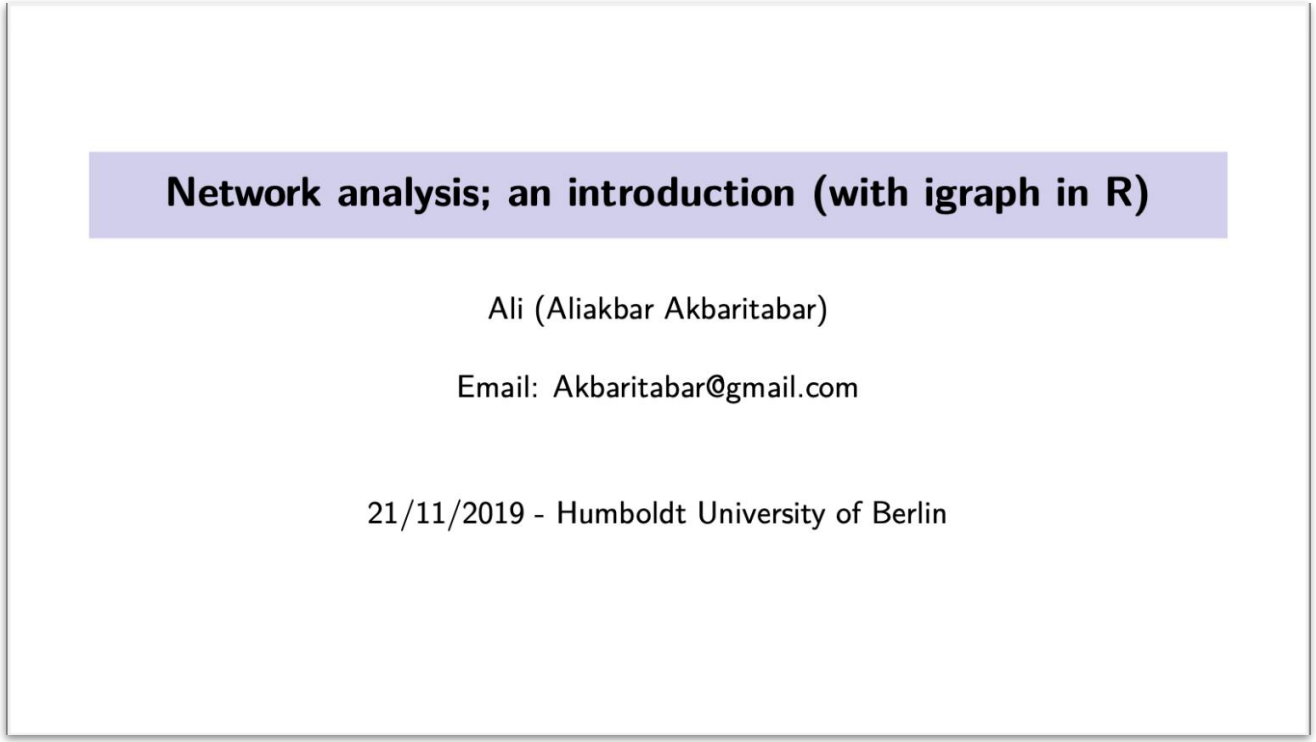
Materials under code, data and output folders of workshop repository, R users see:
https://github.com/akbaritabar/bibliometric_data_for_demographic_research/blob/main/0_code/05_HU_seminar_network_analysis.Rmd (Python users, can see an example in:
https://github.com/akbaritabar/bibliometric_data_for_demographic_research/blob/main/0_code/06_example_network_igraph_python.py)

Presentation file:

https://github.com/akbaritabar/bibliometric_data_for_demographic_research/blob/main/2_presentations/05_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

