Characterizing High-Skilled Mobility Patterns in Europe from Social Media

Daniela Perrotta[1], Diego Alburez-Gutierrez[1], Carlos Callejo Peñalba[2], Kiran Garimella[3],
Tom Theile[1], Ingmar Weber[4], Emilio Zagheni[1]

1. Max Planck Institute for Demographic Research; 2. Institute for Data, Society and Systems - MIT;
3. Aalto University; 4. Qatar Computing Research Institute

Introduction

International high-skilled migration represents an increasingly large component of global migration streams with a significant impact on the global economy, gender imbalance in employment, and migration policies. Official data on migration are usually collected through census by National Statistics Offices, which provide information on the employment status of the population by age, gender and industry. However, such traditional data are generally costly, coarse-grained, and inconsistent across countries. Novel digital sources of data and innovative techniques are used more and more to detect people's physical movements over time [1]. Some examples include Facebook's advertising platform to estimate stocks of migrants [2], the Web of Science to detect the mobility of researchers [3], LinkedIn to investigate the migration of professionals to the United States [4].

In this study, we focus on migration patterns of high-skilled workers in Europe in terms of age, gender and industry. The main source of data is the social networking service LinkedIn, mainly used for online professional networking. LinkedIn allows users to create personal profiles to describe their work experience, education and training, skills, publications, projects, interests, and other additional information. Specifically, here we use the advertising platform called LinkedIn Ads[1] which enables advertisers to create ads and content intended for a specific audience of LinkedIn users who can be reached by selecting specific targeting criteria, such as gender, age group, city, industry, and others. Given these options, LinkedIn provides an estimate of how many LinkedIn users (i.e. *audience size*) meet these criteria and can be potentially reached by the advertisement.

For the purposes of our research, we are interested in identifying migration flows across countries in Europe by retrieving the audience sizes of high-skilled LinkedIn users by age groups, gender and industrial categories. As an illustrative example, LinkedIn estimates that an advertisement targeted to an audience of females aged between 25 and 34 years old who studied in Italy and now work in Germany in the industrial sector of High Tech has the potential to reach 1,600 people. On the contrary, the same advertisement targeted to the same audience but whose LinkedIn users have studied in Germany and now work in Italy has the potential to reach 630 people. Such estimates allows to detect international migration of LinkedIn users from the country of origin where they have studied to the country of destination where they are currently employed.

Dataset

In this study, we collected data from LinkedIn advertising platform to gather audience sizes for various combinations of targeting criteria of interest, including the following options: 1) location where member users have studied, 2) location where member users are currently

[1] https://business.linkedin.com/marketing-solutions/ads

employed, 3) gender, 4) age, and 5) industry. For each combination of these five variables, we queried LinkedIn Ads to obtain the estimated size of the LinkedIn audience matching the given criteria. Targeting options provided by the advertising service are standardized in terms of variables that can be chosen to select the corresponding audience. Specifically, targeting options for education include school, college, university, or other learning institution, including kindergarten, elementary school, high school, etc. Since we focus on highly skilled migrants, here we used the list of all European universities reported on UniRank[2] in order to select only LinkedIn users having at least a BA, MA or PhD. Member age and gender are estimated (or inferred) based on their profile information. Age is aggregated in age groups, 18-24, 25-34, 35-54, and 55+, respectively. The industry indicates where the member is currently employed and include industry groups such as Agriculture, Arts, Finance, High Tech, Manufacturing, to name a few.

As a result of this process, we collected data for all combinations of 47 European countries, 2 genders, 4 age groups and 17 industries, for a total of 842,400 data instances, out of which 29,812 are non-zero. We collected data from LinkedIn during two consecutive years, in November 2018 and October 2019, respectively.

Preliminary Analysis

In 2018 approximately 135 million users were registered on LinkedIn in the 47 European countries and this number has increased to 148 million in 2019. LinkedIn users are mostly distributed in the United Kingdom (about 18%), France (about 13%), Italy (about 9%), Spain (about 8%), Germany (about 7%), the Netherlands and Turkey (about 6%), Russia (about 5%).

Although the sex ratio for the population in Europe is predominantly female, with the exception of a few countries such as Iceland and Norway, the sex ratio of LinkedIn users is unbalanced towards female member users. However, we observe a slight increase in the proportion of female users, from 44.5% in 2018 to 44.6% in 2019, and a slight decrease in the proportion of male users, from 55.5% in 2018 to 55.4% in 2019. Only 13 countries out of 47 show a higher percentage of LinkedIn female users, namely Lithuania, Latvia, Georgia, Moldova, Finland, Montenegro, Slovenia, Macedonia, Romania, Armenia, Republic of Serbia, Bulgaria, and Estonia.

