9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Scholarly migration and collaboration worldwide: A word embedding representation

Keywords: Migration of scholars, Scientific collaboration, Bibliometric data, Word embedding vectors

Extended Abstract

There is a global competition to attract the highly-skilled and talented[1], as they are considered innovation powerhouses[2]. Academics as a subset of the highly-skilled population are highly mobile, even called globetrotters[3], making their mobility the focus of recent literature[4]. Migrant academics contribute to innovation both in home[5] and host[6] countries. Modeling past trajectories of mobility and factors affecting it enables both an explanation and speculation for forecasting future mobility events[7] and global talent circulation[8].

Different factors may affect the decision to emigrate and in the case of academics one important and influential factor is scientific collaboration[9]. Scholars can form collaboration ties before, or during their mobility experience. Some of these formed ties persist, even after the mobility event. Nevertheless, the sequence of events does not always follow a defined order of mobility and collaboration. A theoretical framework considering the effect of network tie formation in migration[10] would help in identifying the influence of collaboration ties on scholarly mobility. An intertwined study of scholarly migration and collaboration is necessary to disentangle this sequence of events, but is rarely done[9]. The few studies considering scholarly migration and collaboration simultaneously report paradoxical findings on the direction of the effect and causation.

To fill this gap, we selected a sample of 10,963 authors from a 2020 snapshot of Scopus data. These authors were chosen based on criteria proposed by Bornmann and Haunschild[11] (N. of publications (o), N. as corresponding author (c), N. in Q1 (top ranked) journals (q1)) to construct a control and observation group. The observation group includes 3,564 authors considered potential talents and the control group includes 7,399 authors. We also consider the author's mobility status in our selection including internal mobility (between sub-national regions of one country), international mobility, both types of mobility, and immobility. We modeled the distance between organizations using word2vec word embeddings[12], where each 'sentence' was an individual author's affiliation or collaboration trajectory[13]. When an individual had multiple affiliations or collaborations in a given year we updated the model five times incorporating random shuffling of the organizations. We used UMAP reduction[14] with cosine similarity to visualize the results in 2D. Cosine similarity, which measures the distance between word vectors, is the most common measurement for comparison between words in word2vec.

We chose a similar number of authors globally for our control and observation groups (see Figure 1 and Table 1). We find that collaboration and affiliation trajectories are highly similar, though embedding representations of collaboration are more densely packed than mobility. This means that the cost of scientific collaboration tie formation between organizations located in geographically distant countries tends to be less than that for mobility (see Figure 2). Nevertheless, authors who are mobile or potential talents (top 1%

based on bibliometric criteria) are more likely to have a higher number of collaborators (see Table 2).

Hierarchical clustering of the countries based on collaboration and affiliation trajectories reveal interesting trends. In mobility (Figure 3) six clusters emerge: 1) the Nordic countries, 2) the continental Europe in addition to Greece and South Africa, 3) the UK, Ireland, Australia and New Zealand, in addition to Israel, 4) a larger cluster spanning eastern and south eastern Asian countries, Middle East and North Africa, 5) Eastern European countries in addition to Russia and Turkey and 6) Americas in addition to Spain and Portugal. These clusters seem to be formed under the influence of geographic distance and to a lesser extent language family and colonial histories. In collaboration (Figure 4), six clusters emerge: 1) a large cluster dominated by European countries, 2) Eastern Asia, 3) Middle East and Turkey, 4) Countries from Italic language family in Americas, 5) a not so clear mix of countries spanning South America, Africa, and East Asia, and 6) Northern America, Australia and New Zealand in addition to Israel. The role of the European Union and other agreements such as the Five Eyes is clear in the collaboration clustering, where countries entered in these agreements have a high level of collaboration even if they are not as close in terms of geographic distance. Figure 5 summarizes these trends and shows that colonial histories are more explanatory at higher levels of clustering than region or language family.

Our intertwined framework in considering scholarly migration and collaboration, and our spatial approach allowed modeling mobility and collaboration trajectories simultaneously while controlling for linguistic similarity and colonial history. Our methodological framework opens up promising avenues for future research on individual level forecasting of scholarly migration and on global dynamics of academic talent circulation.