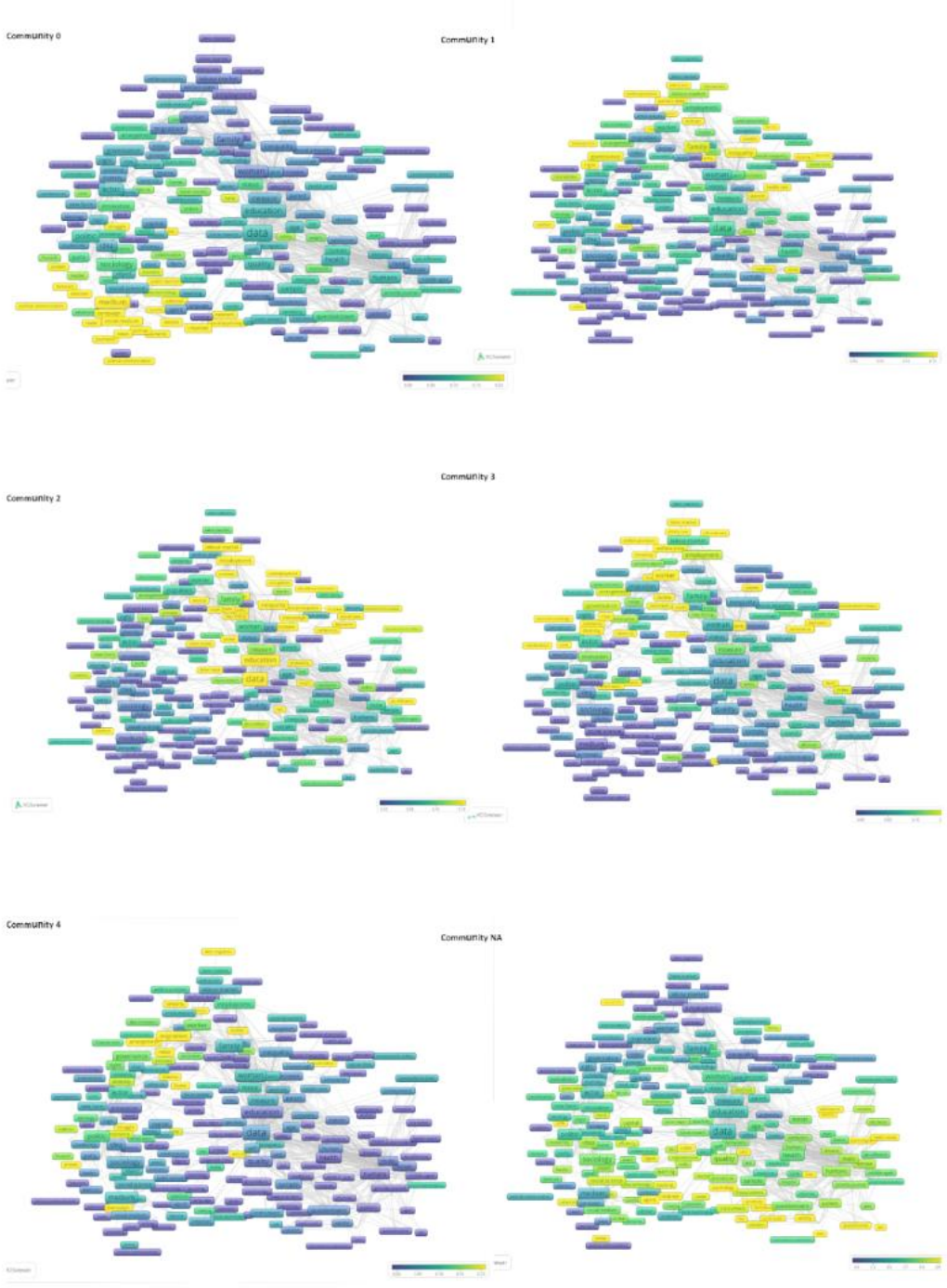
Andrea Miranda-González[1,2], Samin Aref[1], Tom Theile[2] and Emilio Zagheni[2]