For each combination of targeting criteria, the estimates obtained for the audience sizes allows to reconstruct a *mobility network* in which the nodes correspond to the countries and the links corresponds to the connection between country of origin (i.e. place where LinkedIn users have studied) and country of destination (i.e. place where LinkedIn users are currently employed).

Given a country of origin *i* and a country of destination *j*, the flow w_{ij} corresponds to the number of LinkedIn users who moved from *i* to *j* for work reasons. The resulting network is: 1) directed as the users can move in both directions (from *i* to j or from *j* to *i*), 2) not fully connected as not all connections are present in the network, and 3) a multigraph as it allows self-loops, i.e. LinkedIn users who have studied and now work in the same country.

For the sake of simplicity, here we limit our analysis to the mobility network of LinkedIn users, regardless of age, gender and industry category. Figure 1 shows the origin-destination matrix for the top 30 countries in terms of number of links (i.e. *degree*) in the network. The rows correspond to countries of origin and the columns to countries of destination, while blanks indicate no connection between the two countries. Most of the countries show a variable

[2] https://www.4icu.org

number of incoming and outgoing links. For example, Latvia has 15 connections if it is considered as a country of origin, while 8 if considered as a country of destination, self-loop excluded. Consequently, this means that more people leave Latvia than those who enter the country for job purposes. On the contrary, more people enter Switzerland than those who leave the country for job purposes. The United Kingdom has the highest number of high-skilled users who left the country to go to work elsewhere.