References

[1] OECD. The Global Competition for Talent Mobility of the Highly Skilled: Mobility of the Highly Skilled. OECD Publishing, September 2008.

[2] Jennifer Hunt. Immigrant patents boost growth. Science, 356(6339):697–697, May 2017.

[3] John Bohannon. Restless minds. Science, 356(6339):690–692, May 2017.

[4] Cassidy R. Sugimoto, Nicolas Robinson-Garcia, Dakota S. Murray, Alfredo Yegros-Yegros, Rodrigo Costas, and Vincent Larivière. Scientists have most impact when they're free to move. Nature, 550(7674):29–31, 2017.

[5] Stefano Sbalchiero and Arjuna Tuzzi. Italian Scientists Abroad in Europe's Scientific Research Scenario: High skill migration as a resource for development in Italy. International Migration, 55(4):171–187, August 2017.

[6] Giuseppe Scellato, Chiara Franzoni, and Paula Stephan. Migrant scientists and international networks. Research Policy, 44(1):108–120, February 2015.

[7] Giacomo Vaccario, Luca Verginer, and Frank Schweitzer. Reproducing scientists' mobility: a data-driven model. Scientific Reports, 11(1):10733, May 2021.

[8] Torsten Wiesel. Fellowships: Turning brain drain into brain circulation. Nature, 510(7504):213–214, June 2014.

[9] Silvia Appelt, Brigitte van Beuzekom, Fernando Galindo-Rueda, and Roberto de Pinho. Chapter 7 - Which Factors Influence the International Mobility of Research Scientists? In Aldo Geuna, editor, Global Mobility of Research Scientists, pages 177–213. Academic Press, San Diego, January 2015.

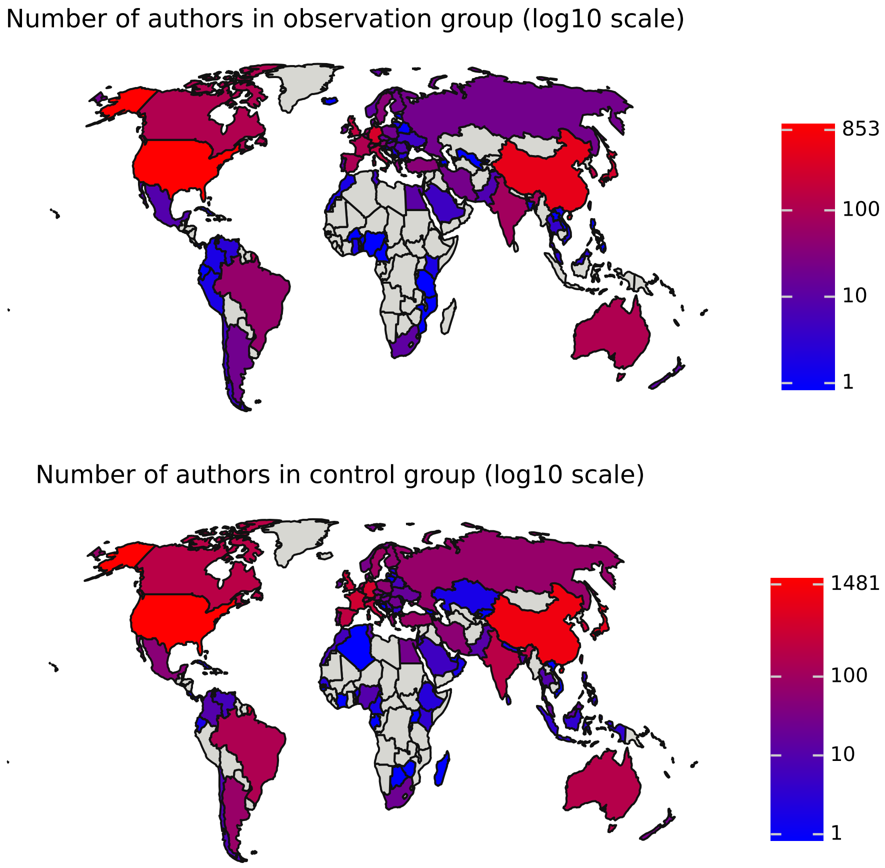
[10] Douglas S. Massey, Joaquin Arango, Graeme Hugo, Ali Kouaouci, Adela Pellegrino, and J. Edward Taylor. Theories of International Migration: A Review and Appraisal. Population and Development Review, 19(3):431, September 1993.