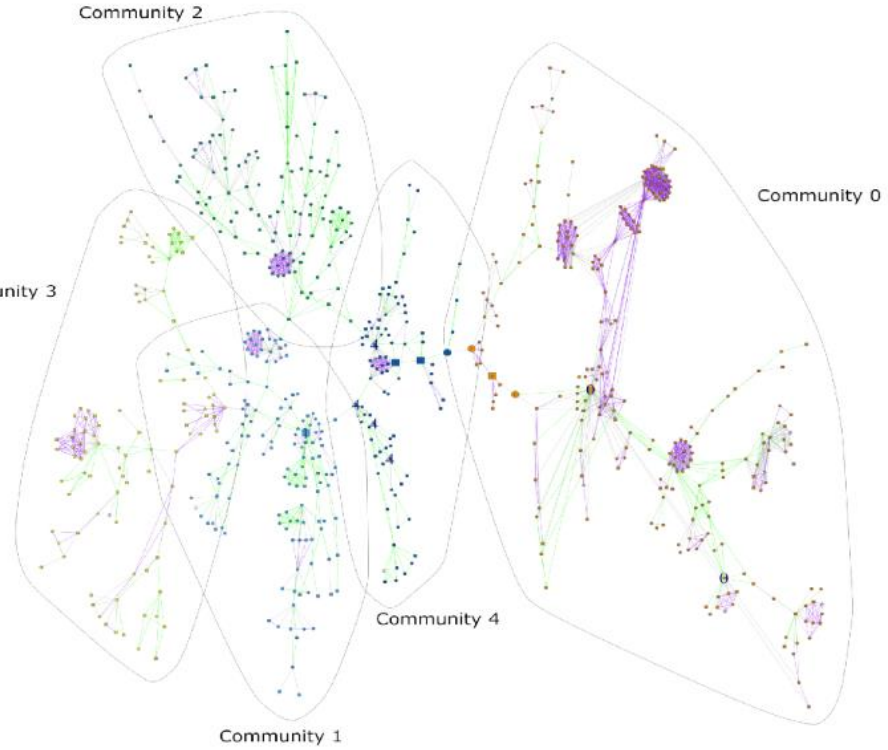
*Correspondence:
aref@demogr.mpg.de
[2]Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research, Rostock, Germany
Full list of author information is available at the end of the article

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

Q&A

Please raise any comments, [clarification or else] questions, or points on the slides and content that were presented!

[Brief] Section on text analysis

Code:

https://github.com/akbaritabar/bibliometric_data_for_demographic_research/blob/main/0_code/08_text_analysis_exact_and_anchor_words.py

https://github.com/akbaritabar/bibliometric_data_for_demographic_research/blob/main/0_code/10_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 01PQ17001. 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!)

Example 2 of text analysis using exact word pairs

PK_ITEMS	SOURCETITLE	UT_EID	PUBYEAR	DOCTYPE	SC_DESCRIPTION	OECD_DESCRIPTION
6980386	NEUROLOGICAL SCIENCES	000372330000006	2016	Article	NEUROSCIENCES	MEDICAL AND HEALTH SCIENCES

Exact word pairs

(None identified)

Title

Biological factors and age-dependence of primary motor cortex experimental plasticity

Abstract

To evaluate whether the age-dependence of brain plasticity correlates with the levels of proteins involved in hormone and brain functions we executed a paired associative stimulation (PAS) protocol and blood tests. We measured the PAS-induced plasticity in the primary motor cortex. Blood levels of the brain-derived neurotrophic factor (BDNF), estradiol, the insulin-like growth factor (IGF)-1, the insulin-like growth factor binding protein (IGFBP)-3, progesterone, sex hormone-binding globulin (SHBG), testosterone, and the transforming growth factor beta 1 (TGF-beta 1) were determined in 15 healthy men and 20 healthy women. We observed an age-related reduction of PAS-induced plasticity in females that it is not present in males. In females, PAS-induced plasticity displayed a correlation with testosterone (p = 0.006) that became a trend after the adjustment for the age effect (p = 0.078). In males, IGF-1 showed a nominally significant correlation with the PAS-induced plasticity (p = 0.043). In conclusion, we observed that hormone blood levels (testosterone in females and IGF-1 in males) may be involved in the age-dependence of brain plasticity.