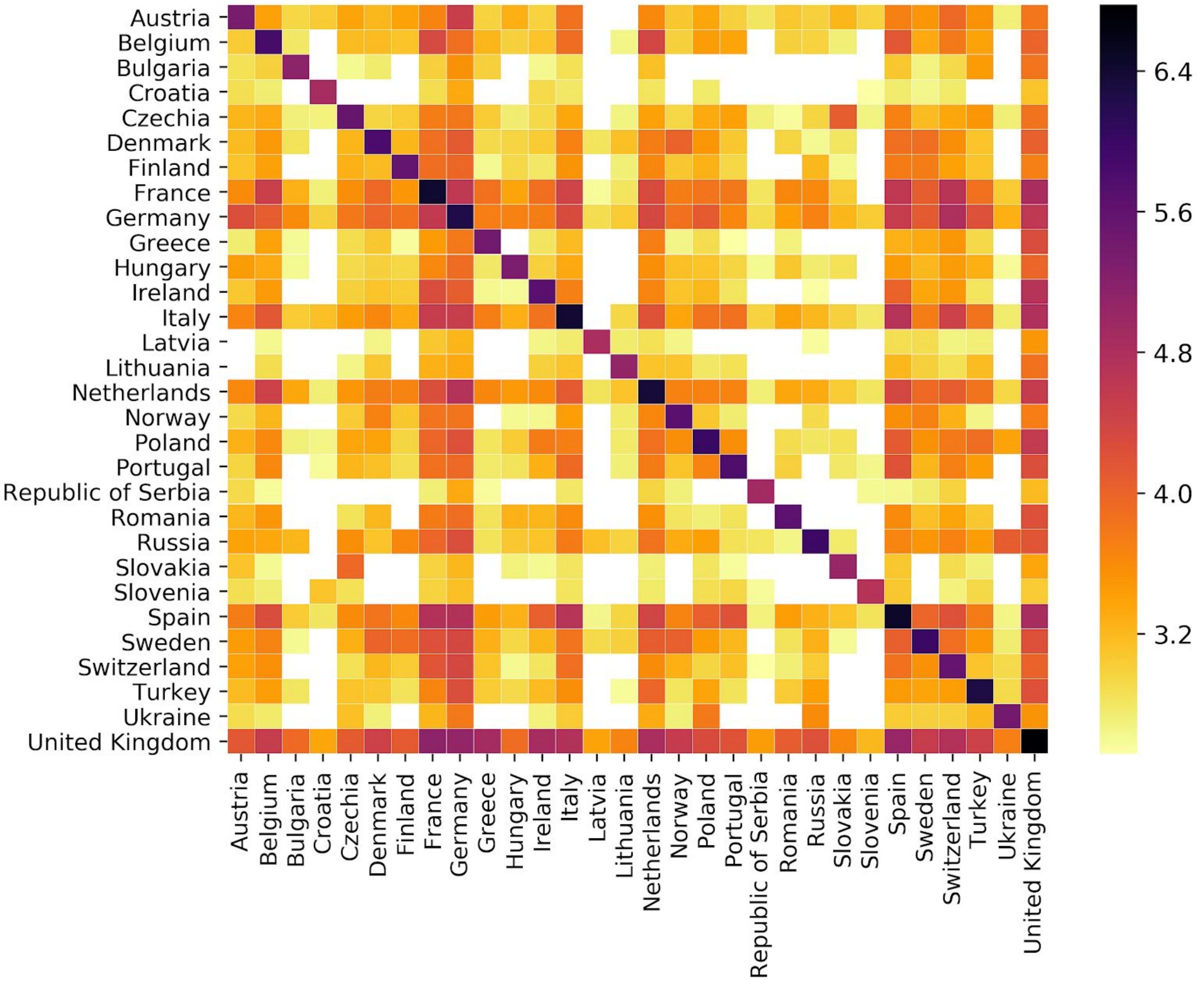

Figure 1. The heatmap shows the origin-destination matrix of the migration of high-skilled LinkedIn users from the country of origin where they have studied to the country of destination where they are currently employed. Flows are reported in logarithmic scale.

Discussion and Future Work

Here we presented a small glimpse of the international migration of high-skilled people as detected from LinkedIn. Indeed, digital data from services like LinkedIn represent a relatively low-cost resource to identify migration patterns and draw a high-resolution picture of human mobility patterns at an unprecedented scale. Compared to traditional data on migration, LinkedIn data have the advantages of: 1) high temporal resolution as data is always available and changes can be detected every time that LinkedIn users update their employment status or every time that new members join LinkedIn, 2) high spatial resolution up to the level of metropolitan areas, 3) availability of additional details, such as skills, interests or years of experience accumulated over the entire career, and 4) common cross-country definitions of the targeting options, such as for example the same definition is used for industrial categories or skills and allows the comparison across different countries.

On the other hand, there are some systematic biases that we must take into account. Firstly, multiple counting of the same user can occur whenever he/she moved multiple times during his/her career. For example, if a LinkedIn user obtained the BA from the Sorbonne in France and the MA from the Imperial College London in the United Kingdom and then moved to Germany for work reasons, in this case the user is counted twice, respectively in the audience size of LinkedIn users who have studied in the United Kingdom and currently work in Germany, and in the audience size of LinkedIn users who have studied in France and currently work in Germany.

Secondly, the audience sizes by age and gender are systematically smaller due to the fact that age and gender are inferred from the profile information and may not be retrieved for all users. Thirdly, the nationality of users is unknown, thus limiting the interpretation of the direction of migration patterns, if towards the home country or towards a foreign country. Lastly, the usage patterns and hence the self-selection biases are heterogeneous across countries, thus migration flows must be adequately weighted in order to take into account the different penetration rate of LinkedIn. Indeed, some of the main challenges of this research include data calibration in order to correct for biases, and data validation against the traditional migration data. As future work, we plan to perform these two steps, as well as a more in-depth exploration of our datasets, including the socio-demographic features, in order to better characterize the migration patterns based on different layers of information that we can gather from LinkedIn data. Furthermore, given these additional features, we plan to develop a modelling approach based on gravity-type migration model [5], to assess the extent to which there are imbalances in migration flows across countries and by socio-demographic characteristics, including age and gender, or by industry of occupation. As we develop our model we expect to be able to evaluate the relative importance of socio-economic, political, and geographical factors in shaping flows of professionals in Europe.

References

[1] Spyratos S, Vespe M, Natale, F, Weber I, Zagheni E, Rango M. *Migration data using social media: a European perspective.* Publications Office of the European Union (2018)

[2] Zagheni E, Weber I, Gummadi K. *Leveraging Facebook's advertising platform to monitor stocks of migrants. Population and Development Review*, 43(4), 721-734 (2017)

[3] Aref S, Zagheni E, West J. *The demography of the peripatetic researcher: Evidence on highly mobile scholars from the Web of Science*. *arXiv preprint arXiv:1907.1341* (2019)

[4] State B, Rodriguez M, Helbing D, Zagheni E. *Migration of professionals to the US. Evidence from LinkedIn data.* In *International Conference on Social Informatics* (pp. 531-543). Springer, Cham. (2014)

[5] Cohen JE, Roig M, Reuman DC, and GoGwilt C. International migration beyond gravity: A statistical model for use in population projections. PNAS, 105 (40) 15269-15274 (2008)