[11] Lutz Bornmann and Robin Haunschild. Identification of young talented individuals in the natural and life sciences using bibliometric data, September 2022.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[13] Dakota Murray, Jisung Yoon, Sadamori Kojaku, Rodrigo Costas, Woo-Sung Jung, Staša Milojević, and Yong-Yeol Ahn. Unsupervised embedding of trajectories captures the latent structure of mobility. arXiv:2012.02785 [physics], June 2021.

[14] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. Journal of Open Source Software, 3(29):861, 2018.



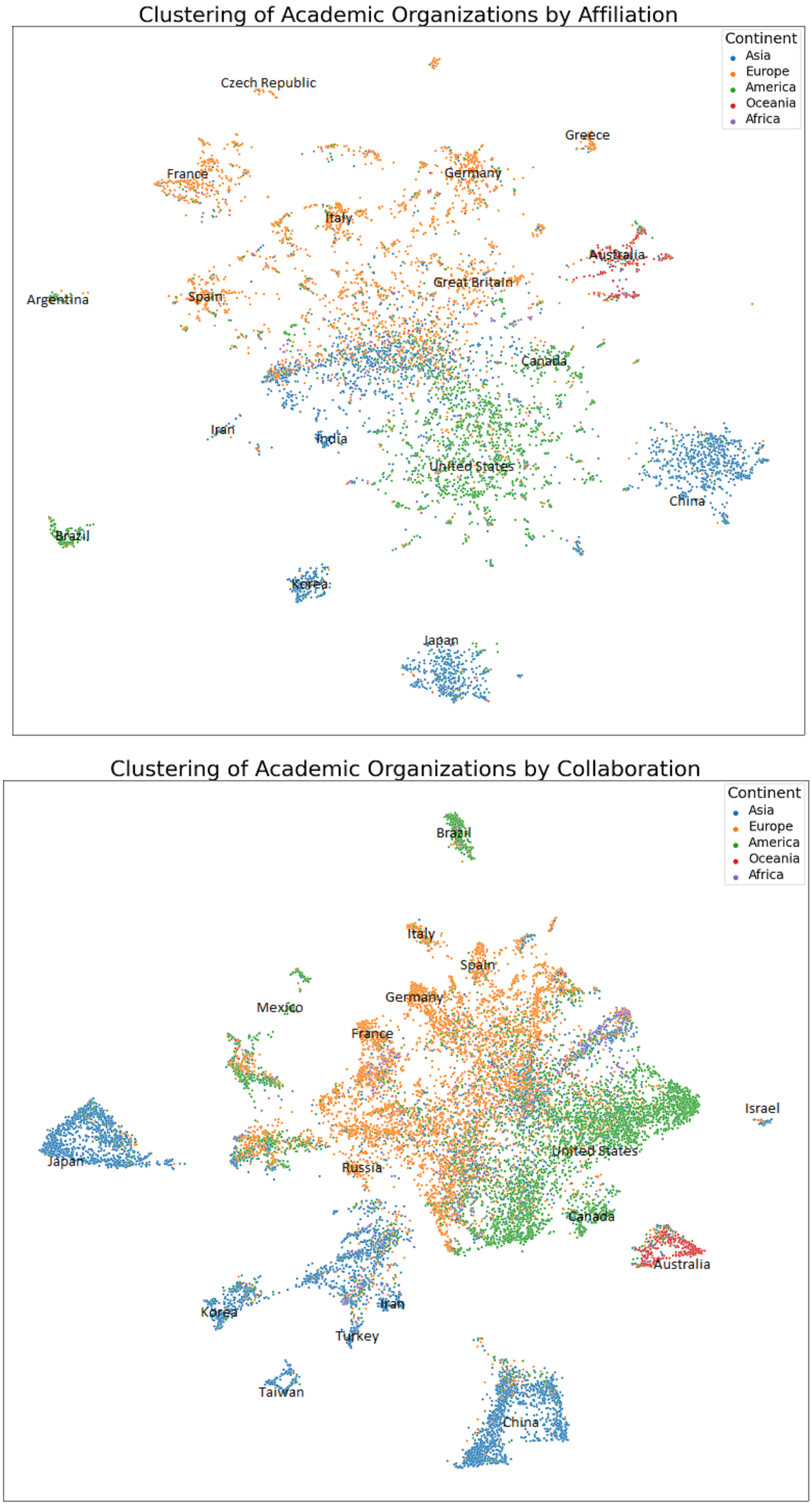
Number of authors in observation group (log10 scale)