Example 2 of text analysis using anchors

PK_ITEMS	SOURCETITLE	UT_EID	PUBYEAR	DOCTYPE	SC_DESCRIPTION	OECD_DESCRIPTION
6980386	NEUROLOGICAL SCIENCES	000372330000006	2016	Article	NEUROSCIENCES	MEDICAL AND HEALTH SCIENCES

Word used before

Anchor

Word used after

Title

Biological factors and age-dependence of primary motor cortex **experimental plasticity**

Abstract

To evaluate whether the age-dependence of brain plasticity correlates with the levels of proteins involved in hormone and brain functions we executed a paired associative stimulation (PAS) protocol and blood tests. We measured the PAS-induced plasticity in the primary motor cortex. Blood levels of the brain-derived neurotrophic factor (BDNF), estradiol, the insulin-like growth factor (IGF)-1, the insulin-like growth factor binding protein (IGFBP)-3, progesterone, sex hormone-binding globulin (SHBG), testosterone, and the transforming growth factor beta 1 (TGF-beta 1) were determined in 15 healthy men and 20 healthy women. We observed an age-related reduction of PAS-induced plasticity in females that it is not present in males. In females, PAS-induced plasticity displayed a correlation with testosterone (p = 0.006) that became a trend after the adjustment for the age effect (p = 0.078). In males, IGF-1 showed a nominally significant correlation with the PAS-induced plasticity (p = 0.043). In conclusion, we observed that hormone blood levels (testosterone in females and IGF-1 in males) may be involved in the age-dependence of brain plasticity.

Example 2 of text analysis using exact Noun-Phrase-Clauses

PK_ITEMS	SOURCETITLE	UT_EID	PUBYEAR	DOCTYPE	SC_DESCRIPTION	OECD_DESCRIPTION
6980386	NEUROLOGICAL SCIENCES	000372330000006	2016	Article	NEUROSCIENCES	MEDICAL AND HEALTH SCIENCES

Relevant noun-phrase clauses found

Title

Biological factors and age-dependence of primary motor cortex experimental plasticity

Abstract

To evaluate whether the age-dependence of brain plasticity correlates with the levels of proteins involved in hormone and brain functions we executed a paired associative stimulation (PAS) protocol and blood tests. We measured the PAS-induced plasticity in the primary motor cortex. Blood levels of the brain-derived neurotrophic factor (BDNF), estradiol, the insulin-like growth factor (IGF)-1, the insulin-like growth factor binding protein (IGFBP)-3, progesterone, sex hormone-binding globulin (SHBG), testosterone, and the transforming growth factor beta 1 (TGF-beta 1) were determined in 15 healthy men and 20 healthy women. We observed an age-related reduction of PAS-induced plasticity in females that it is not present in males. In females, PAS-induced plasticity displayed a correlation with testosterone (p = 0.006) that became a trend after the adjustment for the age effect (p = 0.078). In males, IGF-1 showed a nominally significant correlation with the PAS-induced plasticity (p = 0.043). In conclusion, we observed that hormone blood levels (testosterone in females and IGF-1 in males) may be involved in the age-dependence of brain plasticity.