Number of authors in control group (log10 scale)

Figure 1: Geographical distribution of authors in observation (top) and control (bottom) groups. Selected authors are affiliated with similar countries worldwide (in case of most countries, and especially in larger well established science systems) with a few countries missing from either observation or control group (e.g., see the cases of Peru, Algeria, Kazakhstan and Indonesia which are only present in the control group).

Table 1: The sample of 10,963 authors with composition of the control (7,399 authors) and observation (3,564) groups (numbers printed in the mobility status column are the sum of authors per mobility type over control and observation groups. N. of publications = o, N. as corresponding author = c, N. of publications in Q1 (top ranked) journals = q1. Interaction between these criteria is indicated with an x).

Group	Mobility status	Indicator combination	Count of unique authors
Control group	non-mobile (total: 2k)	top5-10% c	500
		top5-10% o	499
		top5-10% oxc	490
		top5-10% oxq1	496
		top5-10% oxq1xc	15
Control group	mobile internal (3k)	top5-10% c	300
		top5-10% o	296
		top5-10% oxc	277
		top5-10% oxq1	287
		top5-10% oxq1xc	211
		top5-10% q1	264
		top5-10% q1xc	196
Control group	mobile international (3k)	top5-10% c	300
		top5-10% o	298
		top5-10% oxc	283
		top5-10% oxq1	275
		top5-10% oxq1xc	184
		top5-10% q1	270
		top5-10% q1xc	177
Control group	mobile both (2,963)	top5-10% c	686
		top5-10% o	774
		top5-10% oxq1	98
		top5-10% q1	223
Potential talents	mobile internal (3k)	top1% c	288
		top1% o	284
		top1% oxc	253
		top1% oxq1	253
		top1% oxq1xc	91
Potential talents	mobile international (3k)	top1% c	284
		top1% o	277
		top1% oxc	235
		top1% oxq1	243
		top1% oxq1xc	174
Potential talents	mobile both (2,963)	top1% c	572
		top1% o	151
		top1% oxq1	153
		top1% q1	306


Clustering of Academic Organizations by Affiliation

Clustering of Academic Organizations by Collaboration

Figure 2: Clustering of organizations based on affiliation trajectories for mobility (top) and collaboration trajectories (bottom). The collaboration graph is much more dense, and many of the European countries lose their distinct clusters. As an illustrative example, despite the geographic distance between Australia, Europe and North America, on both clustering maps, it is located closer to them. In mobility, Australia is closer to Great Britain and Europe while in collaboration, it is closer to Canada and the US. In contrast, China is located slightly farther away from Asian countries.

Table 2: Shows the mean similarity between affiliation and collaboration trajectories for authors in the dataset separated by movement status, top, and talent status, bottom, as well as the average number of institutions collaborated with.