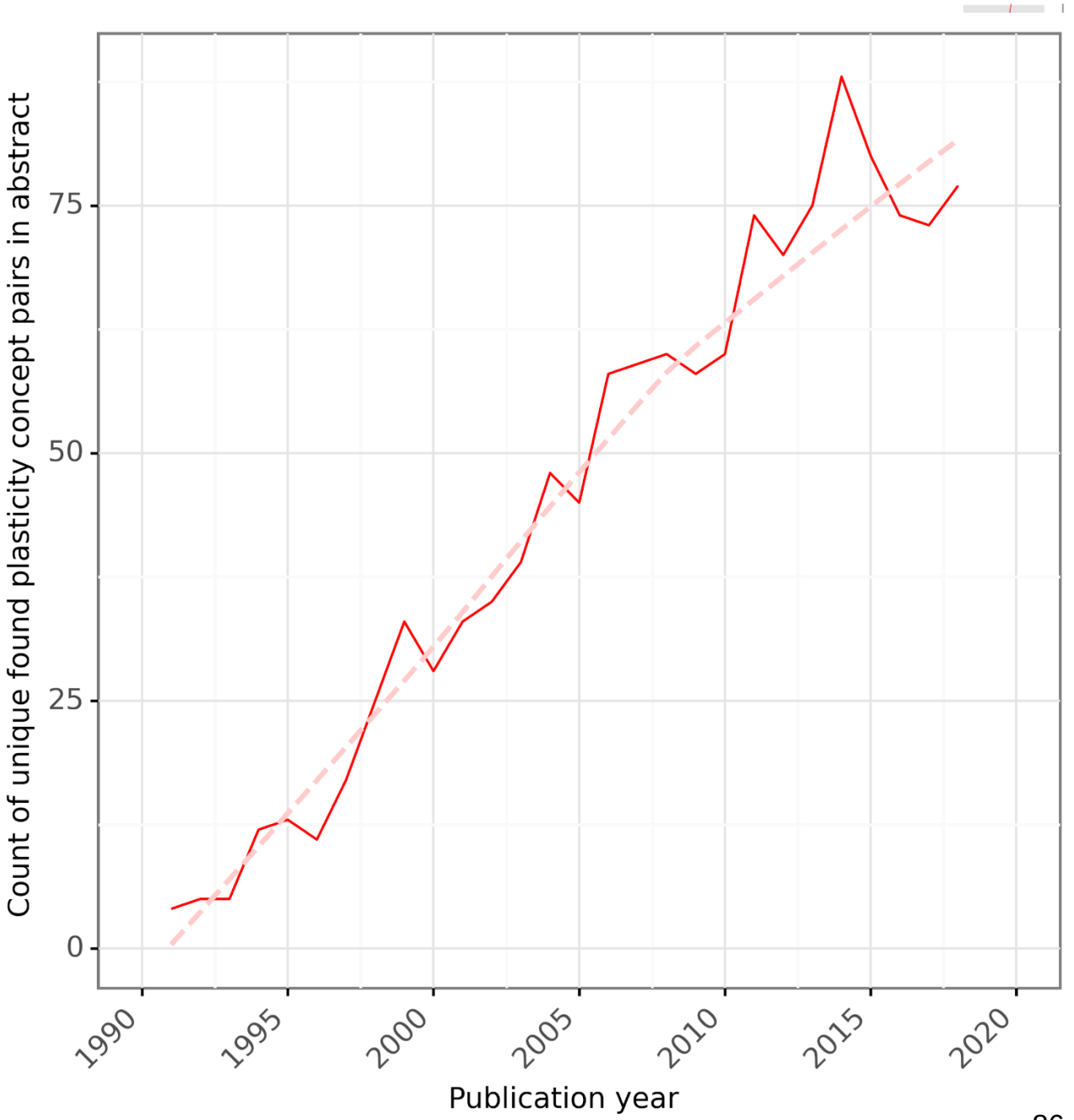
Plasticity as a concept in the neurosciences, a word pair analysis

- 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.



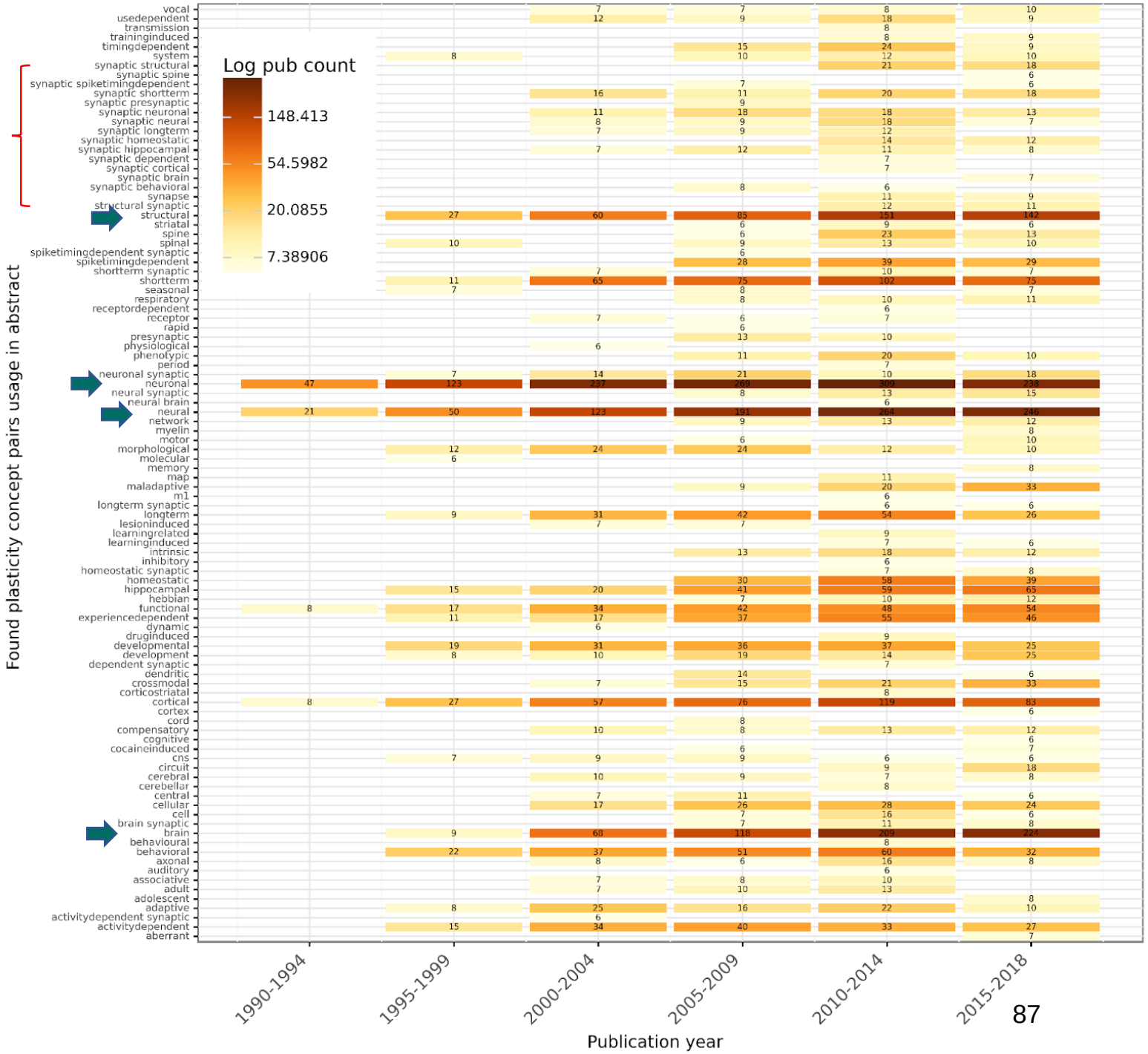
Plasticity as a concept in the neurosciences, a word pair analysis