Group	Similarity	# Collabs
non-mobile (2k)	0.93	39.32
mobile internal (3k)	0.94	99.00
mobile international (3k)	0.92	129.12
mobile both (2,963)	0.94	121.00
Control group	0.94	62.71
Potential talents	0.92	184.50

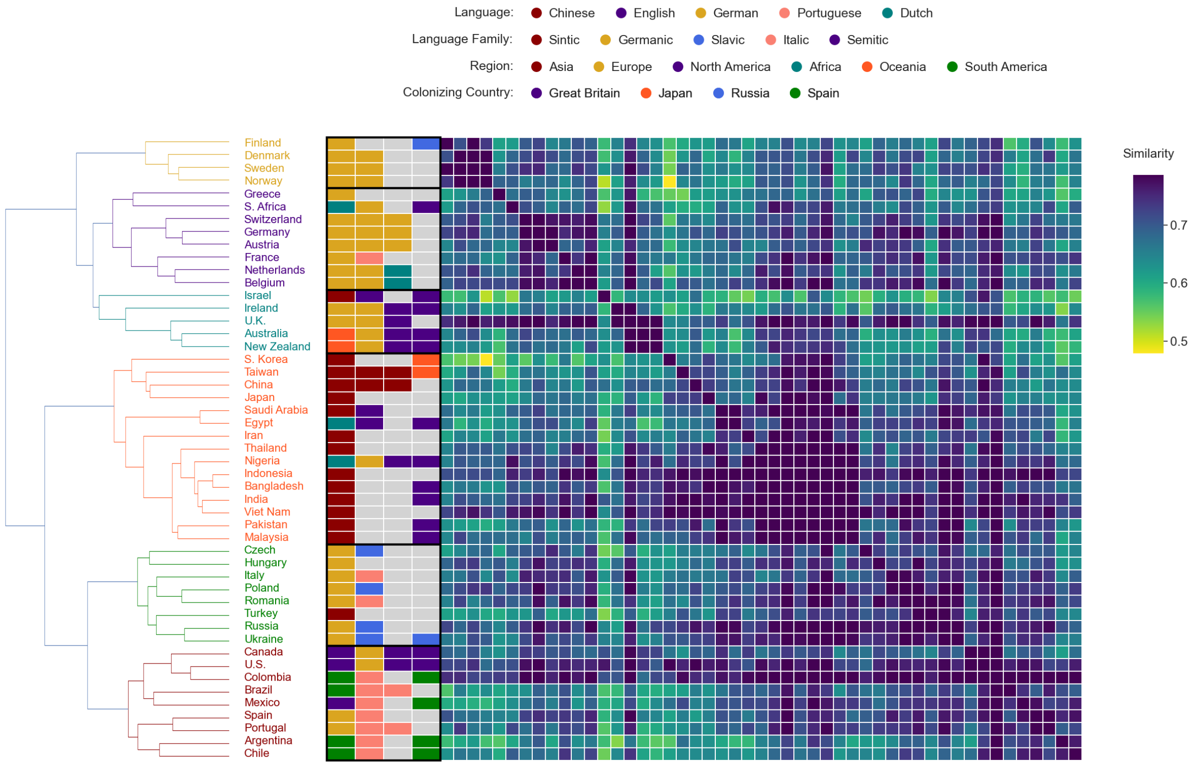


Figure 3: Clustering based on affiliation trajectories of individuals and mobility between institutions in the given countries. Colors in the heatmap represent the cosine similarity between country vectors. Boxes on the left indicate, from left to right, region, language family, language, and colonizing country.