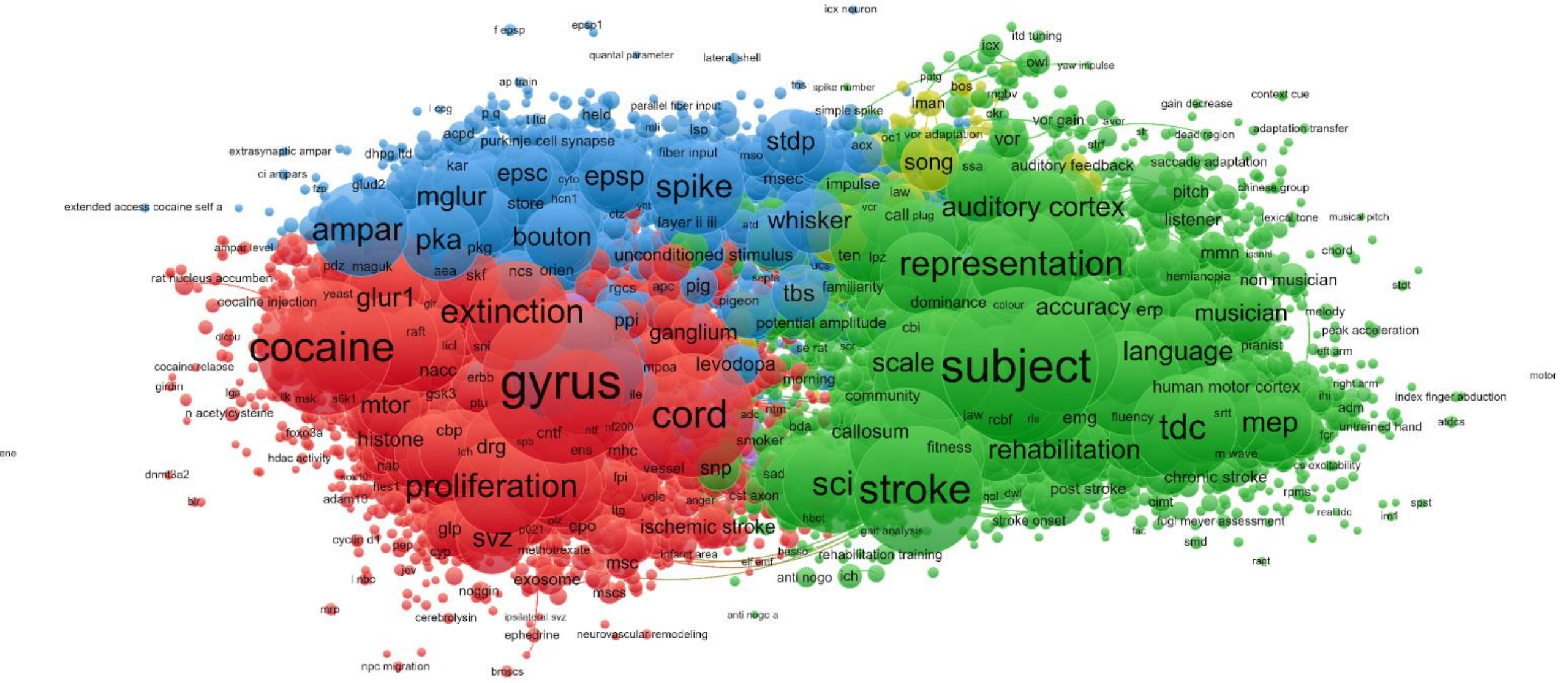
- Synaptic plasticity is the central anchor concept

- 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.

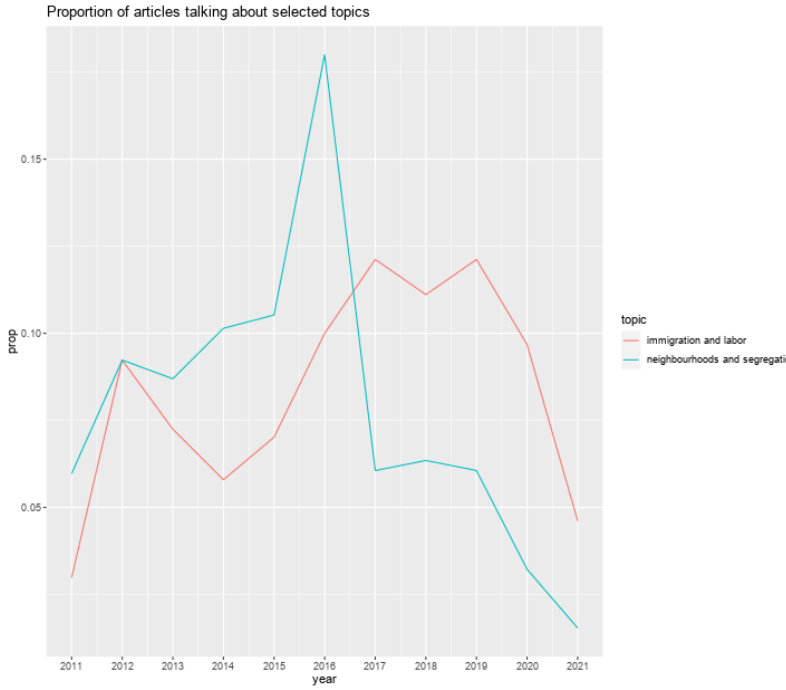
```
r$> bigrams_separated <- bigrams %>%
      separate(bigram, c("word1", "word2"), sep = " ")

    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/bibliometric_data_for_demographic_research/blob/main/2_presentations/06_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**

Q&A

Please raise any comments, [clarification or else] questions, or points on the slides and content that were presented!

5. Conclusion and discussion of potential limitations and future directions of using bibliometric data in demographic research (15 minutes)

Limitations in use of bibliometric data (for demographic research)

- Data quality

    - e.g., Scientific entity (e.g., authors, or institutions) **name disambiguation** (Tekles & Bornmann, 2020; Wu & Ding, 2013, Akbaritabar, 2021).

- Higher level **epistemic questions** need be addressed while repurposing these data for demographic research (Laudel, 2003; Moed et al., 2013; Moed & Halevi, 2014)

    - e.g., assigning the **country of affiliation in the first publication as the country of origin** for academic mobility is prone to error since that could simply be the country of graduation.

    - First publication year as **academic birth**

- There is a **publication delay** that can hinder proper identification of the mobility period.

- These data are limited to only those scholars who have **actively published**

- In indexed scholarly journals, so **coverage** may be incomplete (and over-represented by **WEIRD** countries).

Future directions of research

- New services and methods to prepare **cleaner data**
- Increased availability through initiatives for **open access** to data.

Potential future questions:

- How much of the talent circulation has happened "**within**" national borders versus "**between**" nations?
- Are there **migration corridors** connecting specific regions globally, for example between two specific regions across countries or in the same country, or systems of circulation that involve several countries or subregions?
- Do **scientific collaborations** among scholars facilitate their future mobilities?
- Do scholars have **different probabilities of being mobile** based on the trajectory of their collaborations during their scientific career?
- Complex interactions between processes related to migration of scientists and scholarly collaborations as well as **institutional settings and policies**.
- Finding **migration hubs or regions** with high concentration of academic labour or high attractiveness for future mobility that can inform policy.
- Evaluate theories **explaining migration through network tie formation** (Massey et al., 1993).

IN THE PIPELINE...

- We will announce SMD's beta website using participants' emails
  - So please share your email address with us (write to: scholarlymigration@demogr.mpg.de and include "[sign-up for SMDatabase]" in subject line), if we do not have it yet!

Final feedback!

Please raise any comments, [clarification or else] questions, or points on the slides and content that were presented!

Please tweet with hashtag
#BiblioDemography,
a tribute to James W. Vaupel (1945-2022).
Thanks Ilya and Jonas for bringing up Jim's
labeling idea!
and #EPC2022

Aliakbar Akbaritabar

Research Scientist

Akbaritabar@demogr.mpg.de

https://akbaritabar.github.io/

https://twitter.com/Akbaritabar

Xinyi Zhao

PhD candidate

Zhao@demogr.mpg.de

https://www.demogr.mpg.de/en/about_us_6113/staff_directory_1899/xinyi_zhao_4083/

https://twitter.com/XinyiZhao16

THANK YOU!