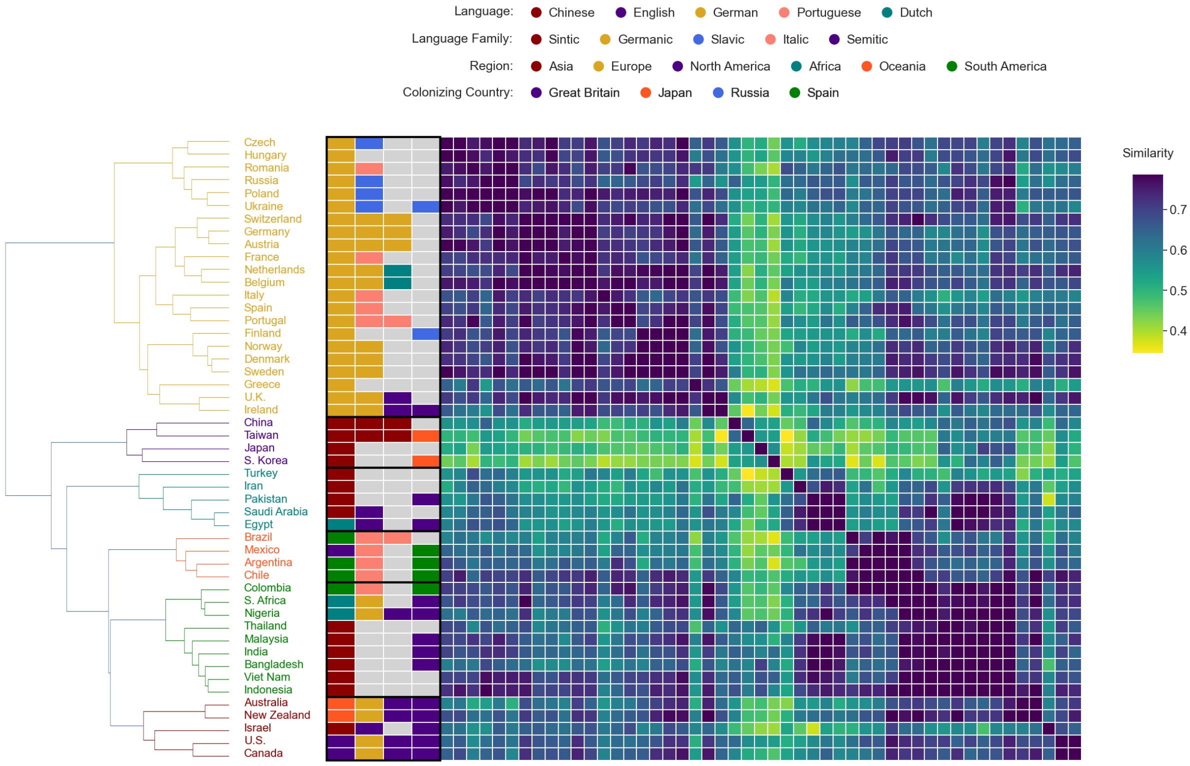
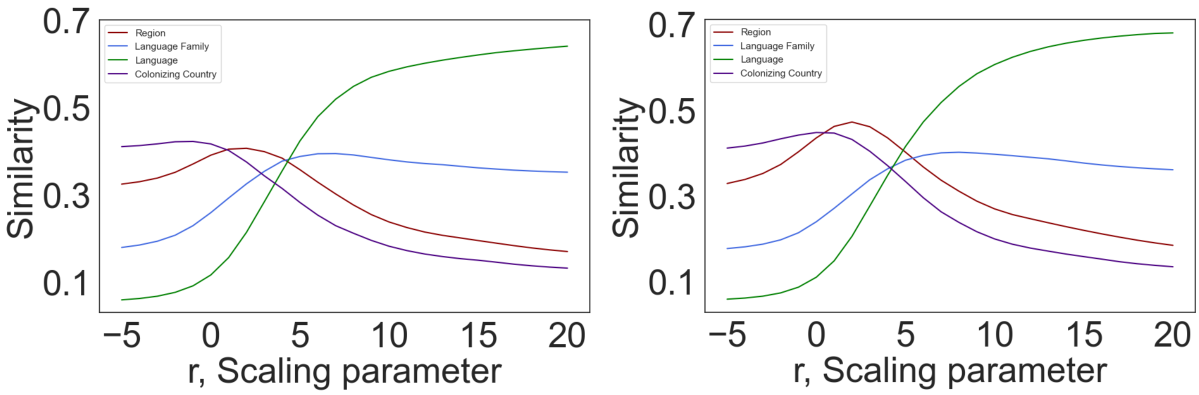


Figure 4: Clustering based on collaboration trajectories of individuals between institutions in the given countries. Colors in the heatmap represent the cosine similarity between country vectors. Boxes on the left indicate, from left to right, region, language family, language, and colonizing country.



(a) Affiliation (b) Collaboration

Figure 5: Factors influencing hierarchical clustering at lower and higher levels for affiliation and mobility of scholars, left, and collaboration, right. Colonizing country is the most influential at high levels, while language is the most influential at lower levels.